République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université 8 Mai 1945 Guelma

Faculté des Sciences et de la Technologie
Département de Génie des Procédés
Laboratoire des Silicates, Polymères et Nanocomposites (LSPN)

THÈSE

EN VUE DE L'OBTENTION DU DIPLOME DE DOCTORAT EN 3ème CYCLE

Domaine : Sciences et Technologie Filière : Génie des Procédés

Spécialité : Génie des Matériaux

Présentée par

ROUBEHIE FISSA Mohamed

Intitulée

Estimation de quelques propriétés physico-chimiques des hydrocarbures en fonction de leurs Propriétés (RQSP) en utilisant la méthode de Réseaux de Neurone Artificiel (RNA)

Soutenue le 19/05/2021 Devant le Jury composé de :

Mr SATHA Hamid	Professeur	Univ. 8 Mai 1945 Guelma	Président
Mme LAHIOUEL Yasmina	Professeur	Univ. 8 Mai 1945 Guelma	Encadreur
Mr HANINI Salah	Professeur	Univ. Yahia Farès de Médéa	Co-encadreur
Mr LAIDI Maamar	M.C.A	Univ. Yahia Farès de Médéa	Examinateur
Mr BELAIDI Salah	Professeur	Univ. Mohamed Khider Biskra	Examinateur
Mme SERIDI Lynda	Professeur	Univ. 8 Mai 1945 Guelma	Examinateur

Année Universitaire : 2020/2021

الحمد لله كما ينبغي لجلال وجهه وعظيم سلطانه

Dédicaces

Je dédie ce modeste travail avec ma profonde conviction, à tous ceux qui ont toujours cru et sans doute à la Science. Cette lumière qui éclaire les esprits et leurs permet de transcender les limites installées par les sociétés et les cultures dans leurs périodes de dégénérescence.

À ceux qui combattaient silencieusement pour les meilleurs jours ;

À ceux qui ne portent que du bien pour les autres ;

À ceux qui ne vivent que dans et par la vérité ;

À ceux qui savent concrétiser leurs idées et savent réussir ;

À ceux qui m'ont appris la rigueur et la patientes.

Je dédie ce travail ***à mes chers parents*** *pour leurs* ***soutiens,*** *leurs* ***patientes*** *et leurs* ***sacrifices*** *par tous les temps pour me voir un* ***homme bien.***

À ma ***Grand famille*** *paternelle* ROUBEHIE FISSA *et maternelle* ABDERRAHMANE *sans exception.*

À mes frères ***Abdelmalek*** *et* ***Hossine.***

À ma future femme ***Inconnue*** *!*

À mes amies proches ***Ahmed, Ali, Zakaria, Abdelkader, Mohamed, Abdellah, Yassine, Soufiane, Samir, Ammar, Rabah, ...*** *Surtout pour leur soutien moral et pour tous les bons moments que nous avons passés ensemble tout au long des années d'études.*

À tous mes amies et collègues de laboratoire et doctorants de ***l'Université 8 Mai 1945 de Guelma*** *et de* ***l'Université Yahia FARES de Médéa*** *et aussi À tous ceux qui me connaissent de près ou de loin.*

À ma promotion ***(2014-2015)*** *de* ***Génie des procédés****, option* ***Bioprocédés Pharmaceutiques*** *sans exception.*

À toutes les personnes que je n'ai pas citées mais que je porte dans mon cœur et À tous ceux qui m'ont soutenu de près ou de loin.

ROUBEHIE FISSA Mohamed

Remerciements et gratitude

À la lumière de ce travail, je tiens en premier lieu à remercier **ALLAH Tout-Puissant** *pour toutes les merveilles qu'il fait dans ma vie et pour m'avoir donné le courage, la force, la volonté et surtout la patience d'achever cette modeste contribution scientifique.*

Tout d'abord, parce qu'elle le mérite vraiment, je tiens à remercier ma directrice de thèse, Madame **Yasmina LAHIOUEL**, *Professeur à l'Université 8 Mai 1945 Guelma, qui m'a guidé avec grande patience tout au long de l'élaboration de ce travail, pour sa gentillesse et sa compréhension, pour ses précieuses aides qui ont judicieusement éclairé mon chemin vers l'aboutissement et la réussite à la concrétisation de ma thèse de Doctorat.*

Je tiens aussi à exprimer ma gratitude et mes sincères remerciements à mon co-directeur de thèse, Monsieur **Salah HANINI**, *Professeur à l'UYFM, qui m'a guidé avec grande patience tout au long du chemin d'élaboration de ce travail, pour ses précieux conseils, sa générosité, pour tous les services qu'il m'a offerts, pour avoir suivi attentivement la progression de ce travail, et pour ses orientations qui ont éclairé mon chemin vers la réussite dans ma thèse de Doctorat.*

Je suis profondément reconnaissant au Madame **Latifa KHAOUANE**, *Maître de Conférences (A) à l'UYFM, de m'avoir donné la chance de travailler avec elle en tant qu'enseignant vacataire à l'UYFM, pour son aide, ses remarques et ses conseils précieux, ses accompagnements et ses éclairages techniques. Je lui dois mes sincères remerciements.*

Je tiens également à exprimer mes plus vifs remerciements aux : Monsieur **Mabrouk HAMADACHE** *Maître de Conférences (A) à l'UYFM, Monsieur* **Othmane BENKORTBI** *Professeur à l'UYFM, et Monsieur* **Chérif SI-MOUSSA**, *Maître de Conférences (A) à l'UYFM, pour leur aide et surtout pour leurs encouragements.*

Je tiens aussi à exprimer mes remerciements à Monsieur le Professeur **SATHA Hamid** *directeur du laboratoire de recherche* **LSPN** *et Responsable de la formation doctorale ainsi que tous les membres du laboratoire de recherche LSPN de l'Université 8 Mai 1945 Guelma. Je lui remercier encore pour avoir accepté de présider le jury de soutenance.*

Je tiens également à présenter ma profonde gratitude aux **membres du jury** *: Monsieur* **Maamar LAIDI**, *Maître de Conférences (A) à l'UYFM, Monsieur* **Salah BELAIDI**, *Professeur à l'UMKB, et Madame* **SERIDI Lynda**, *Professeur à l'Université 8 Mai 1945 Guelma, d'avoir accepté de juger ce modeste travail.*

Merci à tous ceux qui m'ont aidé de près ou de loin durant ces années de recherche et d'études.

Je ne peux pas conclure sans un mot de remerciement à ***tous mes enseignants*** *et à* ***tous mes collègues****.*

Résumé : Estimation de quelques propriétés physico-chimiques des hydrocarbures en fonction de leurs Propriétés (RQSP) en utilisant la méthode de Réseaux de Neurone Artificiel (RNA).

L'objectif de cette thèse vise à développer des modèles mathématiques basés sur les structures moléculaires des composées, pour prédire quelques paramètres physico-chimiques des hydrocarbures purs, par le biais d'un Réseau de Neurone Artificiel-Perceptron Multi Couches (RNA-PMC) basé sur le concept des Relations Quantitatives Structure-Propriétés (RQSP).

Les modèles issus de la combinaison entre le concept RQSP et le RNA appelés dans cette thèse ; modèles de type Relations Quantitative Structure Propriété-Réseau de Neurone Artificiel (RQSP-RNA). Cinq modèles de type RQSP-RNA ont été élaborés pour prédire cinq paramètres physicochimiques des hydrocarbures purs : la température du point d'ébullition normale (*Tb*), la densité relative du liquide (*d20*) et les propriétés critiques {température critique (*Tc*), pression critique (*Pc*) et volume critique (*Vc*)}. Une base de données d'ensemble de 223, 222 et 221 des différents points de données pour (*Tb*, *Tc* et *Vc*), (*d20*) et (*Pc*), ont été respectivement utilisés pour créer cinq modèles de type RQSP-RNA. Pour chaque modèle, la base de données totale a été divisée de manière aléatoire en deux sous-ensembles : 80% pour l'apprentissage et 20% pour le test. 1666 descripteurs moléculaires sont issus de 20 classes différentes, ont été calculés à partir du logiciel en ligne E-Dragon, et ensuite une méthodologie statistique de réduction basée sur la méthode de la Régression Linéaire Multiple (RLM) a été adoptée, afin de réduire et de sélectionner ce grand nombre de descripteurs moléculaires, et de les conserver à un nombre acceptable de descripteurs pertinents. Le pourcentage de descripteurs moléculaires primaires exclu était approximativement supérieur à 98% dans tous les cinq modèles développés. L'algorithme BFGS (Propagation en arrière quasi-newtonienne) a été appliqué afin d'entraîner le RNA. L'architecture des meilleurs modèles de type RQSP-RNA était : (16-12-1), (16-10-1), (10-16-1), (17-11-1) et (12-17-1), respectivement pour les modèles de paramètres : *Tb*, *d20*, *Tc*, *Pc* et *Vc*.

Les résultats des cinq meilleurs modèles de type RQSP-RNA obtenus, ont montré une bonne précision confirmée par les valeurs des coefficients de régression (R) et de détermination (R^2) élevées, qui sont respectivement comprises entre 0,9965–0,9999, et 0,9931–0,9999 pour les trois sous-ensembles de base des données (l'ensemble d'apprentissage, de test, et entier). En outre les petites valeurs de la Racine d'Erreur Quadratique Moyenne (REQM) qui égale à : 2,0168 (K), 0,0060 (-), 6,9672 (K), 4,8231 10^{+4} (Pa) et 10,6349 (cm^3/mol), respectivement pour *Tb*, *d20*, *Tc*, *Pc* et *Vc*. En plus des faibles valeurs de trois types d'erreurs relatives différentes, qui sont : l'Erreur Standard de Prédiction (ESP), l'Erreur Relative Moyenne (ERM), et l'Erreur Relative Absolu Moyenne (ERAM), comprises respectivement entre 0,3255–2,4024%, 0,2613–2,2683%, et 0,2600–2,2497%. Ces résultats concernent les cinq modèles développés pour les trois sous-ensembles de la base de données.

La méthode d'analyse de sensibilité (méthode de poids) a été utilisé dans cette thèse, afin d'apprécier individuellement la contribution des descripteurs d'entrée sur chaque modèle de type RQSP-RNA développée, les résultats ont montré l'importance de chaque descripteur utilisé, dans lequel la majorité des descripteurs ont une grande ou moyenne contribution par rapport à la valeur de la contribution moyenne (VCM) de chaque modèle, ce qui confirme l'efficacité de la méthodologie choisie dans cette thèse pour la sélection des descripteurs pertinents. Afin de connaitre le domaine d'applicabilité de nos modèles développés, la méthode basée sur la distance a été appliquée dans cette thèse avec des valeurs strictes de résidus standardisés (d_i^*) prises à (±2), l'applicabilité des modèles a été quantifié, en le présentant sur un graphique visuel appelé le diagramme de Williams (Williams plot), les résultats montrent qu'il y a entre 87%– 92% de l'ensemble des substances entières dans le domaine d'applicabilité dans les cinq modèles développés, et ceci, indépendamment des valeurs de résidus standardisées pris (±2), ce qui confirme encore l'efficacité des modèles développés dans cette étude.

Une comparaison a été faite entre les résultats obtenus par les modèles de type RQSP-RNA et d'autres modèles dans certaines études publiées dans des articles scientifiques qui ont adoptées les mêmes approches (RQSP, RNA, RLM, …). La comparaison a montré que les modèles de type RQSP-RNA ont fourni de meilleurs résultats en termes : des valeurs des coefficients de régression élevées, des faibles valeurs de différents erreurs calculées, ainsi que les valeurs du vecteur de régression qui sont considérées proches de l'idéal, et qui sont encore prouvées en très bon accord entre les valeurs expérimentales de la base de données et les valeurs calculées obtenues par les cinq modèles de type RQSP-RNA développés dans cette thèse.

Cette approche computationnelle peut être appliquée dans l'ingénierie pétrolière ou d'autres domaines reliés pour une détermination précise des propriétés : *Tb*, *d20*, *Tc*, *Pc* et *Vc* des hydrocarbures purs uniquement par l'utilisation des informations incorporées dans leurs structures chimiques.

Mots clés : Estimation, Propriétés physicochimiques, Hydrocarbures, Modélisation, RQSP-RNA, Descripteurs.

Abstract: Estimation of some physicochemical properties of hydrocarbons as a function of their Properties (QSPR) using the Artificial Neural Networks (ANN) method.

The objective of this thesis aims to develop mathematical models based on the molecular structures of compounds, in order to predict some physicochemical parameters of pure hydrocarbons, through a Multi Layer Perceptron-Artificial Neural Network (MLP-ANN) based on the concept of Quantitative Structure-Property Relationships (QSPR).

The resulting models from the combination between the QSPR concept and the ANN called in this thesis; Quantitative Structure Property Relationship -Artificial Neural Network (QSPR-ANN) models. Five QSPR-ANN model's type have been developed to predict five physicochemical parameters of pure hydrocarbons: normal boiling point temperature (*Tb*), relative liquid density (*d20*) and the critical properties {critical temperature (*Tc*), critical pressure (*Pc*) and critical volume (*Vc*)}. A complete database of 223, 222 and 221 of different data points for (*Tb*, *Tc* and *Vc*), (*d20*) and (*Pc*) were respectively used to create five QSPR-ANN model's type. For each model, the total database was randomly divided into two subsets: 80% for the training set and 20% for the test set. 1666 molecular descriptors come from 20 different classes, were calculated using the online software E-Dragon, and then a statistical methodology of reduction based on the Multiple Linear Regression (MLR) method, was adopted, in order to reduce and select these large numbers of molecular descriptors, and retain them to an acceptable number of relevant descriptors. A percentage of the excluded primary molecular descriptors was approximately higher than 98% in the all five models developed. The BFGS (Quasi-Newton Back Propagation) algorithm has been applied to train the ANN. The architecture of the best QSPR-ANN model's types was: (16-12-1), (16-10-1), (10-16-1), (17-11-1) and (12-17- 1), respectively for the model's parameters: *Tb*, *d20*, *Tc*, *Pc* and *Vc*.

The results of the best five QSPR-ANN model's types obtained, showed good precision confirmed by the high values of the regression (R) and the determination (R^2) coefficients, which are respectively ranging between 0.9965-0.9999, and 0.9931-0.9999 for the three basic subsets of the database (the training set, test, and the whole dataset). In addition to the small values of the Root Mean Squared Error (RMSE) which equals: 2.0168 (K), 0.0060 (-), 6.9672 (K), 4.8231 10^{+4} (Pa) and 10.6349 (cm^3/mol), respectively for *Tb*, *d20*, *Tc*, *Pc* and *Vc*. In addition to the low values of three different relative errors types, which are : The Standard Error of Prediction (SEP), the Mean Percentage Error (MPE) and the Mean Absolute Percentage Error (MAPE), which are respectively ranging between 0.3255-2.4024%, 0.2613-2.2683%, and 0.2600-2.2497%. These results are concerning the five models developed for the three subsets of the database.

The sensitivity analysis method (weight method) was used in this thesis, in order to appreciate individually the contribution of the input descriptors on each QSPR-ANN model type developed, the results showed the importance of each used descriptor, in which the majority of descriptors has high or medium contribution compared to the average contribution value (ACV) of each model, which confirms the effectiveness of the methodology chosen in this thesis for the selection of relevant descriptors. In order to know the applicability domain of our developed models, the distance-based method was applied in this thesis with strict standardized residual values (*di**) taken to (±2), the applicability of the models has been quantified by presenting it on a visual graph called the diagram of Williams (Williams plot), the results showed that between 87%– 92% of all substances in the entire data are within the applicability domain of the five models developed, regardless of the standardized residual values taken (±2), which further confirms the effectiveness of the models developed in this study.

A comparison was made between the results obtained by the QSPR-ANN models type and other models of some studies published in scientific articles that have adopted the same approaches (QSPR, ANN, MLR ...). The comparison showed that the QSPR-ANN models type provided better results in terms of: the high regressions coefficients values, the low values of different calculated errors, as well as the regression vector values that are considered close to the ideal, which are also proved very good agreements between the experimental values of the database and the values obtained by the five QSPR-ANN models type developed in this thesis.

This computational approach can be applied in petroleum engineering or other related fields for accurate determination of the properties: *Tb*, *d20*, *Tc*, *Pc* and *Vc* of pure hydrocarbons only by using the information incorporated in their chemical structures.

Keywords: Estimation, Physicochemical properties, Hydrocarbons, Modelling, QSPR-ANN, Descriptors.

الملخص : تقدير بعض الخواص الفيزيوكيميائية للهيدروكربونات بناءً على خصائصها (RQSP) باستخدام طريقة الشبكات العصبونية الاصطناعية (RNA).

الهدف من هذه الأطروحة هو تطوير نماذج رياضية تعتمد على الهياكل الجزيئية للمركبات، من أجل التنبؤ ببعض المعلمات الفيزيوكيميائية للهيدروكربونات النقية، من خلال شبكة عصبونية اصطناعية - ذات مستقبلات متعددة الطبقات (RNA-PMC) بناءً على مفهوم العلاقات الكمية بين الهياكل-والخصائص (RQSP).

النماذج الناتجة عن الجمع بين مفهوم RQSP و RNA تسمى في هذه الأطروحة؛ نماذج من نوع العلاقات الكمية بين الهيكل والخاصية- شبكة عصبونية اصطناعية (RQSP-RNA). تم تطوير خمسة نماذج من نوع RQSP-RNA للتنبؤ بخمسة معلمات فيزيوكيميائية للهيدروكربونات النقية: درجة حرارة نقطة الغليان العادية (*Tb*)، الكثافة النسبية للسائل (*d20*) والخصائص الحرجة {درجة الحرارة الحرجة (*Tc*)، الضغط الحرج (*Pc*) والحجم الحرج (*Vc*)}. تم استخدام قاعدة بيانات تضم 223، 222 و 221 من نقاط البيانات المختلفة لكل من (*Tb*، *Tc* و *Vc*)، (*d20*) و (*Pc*)، على التوالي لإنشاء خمسة نماذج من نوع RQSP-RNA. لكل نموذج ، تم تقسيم قاعدة البيانات الإجمالية بشكل عشوائي إلى مجموعتين فرعيتين: 80٪ لمجموعة التدريب و 20٪ لمجموعة الاختبار. تم حساب 1666 من الواصفات الجزيئية و المستمدة من 20 فئة مختلفة، باستخدام برنامج E-Dragon عبر الإنترنت، ثم تم اعتماد منهجية إحصائية للتخفيض تعتمد على طريقة الانحدار الخطي المتعدد (MLR)، من أجل تقليل واختيار هذه الأعداد الكبيرة من الواصفات الجزيئية، والاحتفاظ بها لعدد مقبول من الواصفات ذات الصلة. بلغت النسبة المئوية للواصفات الجزيئية الأولية المستبعدة أكثر من 98٪ تقريبًا في جميع النماذج الخمسة المطورة. تم تطبيق خوارزمية BFGS (الانتشار الشبه-نيوتني للخلف) لتدريب الشبكة العصبونية الاصطناعية RNA. بحيث كانت بنية أفضل النماذج من نوع RQSP-RNA كالتالي: (16-12-1)، (16-10-1)، (10-16-1)، (17-11-1) و (12-17-1)، على التوالي لنماذج المعلمات: *Tb*، *d20*، *Tc*، *Pc* و *Vc*.

أظهرت نتائج أفضل خمسة نماذج من نوع RQSP-RNA التي تم الحصول عليها، دقة جيدة أكدتها قيم معاملات الانحدار (R) والتحديد (R^2) المرتفعة، والتي تتراوح بين 0.9965-0.9999 و 0.9931-0.9999 على التوالي، للمجموعات الفرعية الثلاثة الأساسية لقاعدة البيانات (مجموعة التدريب، الاختبار والمجموعة الكاملة). بالإضافة إلى القيم الصغيرة لمتوسط خطأ الجذرالتربيعي (REQM) الذي يساوي : 2.0168 درجة كلفن (K)، 0.0060 (-)، 6.9672 درجة كلفن (K)، $4.8231\ 4^{+}10^{4}$ باسكال (Pa) و 10.6349 سم^3/مول (cm^3/mol) ، على التوالي من أجل *Tb*، *d20*، *Tc*، *Pc* و *Vc*. بالإضافة إلى القيم المنخفضة لثلاثة أنواع مختلفة من الأخطاء النسبية، وهي: خطأ التنبؤ المعياري (ESP)، متوسط الخطأ النسبي (ERM) و متوسط الخطأ النسبي المطلق (ERAM)، والتي تتراوح على التوالي بين 0.3255-2.4024٪، 0.2613-2.2683٪ و 0.2600-2.2497٪. وتتعلق هذه النتائج بالنماذج الخمسة التي تم تطويرها للمجموعات الثلاثة الفرعية من قاعدة البيانات.

تم استخدام طريقة تحليل الحساسية (طريقة الوزن) في هذه الأطروحة، من أجل التقييم الفردي لمساهمة واصفات الإدخال على كل نموذج تم تطويره من نوع RQSP-RNA، وقد أظهرت النتائج أهمية كل واصف مستعمل، بحيث أن غالبية الواصفات لها مساهمة كبيرة أو متوسطة مقارنة بقيمة المساهمة المتوسطة (VCM) لكل نموذج، والتي تؤكد فعالية المنهجية المختارة في هذه الأطروحة للاختيار الواصفات ذات الصلة. من أجل معرفة مجال قابلية التطبيق لنماذجنا المطورة، تم تطبيق الطريقة القائمة على المسافة في هذه الأطروحة مع قيم بقايا موحدة صارمة (d_i^*) أخذت ب (±2)، وقد تم قياس قابلية تطبيق النماذج كميا، من خلال تجسيدها على رسم بياني مرئي يسمى الرسم البياني لوليامز (Williams plot)، أظهرت النتائج أن هناك ما بين 87٪ –92٪ من مجموع المواد الكلية في نطاق مجال قابلية التطبيق للنماذج الخمسة التى تم تطويرها، وهذا بغض النظر عن قيم البقايا الموحدة المأخوذة (±2)، مما يؤكد بشكل واضح فعالية النماذج المطورة في هذه الدراسة.

تم إجراء مقارنة بين النتائج المتحصل عليها بواسطة النماذج من نوع RQSP-RNA وغيرها من النماذج لبعض الدراسات التي نشرت في المقالات العلمية والتي إعتمدت نفس المقاربات (RQSP، RNA، RLM ...). أوضحت المقارنة أن النماذج من نوع RQSP-RNA قد قدمت نتائج أفضل من حيث: قيم معاملات الانحدار العالية، القيم المنخفضة لمختلف الأخطاء المحسوبة، وكذا قيم سهم توجيه الانحدار التي تعتبر قريبة من المثالي والتي أثبتت أيضا اتفاق جيد للغاية بين القيم التجريبية لقاعدة البيانات والقيم المحسوبة التي تم الحصول عليها بواسطة نماذج RQSP-RNA الخمسة التي تم تطويرها في هذه الأطروحة.

يمكن تطبيق هذه المقاربة الحسابية في الهندسة البترولية أو غيرها من المجالات الأخرى ذات الصلة من أجل التحديد الدقيق للخصائص: *Tb*، *d20*، *Tc*، *Pc* و *Vc* للهيدروكربونات النقية فقط باستخدام المعلومات المدمجة في هياكلها الكيميائية.

كلمات مفتاحية: تقدير، خواص الفيزيوكيميائية، هيدروكربونات، نمذجة، RQSP-RNA، واصفات.

Table des matières

Résumé

Table des matières *i*
Abréviations *v*
Nomenclature et symboles *viii*
Liste des tableaux *xii*
Liste des figures *xvi*
Liste des travaux et activités de recherches *xx*

Introduction générale 01
Parti I. : Revue bibliographique liée à la thèse

Chapitre I. Génialités sur les hydrocarbures et les fractions pétrolières

I.1. Introduction 06
I.2. Généralités sur le pétrole bruts et ses différents coupes 06
I.2.1. Bref historique du pétrole 06
I.2.2. Description générale du pétrole brut 07
I.2.3. Description générale des coupes pétrolières 08
I.3. Composition chimique des pétroles bruit et des fractions pétrolières 10
I.3.1. Les hydrocarbures 11
I.3.1.1. Les paraffines 13
I.3.1.2. Les naphtènes 13
I.3.1.3. Les oléfines 13
I.3.1.4. Les aromatiques 14
I.3.1.5. Autres hydrocarbures 15
I.3.2. Les hétéroatomes 16
I.3.2.1. Les composés soufrés 16
I.3.2.2. Les composés azotés 18
I.3.2.3. Les composées oxygénés 19
I.3.2.4. Les organométalliques 21
I.3.3. Les composés asphalténiques 21
I.4. Propriétés usuelles des hydrocarbures et des fractions pétrolières 23
I.4.1. Propriétés liées aux fluides 23
I.4.1.1. Masse volumique 23
I.4.1.2. Densité relative 24
I.4.1.3. Gravité de l'API 24
I.4.1.4. Viscosités 24
I.4.2. Propriétés critiques 25
I.4.3. Autres Propriétés physicochimiques 28
I.4.3.1. Poids moléculaire 28

I.4.3.2. Températures d'ébullition 28
I.4.3.3. Pression de vapeur 28
I.4.3.4. Points d'aniline 28
I.4.4. Propriétés thermodynamiques 29
I.4.4.1. Enthalpie standard de formation 29
I.4.4.2. Enthalpie de vaporisation 29
I.4.5. Facteurs et indices spécifiques 29
I.4.5.1. Facteur de caractérisation de Watson 29
I.4.5.2. Facteur acentrique 30
I.4.5.3. Facteur de compressibilité 30
I.4.5.4. Indice de réfraction 31
I.4.5.5. Indice d'octane 31
I.4.6. Propriétés liées à la sécurité et le risque d'explosion 32
I.4.6.1. Points éclaire 32
I.4.6.2. Température d'auto-inflammation 33
I.5. Raffinage du pétrole et traitement des hydrocarbures 33
I.5.1. Procédés de fractionnement 34
I.5.1.1. Distillation atmosphérique du pétrole bruit 34
I.5.1.2. Distillation sous vide du résidu atmosphérique 34
I.5.2. Procédés de transformation et traitement des coupes pétrolières 35
I.6. Aperçu sur les méthodes de la détermination des propriétés des hydrocarbures 40
I.7. Conclusion 41

Chapitre II. Approches de base, méthodologie de développement, validation et application des méthodes RQSP/RQSA

II.1. Introduction 42
II.2. Approche relations quantitatives structures propriétés/activités RQSP/RQSA 42
II.2.1. Historique 42
II.2.2. Domaines d'application et utilité des méthodes RQSP/RQSA 43
II.2.3. Principe et méthodologie de base des études RQSP/RQSA 44
II.3. Codage et représentation des structures chimiques 45
II.3.1. Utilisation de la théorie des graphes pour le codage des structures chimiques 45
II.3.1.1. Représentations d'une structure par quelques éléments basés sur la théorie des graphes chimiques 46
II.3.2. Notification SMILES et codage des structures chimiques 47
II.4. Notion de base sur les descripteurs 48
II.4.1. Définition 48
II.4.2. Classification 48
II.4.2.1. Descripteurs 0-D 49
II.4.2.2. Descripteurs 1-D 49
II.4.2.3. Descripteurs 2-D 49
II.4.2.4. Descripteurs 3-D 50
II.4.2.5. Descripteurs 4-D 51
II.4.3. Généralisation des descripteurs 51
II.5. Méthodes de sélection des descripteurs 52
II.5.1. Sélection objective 53
II.5.2. Sélection subjective 53
II.6. Méthodes d'apprentissage statistiques 54

II.6.1. Régression linéaire multiple 55
II.6.2. Réseaux de neurones artificiels 55
II.6.2.1. Du neurone biologique au neurone artificiel formel 55
II.6.2.2. Classification des réseaux de neurones selon leurs architectures 60
II.6.2.3. Classification des réseaux de neurones selon leurs modes d'apprentissage 61
II.6.2.4. Phénomènes de sous-apprentissage et sur-apprentissage 62
II.6.2.5. Algorithme d'apprentissage 65
II.6.2.6. Réseaux de neurones perceptron multicouche 66
II.6.3. Autres méthodes d'apprentissage 68
II.7. Analyses statistiques, validation et comparaison 70
II.8. Conclusion 71

Chapitre III. Etat de l'art et revue critique de la littérature sur la détermination des propriétés des hydrocarbures

III.1. Introduction sur l'état statistique des références bibliographiques 73
III.2. Revue des travaux de la modélisation des hydrocarbures effectués par la méthode QSPR 76
III.2.1. Travaux réalisés sur la modélisation de la température normale d'ébullition 76
III.2.2. Travaux réalisés sur la modélisation de la densité relative du liquide 80
III.2.3. Travaux réalisés sur la modélisation des trois propriétés critiques 81
III.3. Revue des travaux de la prédiction des cinq propriétés d'intérêts des hydrocarbures réalisés par d'autres méthodologies 84
III.3.1. Travaux réalisés par des méthodes computationnelles 84
III.3.2. Travaux réalisés par d'autres méthodes 86
III.4. Conclusion 93

Parti II. : Travail pratique de la thèse

Chapitre IV. Matériels et méthodes

IV.1. Introduction 94
IV.2. Outils informatique utilisés 94
IV.3. Méthodologie générale de la modélisation RQSP utilisée 97
IV.3.1. Collection et traitement de la base des données 97
IV.3.2. Codage SMILES des structures chimiques des hydrocarbures 98
IV.3.3. Calcul des descripteurs moléculaires 99
IV.3.4. Procédure de prétraitement pour la sélection des descripteurs pertinents 101
IV.3.5. Sélection des paramètres de choix pour l'apprentissage des modèles de RQSP-RNA 102
IV.3.6. Estimation de la performance des meilleurs modèles de RQSP-RNA développées 104
IV.3.6.1. Analyses statistiques et calcules des erreurs 104
IV.3.6.2. Mesure de la contribution des descripteurs de chaque modèle de RQSP-RNA développé 105

IV.3.6.3. Détermination du domaine d'applicabilité des modèles de RQSP-RNA 109
IV.4. Conclusion 110

Chapitre V. Résultats et discutions

V.1. Introduction 112
V.2. Évaluation du nombre de descripteurs au cours de la procédure de sélection 112
V.3. Caractéristiques des descripteurs pertinent sélectionnés 116
V.4. Architectures des meilleurs modèles RQSP-RNA obtenus 116
V.5. Application des expressions mathématiques des modèles RQSP-RNA obtenus 118
V.6. Prédiction des propriétés physicochimiques des hydrocarbures par les modèles RQSP-RNA 126
V.7. Estimation de la performance des résultats d'estimation issue des modèles RQSP-RNA développés 126
V.7.1. Comparaison graphique entre les valeurs expérimentales et les valeurs calculées à partir des modèles RQSP-RNA 126
V.7.2. Calculs des paramètres de performance et analyse statistique des modèles de type RQSP-RNA obtenus 132
V.7.3. Analyse de sensibilité des descripteurs dans chaque modèle RQSP-RNA par la méthode de poids 134
V.7.3.1. Interprétation et classification des contributions selon la (VMC) et selon chaque groupe 137
V.7.4. Domaine d'applicabilité pour la validation des modèles RQSP-RNA obtenus 143
V.7.5. Comparaisons des performances des modèles de type RQSP-RNA obtenus avec d'autres modèles dans la littérature 148
V.8. Conclusion 162

Conclusion générale et perspectives 163

Références bibliographiques 167

Annexes 183
Annexe A. 183
Annexe B. 192
Annexe C. 202
Annexe D. 207
Annexe E. 212

Article et attestations des participations aux autres travaux de recherche 249
Attestation de publication et support d'un article scientifique classe "A" 249
Attestations, résumes et poster des conférences internationales avec comité de lecture 262
Attestations, résumes, posters et support d'article des conférences et séminaires nationales avec comité de lecture 270
Attestation de comité d'organisation d'un séminaire nationale avec comité de lecture 285

Abréviations

3DMoRSE	***3D-Mo**lecule **R**epresentation of **S**tructures based on **E**lectron diffraction*
AA	***A**lgorithme d'**A**pprentissage*
AAD	***A**verage **A**bsolute **D**eviation*
AAE	***A**verage **A**bsolute **E**rror*
AARD	***A**verage **A**bsolute **R**elative **D**eviation*
AD	***A**bsolute **D**eviation*
ADME	***A**bsorption **D**istribution **M**étabolisme **É**limination*
AG	***A**lgorithme **G**énétique*
AIT	***A**uto-**I**gnition **T**emperature*
ANN	***A**rtificial **N**eural **N**etwork*
API	***A**merican **P**etroleum **I**nstitute*
ARTMAP	***A**daptive **R**esonance **T**heory **MAP***
ARTN	***A**daptive **R**esonance **T**heory **N**etworks*
ASTM	***A**merican **S**ociety for **T**esting and **M**aterials*
BFGS	***B**royden-**F**letcher-**G**oldfarb-**S**hanno algorithm*
BMDRC	***B**ioinformatics and **M**olecular **D**esign **R**esearch **C**enter*
BP	***B**oiling **P**oint*
bpNN	***b**ack-**p**ropagation **N**eural **N**etwork*
C	***C**arbone*
CG	***C**ontribution des **G**roupes*
CIG	***C**ontribution à **I**nteraction des **G**roupes*
CN	***C**ompetitive **N**etworks*
CNN	***C**ompetitive **N**eural **N**etwork*
CoMFA	***Co**mparative **M**olecular **F**ield **A**nalysis*
CoMSIA	***Co**mparative **M**olecular **S**imilarity **I**ndices **A**nalysis*
CPG	***C**hromatographie en **P**hase **G**azeuse*
CPL	***C**hromatographie en **P**hase **L**iquide*
DA	***D**omaine d'**A**pplicabilité*
DAO	***D**e**A**sphalted **O**il*
DIPPR	***D**esign **I**nstitute for **P**hysical **PR**operty Data*
DS	***D**éviation **S**tandard (Écart type)*
DSR	***D**éviation **S**tandard **R**elative (Écart type relatif)*
DSV	***D**istillats **S**ous **V**ide*
E-Dragon	***E**lectronic **Dragon***
EQM	***E**rreur **Q**uadratique **M**oyenne*
ERAM	***E**rreur **R**elative **A**bsolu **M**oyenne*
ERM	***E**rreur **R**elative **M**oyenne*
ESP	***E**rreur **S**tandard de **P**rédiction*
FBR	***F**onction de **B**ase **R**adiale*
FF-ANN	***F**eed **F**orward-**A**rtificial **N**eural **N**etwork*

FP	***F**raction **P**étrolière*
GA	***G**enetic **A**lgorithm*
GC	***G**roup **C**ontribution method*
GETAWAY	***GE**ometry, **T**opology, and **A**tom-**W**eights **A**ssembl**Y***
GIC	***G**roup **I**nteraction **C**ontribution method*
GMDH	***G**roup **M**ethod of **D**ata **H**andling*
GNL	***G**az **N**aturel **L**iquéfié*
GPL	***G**az de **P**étrole **L**iquéfié*
GVS	***G**roup **V**ector **S**pace*
H	***H**ydrogène*
HC	***H**ydro**C**arbures*
HOMO	***H**ighest **O**ccupied **M**olecular **O**rbital*
IAM	***I**nformatics and **A**pplied **M**athematics*
IR	***I**mportance **R**elative*
Ir	***I**nfra **r**ouge*
ISO	***I**nternational **S**tandard **O**rganization*
KSOM	***K**ohonen's **S**elf **O**rganising **M**ap*
LCO	***L**ight **C**ycle **O**il*
LFER	***L**inear **F**ree **E**nergy **R**elationships*
LM	***L**evenberg–**M**arquardt*
LSSVM	***L**east **S**quares **S**upport **V**ector **M**achine*
LUMO	***L**owest **U**noccupied **M**olecular **O**rbital*
MAPE	***M**ean **A**bsolute **P**ercentage **E**rror*
MD	***M**ean **D**eviation*
MLP	***M**ulti **L**ayer **P**erceptron*
MLR	***M**ultiple **L**inear **R**egression*
MON	***M**otor **O**ctane **N**umber*
MPE	***M**ean **P**ercentage **E**rror*
MS	***M**ass **S**pectrometry*
MTBE	***M**éthyle **T**ertio**B**utyle **É**ther*
MVS	***M**achines à **V**ecteurs de **S**upports*
MW	***M**olecular **W**eight*
NO	***N**ombre d'**O**ctane*
PCLIENT	***P**arameter **Client***
PLS	***P**artial **L**east **S**quares*
PMC	***P**erceptron **M**ulti **C**ouche*
ppm	***p**arti **p**ar **m**illion*
PSC	***P**erceptron d'un **S**eul **C**ouche*
PVT	***P**ression- **V**olume- **T**empérature*
QSAR	***Q**uantitative **S**tructure **A**ctivity **R**elationship*
QSPR	***Q**uantitative **S**tructure **P**roperty **R**elationship*
RA	***R**ésidu **A**tmosphérique*

RBFN	***R**adial **B**asis **F**unction **N**etwork*
RBF-NN	***R**adial **B**asis **F**unction-**N**eural **N**etwork*
RDF	***R**adial **D**istribution **F**unction*
REQM	***R**acine d'**E**rreur **Q**uadratique **M**oyenne*
RFBR	***R**éseau à **F**onctions de **B**ase **R**adiale*
RLM	***R**égression **L**inéaire **M**ultiple*
RMN	***R**ésonance **M**agnétique **N**ucléaire*
RMSE	***R**oot **M**ean **S**quared **E**rror*
RN	***R**éseau de **N**eurone*
RNA	***R**éseau de **N**eurone **A**rtificiel*
RNAs	***R**éseaux de **N**eurones **A**rtificiel**s***
RN-FBR	***R**éseaux de **N**eurones à **F**onction de **B**ase **R**adiale*
RNFF	***R**éseau de **N**eurones de type **F**eed **F**orward*
RON	***R**esearch **O**ctane **N**umber*
RQSA	***R**elation **Q**uantitative **S**tructure **A**ctivité*
RQSP	***R**elation **Q**uantitative **S**tructure **P**ropriété*
RSD	***R**elative **S**tandard **D**eviation*
RSV	***R**ésidu **S**ous **V**ide*
RVS	***R**égression à **V**ecteur de **S**upport*
SARA	***S**aturés- **A**romatiques- **R**ésines- **A**sphaltènes*
SDC	***S**omme **D**es **C**arrés*
SEP	***S**tandard **E**rror of **P**rediction*
SG	***S**pecific **G**ravity*
SI	***S**ystème **I**nternational*
SIGMA	***SIG**naux, **M**odèles, **A**pprentissage*
SLP	***S**ingle **L**ayer **P**erceptron*
SMILES	***S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem*
SOS	***S**um **O**f **S**quares*
SR	***S**traight **R**un*
SVM	***S**upport **V**ector **M**achine*
SVR	***S**upport **V**ector **R**egression*
TAME	***T**ertio-**A**myle **M**éthyle **É**ther*
TRC	***T**hermodynamic **R**esearch **C**enter*
UNIFAC	***UNI**versal **F**unctional **A**ctivity **C**oefficient*
UNIQUAC	***UNI**versal **QUA**si **C**hemical*
UV	***U**ltra-**V**iolet*
VCCLAB	***V**irtual **C**omputational **C**hemistry **Lab**oratory*
VCM	***V**aleur de **C**ontribution **M**oyen*
WHIM	***W**eighted **H**olistic **I**nvariant **M**olecular*
WMSE	***W**eighted **M**ean **S**quare **E**rror*

Nomenclatures et symboles

Lecture	Désignation	Unité
P_r^{vap}	*Pression de vapeur réduite*	**(-)**
$SpGr_{60F/60F}$	*Spécifique Gravité 60°F sur 60°F*	**(-)**
a_i	*Potentiel du neurone*	**(-)**
b_j	*Biais d'un neurone artificiel (j)*	**(-)**
d_4^T	*Densité relative ou Gravité spécifique*	**(-)**
w_{ij}	*Poids synaptique des connexions inter-neuronales (ij)*	**(-)**
$\Delta H°_f$	*Enthalpie standard de formation*	**J/mole**
ΔH_{vap}	*Enthalpie de vaporisation*	**J/mole**
0-D	*0 Dimension*	**(-)**
1-D	*1 Dimension*	**(-)**
2-D	*2 Dimension*	**(-)**
3-D	*3 Dimension*	**(-)**
4-D	*4 Dimension*	**(-)**
Alg	*Algorithme*	**(-)**
d20	*Densité relative de liquide*	**(-)**
d_i	*Résidus standardisés d'un éliment (i) (en anglais : standardized residuals of (i) element*)	**(-)**
DM	Moment dipolaire électrostatique	**C.m**
E_i	*Neurones d'entrée (i)*	**(-)**
Eq	*Équation*	**(-)**
Eqs	*Équations*	**(-)**
E_{tot}	Energie totale	**J**
F ou f	*Fonction : Objectif, Object, Coût, Fitness …*	**(-)**
G	*Graphe moléculaire*	**(-)**
h	*Levier (leverage)*	**(-)**
h*	*Levier standard (standard leverage)*	**(-)**

I ou ***η***	*Indice de réfraction*	**(-)**
Iter	*Nombre d'itération actuel*	**(-)**
$\boldsymbol{Iter_{max}}$	*Nombre d'itération maximal*	**(-)**
$\boldsymbol{K_W}$ ou ***Kuop***	*Facteur de caractérisation de Watson*	**(-)**
logP	*Coefficient de partage (Octanol/Eau)*	**(-)**
M	*Poids moléculaire*	**Kg/Kmol**
N *ou* ***n***	*Nombre des points expérimentaux ou composés dans l'ensemble des données*	**(-)**
Nb	*Nombre*	**(-)**
Pc	*Pression critique*	**Pa**, **bar** ou **atm**
$\boldsymbol{P^{vap}}$	*Pression de vapeur*	**Pa**, **bar** ou **atm**
R	*Coefficient de corrélation ou de régression*	**(-)**
$\boldsymbol{R^2}$	*Coefficient de détermination*	**(-)**
S	*Sortie de chaque modèle*	**(-)**
$\boldsymbol{S_t}$	Surface totale de Van der Waals	**Å²**
Std	*Standard*	**(-)**
t	*Temps*	***(S)***
T *ou* ***T***°	*Température*	**K** ou °***C***
Tb	*Température du point d'ébullition normal*	**K** ou °***C***
Tc	*Température critique*	**K** ou °***C***
Valeur-P	*Valeur de probabilité (en anglais : P-value Probability value)*	**(-)**
Valeur-t	*Valeur de Student (en anglais : t-value Student's t-test)*	**(-)**
Vc	*Volume critique*	$\boldsymbol{Cm^3/mol}$
W	*Indice de Wiener*	**(-)**
X	*Entrée de réseau (Variables d'entrée)*	**(-)**
Y	*Sortie de réseau*	**(-)**
Z	*Facteur de compressibilité*	**(-)**
$\boldsymbol{Z_j}$	$j^{ème}$ *sortie de la couche cachée*	**(-)**
$\boldsymbol{A(G)}$	*Matrice adjacence des sommets d'un graphe moléculaire G*	**(-)**
$\boldsymbol{D(G)}$	*Matrice de distance d'un graphe moléculaire G*	**(-)**

J	*Indice de Balaban*	**(-)**
$V(G)$	*Vecteur des valences d'un graphe moléculaire G*	**(-)**
Z	*Indice de Hosoya*	**(-)**

Lettres Grecques

Lecture	Désignation	Unité
α	*Pente de la droite du meilleur ajustement linéaire*	**(-)**
β	*Interception ou Ordonnée à l'origine de la droite du meilleur ajustement linéaire*	**(-)**
η	*Indice de réfraction*	**(-)**
μ	*Viscosité dynamique*	**Pa.S**
v	*Viscosité cinématique*	**m^2/s**
ρ	*Masse volumique*	**Kg/m^3** ou **lb/ft^3**
ρ_{20}	*Masse volumique à la température de référence à 20°C*	**Kg/m^3** ou **lb/ft^3**
ω	*Facteur acentrique*	**(-)**
χ	*Indice de Randic*	**(-)**

Exposants

Lecture	Désignation
App	*Apprentissage*
cal	*Calculée*
Ent	*Entière*
exp	*Expérimentale*
H	*Cachée (Hidden)*
I	*Entrée (Input)*
O	*Sortie (Output)*
T	*Transposé*
Test	*Test*
Vap	*Vapeur*

Indices

Lecture	Désignation
adj	*Ajusté*
Avr	*Average*
i	*Elément **i** d'un vecteur*
ij	*Croisement de la ligne **i** avec la colonne **j** d'une matrice*
j	*Elément ou dimension **j** d'un vecteur*
max	*Maximum*
min	*Maximum*
moy	*Moyen*
r	*Réduite*

Liste des tableaux

Chapitre I. Généralités sur les hydrocarbures et les fractions pétrolières		
Tableau I.1	Composition élémentaire des principaux bruts et leurs densités à 15°C	08
Tableau I.2	Classification des hydrocarbures en fonction de leurs familles	12
Tableau I.3	Structures de quelques hydrocarbures aromatiques	14
Tableau I.4	Exemples de quelques structures d'autres hydrocarbures insaturés	16
Tableau I.5	Exemples de quelques structures de composés soufrés	17
Tableau I.6	Exemples de quelques structures de composés azotés	19
Tableau I.7	Exemples de quelques structures de composés oxygénés	20
Tableau I.8	Principaux procédés de transformation appliqués aux coupes pétrolières légères ou moyennes	36
Tableau I.9	Principaux procédés de conversion appliqués aux coupes pétrolières lourdes	38
Tableau I.10	Principaux procédés de traitement appliqués aux produits pétroliers finis	39

Chapitre II. Approches de base, méthodologie de développement, validation et application des méthodes RQSP/RQSA		
Tableau II.1	Notions correspondant à la théorie des graphes et la théorie des graphes chimique	46
Tableau II.2	Exemple de calcul des matrices et vecteur basé sur la théorie des graphes chimique corresponds à la structure du *2.3.4-triméthylhexane*	47
Tableau II.3	Fonctions de transfert (d'activation) usuelles dans logiciel *Statistica*	58
Tableau II.4	Fonctions de transfert usuelles dans le logiciel *Matlab*	59

Chapitre IV. Matériels et méthodes		
Tableau IV.1	Nombres de descripteurs initialement calculés par chaque classe pour tous les paramètres étudiés	100
Tableau IV.2	Exemple illustrant la méthode de la sensibilité basée sur la partition des poids de connexion (méthode de Poids)	107

Chapitre V. Résultats et discussions		
Tableau V.1	Évolution du choix du nombre de descripteurs selon la procédure de prétraitement par classe pour chaque paramètre étudié	115
Tableau V.2	Architectures des meilleurs modèles RQSP-RNA obtenus pour chaque paramètre	117
Tableau V.3	Fonctions d'activation apparues dans cette étude	118
Tableau V.4	Poids et biais du meilleur modèle RQSP-RNA pour *Tb* (16-12-1)	121
Tableau V.5	Poids et biais du meilleur modèle RQSP-RNA pour *d20* (16-10-1)	122
Tableau V.6	Poids et biais du meilleur modèle RQSP-RNA pour *Tc* (10-16-1)	123
Tableau V.7	Poids et biais du meilleur modèle RQSP-RNA pour *Pc* (17-11-1)	124
Tableau V.8	Poids et biais du meilleur modèle RQSP-RNA pour *Vc* (12-17-1)	125
Tableau V.9	Paramètres statistiques des meilleurs modèles RQSP-RNA pour les cinq propriétés	132
Tableau V.10	Comparaison entre le modèle de RQSP-RNA du paramètre *Tb* développé dans cette étude et d'autres modèles présents dans la littérature	157
Tableau V.11	Comparaison entre le modèle de RQSP-RNA du paramètre *d20* développé dans cette étude et d'autres modèles présents dans la littérature	158
Tableau V.12	Comparaison entre le modèle de RQSP-RNA du paramètre *Tc* développé dans cette étude et d'autres modèles présents dans la littérature	159

Tableau V.13	Comparaison entre le modèle de RQSP-RNA du paramètre *Pc* développé dans cette étude et d'autres modèles présents dans la littérature	160
Tableau V.14	Comparaison entre le modèle de RQSP-RNA du paramètre *Vc* proposé dans cette étude et d'autres modèles présents dans la littérature	161

Annexes

	Annexe A.	
Tableau A.1	Valeurs expérimentales et prédites par les modèles RQSP-RNA des cinq propriétés physicochimiques des hydrocarbures ainsi leurs ordres de partition dans les bases d'apprentissage et de test pour chaque paramètre	183
	Annexe B.	
Tableau B.1	Écriture *SMILES* des molécules des hydrocarbures utilisés dans cette étude	192
	Annexe C.	
Tableau C.1	Caractéristiques des descripteurs pertinentes sélectionnés pour *Tb*	202
Tableau C.2	Caractéristiques des descripteurs pertinentes sélectionnés pour *d20*	203
Tableau C.3	Caractéristiques des descripteurs pertinentes sélectionnés pour *Tc*	204
Tableau C.4	Caractéristiques des descripteurs pertinentes sélectionnés pour *Pc*	205
Tableau C.5	Caractéristiques des descripteurs pertinentes sélectionnés pour *Vc*	206
	Annexe D.	
Tableau D.1	Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Tb*	207
Tableau D.2	Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *d20*	208

Tableau D.3	Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Tc*	209
Tableau D.4	Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Pc*	210
Tableau D.5	Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Vc*	211
Annexe E.		
Tableau E.1	Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Tb ainsi leurs ordres de partition dans les bases d'apprentissage et de test	212
Tableau E.2	Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour d20 ainsi leurs ordres de partition dans les bases d'apprentissage et de test	221
Tableau E.3	Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Tc ainsi leurs ordres de partition dans les bases d'apprentissage et de test	229
Tableau E.4	Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Pc ainsi leurs ordres de partition dans les bases d'apprentissage et de test	235
Tableau E.5	Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Vc ainsi leurs ordres de partition dans les bases d'apprentissage et de test	243

Liste des figures

Chapitre I. Généralités sur les hydrocarbures et les fractions pétrolières		
Figure I.1	Classification des principales coupes pétrolières selon les intervalles de température d'ébullition et le nombre d'atomes de carbone équivalent des paraffines	09
Figure I.2	Structure des rendements en coupes pétrolières de quelques pétroles bruts	10
Figure I.3	Représentation de la structure du motif *Porphyrine* ou *Porphine* (**$C_{20}H_{14}N_4$**)	21
Figure I.4	Représentation des deux molécules asphalténiques possibles : **A)** Type ***Continental***, **B)** Type ***Archipel***	22
Figure I.5	Diagramme de coordonnées (P-T) des trois équilibres d'un corps pur	26
Figure I.6	Représentation de la chaleur de vaporisation pour un mélange de cops purs	26
Figure I.7	Indices d'octane des différentes familles chimiques d'hydrocarbures	32

Chapitre II. Approches de base, méthodologie de développement, validation et application des méthodes RQSP/RQSA		
Figure II.1	Principe et méthodologie de base des modèles RQSP/RQSA	44
Figure II.2	Graphe moléculaire (***G***) correspondant à la structure du *2.3.4-triméthylhexane*	47
Figure II.3	Conception générale de neurones biologiques et artificiels	56
Figure II.4	Schématisation distinctive d'un neurone artificiel formel	57
Figure II.5	Classification des réseaux de neurones artificiels selon les architectures d'interconnexions entre neurones	61
Figure II.6	Phénomènes d'apprentissage optimale, sous-apprentissage et sur-apprentissage	64
Figure II.7	Structure d'un perceptron multi couches non-bouclé	67

Chapitre III. Etat de l'art et revue critique de la littérature sur la détermination des propriétés des hydrocarbures		
Figure III.1	Nombre de publications pour : (■) les études RQSP, (■) les études de RNA et (■) la détermination des propriétés des hydrocarbures par différentes méthodes publiées dans la plateforme *Sciencedirect*	74
Figure III.2	Nombre de publications traitant la détermination des propriétés des hydrocarbures par les méthodes : (■) RNA, (■) RQSP et (■) RNA-RQSP publiées dans la plateforme *Sciencedirect*	75

Chapitre IV. Matériels et méthodes		
Figure IV.1	Capture d'écran des logiciels utilisés dans cette étude : ChemDraw Office, MATLAB, STATISTICA, Boîte à outils de DA, Microsoft Office Excel, et E-Dragon 1.0	96
Figure IV.2	Organigramme de la méthodologie suivie dans cette étude	97
Figure IV.3	Capture d'écran des notations SMILES des molécules organisées dans des fichiers de format ".txt"	99
Figure IV.4	Représentation graphique des pourcentages de division de la base de données	103
Figure IV.5	Capture d'écran pour le choix des fonctions d'activation sur STATISTICA	103
Figure IV.6	Organigramme illustrant l'analyse de sensibilité par la méthode de la partition des poids (méthode de poids)	108

Chapitre V. Résultats et discussions		
Figure V.1	Représentation schématique d'un réseau du modèle de type RQSP-RNA	118
Figure V.2	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la température du point d'ébullition normal Tb pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test	127
Figure V.3	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la densité relative liquide d20 pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test	128
Figure V.4	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la température critique Tc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test	129
Figure V.5	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la pression critique Pc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test	130
Figure V.6	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA du volume critique Vc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test	131
Figure V.7	Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la température du point d'ébullition normal Tb	135
Figure V.8	Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la densité relative du liquide d20	135
Figure V.9	Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la température critique Tc	136
Figure V.10	Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la pression critique Pc	136
Figure V.11	Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA du volume critique Vc	137
Figure V.12	Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la température du point d'ébullition normal Tb	139

Figure V.13	Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la densité relative du liquide d20	140
Figure V.14	Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la température critique Tc	141
Figure V.15	Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la pression critique Pc	142
Figure V.16	Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA du volume critique Vc	143
Figure V.17	Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la température du point d'ébullition normal Tb	144
Figure V.18	Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la densité relative de liquide d20	145
Figure V.19	Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la température critique Tc	146
Figure V.20	Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la pression critique Pc	147
Figure V.21	Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA du volume critique Vc	148

Liste des travaux et activités de recherches

Publications dans des revues internationales avec comité de lecture

1. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120. ISSN: 1093-3263. Lien : https://www.sciencedirect.com/science/article/pii/S1093326318306673.

Communications internationales avec comité de lecture

2. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR-ANN Estimation Models for Critical Pressure and Critical Volume of Pure Hydrocarbons Using MLR and MLP-ANN Methods. *The 8th International Symposium on Hydrocarbons and Chemistry (ISHC8) 8- 10 April 2019*. Faculté des Hydrocarbures et de la Chimie, Université M'hamed Bougara (UMBB), Boumerdès, Algérie. Participation par Présentation **Orale**, *Paper ID Number : PPE-47, 08, 42* (*ISHC8_BOOK OF ABSTRACTS)*.

1. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR-ANN And QSPR-MLR Estimation Models for Normal Boiling Point Temperature and Critical Temperature of Pure Hydrocarbons and Petroleum Fractions. *The 8th International Symposium on Hydrocarbons and Chemistry (ISHC8) 8- 10 April 2019*. Faculté des Hydrocarbures et de la Chimie, Université M'hamed Bougara (UMBB), Boumerdès, Algérie. Participation par Présentation **Poster**, *Paper ID Number : PPE-47, 08, 58*-59 (*ISHC8_BOOK OF ABSTRACTS)*.

Communications nationales avec comité de lecture

3. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). Comparative study between QSPR-MLP, QSPR-SVR and QSPR-MLR models for estimation of critical temperature of pure hydrocarbons and petroleum fractions. *2nd Conference on Informatics and Applied Mathematics, (IAM'2019)*, on 12-13 June 2019, Université 08 Mai 1945 Guelma, Guelma, Algérie. Participation par Présentation **Poster**, *Paper-ID : 110*.

2. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR based models for estimation of normal boiling point temperature of pure hydrocarbons and petroleum fractions using RBF-ANN, MLP-ANN and MLR Methods. *2nd Conference on Informatics and Applied Mathematics, (IAM'2019)*, on 12-13 June 2019, Université 08 Mai 1945 Guelma, Guelma, Algérie. Participation par Présentation **Poster**, *Paper-ID : 109*.

1. Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2018). Robust QSPR-ANN Estimation Model for Normal Boiling Point Temperature of Pure Hydrocarbons and Petroleum Fractions. *1er Séminaire National de Physique et Chimie des Matériaux (SNPCM1'2018)* - Novembre, 28-29, 2018 ENSET Skikda -Algérie. Participation par Présentation **Orale**, *Paper-ID : 73*, 142-148 (PROCEEDINGS de *SNPCM1'2018*).

Projet de recherche

1. Membre du projet **PRFU** intitulé "Modélisation des propriétés physiques et thermodynamiques de fluides et Simulation du transport des fluides par les algorithmes génétiques et les réseaux neurones artificiels" (MESRS réf. **A16N01UN240120200003** à partir du **01/01/2020**), Université 08 Mai 1945 Guelma (**Chargé de recherche**).

Introduction générale

Le pétrole est l'une des sources énergétiques non renouvelable, qui a connu une augmentation remarquable de la demande mondiale au cours des cinquante dernières années **(Apergis & Payne 2012 ; Pao & Fu 2013 ; Amri 2017 ; Dudley 2018)**, à cause de son importance en tant que matière première utilisée dans de nombreuses industries chimiques. Il est considéré comme une pierre angulaire de plusieurs produits fabriqués, tels que les lubrifiants, les polymères, les plastiques, les textiles, les engrais et même les médicaments **(Guthrie & Guthrie 1960 ; Speight 2015)**. Le pétrole a toujours été le principal carburant concurrent sur le marché énergétique **(Rebas 2006 ; Belghit 2019)**, il reste la ressource la plus consommée au monde avec un taux moyen égal à (34%) avant le charbon, ce qui équivaut à (30%), (24%) par rapport au gaz, et (12%) relativement à la consommation d'électricité **(BP 2012)**.

L'industrie pétrolière est l'un des principaux secteurs qui ont contribué au développement de la richesse de nombreux pays à travers le monde **(Masseron 1990 ; Dudley 2018)**, y compris la nôtre où il représente la plus grande part de notre revenu national **(Amri 2017 ; Dudley 2018)**. Les hydrocarbures et les fractions pétrolières sont considérés comme l'un des matériaux énergétiques qui sont souvent traités dans ce domaine. Le pétrole piégé naturellement dans les roches réservoirs est produit de diverses manières selon la profondeur, les caractéristiques du pétrole et la pression du fluide dans le bassin **(Wauquier 1998 ; Arab *et al*. 2015)**. Le pétrole brut et les fractions pétrolières consistent en un mélange de molécules et de structures complexes, dans lesquelles les principaux constituants sont les hydrocarbures **(Wauquier 1998)**.

La connaissance des propriétés physicochimiques et thermodynamiques joue un rôle fondamental dans l'ingénierie et l'industrie pétrochimique et d'autres secteurs liés. Car ils peuvent être considérés comme l'un des paramètres clés impliqués dans presque toutes les opérations unitaires et les procédés de fabrication, de transformation, de traitements et même de stockage des matières énergétiques, tels que les hydrocarbures et les fractions pétrolières **(Chitour 1999 ; Riazi 2005 ; Fahim *et al*. 2009 ; Rebas 2016)**. Ils peuvent également expliquer de nombreux phénomènes qui peuvent aider aussi bien les industriels que les scientifiques, afin de se faire une idée sur la nature de leurs matériaux, en vue de trouver et de mettre en œuvre des méthodes ciblées pour sa transformation et son transfert, à moindre coût, avec des qualités et sécurités supérieures.

Le calcul des propriétés du mélange pétrolier dépend des propriétés de ses composants purs. Certaines propriétés des mélanges peuvent être facilement mesurées. En revanche d'autres caractéristiques, telles que les propriétés critiques, ne peuvent pas être mesurées facilement

(Constantinou & Gani 1994 ; Riazi 2005 ; Roubehie fissa *et al*. 2019). Cependant, malgré les progrès réalisés dans les études observées dans ce secteur, les données sur ces propriétés ne sont pas toujours disponibles à hauteur du besoin pour plusieurs raisons ; dans la majeure partie des cas, les tests au laboratoire longs, coûteux et fastidieux sont les principales causes **(Roubehie fissa *et al*. 2019)**. Dans cette étude nous avons essayé de remédier à ces problèmes en proposant de modèles mathématiques pour l'estimation de ces propriétés.

La modélisation moléculaire est une technique qui permet non seulement de représenter les structures en deux ou trois dimensions, mais aussi d'aider tout chimiste théorique, de mener des études approfondies et précises des systèmes moléculaires et de les exprimer explicitement avec des modèles mathématiques **(Jensen 2007 ; Engel & Gasteiger 2018)**. La modélisation actuelle basée sur les méthodes computationnelles d'intelligence artificielle est largement utilisée dans de nombreux domaines, en particulier dans le secteur pétrolier, afin d'élaborer des modèles fiables capables de prédire différentes propriétés et activités liées aux phénomènes physiques, chimiques et biologiques **(Minsky 1961 ; Duch *et al*. 2007 ; Ghahramani 2015 ; Roubehie fissa *et al*. 2019)**. L'une de ces techniques est la modélisation de Relation Quantitative Structure Propriété/Activités (RQSP/RQSA) {*en anglais* : Quantitative Structure Property/Activity Relationships (QSPR/QSAR)}, qui permet de prédire les propriétés et/ou les activités des systèmes chimiques et/ou biologiques en fonction de leurs structures moléculaires **(Charton 2002)**. Les relations entre les structures moléculaires, leurs propriétés et/ou leurs activités sont généralement établies à l'aide des méthodes de la modélisation d'apprentissage statistiques, telles que les Réseaux de Neurone Artificiels (RNAs) {*en anglais* : Artificial Neural Networks (ANN)} **(Hamadache 2014 ; Roubehie fissa *et al*. 2019)**.

L'objectif principal de cette thèse est de développer des modèles mathématiques à partir des informations structurelles incorporées dans les molécules des hydrocarbures purs, basées sur des méthodes computationnelles d'intelligence artificielle afin de prédire cinq propriétés physicochimiques : la température du point d'ébullition normale (Tb), la densité relative du liquide (d20) et les propriétés critiques ; la température critique (Tc), la pression critique (Pc), et le volume critique (Vc). Par conséquent, cinq modèles ont été développés, utilisant l'approche de Relation Quantitative Structure Propriété (RQSP) et la méthode de Réseau de Neurones Artificiels-Perceptron Multi Couches (RNA-PMC) ; les modèles résultant de la combinaison de ces deux techniques sont appelés dans cette thèse : modèles de RQSP-RNA.

Plusieurs étapes ont été utilisées lors de l'élaboration des modèles RQSP-RNA, les principales sont : (1) la collecte et le traitement de la base de données, (2) le calcul et la sélection des descripteurs moléculaires, et (3) l'élaboration et la sélection des meilleurs modèles RQSP-ANN. L'efficacité et la précision de ces modèles ont été testées à l'aide de plusieurs paramètres statistiques tels que : le coefficient de détermination (R^2), la Racine d'Erreur Quadratique Moyenne (REQM) et l'Erreur Relative Moyenne (ERM). La méthode d'analyse de la sensibilité au poids a été utilisée afin d'évaluer l'impact de la contribution de chaque descripteur d'entrée sur chaque modèle RQSP-RNA. Les modèles ont également été validés à l'aide de la méthode de Domaine d'Applicabilité (DA). Les résultats ont montré que les modèles de type RQSP-RNA prédisent les paramètres Tb, d20, Tc, Pc et Vc des hydrocarbures avec une précision appropriée par rapport aux autres modèles de la littérature.

La présente thèse s'articule principalement autour de cinq (05) chapitres et une conclusion générale :

- **Le premier chapitre :** comprend une étude bibliographique sur le pétrole brut, ses fractions et les hydrocarbures purs, où j'ai parlé de généralités et de notions de base liées au sujet, telles que : la définition, les origines et la composition du pétrole, les critères de classification du pétrole brut, les caractéristiques des hydrocarbures, les domaines d'utilisation, la présentation des paramètres essentiels dans l'industrie pétrolière et les méthodes classiques de détermination de ces paramètres.
- **Le second chapitre :** est consacré à l'étude des méthodes de modélisation, en particulier le concept des Relations Quantitative Structure Propriété/Activités (RQSP/RQSA), et aux techniques de modélisation basées sur les apprentissages statistiques, telles que les Réseaux de Neurone Artificielles (RNAs) comme étant une technique d'intelligence artificielle. J'ai donné un aperçu général sur ces deux approches de modélisation, des différents types possibles, des éléments nécessaires au développement de leurs modèles et de leurs applications, de la notion des descripteurs moléculaires et les outils de calculs, des étapes de sélection et de traitement. Enfin j'ai parlé de façon générale, de quelques méthodes statistiques d'analyse et de validation de la modélisation.
- **Le troisième chapitre :** contient une analyse de certaines études pertinentes trouvées dans la littérature. J'ai donné brièvement une explication sur l'état de l'art et revue critique de la littérature. Ce chapitre est consacré aux travaux réalisés sur la modélisation des paramètres physicochimiques et thermodynamiques des hydrocarbures,

particulièrement sur la modélisation utilisant les méthodes RQSP/RQSA et sur les travaux ayant adopté la méthode de RNA ou d'autres méthodes similaires.

- **Le quatrième chapitre :** est dédié à l'explication des matériels et des méthodes utilisées pour la réalisation de ce travail. Je me suis appuyé particulièrement sur les outils de développement des modèles de type RQSP-RNA pour prédire les cinq paramètres physicochimiques des hydrocarbures purs, tels que la description quantitative et qualitative de la base de données, l'ensemble des étapes utilisées lors de l'extraction des informations structurales du calcul et de prétraitement des descripteurs moléculaires, la sélection des paramètres utilisés dans le développement, et enfin les critères d'évaluation et d'analyse de ces modèles. J'ai également expliqué la méthode d'analyse de sensibilité, plus précisément la méthode de poids pour évaluer l'effet des descripteurs d'entrée sur les sorties de chaque modèle. Parallèlement, j'ai déterminé le domaine de chaque modèle, par la méthode du domaine d'applicabilité afin de confirmer la concordance et de mieux valider les modèles de type RQSP-RNA développés dans cette thèse.
- **Le dernier chapitre :** inclut l'ensemble des résultats de la modélisation des cinq propriétés physicochimiques des hydrocarbures purs avant, au cours et après le développement. J'ai présenté puis discuté tous ces résultats issus des modèles de type RQSP-RNA développés dans cette thèse, tels que les résultats de calcul des descripteurs moléculaires, l'évolution du nombre des descripteurs au cours des étapes de prétraitement, l'analyse qualitatives des descripteurs pertinents, l'architecture des meilleurs modèles de type RQSP-RNA obtenus, l'examen et l'analyse statistique des modèles à l'aide de différents types d'erreurs et coefficients statistiques pour déterminer la précision de chaque modèle lors de l'estimation des cinq propriétés physicochimiques des hydrocarbures, l'appréciation des descripteurs d'entrée de chaque modèle par la méthode d'analyse de sensibilité qui se traduit par la classification de ces descripteurs selon leurs importances individuellement et dans leur catégorie respective. J'ai également discuté les résultats issus de la détermination du domaine d'applicabilité, et j'ai finalisé ce chapitre par une comparaison entre les modèles développées dans cette thèse et d'autres modèles dans des travaux semblables trouvés dans la littérature, afin de décrire et juger d'une façon globale mon travail et voir son importance entre les travaux publiés dans ce domaine de recherche.

À la fin de cette thèse, j'ai terminé par une conclusion générale récapitulant les principaux résultats obtenus tout au long des travaux de cette thèse, puis j'ai présenté quelques

perspectives qui peuvent être considérées comme un guide de référence pour les futurs travaux qui seront réalisés dans cet axe de recherche.

Chapitre I

Généralités sur les hydrocarbures et les fractions pétrolières

I.1. Introduction

Le pétrole ou l'or noir est actuellement considéré comme l'une des matières premières les plus recherchées au monde, en raison de son importance dans le développement économique, énergétique et technologique des pays producteurs (**Tugcu *et al*. 2012 ; Abdel-Aal & Alsahlawi 2013 ; Belaid & Youssef 2017**).

Dès les années cinquante, le pétrole est devenu la première source d'énergie dans le monde, en raison de sa forte densité énergétique qui en a fait la matière première des carburants utilisés dans les moyens de transports courants (voitures, camions, avions, etc.). Il est considéré comme une matière première indispensable, utilisée dans plusieurs industries, y compris l'industrie pétrochimique, pour un nombre considérable des produits de la vie quotidienne, telles que : les matières plastiques, les peintures, les colorants, et les produits cosmétiques. Le pétrole sert aussi comme combustible dans le chauffage domestique et comme source de chaleur dans l'industrie **(Rebas 2006 ; Belghit 2019)**.

La connaissance des propriétés physico-chimiques du pétrole et des fractions pétrolières a toujours été un souci pour l'ingénieur en raffinage et en pétrochimie. La maîtrise de toutes les étapes de séparation et de transformation des produits pétroliers, dès l'acheminement du pétrole des puits jusqu'au stockage des produits finis, nécessite une parfaite connaissance des caractéristiques du pétrole et de ses dérivés **(Rebas 2016)**.

Le but de ce chapitre est d'introduire différents concepts et terminologies de base liés aux domaines pétroliers, qui seront rencontrés au cours de cette thèse à travers plusieurs sections qui présentent et expliquent toutes les notions fondamentales propres aux hydrocarbures et au pétrole, tels que : l'historique et l'origine du pétrole, la description et la composition du pétrole brut et des coupes pétrolières, les critères de classification des différents types de pétrole et ses caractérisations, les paramètres physicochimiques et thermodynamiques des hydrocarbures et des fractions pétrolières, et enfin les différentes méthodes de modélisation pour la détermination de ces paramètres.

I.2. Généralités sur le pétrole brut et ses différentes coupes

I.2.1. Bref historique du pétrole

- En 1852 le pétrole est repéré pour la première fois au Moyen-Orient sur la surface de la terre. La découverte du kérosène a augmenté la consommation du pétrole en raison de son utilisation dans l'éclairage **(Henry 1873 ; Gurbanov 2007)**.

- En 1859, les premiers puits de pétrole ont été forés par le colonel Drake en Pennsylvanie, aux États-Unis **(Henry 1873 ; Masseron 1982)**.
- Au début des années cinquante, le pétrole était la principale source d'énergie dans le monde et pendant la même période, plusieurs champs pétrolifères très riches ont été découverts dans la région du Golfe Persique **(Perrodon 1985)**.
- La consommation mondiale de pétrole en tant que source chimique a augmenté de 11% entre 1970 et 1980 **(Cabet & Lizoret 1974)**.
- Depuis l'antiquité, le pétrole a été utilisé dans divers domaines, tels que les médicaments miracles, les armes de guerre et les peintures hydrofuges. Le pétrole est maintenant largement utilisé sous des fractions pétrolières ou comme produits dérivés, tels que le naphtalène, la super essence, le kérosène et le diesel **(Banque mondiale 2002 ; Belghit *et al*. 2018)**.

I.2.2. Description générale du pétrole brut

Le pétrole brut est un mélange extrêmement complexe de plusieurs millions de molécules différentes **(Rebas 2016)**. Ces molécules sont principalement des **hydrocarbures**, autrement dit des composés qui ne contiennent que des atomes de **Carbone** et d'**Hydrogène**, leurs proportions dans les pétroles bruts sont généralement comprises entre 84 à 87% en masse pour le carbone et entre 11 à 14% en masse pour l'hydrogène. Néanmoins, les molécules pétrolières peuvent également contenir des hétéroatomes qui sont souvent considérés comme des impuretés, comme le Soufre qui est généralement compris entre 0,04 et 6% en masse, l'Azote entre 0,1 et 1,5% et l'Oxygène entre 0,1 et 0,5% **(Dupuy 2015 ; Rebas 2016)**.

En outre, il y a aussi certains métaux comme le Nickel et le Vanadium habituellement compris entre 0,005 et 0,015% en masse soit 50 à 150g dans une (1) tonne de pétrole brut, où le Vanadium représente généralement plus de 75 % de la teneur en métaux. Les pétroles bruts contiennent aussi des éléments étrangers non organiques comme l'eau et les sédiments généralement compris entre 0,1 et 0,6% en volume, et les sels minéraux qui sont habituellement compris entre 20 et 200 g/tonne **(Dupuy 2015)**.

La proportion de chaque élément, surtout des hétéroatomes et des métaux, dépend fréquemment de l'origine géographique du gisement de pétrole, le **tableau I.01** illustre des exemples de la densité à 15 °C et la composition élémentaire des quatre pétroles bruts différents très connue mondialement dans l'industrie pétrolière.

Tableau I.1. Composition élémentaire des principaux bruts et leurs densités à 15°C **(Rebas 2016)**

Nom de brut	Brent	Oural	Maya	Athabasca
Origine Géographique	Royaume-Uni	Russie	Mexique	Canada
Densité à 15°C (g/cm^3)	0,833	0,869	0,927	1,011
Carbone (% poids)	86,1	85,8	84,7	83,9
Hydrogène (% poids)	13,4	12,6	11,6	10,8
Soufre (% poids)	0,4	1,4	3,4	4,8
Azote (ppm)	948	2050	3811	4775
Métaux (ppm)	17,3	75,8	334,5	198,8

NB : (ppm) parti **p**ar **m**illion (10^{-6}) ou (1 ppm = 1 mg/kg et 10000 ppm = 1 %)

En dehors de la variété de la composition élémentaire, les molécules pétrolières ont aussi des masses molaires très diverses pouvant varier entre 16 g/mol de la molécule de méthane et plus de 10000 g/mol pour les molécules asphaltiques **(Rebas 2016)**.

Le pétrole brut est un mélange très complexe composé d'un grand nombre de produits que l'on sépare suivant leurs volatilités dont : des produits volatils ou gazeux comme le gaz et l'essence légère ; des produits légers comme l'essence lourde, le kérosène, et gasoil ; et des produits lourds qui sont généralement des résidus **(Slaoui 1986)**. En plus, selon leur densité à 15°C les pétroles bruts sont fréquemment classés en quatre familles **(Rebas 2016)** :

- Bruts légers : $d < 0,825$ g/cm^3
- Bruts moyens : $0,825$ g/cm^3 $< d < 0,875$ g/cm^3
- Bruts lourds : $0,875$ g/cm^3 $< d < 1,000$ g/cm^3
- Bruts extra lourds : $d > 1,000$ g/cm^3

I.2.3. Description générale des coupes pétrolières

Le pétrole brut soumis au processus de raffinage est séparé en plusieurs sous-mélanges dénommés coupes ou fractions pétrolières. Elles sont traditionnellement caractérisées par un intervalle de distillation, exprimé en température d'ébullition normale ou en nombre d'atomes de carbone équivalent des paraffines. La **figure I.1** donne les principales coupes pétrolières

obtenues dans les raffineries avec leurs frontières typiques au niveau du nombre d'atomes de carbone et de la température d'ébullition **(Rebas 2016)**.

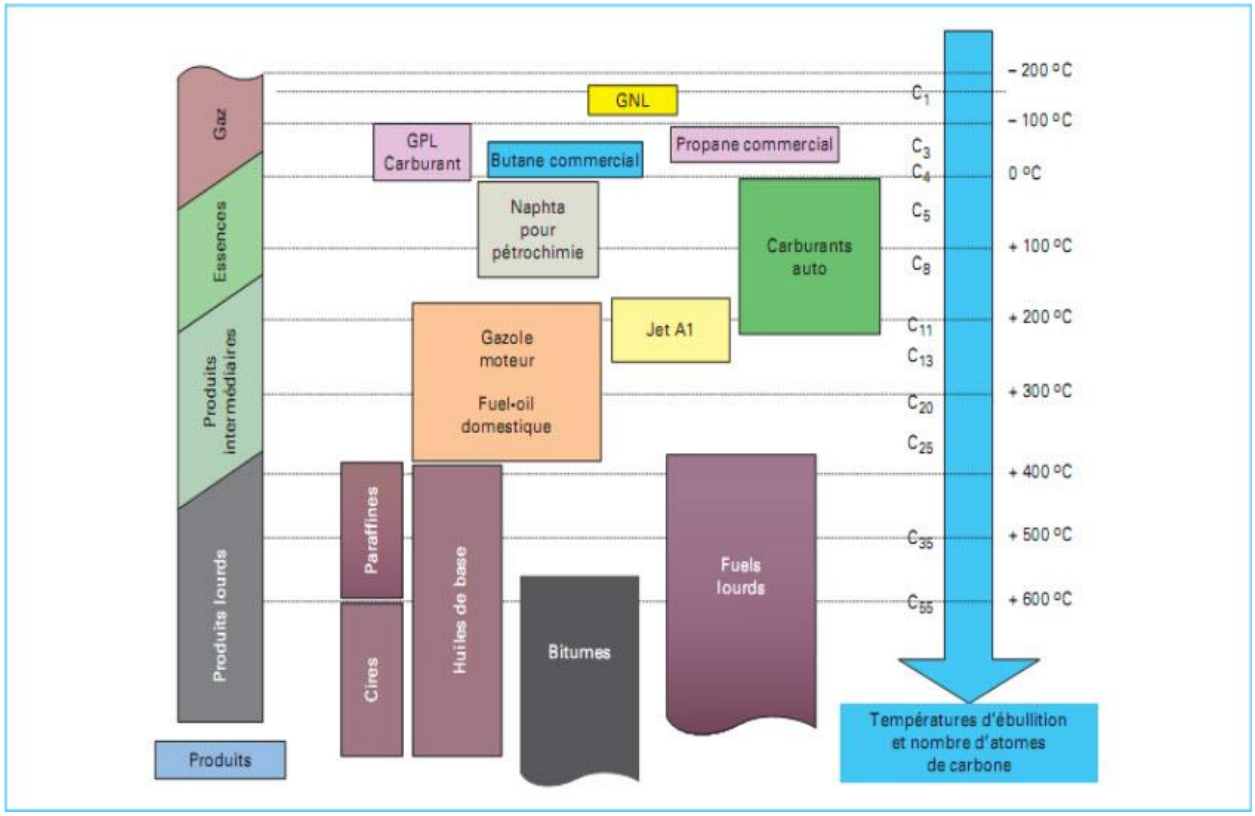


Figure I.1. Classification des principales coupes pétrolières selon les intervalles de la température d'ébullition et le nombre d'atomes de carbone équivalent des paraffines **(Rebas 2016)**.

Le gaz et l'essence sont considérés comme des coupes pétrolières légères ; leurs points de coupes supérieures à 0°C pour le gaz et compris entre 145 et 185°C pour l'essence, leurs molécules contiennent respectivement entre 1 et 4 et entre 5 et 10 atomes de carbone. Le kérosène et les gazoles correspondent aux coupes dites intermédiaires ; le kérosène est obtenu à un point de coupe compris entre 220 et 240°C avec une taille de leurs molécules variant de C9-C10 à C13-C14. La température d'ébullition finale des gazoles est généralement comprise entre 360 et 380°C, ses molécules contiennent 25 atomes de carbone au plus. Les deux dernières coupes sont appelées coupes lourdes et sont normalement récupérées au fond de la colonne de distillation atmosphérique, les distillats sous vide (DSV) contiennent les molécules ayant un point d'ébullition compris entre 360-380°C et 500-600°C et un nombre d'atomes de carbone qui varient entre C25 et C40-C50. Le résidu sous vide (RSV) correspond à la fraction la plus lourde du pétrole et composés de molécules ayant un point d'ébullition supérieur à 500-600°C et un nombre d'atomes de carbone supérieur à C40-C50 **(Rebas 2016)**.

Les coupes pétrolières possèdent des propriétés distinctes selon l'origine du pétrole brut et les étapes de raffinage dont elles sont issues. Par exemple, dans le cas des gazoles Light Cycle Oil (LCO) issus du procédé de craquage catalytique, qui contiennent généralement une basse teneur en paraffines et une haute teneur en aromatiques, tandis que l'inverse est observé dans le cas des gazoles dits Straight Run obtenus de la distillation atmosphérique **(Rebas 2016 ; Belghit 2019)**.

La proportion des fractions pétrolières dans les pétroles bruts aussi appelée structure de rendement, dépend fortement de l'origine du pétrole, comme illustré dans la **figure I.2**. Plus la proportion en coupes lourdes est importante plus la valeur de la densité est élevée. Le même type de relation est observé avec la teneur en hétéroatomes et en métaux, ce qui est attendu car ces éléments se trouvent généralement sur les coupes lourdes **(Rebas 2016 ; Belghit 2019)**.

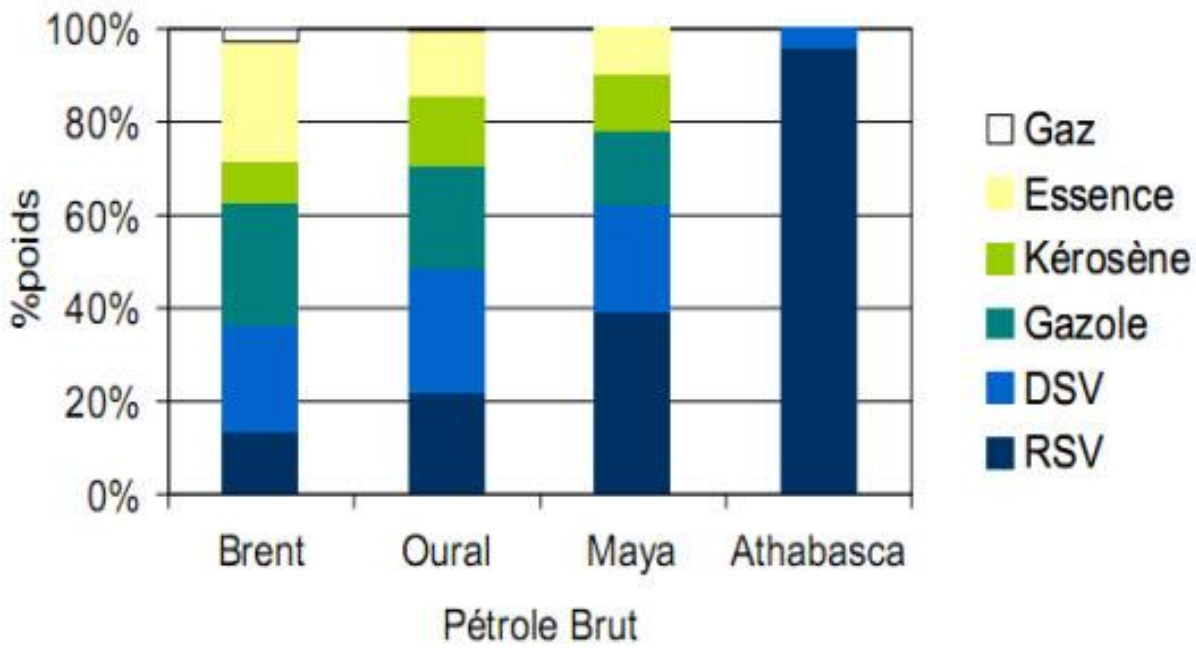


Figure I.2. Structure des rendements en coupes pétrolières de quelques pétroles bruts **(Rebas 2016)**.

I.3. Composition chimique du pétrole brut et des fractions pétrolières

Le pétrole brut et ses fractions ; les coupes pétrolières, sont des mélanges de molécules très complexes et très différentes entre elles. Les molécules sont principalement constituées par des atomes de carbone et d'hydrogène, mais peuvent également contenir des hétéroatomes, comme le soufre, l'azote et l'oxygène, ainsi que certains métaux, comme le nickel et le

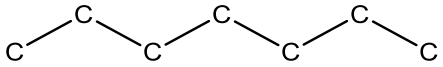
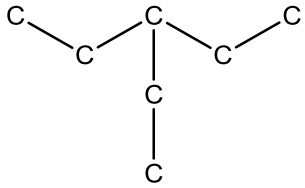
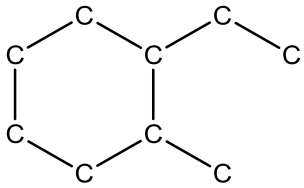
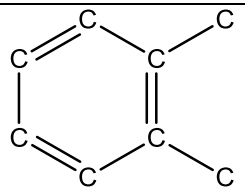
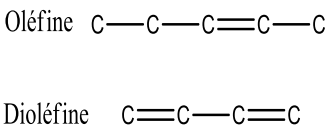
vanadium. On peut trouver des structures linéaires, comme les paraffines et les oléfines, ou cycliques tels que les naphtènes et les aromatiques. Les molécules peuvent contenir un seul atome de carbone, comme le méthane, ou plusieurs centaines d'atomes de carbone, comme les molécules asphaltiques. Les coupes pétrolières peuvent contenir plusieurs centaines à plusieurs millions de molécules différentes **(Rebas 2016)**. Dans ce qui suit, tous ces grands constituants seront présentés puis expliqués en détails de manière indépendante.

I.3.1. Les hydrocarbures

Les hydrocarbures constituent la majeure partie des pétroles bruts et des coupes pétrolières. Au sens strict, les hydrocarbures sont des molécules organiques contenant exclusivement des atomes de carbone et d'hydrogène, elles se divisent en plusieurs familles chimiques selon leurs structures qui sont basées principalement sur la tétravalence du carbone. Les enchaînements moléculaires carbone-carbone peuvent être : **(1)** soit réunis par une simple liaison (suffixe **ane**) ; les hydrocarbures saturés, **(2)** soit par des liaisons multiples doubles (suffixe **ène**) ou triples (suffixe **yne**) ; les hydrocarbures insaturés. Ces enchaînements moléculaires constituent soit des chaines linéaires soit des cycles, sachant que les deux derniers peuvent former soit des structures simples ou des structures ramifiées **(Vidal 1997 ; Wauquier 1998 ; Rebas 2016)**.

Dans le secteur pétrolier, il existe plusieurs méthodes de classification hiérarchique des hydrocarbures. Cependant, les hydrocarbures sont souvent classés en : **(1)** hydrocarbures saturés et **(2)** hydrocarbures insaturés ou non-saturés. Sur la base de ces derniers et selon leurs structures moléculaires, on peut également distinguer quatre familles chimiques : les paraffines, les naphtènes, les aromatiques et les oléfines. Le **tableau I.2** récapitule chaque famille dans sa classification possible et donne certaines de ses principales caractéristiques.

Tableau I.2. Classification des hydrocarbures en fonction de leurs familles (**Dupuy 2015**).

Principales familles d'Hydrocarbures	Hydrocarbures **saturés** (Sans doubles liaisons)			Hydrocarbures **insaturés** (Avec doubles liaisons)	
	Atomes de carbone en ---**Chaines**---		Molécules comprenant des ---**Cycles**---	Hydrocarbures comprenant des ---**Cycles benzéniques**---	Atomes de carbone en ---**Chaines contenant des doubles liaisons**---
	Paraffines		**Naphtènes**	**Aromatiques**	**Oléfines** (une seul double liaison)
	En chaines droites (**Paraffines normale** ou bien **n-Paraffines**)	En chaines ramifiées (**Iso-Paraffines**)	(**Cyclo-Paraffines**)		Dioléfines (deux doubles liaisons)
	←Hydrocarbures présents dans le pétrole brut→				
	Exemples				
					
Stabilité	Bonne	Bonne	Bonne	Bonne	Faible
Densité liquide	Faible	Faible	Moyenne	Élevée	Faible
Teneur en H. Rapport H/C	Élevée	Élevée	Moyenne	Faible	Moyenne

I.3.1.1. Les paraffines

Le terme paraffine fait référence à toutes les molécules classiquement nommées en chimie organique sous le nom d'hydrocarbures aliphatiques saturés ou simplement d'alcanes. Les paraffines sont des enchaînements non cycliques d'atomes de carbone liés exclusivement par des liaisons simples. Chaque atome de carbone est toujours lié à quatre autres atomes, soit de carbone soit hydrogène, sa formule chimique générale est $\boldsymbol{C_nH_{2n+2}}$. La chaîne des paraffines peut également être linéaire, appelées paraffines normales ou n-paraffines, autrement elle peut être de forme ramifiée et dans ce cas, on parle de paraffines ramifiées ou iso-paraffines **(Rebas 2016 ; Belghit 2019)**. Le **tableau I.2** donne un exemple des deux molécules qui peuvent être rencontrées dans un pétrole brut, on parle de l'*heptane* pour les paraffines linéaires et la *3-ethylpentane* pour les paraffines ramifiées.

I.3.1.2. Les naphtènes

Les naphtènes sont généralement dénommés en chimie organique par le nom des cycloalcanes. Ils appartiennent aux familles des hydrocarbures saturés comprenant au moins une structure cyclique. La formule chimique générale est $\boldsymbol{C_nH_{2(n-c+1)}}$, où $\boldsymbol{c}$ représente le nombre de cycles dans la molécule. Les structures cycliques les plus souvent rencontrées dans le pétrole brut sont celles à cinq atomes, comme le cyclopentane, ou à six atomes, comme le cyclohexane, en outre, les molécules naphténiques sont entièrement dérivées de ces deux structures : **(1)** par addition de chaînes paraffiniques latérales comme l'*éthyle-cyclopentane*, ou **(2)** par fusion de plusieurs cycliques saturées comme l'exemple de *décaline* **(Wauquier 1998 ; Rebas 2016)**. Le **tableau I.2** montre l'exemple d'une molécule de *1-éthyl-2-méthylcyclohexane* qui est l'une des molécules naphténiques qui peuvent être rencontrées dans le pétrole brut.

I.3.1.3. Les oléfines

Traditionnellement, les oléfines reconnues par le nom alcène sont des hydrocarbures aliphatiques insaturées qui possèdent une ou plusieurs double liaisons carbone-carbone insaturées, ils peuvent être structurées soit en chaîne linéaire, dénommées oléfines normales, soit en chaîne ramifiée, dénommées iso-oléfines. Les oléfines ne se rencontrent pas ou très peu naturellement présentes dans les pétroles bruts en raison de leur forte réaction des doubles liaisons. En revanche, lorsque le pétrole est soumis aux procédés de raffinage, particulièrement les procédés en absence d'hydrogène ont permis de former ce type des molécules. La formule chimique générale des oléfines est $\boldsymbol{C_nH_{2(n-d+1)}}$, où $\boldsymbol{d}$ représente le nombre de doubles liaisons

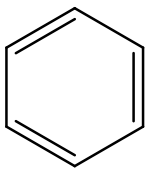
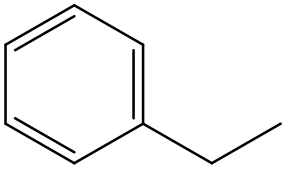
dans la molécule **(Rebas 2016 ; Belghit 2019)**. Le **tableau I.2** donne deux exemples de molécules oléfiniques qui peuvent être rencontrées dans les coupes pétrolières, la *(2E)-2-Pentène* qui représente une oléfine simple et le *1,3-Butadiène* qui représente une dioléfine simple.

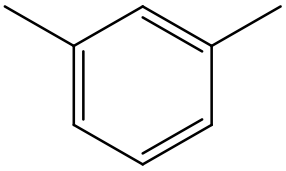
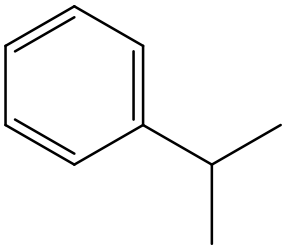
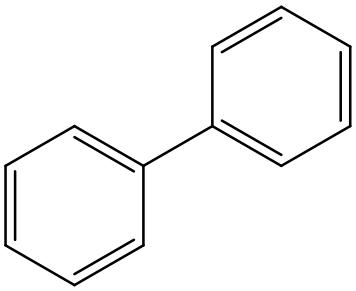
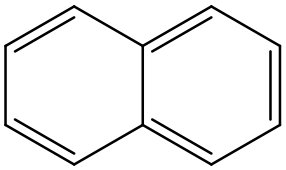
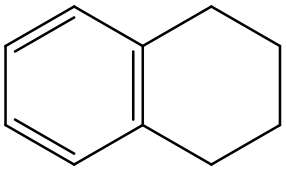
I.3.1.4. Les aromatiques

Les hydrocarbures aromatiques sont des polycycliques insaturés qui contiennent au moins une structure benzénique, composée de six atomes de carbone reliés entre eux par six liaisons en alternance entre les simples et les doubles liaisons issues de 3 doubles liaisons conjuguées. Cette conjugaison dans un cycle à 6 atomes le rend extrêmement stable. Les autres structures aromatiques peuvent donc se dériver à partir de la structure benzénique par : **(1)** addition des chaînes paraffiniques latérales comme le cas de l'*éthylbenzène*, **(2)** jonction entre deux cycles benzéniques comme la formation de *diphényle*, **(3)** fusion de cycles benzéniques comme l'exemple de la molécule *naphtalène*, et enfin **(4)** fusion de cycles saturés comme la molécule de *tétraline* **(Wauquier 1998 ; Rebas 2016)**.

La formule générale des aromatiques dépend fortement de leur dérivation par rapport à la structure benzénique de base. Pour les alkyles benzènes, la formule chimique générale est $\boldsymbol{C_nH_{2(n-6)}}$, les autres mono-aromatiques contenant des cycles naphténiques ont comme formule brute $\boldsymbol{C_nH_{2(n-c)-6}}$, où $\boldsymbol{c}$ désigne le nombre de cycles naphténiques dans la molécule. Les di-aromatiques de type bi-phényle alkylé ont comme formule brute $\boldsymbol{C_nH_{2n-14}}$, tandis que la formule chimique générale est $\boldsymbol{C_nH_{2n-12}}$ pour les di-aromatiques alkylés avec deux cycles benzéniques fusionnés. Le **tableau I.3** montre quelques exemples des molécules aromatiques rencontrées dans des coupes pétrolières **(Rebas 2016 ; Belghit 2019)**.

Tableau I.3. Structures de quelques hydrocarbures aromatiques **(Rebas 2016)**.

Nom	Formule chimique	Structure moléculaire
Benzène	$\boldsymbol{C_6H_6}$	
Ethylbenzène	$\boldsymbol{C_8H_{10}}$	

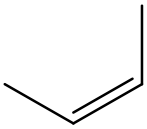
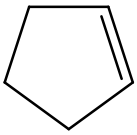
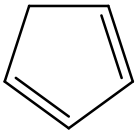
m-xylène	C_8H_{10}	
i-propylbenzène	C_9H_{12}	
Diphényle	$C_{12}H_{10}$	
Naphtalène	$C_{10}H_8$	
Tétraline	$C_{10}H_{12}$	

I.3.1.5. Autres hydrocarbures

En plus des paraffines, naphtènes, oléfines et aromatiques, il est possible de trouver à l'état de traces dans les mélanges pétroliers d'autres hydrocarbures insaturés, en particulier des alcynes et des cyclo-oléfines. Les alcynes sont des hydrocarbures aliphatiques contenant des triples liaisons carbone-carbone, tandis que les cyclo-oléfines possèdent une ou plusieurs doubles liaisons carbone-carbone sous forme cyclique. Les hydrocarbures alcyniques sont très réactifs, c'est la raison qu'ils ne sont pas détectés dans les pétroles bruts et leurs teneurs dans les coupes pétrolières sont très faibles. En outre, il existe plusieurs isomères de position à l'état de traces dans les coupes pétrolières qui sont issues des effluents de procédés de conversion tels que les isomères *cis*, *trans*, *para*, *méta* et *ortho* **(Wauquier 1998 ; Rebas 2016 ; Belghit 2019)**. Le

tableau I.4 montre quelques exemples de ces molécules insaturés rencontrées dans les coupes pétrolières.

Tableau I.4. Exemples de quelques structures d'autres hydrocarbures insaturés (**Rebas 2016 ; Belghit 2019**).

Nom	Formule chimique	Structure moléculaire
Éthylène	C_2H_4	
2-Butyne	C_4H_6	
Cis-2-butène	C_4H_8	
Cyclopentène	C_5H_8	
Cyclopentadiène	C_5H_6	

I.3.2. Les hétéroatomes

Les hétéroatomes sont ceux qui, en dehors des atomes de carbone et d'hydrogène, les composés hétéroatomiques constitués principalement par le soufre, l'azote ou l'oxygène qui sont inclus dans un composé organique. Pour cela ils peuvent être également considérés en raffinage comme des hydrocarbures renfermant des impuretés. Ces hétéroatomes se trouvent sous forme de fonctions chimiques telles que des sulfures, des amines, des alcools, et des acides carboxyles. Les métaux, comme le nickel et le vanadium, sont considérés comme autres types d'hétéroatomes qui peuvent être présents dans les coupes pétrolières et connus sous le terme organométallique (**Wauquier 1998 ; Rebas 2016**).

I.3.2.1. Les composés soufrés

Le soufre est généralement l'hétéro élément le plus répandu dans le pétrole brut et les coupes pétrolières. Il peut se trouver sous deux formes différentes : **(1)** sous forme inorganique comme un soufre élémentaire ou comme un sulfure de di-hydrogène (H_2S), **(2)** incorporé dans des molécules organiques (**Rebas 2016**).

Le soufre peut être présent sous forme de trois structures différentes dans les molécules organiques **(Wauquier 1998 ; Rebas 2016 ; Belghit 2019)** :

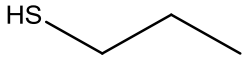
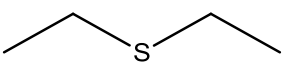
1) Les mercaptans également appelés thiols, sont des molécules dont une chaîne hydrocarbonée possède un groupement **–SH** comme atome terminal de la chaîne alkyle ($\boldsymbol{C_nH_{2n+1}-SH}$) ou d'une paraffine (ex : *Propanethiol*). Ces composés sont principalement détectés dans les coupes pétrolières les plus légères, autrement dit sur des molécules avec un bas point d'ébullition.

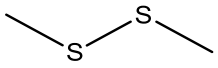
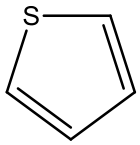
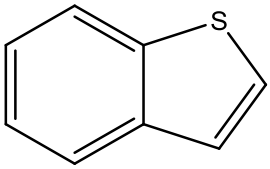
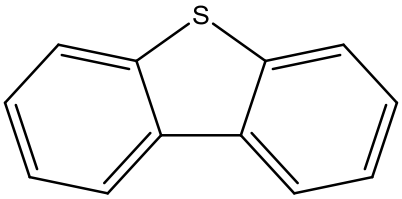
2) Les sulfures sont également des molécules possédant des chaînes hydrocarbonées avec un ou plusieurs atomes de soufre à l'intérieur de la chaîne. Selon le nombre d'atomes de soufre reliés en série, on distingue les sulfures, les disulfures et les polysulfures. Dans le cas des sulfures, les atomes de soufre s'intercalent dans la chaîne (ex : *diéthylsulfure*). D'autre part, lorsque qu'un atome de soufre est relié à un ou deux autres soufres le composé est appelé respectivement disulfure et polysulfure, (ex : *diméthyldisulfure*). Ces composés sont également détectés plus fréquemment dans les coupes légères.

3) Les thiophènes sont des composés dérivés de la molécule de thiophène. Le thiophène est une molécule cyclique à cinq atomes avec un atome de soufre, quatre atomes de carbone et deux doubles liaisons. Elle possède des propriétés proches de celles des hydrocarbures aromatiques grâce à la présence de l'atome de soufre. En effet, la conjugaison des électrons des deux doubles liaisons et le doublet de l'atome de soufre permet de créer une orbitale moléculaire très stable, comme pour les aromatiques. C'est pourquoi les soufres thiophéniques sont plus difficiles à éliminer. De par leur point d'ébullition, ces molécules se trouvent généralement dans les coupes moyennes et lourdes.

Le **tableau I.5** montre quelques exemples de composés soufrés rencontrés dans les pétroles bruts et dans les coupes pétrolières.

Tableau I.5. Exemples de quelques structures de composés soufrés **(Rebas 2016 ; Belghit 2019)**.

Nom	Formule chimique	Structure moléculaire
Propanethiol	$\boldsymbol{C_3H_8S}$	
Diéthylsulfure	$\boldsymbol{C_4H_{10}S}$	

Diméthyldisulfure	$C_2H_6S_2$	
Thiophène	C_4H_4S	
Benzothiophène	C_8H_6S	
Dibenzothiophène	$C_{12}H_8S$	

I.3.2.2. Les composés azotés

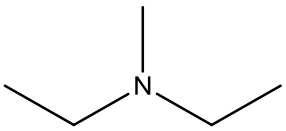
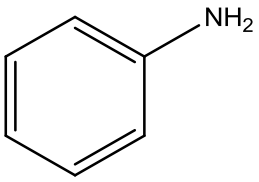
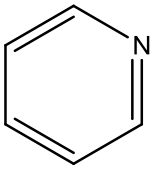
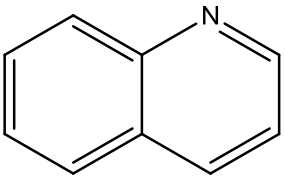
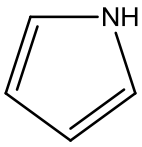
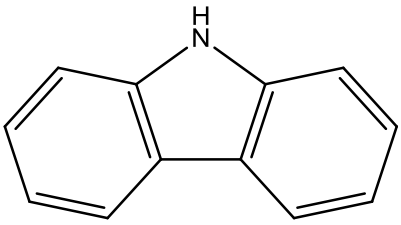
Les composés azotés sont présents dans une proportion plus faible que les composés soufrés. L'azote se trouve particulièrement dans les fractions lourdes du pétrole brut sous forme de deux structures cycliques : pyrrole et pyridine **(Wauquier 1998 ; Rebas 2016 ; Belghit 2019)**.

- **Le pyrrole** est une structure cyclique analogue au thiophène, mais elle contient un groupe (***NH***) à la place de l'atome de soufre.
- **La pyridine** est une structure cyclique aromatique constituée par cinq atomes de carbone et un atome d'azote.

Ces deux structures donnent des caractéristiques différentes aux molécules. Les molécules dérivées de la pyridine ont un caractère basique, tandis que les dérivés du pyrrole sont plutôt neutres **(Rebas 2016)**.

Il est aussi possible de trouver, en très faibles quantités des structures non-cycliques, comme les amines (***R–NH₂***, ***R–NH–R'*** et ***(R)₃–N***) ou les anilines **(Rebas 2016 ; Belghit 2019)**. Le **tableau I.6** montre quelques exemples de composés azotés rencontrés dans les pétroles bruts et dans les coupes pétrolières.

Tableau I.6. Exemples de quelques structures de composés azotés (**Rebas 2016 ; Belghit 2019**).

Nom	Formule chimique	Structure moléculaire
Diéthylméthylamine	$C_5H_{13}N$	
Aniline	C_6H_7N	
Pyridine	C_5H_5N	
Quinoléine	C_9H_7N	
Pyrrole	C_4H_5N	
Carbazole	$C_{12}H_9N$	

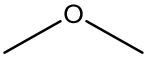
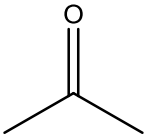
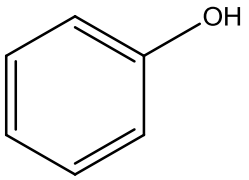
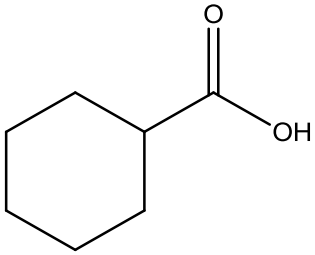
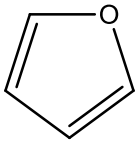
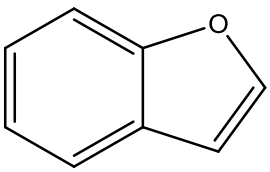
I.3.2.3. Les composés oxygénés

Bien moins abondants que les autres deux hétéroatomes, l'oxygène peut jouer un rôle non négligeable, étant notamment responsable de l'acidité des pétroles. Les composés oxygénés les plus courants sont dérivés de l'acide naphténique et le phénol. Le premier est un acide carboxylique (***R–COOH***) avec un cycle saturé comme substituant hydrocarboné (***R***), tandis que

le phénol est un alcool formé à partir de la substitution d'un hydrogène d'un cycle aromatique par un groupe hydroxyle (***R–OH***). Toutefois, l'oxygène peut aussi se trouver, soit sous forme d'autres structures non-cycliques, comme les éthers (***R–O–R'***), les cétones (***R–CO–R'***), les aldéhydes (***R–CHO***) ou les esters (***R–COO–R'***), intégré dans une structure dérivée du furane. Le furane est un composé cyclique oxygéné analogue au thiophène et au pyrrole **(Wauquier 1998 ; Rebas 2016 ; Belghit 2019**).

Le **tableau I.7** montre quelques exemples de composés oxygénés rencontrés dans les pétroles bruts et dans les coupes pétrolières.

Tableau I.7. Exemples de quelques structures des composés oxygénés **(Rebas 2016 ; Belghit 2019**).

Nom	Formule chimique	Structure moléculaire
Diméthyléther	C_2H_6O	
Acétone	C_3H_6O	
Phénol	C_6H_6O	
Acide Cyclohexanecarboxylique	$C_7H_{12}O_2$	
Furane	C_4H_4O	
Benzofurane	C_8H_6O	

I.3.2.4. Les organométalliques

Les composées organométalliques sont des métaux sous forme complexée. Ils se trouvent dans les coupes pétrolières les plus lourdes. Le nickel et le vanadium sont les atomes les plus rencontrés. Ces composés organométalliques possèdent une structure de base Porphyrinique composée de 4 cycles pyrroliques, le métal étant au centre de ce motif sous forme Ni^{++} ou VO^{+} complexée **(Rebas 2016 ; Belghit 2019)**. La **figure I.3** montre la structure du motif Porphyrine ou Porphine ($C_{20}H_{14}N_4$).

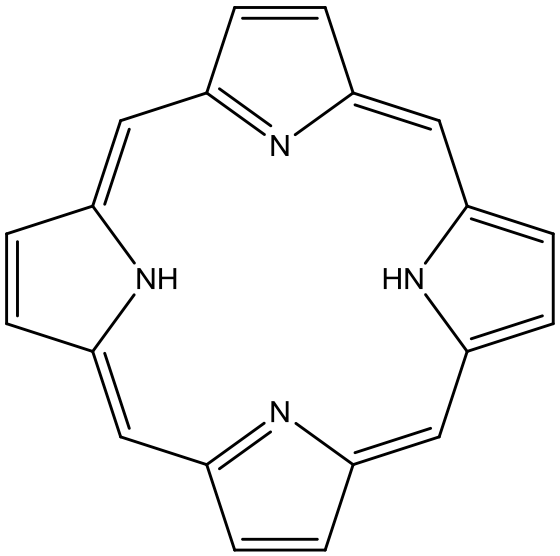


Figure I.3. Représentation de la structure du motif *Porphyrine* ou *Porphine* ($C_{20}H_{14}N_4$).

I.3.3. Les composés asphalténiques

Les composées asphalténiques sont des composées mal définis de type colloïdal, de couleur noirâtre trouvés dans les coupes pétrolières les plus lourdes. Contrairement aux composés décrits ci-dessus, les asphaltènes ne sont pas définis en termes de caractéristiques structurales, mais par des conditions purement opératoires. Ces composés sont des constituants de la fraction pétrolière qui est insoluble dans un solvant paraffinique peut-être le pentane, l'hexane ou l'heptane, mais soluble dans le toluène **(Rebas 2016 ; Belghit 2019)**. Les composés asphalténiques sont constitués de structures polyaromatiques riches en hétéroatomes (S, N, O et métaux), dénommés noyaux, qui sont liés à des chaînes aliphatiques. Les asphaltènes sont chimiquement très polydispersées. Les masses molaires estimées de ces entités varient entre **500** et **1000** g/mol **(Rebas 2016 ; Groenzin & Mullins 2000)**. Étant donné leur hétérogénéité, il n'existe pas de motif structural unique pour les asphaltènes.

Plusieurs études ont été menées pour décrire ces structures, qui peuvent être selon le degré de condensation du noyau : de type *continental* et/ou de type *archipel*. Les asphaltènes du type continental sont constitués par très peu de noyaux, voir un seul, mais très condensés, tandis que ceux du type archipel sont moins condensés, mais plus nombreux et reliés entre eux par des chaînes alkyles **(Rebas 2016)**. La **figure I.4** donne l'exemple de deux molécules asphalténiques possibles, un de type continental et l'autre de type archipel.

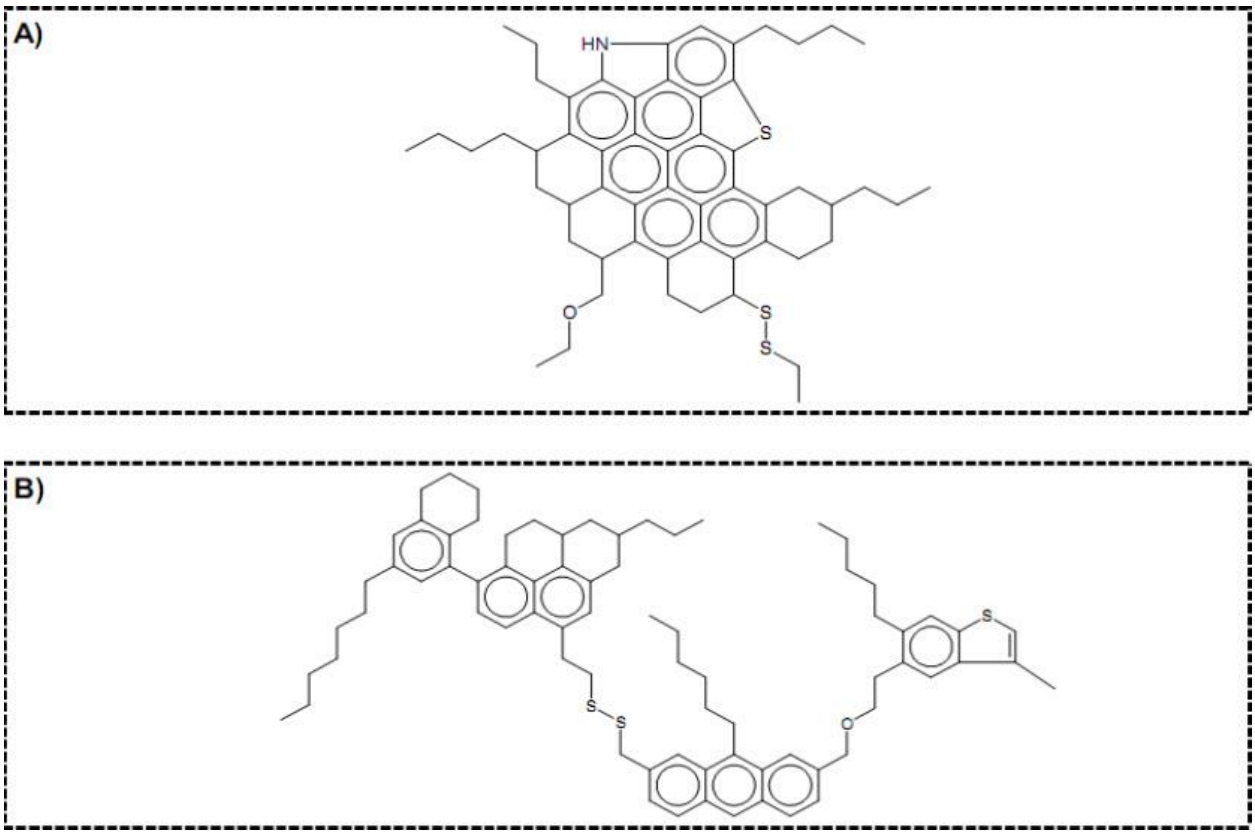


Figure I.4. Représentation des deux molécules asphalténiques possibles : **A)** Type ***Continental***, **B)** Type ***Archipel*** **(Rebas 2016)**.

À cause de sa forte aromaticité existée, les molécules asphalténiques ont tendance à former des agrégats qui peuvent exister sous deux formes différentes ; selon la teneur en asphaltènes, la température du milieu et le type de solvant. La première forme correspond à un empilement de plusieurs noyaux aromatiques appelés aussi "feuillets aromatiques" de molécules d'asphaltènes. Ces empilements sont nommés particules ou nano-agrégats. Tant pour les asphaltènes du type continental comme pour ceux du type archipel, l'empilement se fait via des interactions **(π–π)** entre les cycles aromatiques de différents noyaux et des liaisons par ponts d'hydrogène **(Rebas 2016)**.

La deuxième forme d'agrégats existe seulement à de fortes concentrations des asphaltènes dans le milieu. Cette forme, nommé agrégat colloïdal, résulte de l'autoassociation des nano-agrégats via des interactions de Van der Waals ou des liaisons par ponts d'hydrogène. Ces agrégats colloïdaux peuvent être constitués de **4** à **100** nano-agrégats, avec au moins **10** nano-agrégats, ce qui conduit alors à des masses molaires de $\mathbf{4 \times 10^{+4}}$ à $\mathbf{1 \times 10^{+6}}$ g/mol avec des rayons de giration entre **40** à **400** Å, avec une moyenne de **70** Å **(Eyssautier 2012 ; Rebas 2016)**.

La différence entre les deux modèles réside dans le type d'association impliquée dans la formation des nano-agrégats. Les molécules du type continental s'associent principalement par des mécanismes intermoléculaires, tandis que celles du type archipel peuvent s'associer, à travers des mécanismes intramoléculaires par pliage de leurs noyaux. Des travaux récents effectués sur les asphaltènes C5 du pétrole brut suggèrent que les nano-agrégats sont constitués par au maximum dix feuillets, conduisant à une masse molaire estimée autour de $\mathbf{1 \times 10^{+4}}$ g/mol et un rayon de giration (taille) d'environ **20** à **30** Å **(Eyssautier 2012 ; Rebas 2016)**.

I.4. Propriétés usuelles des hydrocarbures et des fractions pétrolières

Les pétroles bruts ont des caractéristiques physiques et chimiques très variables d'un champ de production à un autre et même à l'intérieur du même gisement. Les propriétés physicochimiques, thermodynamiques, les facteurs et les indices spécifiques des hydrocarbures, du pétrole brut et ses fractions jouent un rôle fondamental dans la caractérisation, l'identification et l'analyse de leurs qualités. Ils sont considérés comme des outils analytiques permettant de déterminer d'autres propriétés des pétroles bruts et des mélanges pétroliers. Ces outils doivent non seulement être simples à mettre en place, mais également être parfaitement reproductibles. De plus, ils doivent fournir des informations intéressantes aux raffineurs, comme la densité, l'indice de réfraction, la viscosité et l'indice d'octane ou de cétane **(Rebas 2016)**. Dans cette partie, je vais juste définir les propriétés les plus repandues et les organiserai de manière innovante pour faciliter la compréhension de chaque propriété.

I.4.1. Propriétés liées aux fluides

I.4.1.1. Masse volumique

La masse volumique d'une substance, aussi appelée densité volumique de masse, est une grandeur physique qui caractérise la masse de cette substance par unité de volume, généralement

notée par la lettre grecques $\boldsymbol{\rho}$, dépendant de la température et de la pression où elle décroit par l'augmentation de la température et vice-versa, elle est exprimée par l'unité de **Kg/m^3**, **g/cm^3** ou par **lb/ft^3 (Daubert & Denner 1997)**.

I.4.1.2. Densité relative

La densité relative du liquide noté $\boldsymbol{d_4^T}$, connue sous le nom anglo-saxon "Spécifique Gravity" ou **SG**, c'est un paramètre physique sans dimension spécifique à chaque produit **(Daubert & Denner 1997)**. Leur définition mathématique est indiquée par l'**Eq (I.1)**.

$$\mathbf{SG} = \boldsymbol{d_4^T} = \frac{\boldsymbol{Masse\ du\ volume\ de\ corps}\ \text{à}\ \boldsymbol{T°}}{\boldsymbol{Masse\ du\ volume\ d'eau}\ \text{à}\ \mathbf{4°C}} \qquad \textbf{(I.1)}$$

La densité à **15,56°C** est habituellement utilisée comme paramètre d'entrée avec le point d'ébullition **Tb** pour prédire d'autres propriétés physiques des hydrocarbures purs et des fractions pétrolières. Elle est définie aussi par l'**Eq (I.2)** :

$$\mathbf{SG_{(60°F/60°F)}} = \frac{\boldsymbol{Densité\ de\ liquide}\ \text{à}\ \mathbf{60°F}}{\boldsymbol{Densité\ d'eau}\ \text{à}\ \mathbf{60°F}} \qquad \textbf{(I.2)}$$

I.4.1.3. Gravité de l'API

Densité **API** (*en anglais* : API Gravity), c'est l'abréviation de "American Petroleum Institute Gravity", est une échelle employée pour calculer la densité des produits pétroliers liquides. La densité API s'exprime en degrés API comprise entre 0 équivalent à une densité de 1,076 et 100 équivalent à une densité de 0,6112. Leur définition mathématique est indiquée dans l'**Eq (I.3)**.

$$\mathbf{API} = \frac{\mathbf{141{,}5}}{\mathbf{SG_{(60°F/60°F)}}} - \mathbf{131{,}5} \qquad \textbf{(I.3)}$$

I.4.1.4. Viscosité

La viscosité est l'une des propriétés de transport utilisées pour la caractérisation des hydrocarbures. La viscosité dynamique usuellement notée "$\boldsymbol{\mu}$", est une grandeur physique qui caractérise la résistance à l'écoulement laminaire d'un fluide incompressible, elle s'exprime en (**Pa.s**) dans le système SI. Elle est déterminée par la mesure du temps d'écoulement d'un fluide dans un tube capillaire de longueur donnée à une température bien déterminée. La viscosité cinématique notée par "$\boldsymbol{\nu}$" est le quotient de la viscosité dynamique par la masse volumique du

fluide. Elle représente la capacité de rétention des particules du fluide et quantifie sa capacité à s'épancher, elle est exprimée en (**m^2/s**) (**Wauquier 1998**).

I.4.2. Propriétés critiques

Le terme "critique" a été proposé par Andrews entre 1869 et 1876 dans ses travaux sur les propriétés du dioxyde de carbone. Les travaux de Van der Waals en 1873 et le développement de la loi des états correspondants furent le point de départ d'une multitude de méthodes expérimentales visant à mesurer les propriétés critiques des corps purs et des mélanges (**Riazi 2005**).

La connaissance des propriétés critiques est indispensable dans le traitement des mélanges d'hydrocarbures où la coexistence d'une phase gazeuse et liquide est fréquente. Ils permettent de prévoir le comportement des fluides dans un large intervalle de températures et de pressions et définir ainsi les conditions d'existence de leurs différentes phases (**Rebas 2016**).

Les propriétés critiques sont des caractéristiques physiques des corps purs. La température et la pression critiques définissent le point limite supérieur de la courbe de tension de vapeur, au-delà duquel le changement de phase n'est plus observé. En ce point, la densité et la composition des deux phases liquide et vapeur sont identiques ; la séparation d'un tel mélange et donc impossible en ce point. La connaissance du point critique permet de prévoir le comportement des substances dans un large domaine de températures et de pressions par le biais de la loi des états correspondants ou des équations d'état (**Rebas 2006**).

La **figure I.5**, montre le diagramme des domaines de coordonnées pression-température, le point de convergence des trois courbes d'équilibres de sublimation, de fusion et de vaporisation d'un corps pur est appelé "le point triple", ce point permet de délimiter les domaines relatifs aux états liquide, solide et gaz. Comme indique la **figure I.6**, dans le cas d'un mélange, la courbe d'équilibre de vaporisation se décompose en deux courbes : courbe de bulle et courbe de rosée (**Rebas 2006 ; Belghit 2019**).

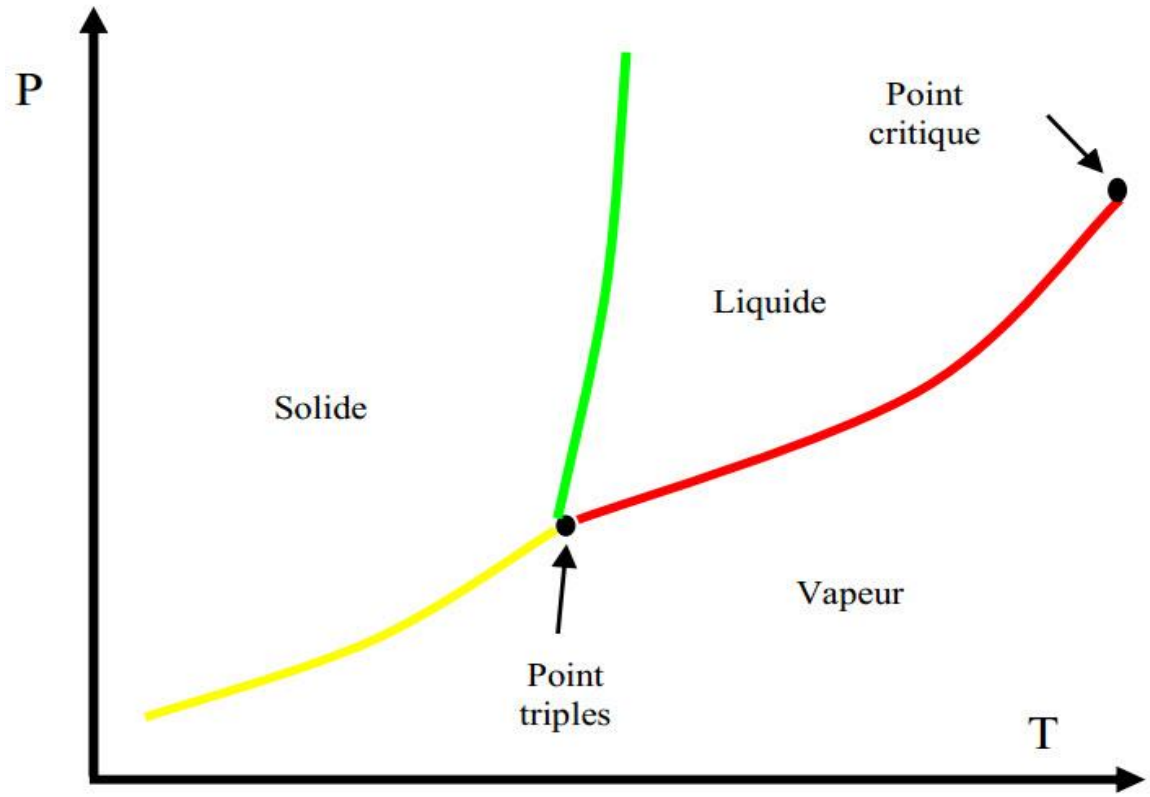


Figure I.5. Diagramme de coordonnées (P-T) des trois équilibres d'un corps pur.

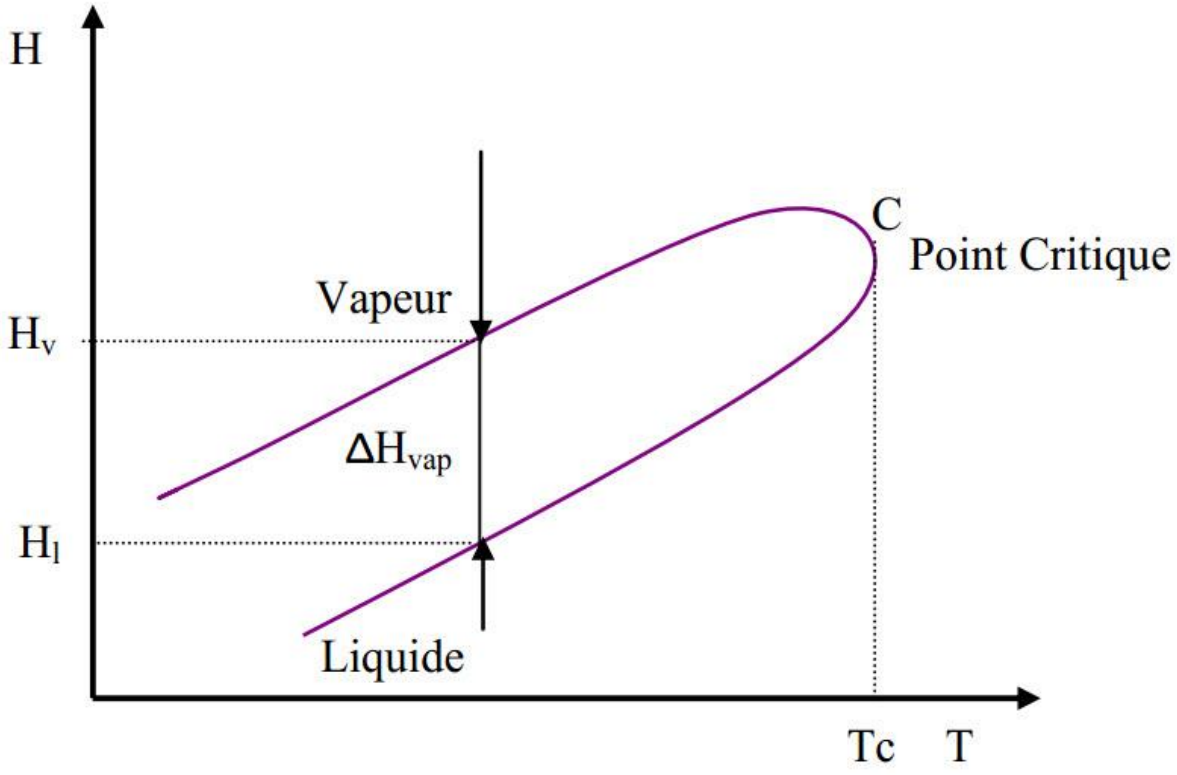


Figure I.6. Représentation de la chaleur de vaporisation pour un mélange de corps purs.

Cependant, malgré les différences entre les courbes P-V-T des corps purs et des mélanges de corps purs, cas de l'exemple du diagramme P-T de la **figure I.5** ; les deux cas d'équilibre entre la phase liquide et la phase gazeuse sont impossible au-delà d'un point "C" appelé point critique. Ce point, qui a pour coordonnées les trois propriétés critiques : température critique (**Tc**), pression critique (**Pc**) et volume critique (**Vc**), qui constitue la limite au-dessus de laquelle la coexistence des deux phases liquide et vapeur n'est plus possible et le changement de phase n'est plus observé. Il correspond à l'identité parfaite des propriétés du liquide et de la vapeur : densité, indice de réfraction, etc., en particulier la chaleur de vaporisation devient nulle **(Rebas 2006)**. Les gaz au-dessus de sa température critique ne peuvent pas être liquéfiés avec un simple changement de pression. En outre, les liquides au-dessus de sa pression critique ne peuvent pas être transformés en phase gazeuse, quelle que soit la température. Le fluide au-dessus de son point critique est appelé fluide supercritique. Le volume critique (**Vc**) est le volume molaire à sa température et pression critiques (**Tc** et **Pc**) **(Riazi, 2005)**.

Les propriétés critiques des hydrocarbures ou des fractions pétrolières sont très importantes pour la détermination des conditions de phase existantes, des plages de fonctionnement admissibles des réacteurs et des équipements de transfert de matière tels que les colonnes de distillation, les absorbeurs et les colonnes d'extraction **(Daubert & Denner 1997)**. Ces propriétés sont également utilisées pour calculer certaines propriétés physiques, thermodynamiques et de transport des hydrocarbures et des fractions pétrolières.

Deux techniques sont généralement utilisées pour la mesure des propriétés critiques, à savoir : la méthode d'écoulement et la méthode d'ampoule. Ces deux méthodes sont basées sur le principe de l'opalescence critique, qui consiste à évaluer la température, la pression et le volume critiques d'une substance lorsque le ménisque séparant les phases liquide et vapeur disparaît, laissant place à une bande d'un brouillard opalescent. Une technique de vidéo a permis de visualiser l'image de la cellule et un ordinateur enregistrait la propriété critique mesurée. Pour les substances instables thermiquement, il faut effectuer les mesures de température et de pression lorsque le temps de séjour dans la cellule est très faible et ceci afin de limiter les réactions de décomposition **(Rebas 2006)**. Comme la détermination expérimentale de ces propriétés nécessite un appareillage coûteux et sophistiqué ainsi qu'une manipulation longue et délicate, il est souvent nécessaire de faire appel à des techniques de modélisation pour les estimer.

I.4.3. Autres propriétés physicochimiques

I.4.3.1. Poids moléculaire

Le poids moléculaire d'un composé désigné par **M**, est la somme des poids atomiques des différents éléments contenus dans le composé. Il est utilisé pour convertir des quantités molaires en quantités massiques pour des applications pratiques. C'est l'une des propriétés importantes utilisée pour la caractérisation des hydrocarbures purs et des fractions pétrolières. Il est généralement exprimé en **Kg/Kmol**, en **g/mol** ou en **lb/mol**.

I.4.3.2. Température d'ébullition

Le point d'ébullition d'une fraction pétrolière désignée par **Tb**, est la température à laquelle elle commence à bouillir et passer de l'état liquide à l'état de vapeur sous pression atmosphérique **(Wauquier 1998)**. En d'autres termes pour une fraction pétrolière, la température d'ébullition est la température à laquelle la pression de vapeur est égale à la pression atmosphérique. Il est utilisé pour la détermination des autres paramètres ou propriétés physiques tels que la volatilité, le poids moléculaire, le facteur acentrique (**ω**), les propriétés critiques, et l'indice de réfraction (**η**).

I.4.3.3. Pression de vapeur

A l'état d'équilibre thermodynamique la pression de vapeur (P^{vap}) d'une substance particulière est la force exercée par unité de surface de vapeurs, de liquide ou de solide présentée dans un système **(Daubert & Danner 1997)**. La pression de vapeur est la propriété qui indique la volatilité d'un liquide. Elle est généralement utilisée dans le calcul des propriétés physiques, la perte de vapeur et la plage d'inflammabilité des liquides. Plus la pression de vapeur d'un liquide est élevée plus le liquide est volatile. Généralement les liquides volatiles ont une pression critique plus élevée **(Riazi 2005)**.

I.4.3.4. Point d'aniline

Le point d'aniline est la température la plus basse à laquelle des volumes égaux d'aniline et de produit à analyser, par exemple une fraction pétrolière, sont complètement miscibles. La rupture de la miscibilité se manifeste par l'apparition d'un trouble net **(Rebas 2006)**. Le point d'aniline d'un hydrocarbure dépend du poids moléculaire et du nombre de carbone**,** la connaissance de ce paramètre est utilisée pour estimer la part des aromatiques dans les fractions pétrolières. Il nous permet de classer les hydrocarbures en différentes familles. En effet, un point

d'aniline élevé dénote une nature paraffinique, par contre un point d'aniline bas dénote une nature aromatique, les naphténiques se classent intermédiairement **(Daubert & Denner 1997 ; Rebas 2006 ; Belghit 2019)**.

I.4.4. Propriétés thermodynamiques

I.4.4.1. Enthalpie standard de formation

L'enthalpie standard de formation ou chaleur standard de formation, $(\boldsymbol{\Delta H°_f})$ est l'une des propriétés thermodynamiques fondamentales importantes des composés définie comme un changement d'enthalpie qui accompagne la formation de **1 mole** de composé dans son état standard à partir de ses éléments constitutifs dans leurs états standards (la forme la plus stable de l'élément à **1 atm** de pression et à la température spécifiée généralement **298 K** ou **25°C**). Le changement de $(\boldsymbol{\Delta H°_f})$ est principalement utilisé en thermochimie pour trouver le changement $(\boldsymbol{\Delta H_{reaction}})$. Elle est mesurée expérimentalement en soustrayant la somme des enthalpies standards de formation des réactifs de la somme des enthalpies standards de formation des produits. Cependant, il existe de nombreuses méthodes de calcul de $(\boldsymbol{\Delta H°_f})$ dans la littérature, parmi elles on trouve les modèles basés sur les méthodes de contribution des groupes qui sont largement utilisées **(Vatani *et al*. 2007)**.

I.4.4.2. Enthalpie de vapeur

Les attractions entre dipôles permanents ont une grande influence sur l'évaporation d'un liquide ou la condensation d'un gaz, les deux changements d'état sont toujours accompagnés par une variation d'énergie. La quantité de chaleur nécessaire pour faire passer une mole de substance de l'état liquide à l'état de vapeur (à la température d'ébullition et à une pression constante) est appelée enthalpie de vaporisation $(\boldsymbol{\Delta H°_{vap}})$.

I.4.5. Facteurs et indices spécifiques

I.4.5.1. Facteur de caractérisation de Watson

En 1933, Watson et Nelson ont proposé un paramètre de caractérisation appelé plus tard le facteur de caractérisation de Watson noté ($\mathbf{K_w}$), qui met deux propriétés physiques en corrélation comme indique l'**Eq (I.4)**, à savoir le point d'ébullition moyen **Tb** en **K** et la densité **SG** à 60 °F (15,56°C) **(Riazi 2005)**.

$$K_w = \left(\frac{1{,}8 \times T_b}{SG}\right)^{1/3} \quad \textbf{(I.4)}$$

Le facteur de caractérisation de Watson est également utilisé pour prédire différentes propriétés physiques des fractions pétrolières, telles que la viscosité cinématique, la densité API et le poids moléculaire.

I.4.5.2. Facteur acentrique

Le facteur acentrique désigné par (**ω**) est l'un des paramètres de caractérisation développé par Pitzer, qui est utilisé pour améliorer la précision des équations d'états correspondants pour les hydrocarbures lourds, il mesure la déviation des propriétés thermodynamiques d'un fluide particulier aux propriétés calculées à partir du principe des états correspondants (**Riazi 2005**), leur définition mathématique est indiquée par l'**Eq (I.5)**.

$$\omega = -\log_{10}\left(P_r^{vap}\right) - 1 \quad \textbf{(I.5)}$$

Où P_r^{vap} la pression de vapeur réduite définie par l'équation :

$$P_r^{vap} = \frac{P^{vap}}{Pc} \quad \textbf{(I.6)}$$

I.4.5.3. Facteur de compressibilité

En thermodynamique, le facteur de compressibilité noté (**Z**) est une grandeur intensive adimensionnelle, qui désigne le rapport du volume molaire d'un mélange ou d'un corps pur réel, liquide ou gazeux, au volume molaire du gaz parfait correspondant aux mêmes pressions, températures et compositions. Pour un gaz parfait, le facteur de compressibilité vaut **1**, quelle que soit la nature du gaz et les conditions de pression, température et composition. Pour un fluide réel, le facteur de compressibilité est généralement compris entre **0,2** et **1,2**, il dépend de la nature du fluide et des conditions opératoires (**Sutton 1985 ; Elsharkawy & Elkamel 2000**).

Le facteur de compressibilité est une variable utilisée notamment dans les équations d'état telles que les équations d'états cubiques ou l'équation de Viriel. Il est nécessaire dans de nombreux calculs d'ingénierie pétrolière. Il est, par exemple utilisé pour calculer la chute de pression due à l'écoulement de gaz à travers une conduite (**Elsharkawy *et al*. 2001**).

I.4.5.4. Indice de réfraction

L'indice de réfraction désigné par (**I**) parfois par (**η**) est le rapport entre la vitesse de la lumière (à une longueur d'onde spécifiée) dans l'air à une température et une pression donnée, et sa vitesse dans la substance (dans le fluide) soumise à l'essai, pour lequel l'indice de réfraction doit être mesuré **(Riazi 2005)**.

L'indice de réfraction peut également être défini comme étant le rapport entre le sinus de l'angle d'incidence et le sinus de l'angle de réfraction quand la lumière passe de l'air dans la substance soumise á l'essai. Il s'agit de l'indice de réfraction relatif. Si l'on souhaite avoir l'indice de réfraction absolu (c'est-à-dire par rapport au vide), il faut multiplier cet indice de réfraction relatif par **1,000 293**, qui est l'indice de réfraction absolu de l'air. La valeur numérique de l'indice de réfraction des liquides varie en sens inverse de la longueur d'onde de la lumière et de la température **(Riazi & Roomi 2001)**.

I.4.5.5. Indice d'octane

La qualité des divers carburants pétroliers dépend de leur composition et les types d'hydrocarbures présents dans le mélange. L'indice d'octane ou le nombre d'octane (NO) est l'une des caractéristiques des carburants pour moteurs à allumage par étincelle comme l'essence et le kérosène. Ce nombre indique la caractéristique antidétonante d'un carburant et dépend fortement du type d'hydrocarbure **(Albahri *et al*. 2002 ; Albahri 2003)**.

Le nombre d'octane d'un produit pétrolier dépend de la structure chimique de ses constituants. Les hydrocarbures ramifiés ont un meilleur comportement antidétonant que ceux qui ont une structure linéaire. Les risques de cliquetis du moteur diminuent avec l'augmentation de l'indice d'octane. En effet, les n-paraffines ont des NO faibles, les iso-paraffines, les oléfines et les aromatiques ont des NO élevés. Les naphtènes ont, quant à eux, des valeurs de NO intermédiaires. Ainsi, une essence de distillation directe possède un faible indice d'octane alors qu'une essence issue d'un reforming catalytique ou d'un cracking catalytique a un indice assez élevé **(Rebas 2016 ; Abdul Jameel *et al*. 2018)**.

L'indice que vous trouverez le plus souvent dans les stations de service est le RON (Research Octane Number) il reflète le comportement du carburant dans le cadre d'une utilisation normale. Il existe un autre indice : MON (Motor Octane Number) qui reflète le comportement du carburant en cas d'utilisation plus éprouvante (haut régime, conduite sur autoroute, remorquage, etc.). La **figure I.7** fournit, par ailleurs, une vue d'ensemble des

particularités de chaque famille d'hydrocarbures, en ce qui concerne le RON. Quelques informations sur les relations générales entre structure chimique et indice d'octane **(Rebas 2016)**.

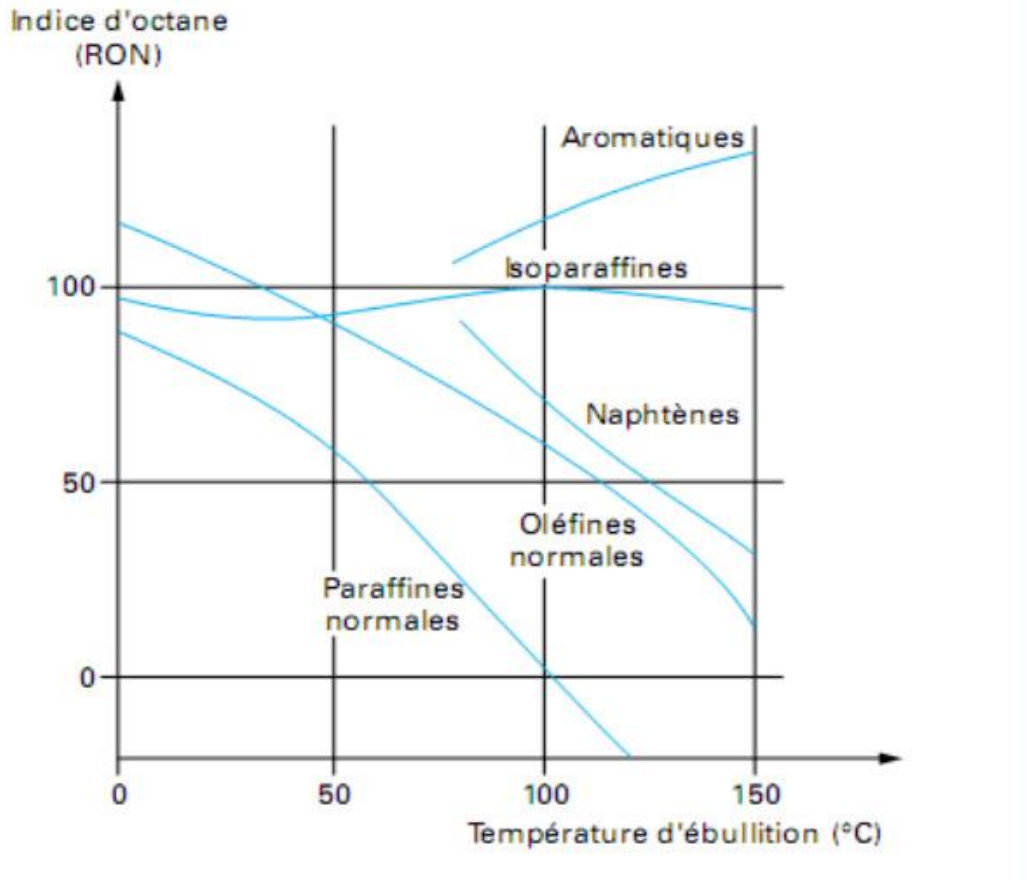


Figure I.7. Indices d'octane des différentes familles chimiques d'hydrocarbures.

I.4.6. Propriétés liées à la sécurité et le risque d'explosion

I.4.6.1. Point éclair

Le point éclair des hydrocarbures et des mélanges pétroliers est une donnée de sécurité importante pour le traitement et la manipulation de ces matériaux. Il est l'une des principales données de sécurité spécifiées dans la fiche de données de sécurité typique. Le point éclair des hydrocarbures ou de leurs mélanges est défini comme étant la température minimale à laquelle la pression de vapeur de l'hydrocarbure est suffisante pour produire la vapeur nécessaire pour enflammer l'hydrocarbure avec l'air en présence d'une source externe, à savoir une étincelle ou une flamme (**Atlas & Bartha 1992 ; Pan *et al*. 2007 ; Alqaheem & Riazi 2017**).

Cette caractéristique est un paramètre important pour les considérations de sécurité, en particulier lors de la production, le traitement, le stockage et le transport des produits pétroliers volatils et des combustibles (c'est-à-dire, GPL, naphta, essence, diesel) ainsi que des carburants non-pétroliers tels que les carburants dérivés du charbon et les biocarburants dans un environnement à haute température **(Atlas & Bartha 1992 ; Alqaheem & Riazi 2017)**.

Il indique le potentiel d'incendie et d'explosion d'un carburant. Un combustible à point éclair bas est un matériau présentant un risque d'incendie plus élevé. Le point éclair ne doit pas être confondu avec d'autres propriétés liées à la volatilité telles que la limite d'inflammabilité, la température d'auto-inflammation, le point de feu et la pression de vapeur **(Alqaheem & Riazi 2017)**.

I.4.6.2. Température d'auto-inflammation

La température d'auto-inflammation ; en anglais : the *Auto-Ignition Temperature* (AIT), est l'un des paramètres cruciaux d'inflammabilité qui nécessitent une attention particulière dans les industries et les processus qui traitent les matériaux dangereux et inflammables **(Borhani *et al.* 2016)**. En plus, elle est parmi les paramètres les plus importants utilisés pour classer les produits chimiques en fonction de leur degré d'inflammabilité et pour évaluer leurs capacités à former une atmosphère explosive **(Bagheri *et al.* 2012)**. La température d'auto-inflammation (AIT) est définie comme la température la plus basse à laquelle une substance s'enflamme dans l'air à la pression atmosphérique **(Bagheri *et al.* 2012 ; Borhani *et al.* 2016)**.

Les données expérimentales d'inflammabilité sont la principale source des informations de sécurité utilisées dans les évaluations des risques. Cependant, les mesures expérimentales des données AIT sont coûteuses, fastidieuses et chronophages, et pour les produits chimiques toxiques, volatils, explosifs et radioactifs, la mesure est plus difficile, voire impossible **(Pan *et al.* 2009 ; Katritzky *et al.* 2001 ; Bagheri *et al.* 2012)**.

I.5. Raffinage du pétrole et traitement des hydrocarbures

Le raffinage consiste à fractionner le pétrole brut en plusieurs coupes pétrolières puis à le transformer en produits à haute valeur commerciale, tels que l'essence, le kérosène, les gazoles ou les fuels domestiques pour les produits finis, l'exemple des naphtas pour les produits intermédiaires utilisées dans les industries pétrochimiques, ou certains produits de spécialité tels que les huiles lubrifiantes, les solvants et le bitume **(Rebas 2016)**.

Ces transformations se font grâce à une multitude de procédés ou opérations qui comprennent des unités de fractionnement tels que la distillation, les unités de procédés de transformation et de conversion, telles que le reformage, l'alkylation, le craquage catalytique ou l'hydrocraquage, et les unités de traitement, comme les procédés d'hydrotraitement ou le désasphaltage. Certains sont des procédés purement physiques comme l'exemple de la distillation et le désasphaltage, d'autres sont plutôt chimiques dans lesquels l'aspect chimique est fondamental, comme l'hydrotraitement ou les procédés de conversion **(Wauquier 1998)**.

I.5.1. Procédés de fractionnement

Les deux méthodes de distillations décrites ci-dessous sont les principaux procédés de fractionnement utilisés dans le raffinage du pétrole. Néanmoins, il existe de nombreuses unités de fractionnement associées aux procédés de transformation de façon à séparer leurs effluents en coupes pétrolières. Les procédés de fractionnement permettent de séparer un mélange pétrolier en fractions plus simples ou bien plus étroites. Dans le raffinage, les principaux procédés de séparation sont la distillation atmosphérique du pétrole brut et la distillation sous vide du résidu atmosphérique **(Wauquier 1998 ; Rebas 2016)**.

I.5.1.1. Distillation atmosphérique du pétrole brut

La distillation atmosphérique appelée parfois distillation primaire est considérée comme la principale et la première étape du raffinage. Une opération de dessalage (lavage à l'eau et à la soude) est effectuée pour enlever les minéraux, cette opération permet de fractionner le pétrole brut en différentes coupes pétrolières, appelées coupes Straight Run (SR). La coupe gaz est récupérée en tête de colonne, tandis que les coupes essence, kérosène et gazole sont obtenues sur les plateaux intermédiaires de la colonne. Ces coupes seront ensuite soumises à des procédés d'amélioration des propriétés et/ou à des procédés de finition pour obtenir les produits pétroliers à haute valeur commerciale. La fraction non distillée du brut, appelé aussi résidu atmosphérique (RA), est utilisée directement comme fuel industriel, ou envoyée à la distillation sous vide **(Wauquier 1998 ; Rebas 2016)**.

I.5.1.2. Distillation sous vide du résidu atmosphérique

La distillation sous vide est une étape complémentaire de la distillation atmosphérique. En effet, le résidu atmosphérique comporte des molécules avec un point d'ébullition supérieur à 340-380°C. En dehors de cette gamme de températures, les molécules commencent à subir des

réactions de craquage thermique avant l'évaporation. C'est la raison pour laquelle ce type de distillation est effectué sous pression réduite **(Rebas 2016)**.

La fraction issue de cette opération récupérée en tête de colonne est appelée le distillat sous vide (DSV), sera ultérieurement soumise à des procédés de conversion afin d'obtenir des produits valorisables. La fraction restante non distillée correspond à la coupe résidu sous vide (RSV), elle contient la majeure partie des impuretés du brut, comme le soufre, l'azote, les métaux et les asphaltènes. Le RSV peut être employé comme charges pour les procédés de conversion, tels que l'hydroconversion, la cokéfaction et la viscoréduction, ou utilisé dans la production des fuels lourds et la fabrication des bitumes **(Wauquier 1998 ; Rebas 2016)**.

I.5.2. Procédés de transformation et traitement des coupes pétrolières

Les procédés de transformation appliqués aux coupes légères ou moyennes consistent surtout à améliorer les propriétés des coupes, notamment des essences. Pour cela, la structure des molécules est réarrangée par isomérisation ou cyclisation ou elle est agrandie par des réactions d'alkylation ou de polymérisation. Les principales unités sont le reformage catalytique, l'isomérisation, l'alkylation, l'oligomérisation et l'éthérification **(Wauquier 1998 ; Rebas 2016)**. Le **tableau I.8** montre les principaux procédés de transformation appliqués aux coupes pétrolières légères ou moyennes.

Tableau I.8. Principaux procédés de transformation appliqués aux coupes pétrolières légères ou moyennes **(Rebas 2016)**.

Procédés	Objectifs	Charges	Produits	Conditions
Reformage catalytique	Augmenter l'indice d'octane des essences. Produire de l'hydrogène qui est essentiel dans les autres procédés de raffinage.	Coupes pétrolières légères contenant une haute teneur en paraffines et naphtènes. Fraction plus lourde des essences SR (C7-C10). Coupes essences provenant des procédés de conversion.	Composés aromatiques.	Basse ou moyenne Pression (4 à 40 bars). Haute Température (510 à 530°C). Présence d'un catalyseur solide.
Isomérisation	Augmenter l'indice d'octane des essences par isomérisation des n-pentanes (C5) et n-hexanes (C6). Isomériser des n-butanes (C4) qui seront utilisés comme charges des autres procédés. Transformer les n-paraffines contenues dans les coupes légères en iso-paraffines.	Provient principalement de la distillation directe du pétrole brut ou du reformage catalytique. Les essences légères issues des procédés de conversion, comme le craquage catalytique, sont aussi traitées.	Composées iso-paraffiniques.	Présence d'un catalyseur. Basse Température (120 à 270°C). Moyenne Pression (15 à 30 bars).
Alkylation	Augmenter l'indice d'octane des essences par production d'iso-paraffines.	Oléfines à 3, 4 ou 5 atomes de carbone provenant surtout du craquage catalytique. Isobutane (C4) issu des gaz de raffinerie ou du procédé d'isomérisation du n-butane.	Composées iso-paraffiniques à 7 et 8 atomes.	Basse Température (0 à 40°C). Basse Pression (1 à 10 bars). Présence d'un catalyseur liquide très acide, comme l'acide sulfurique.
Ethérification	Production des éthers par réaction d'une oléfine sur un mono-alcool aliphatique. Les éthers sont utilisés comme additifs pour augmenter l'indice d'octane des essences.	Mono-alcool aliphatique, comme le méthanol ou l'éthanol. Composés oléfiniques issus du craquage catalytique comme l'isobutène et l'isopentène.	Les éthers comme : Méthyl tert-butyl éther (MTBE) et Tert-amyl méthyl éther (TAME).	Basse Température (40 à 90°C). Moyenne Pression (7 à 20 bars). Présence d'une résine échangeuse d'ions qui catalyse la réaction.
Oligomérisation	Polymérisation qui forme des polymères avec un nombre limité de monomères appelés oligomères. Produire des essences oléfiniques à partir des oléfines à 2 et 3 atomes de carbone.	Composés oléfiniques à 2 et 3 atomes de carbone provenant essentiellement du craquage catalytique.	Essences oléfiniques.	Température variant entre 50 et 200°C. Moyenne Pression (30 à 40 bars). Présence d'un catalyseur qui peut être solide ou liquide.

Les procédés de conversion visent à transformer les coupes lourdes du pétrole en produits dont la demande du marché est forte, comme l'essence, le kérosène ou gazole. Pour cela, les molécules de coupes situées au-dessus de 350°C sont fragmentées en molécules de plus faible taille. Cette transformation peut s'effectuer soit à l'aide de procédés catalytiques, comme

l'hydrocraquage, le craquage catalytique ou l'hydroconversion, soit par une voie thermique à travers des procédés comme la viscoréduction ou la cokéfaction **(Rebas 2016)**. Le **tableau I.9** montre les principaux procédés de conversion appliqués aux coupes pétrolières lourdes.

Tableau I.9. Principaux procédés de conversion appliqués aux coupes pétrolières lourdes **(Rebas 2016)**.

Procédés	Objectifs	Charges	Produits	Conditions
Hydrocraquage	Transforme les coupes de (DSV) en composés de poids moléculaire plus faible en présence d'hydrogène. Coupure de liaisons (carbone–carbone) ou application des réactions d'isomérisation. Eliminer les impuretés de la coupe, comme la désulfurisation, désazotation, et réduire la teneur en aromatiques par les réactions d'hydrotraitement.	Proviennent soit directement de la distillation, soit des procédés de conversion, comme la cokéfaction ou l'hydroconversion.	Kérosène. Gazole. Essences.	Variant entre 350 à 400°C de Température. Et entre 50 à 200 bars de Pression. Présence d'hydrogène. Présence d'un système complexe de catalyseurs d'hydrotraitement et d'hydrocraquage.
Craquage catalytique	Fragmenter des hydrocarbures de coupes lourdes en composés de masse molaire plus faible en absence d'hydrogène.	Coupes DSV. Huiles désasphaltées. Coupes résidues d'hydrocraquage sont également admises.	Essences de bonne qualité. Coupe gazole très oléfiniques (gazole LCO). Formation de coke.	Haute Température (500 à 540°C). Basse pression (2 à 3 bars). Présence d'un catalyseur acide. Absence d'hydrogène.
Hydroconversion	Transforme les coupes de résidu atmosphérique ou résidu sous vide en composés de poids moléculaire plus faible en présence d'hydrogène, en éliminant une grande partie des impuretés, comme le soufre, l'azote ou les métaux.	Composés très réfractaires, comme les asphaltènes.	Coupe DSV. Gazole bien hydrogéné de qualité moyenne. Fraction essence. Résidu non-converti.	Haute Pression (100 à 200 bars). Haute Température (400 à 450°C). Présence d'hydrogène.
Viscoréduction	Réduire la viscosité des coupes lourdes après traitement à cause des composés oléfiniques instables et de forte teneur en impuretés.	Coupe de Résidus atmosphériques. Coupe de résidus sous vide.	Gazole. Essence. Fuels industriels avec une viscosité réduite.	Température entre 430 et 490 °C. Absence de catalyseur.
Cokéfaction	Conversion thermique pour fabriquer des produits légers.	Coupes très lourdes. Pétrole brut lourd.	Produits liquides ; essences et gazoles très oléfiniques avec une haute teneur en impuretés. Quantité importante de gaz. Production de coke qualité électrode ou coke combustible.	Haute Température (480 à 510°C). Pression atmosphérique. Absence de catalyseur.

Les procédés de traitement visent d'une part à améliorer les caractéristiques des produits finis en termes de spécifications de qualité et d'environnement et d'autre part à préparer les mélanges d'hydrocarbures pour d'autres unités de transformation ou de conversion. Le traitement peut inclure l'élimination et/ou la séparation de composés aromatiques et de naphtènes ainsi que des contaminants indésirables, comme le soufre, l'azote, les métaux. Le traitement peut impliquer des réactions chimiques, la séparation chimique ou physique, comme la dissolution, l'absorption ou la précipitation. Les principaux procédés de traitement de mélanges pétroliers sont l'hydrotraitement, le désasphaltage, le procédé d'adoucissement **(Rebas 2016)**. Le **tableau I.10** montre les principaux procédés de traitement appliqués aux produits pétroliers finis.

Tableau I.10. Principaux procédés de traitement appliqués aux produits pétroliers finis **(Rebas 2016)**.

Procédés	Objectifs	Charges	Produits	Conditions
Hydrotraitement	Eliminer des impuretés comme le soufre, l'azote ou les métaux et réduire la teneur en aromatiques.	Appliquer sur toutes les coupes pétrolières de la raffinerie.	Produits finis conforme aux spécifications comme l'hydrotraitement de gazoles et de kérosènes. Produits prétraités utilisés comme charges pour les autres procédés de conversion.	Pression variant entre 10 et 200 bars. Température variant (entre 250 et 450°C). Présence d'hydrogène et d'un catalyseur solide.
Désasphaltage	Récupérer les derniers hydrocarbures encore potentiellement utilisables du RSV par extraction au solvant.	Coupes des RSV. Solvants paraffiniques légères comme le propane et le butane.	Huile désasphaltée (DAO). Asphalte qui constitue le résidu ultime du brut pétrolier.	Varient selon le solvant utilisé.
Adoucissement	Eliminer les mercaptans présents dans les coupes pétrolières par oxydation. Transformations des mercaptans en disulfures par oxydation.	Coupes GPL et Kérosènes. Produits légers de craquage.	Coupes pétrolières sans mercaptans.	Présence d'un catalyseur en milieu solide.

Il existe d'autres procédés de traitement mineurs comme le dessalage, l'extraction des huiles lubrifiantes ou l'adsorption dans le cadre de la purification de l'hydrogène. En dehors des procédés de fractionnement, de transformation et de traitement, il existe de nombreux procédés qui ne sont pas directement liés au raffinage du pétrole, mais qui sont également nécessaires au

bon fonctionnement de la raffinerie. Ces procédés incluent les unités de protection de l'environnement, comme le traitement des gaz et le traitement des eaux résiduaires, l'unité de Claus ou les unités de production d'hydrogène **(Rebas 2016)**.

I.6. Aperçu sur les méthodes de la détermination des propriétés des hydrocarbures

La modélisation d'un processus chimique requière la connaissance de quelques propriétés physicochimiques. Habituellement ces propriétés peuvent être trouvées dans des bases de données ou dans la littérature. Mais, il arrive que nous soyons confrontés à des cas où ces propriétés ne sont pas disponibles. Dans le cas idéal ces propriétés doivent être mesurées dans le laboratoire mais à cause de multiples raisons ; temps de manipulation long, produits chimiques purs, ou bien protocole opératoire indisponible. Ces propriétés ne peuvent être obtenues expérimentalement, en conséquence elles doivent être estimées ou prédites **(Rebas 2016)**.

Il existe différentes méthodes de caractérisation des hydrocarbures, certaines sont concerrnais à la caractérisation des composés chimiques des hydrocarbures comme l'exemple des méthodes physique d'analyse telles que les méthodes chromatographiques (CPG, CPL), les méthodes spectroscopiques (IR, UV, RMN) et les méthodes de couplage chromatographies et méthodes spectrales (GC, MS), sans oublier les essais expérimentaux normalisés **(Rebas 2016)**.

Indépendamment de l'expérience ou du raisonnement thermodynamique, une propriété peut être déterminée par des méthodes d'estimation ou de prédiction, les unes étant plus précises que les autres **(Rebas 2016)**. Elles sont généralement appliquées pour la détermination des propriétés physicochimiques et thermodynamiques, tels que les méthodes basées sur les équations d'état et les méthodes des états correspondants **(Si-Moussa 2010)**, les méthodes empiriques **(Belghit *et al*. 2018)** et les méthodes de contributions de groupes **(Farzi & Esmaeilzadeh 2016 ; Rebas 2016)**. Une attention particulière est accordée aux méthodes des réseaux de neurones artificiels (RNAs) (en anglais : *artificial neural networks* ou ANN), et les méthodes des relations quantitatives structures propriétés (RQSP) (en anglais : *quantitative structure-property relationship* ou QSPR) ; qui seront présentées en détails dans le chapitre suivant, puisque ce sont les principales méthodes de prédiction appliquées dans ce travail de thèse.

I.7. Conclusion

Dans ce chapitre j'ai présenté les concepts de base et des généralités sur les hydrocarbures purs, le pétrole brut et ses fractions pétrolières, comme les définitions importantes, leurs classifications et les différentes compositions. De même, des notions très importantes ont été présentées qui permettent de connaître les différentes caractéristiques physicochimiques des hydrocarbures purs, des pétroles bruts et de leurs fractions pétrolières. L'opération courante de séparation des pétroles bruts, le procédé de raffinage a également été brièvement présenté dans ce chapitre.

Comme il est connu, la détermination des propriétés des hydrocarbures est une tâche difficile en raison du temps et des matériels consommables pour le faire expérimentalement. Cependant, un aperçu sur les méthodes de la détermination des propriétés des hydrocarbures, ses principales techniques et certaines méthodes utilisées pour prédire ces propriétés ont été présentées. En fin de compte, je peux dire que le domaine des hydrocarbures et un domaine très large et qu'il ne peut pas tout expliquer et tout donner en un seul chapitre, mais heureusement il existe des thèses et des livres références qui abordent et expliquent ce domaine d'une manière plus profonde.

Chapitre II

Approches de base, méthodologie de développement, validation et application des méthodes RQSP/RQSA

II.1. Introduction

Les propriétés physicochimiques de base et les caractéristiques structurelles d'une molécule déterminent son comportement dans les processus physiques, chimiques, biologiques ou environnementaux et sont donc intéressantes pour comprendre et modéliser l'action des composés **(Engel & Gasteiger 2018)**.

Une relation quantitative structure-propriété (RQSP) (en anglais : *quantitative structure-property relationship* ou QSPR), aussi bien que la relation quantitative structure-activité (RQSA) (en anglais : *quantitative structure-activity relationship* ou QSAR) est le procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique **(Chtita 2017)**. La méthodologie de la RQSP tente de modéliser la relation entre une structure chimique et diverses propriétés physiques ou chimiques. De manière similaire, la RQSA tente d'établir un lien quantitatif entre la structure moléculaire d'un composé et son activité biologique. À travers des descripteurs moléculaires **(Chtita 2017 ; Engel & Gasteiger 2018)**.

Les descripteurs moléculaires peuvent être dérivés en appliquant les principes de différentes théories, telles que la chimie quantique, l'informatique, la chimie organique et la théorie des graphes (**Todeschini & Consonni 2009 ; Engel & Gasteiger 2018)**. Ils sont utilisés pour modéliser une large gamme de propriétés dans tous les domaines de la chimie et de la biologie, par exemple les sciences de l'environnement, la toxicologie, la chimie analytique, la chimie physique, la chimie médicinale et les sciences de la vie. En plus des descripteurs théoriques classiques (calculés), il existe également des descripteurs dérivés expérimentalement (par exemple, logP, moment dipolaire) qui sont parfois utilisés dans les études RQSA **(Engel & Gasteiger 2018)**. Dans ce qui suit, je m'intéresse uniquement aux descripteurs moléculaires théoriques.

Ce chapitre donne un aperçu sur l'approche RQSP/RQSA. Les différentes méthodologies de développement, de validation et l'application de ces méthodes sont aussi présentées.

II.2. Approche relations quantitatives structures propriétés/activités RQSP/RQSA

II.2.1. Historique

La première formule générale du RQSP est apparue il y a plus d'un siècle et demi, exactement en 1863, quand Cros **(Cros 1863)** a remarqué que le point d'ébullition et le point de fusion des alcanes augmente avec le nombre d'atomes de carbone et la masse moléculaire. Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire. De même, dès 1868, Crum-Brown et Fraser **(Brown & Fraser 1968 ; Brown & Fraser 1969)** ont assumé qu'il y a un lien entre l'activité biologique d'une molécule est sa structure chimique. Peu de temps après, Richet en 1893 (**Richet 1893)**, Meyer en 1899 (**Meyer 1899)**, et Overton en 1901 (**Overton 1901)** ont découvert séparément une corrélation linéaire entre la lipophilie (par exemple, les coefficients de partage huile-eau) et les effets biologiques (par exemple, les effets narcotiques et la toxicité). L'année 1964 est considérée comme l'année de naissance des méthodes RQSP/RQSA modernes, grâce aux travaux de Hansch et Fujita, qui ont établi les premiers modèles mathématiques pour corréler les propriétés physico-chimiques (***logP***, ***pKa***) et/ou l'activité biologique (activité enzymatique, pharmacologique) d'une part, et leurs structures chimiques de l'autre part. Ces méthodes seront appelées par la suite l'analyse de Hansch (**Hansch & Fujita 1964)** et l'analyse de Free Wilson **(Free & Wilson 1964)**. Depuis lors, un grand nombre de modèles a été conçu pour la prédiction de diverses propriétés et/ou activités d'intérêt biologique.

II.2.2. Domaines d'application et utilité des méthodes RQSP/RQSA

Pendant ces dernières décennies, les méthodes RQSP/RQSA peuvent être trouvés dans toutes les disciplines où des produits chimiques sont utilisés ou étudiés ; telles que le domaine pharmaceutique, la pétrochimie, les sciences de l'environnement et d'autres. Les modèles établis permettent de prédire les propriétés ou les activités de nouveaux composés **(Hamadache 2014)**. De plus, les descripteurs moléculaires et les techniques de modélisation appropriées peuvent contribuer à la compréhension et à la rationalisation des relations entre la structure moléculaire et une propriété ou une activité dans une série de molécules **(Dearden *et al*. 2009 ; Engel & Gasteiger 2018)**. En outre, l'utilisation des méthodes RQSP/RQSA à de nombreuses utilités ; nous donnons ci-après, à titre d'exemple, quelques scénarios typiques d'utilisation des modèles RQSP/RQSA dans divers domaines **(Dearden *et al*. 2009 ; Hamadache 2014 ; Engel & Gasteiger 2018)** :

- Prédiction des différentes propriétés physico-chimiques et activités biologiques à partir de leurs structures chimiques.

- Identification des composés dangereux dès les premiers stades de développement de tous types des produits.
- Comprendre et rationaliser les mécanismes d'action dans une série de produits chimiques.
- Réduire le coût de développement des différents produits industriels tels que les produits pharmaceutiques, les hydrocarbures, et les pesticides, …
- Optimisation des propriétés pharmacocinétiques (ADME) des médicaments.
- Diminuer ou même remplacer les tests longs et coûteux sur les animaux.
- Conception et prédiction de la toxicité et des effets secondaires des nouveaux composés pour l'homme et les espèces environnementales.
- Prédire les effets ou la toxicité non ciblés de nouveaux composés.
- Prévision du devenir des composés rejetés dans l'environnement.

II.2.3. Principe et méthodologie de base des études RQSP/RQSA

Les méthodes RQSP/RQSA sont basées sur l'hypothèse que la propriété ou l'activité d'un composé chimique est liée à sa structure. Par définition, Une RQSP/RQSA est un modèle mathématique qui associe un ou plusieurs paramètres quantitatifs dérivés de la structure chimique, à une mesure quantitative d'une propriété chimique ou d'une activité biologique **(Chtita 2017)**.

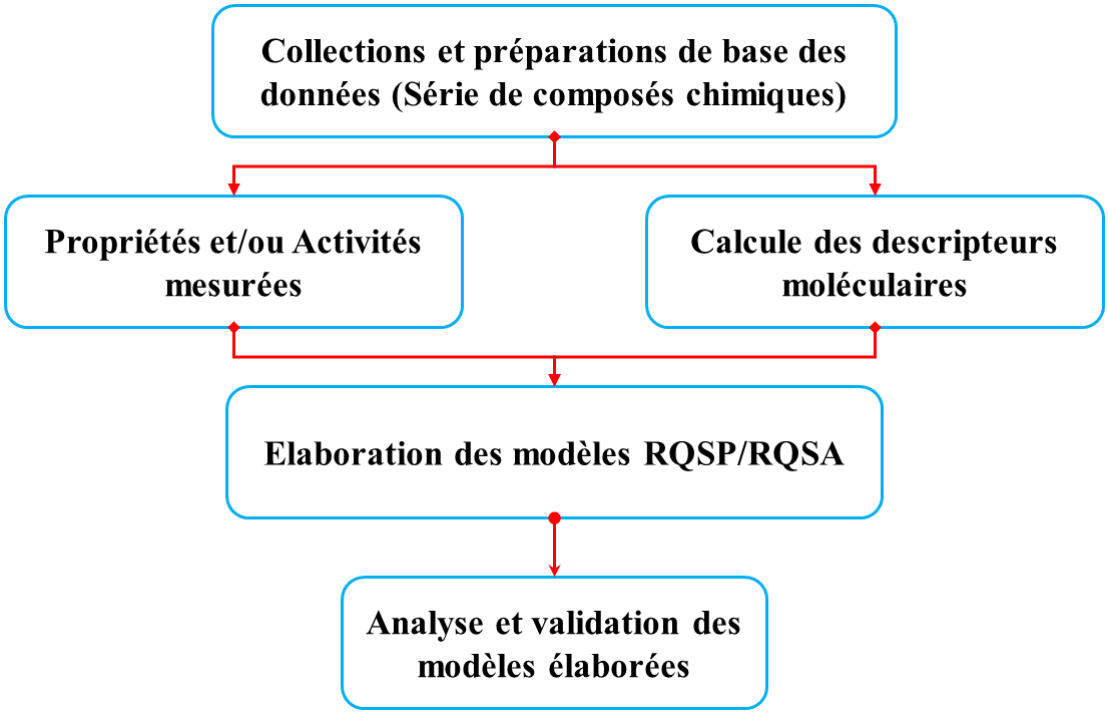


Figure II.1. Principe et méthodologie de base des modèles RQSP/RQSA.

Le principe de base adopté dans toutes études RQSP/RQSA (**Figure II.1**), consiste à trouver une relation mathématique reliant de manière quantitative un effet bien déterminé, comme une propriété chimique, ou une activité biologique, mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques **(Chtita 2017)**.

L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour la propriété ou l'activité étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employés, et sont étudiés dans la section **II.6** de ce chapitre (voir plus loin la partie II.6. Méthodes d'apprentissage statistiques).

L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de la propriété ou l'activité étudiée pour de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas disponibles. Ceci peut être traduit par l'équation suivante :

$$\boldsymbol{Propriété/Activité} = \boldsymbol{f}(\boldsymbol{Descripteurs\ Moléculaires})$$

II.3. Codage et représentation des structures chimiques

Les molécules sont représentées par différentes structures chimiques, ces structures sont figurées par des écritures de forme développées et semi-développées. Pour quantifier ces structures chimiques lors de la modélisation, il est nécessaire de les coder préalablement pour les rendre compréhensibles par les outils informatiques. Par la suite les informations structurelles seront extraites pour être utilisées lors de la construction du modèle de type RQSP/RQSA **(Roubehie Fissa *et al.* 2019)**. Il existe de nombreuses manières ou langages de codage, telles que la théorie des graphes qui est considérée comme une source d'inspiration pour la création de plusieurs types de descripteurs moléculaires, notamment les descripteurs topologiques **(Balaban 1985 ; Bonchev 1991 ; Randić 1991 ; Lahiouel 1995a)**.

II.3.1. Utilisation de la théorie des graphes pour le codage des structures chimiques

La théorie des graphes est une branche des mathématiques discrètes, liée à la topologie et à la combinatoire. Elle traite la façon dont les objets sont connectés et de toutes les conséquences de leur connectivité. La connectivité dans un système est donc une qualité fondamentale de la théorie des graphes **(Balaban 1985 ; Hansen & Jurs 1988 ; Mihalić & Trinajstić 1992)**. La théorie des graphes chimique est une branche de la chimie mathématique, et par conséquent de la

chimie théorique. Il s'agit de manipuler des graphes chimiques, c'est-à-dire des graphes qui représentent des systèmes chimiques. Par conséquent, la théorie des graphes chimiques traite l'analyse de toutes les conséquences de la connectivité dans un système chimique. En d'autres termes, la théorie des graphes chimique concerne tous les aspects de l'application de la théorie des graphes à la chimie **(Randić 1991 ; Mihalić & Trinajstić 1992)**.

Comme l'indique son nom, un graphe structurel correspond à une structure chimique spécifique. Le **Tableau II.1** fournit une liste des concepts de la théorie des graphes et les termes correspondants utilisés en chimie pour les types de graphes chimiques.

Tableau. II.1. Notions correspondant à la théorie des graphes et la théorie des graphes chimiques **(Hansen & Jurs 1988)**.

Théorie des graphes	Théorie des graphes chimiques
Point	Atome
Ligne	Liaison chimique
Chemin	Sous-structure chimique
Degré	Valence de l'atome
Graphique cyclique	Composé cyclique
Arbre	Structure acyclique

II.3.1.1. Représentations d'une structure par quelques éléments basés sur la théorie des graphes chimiques

Plusieurs représentations sont possibles pour les graphes. Un graphe chimique noté ($\boldsymbol{G}$) peut être associé à plusieurs vecteurs ou matrices théoriques très importantes, telles que la matrice d'adjacence des sommets (en anglais : *Vertex-Adjacency Matrix*), la matrice des distances (en anglais : *Distance Matrix*) et le vecteur des valences (en anglais : *Valences Vector*) **(Mihalić & Trinajstić 1992 ; Goulon-Sigwalt-Abram 2008)**. Le **Tableau II.2** montre un exemple de calcul de ces éléments pour un graphe moléculaire ($\boldsymbol{G}$) correspondent à la structure du *2.3.4-triméthylhexane* présenté dans la **Figure II.2**, avec les sommets correspondant aux atomes et les lignes correspondent aux liaisons chimiques.

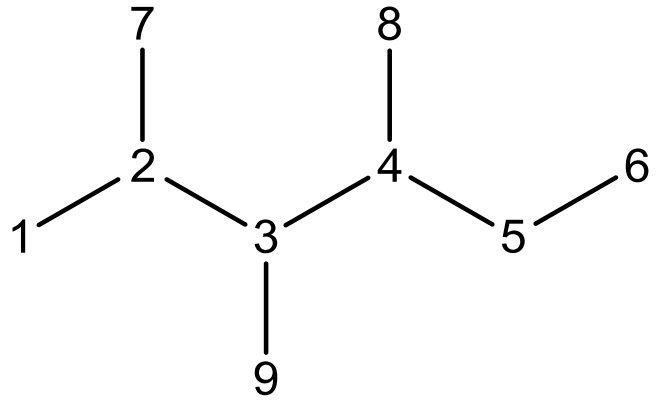


Figure II.2. Graphe moléculaire ($\boldsymbol{G}$) correspondant à la structure du *2.3.4-triméthylhexane*.

Tableau. II.2. Exemple de calcul des matrices et vecteur basé sur la théorie des graphes chimique corresponds à la structure du *2.3.4-triméthylhexane* **(Mihalić & Trinajstić 1992)**.

Matrice d'adjacence des sommets A($\boldsymbol{G}$)	**Matrice de distance D($\boldsymbol{G}$)**	**Vecteur des valences V($\boldsymbol{G}$)**
$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 2 & 4 & 3 \\ 1 & 0 & 1 & 2 & 3 & 4 & 1 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 & 2 & 2 & 1 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 1 & 2 \\ 4 & 3 & 2 & 1 & 0 & 1 & 4 & 2 & 3 \\ 5 & 4 & 3 & 2 & 1 & 0 & 5 & 3 & 4 \\ 2 & 1 & 2 & 3 & 4 & 5 & 0 & 4 & 3 \\ 4 & 3 & 2 & 1 & 2 & 3 & 4 & 0 & 3 \\ 3 & 2 & 1 & 2 & 3 & 4 & 3 & 3 & 0 \end{pmatrix}$	$(1\ 3\ 3\ 3\ 2\ 1\ 1\ 1\ 1)$

II.3.2. Notification SMILES et codage des structures chimiques

La première étape dans la formalisation de la chimie consiste à nommer un composé chimique. Cela nécessite une notation non ambiguë et reproductible pour le plus simple atome à la structure la plus compliquée.

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) est un système de notation chimique conçu pour le traitement de l'information chimique moderne. Sur la base des principes de la théorie des graphes moléculaires, la notation SMILES permet une spécification de structure rigoureuse en utilisant une grammaire simple et très pratique **(Weininger 1988)**. En raison de sa flexibilité, de sa facilité d'utilisation et de sa compatibilité avec les outils informatiques chimiques modernes, le système de notation SMILES est bien adapté au traitement des bases de données dans plusieurs études portant sur la modélisation, qui permettent de concevoir de

nombreuses applications informatiques chimiques hautement efficaces, notamment la génération d'une notation unique, la récupération de bases de données, en plus de contribuer à l'émergence de nombreux descripteurs moléculaires **(Weininger 1988)**.

Pour plus d'information sur la méthodologie et les règles de codage utilisées par le système de notation SMILES, on peut se référer aux travaux de Weininger et ses collaborateurs qui les ont introduits et expliqués en détails **(Weininger 1988 ; Weininger *et al*. 1989 ; Weininger 1990)**.

Comme montré dans cette partie, la théorie des graphes chimiques et divers systèmes de notations tels que la notation SMILES ont efficacement contribué à l'évolution de l'apparition et l'utilisation des descripteurs moléculaires dans le domaine de la modélisation, notamment la modélisation de type RQSP/RQSA. Dans ce qui suit, quelques concepts de base sur les descripteurs moléculaires permettant de clarifier et de comprendre la signification d'un descripteur moléculaire, vont être donnés.

II.4. Notions de base sur les descripteurs

II.4.1. Définition

Un descripteur moléculaire est défini comme le résultat final d'une procédure logique et mathématique ou d'une expérience normalisée, qui transforme l'information chimique codée par une représentation symbolique d'une molécule en un nombre utile afin de pouvoir quantifier la molécule **(Todeschini & Consonni 2008 ; Mauri *et al*. 2017 ; Roubehie Fissa *et al*. 2019)**.

II.4.2. Classification

Les descripteurs peuvent être de différents types en fonction de leur méthode de calcul ou de détermination **(Roy *et al*. 2015)**. Historiquement, deux grands schémas pour la classification des descripteurs moléculaires ont été établis : l'un en fonction de leur origine (constitutionnel, topologique, géométrique, quantique, thermodynamique...), et l'autre basée sur leurs dimensionnalités (0D, 1D, 2D, 3D ou 4D) **(Karelson 2000 ; Dudek *et al*. 2006)**. Dans ce qui suit on donnera un aperçu sur ces catégories de descripteurs afin de bien clarifier les notions liées à chaque type.

II.4.2.1. Descripteurs 0-D

Les descripteurs moléculaires de classe ***0-D*** ne nécessitent aucune information structurelle liée à la connectivité des atomes, ils contiennent seulement quelques informations simples telles que le nombre d'atomes et de liaisons ainsi que la somme ou la moyenne des propriétés atomiques. Les descripteurs de cette classe peuvent toujours être facilement calculés sans utiliser les outils informatiques modernes, ne nécessitent pas d'optimisation des structures et sont aisément interprétables, leur contenu informationnel est faible, mais ils peuvent néanmoins jouer un rôle important dans la modélisation de plusieurs propriétés physico-chimiques ou prendre part à des modèles plus complexes **(Bonachera 2011 ; Chtita 2017)**.

II.4.2.2. Descripteurs 1-D

Les descripteurs moléculaires de classe ***1-D*** nécessitent quelques informations sur la structure moléculaire en ***1-D***, comme le dénombrement des groupes fonctionnels et des fragments de sous-structure. Pareillement que les descripteurs de classe ***0-D***, ces descripteurs ont une caractéristique similaire compte tenu de la simplicité du calcul, mais ils peuvent nécessiter des outils informatiques pour accélérer et faciliter la tâche de calcul, leur contenu informationnel est plus ou moins acceptable par rapport aux descripteurs de classe ***0-D***. Ils montrent, généralement, une dégénérescence moyenne ou élevée et sont souvent très utiles dans la modélisation à la fois des propriétés physico-chimiques et biologiques **(Tauler *et al*. 2009 ; Nieto-Draghi *et al*. 2015)**.

II.4.2.3. Descripteurs 2-D

A l'opposé des deux classes précédentes, les descripteurs ***2-D*** nécessitent des informations structurelles sur les molécules notamment ceux qui sont liées à la connectivité des atomes et certains fragments moléculaires. De ce fait, toutes les molécules sont représentées sous formes des graphes chimiques, ensuite les informations nécessaires pour calculer les descripteurs de ce type sont extraites. C'est à partir de ce niveau que l'on peut espérer capturer des informations pertinentes pour la prédiction de la majorité des propriétés moléculaires **(Karelson 2000 ; Chtita 2017)**.

Parmi les descripteurs les plus importants appartenant à cette classe, on trouve les indices topologiques qui ont été développé sur la base de la théorie des graphes chimiques, comme

l'exemple de l'indice de Wiener "***W***" fondé en 1947 **(Wiener 1947)**, qui compte le nombre total des liaisons dans les chemins les plus courts entre les paires d'atomes **(Wiener 1947 ; Mihalić & Trinajstić 1992)**, l'exemple de l'indice de Randić "χ" fondé en 1975 **(Randić 1975)**, qui se base sur la somme des bords pondérés dans un graphe moléculaire **(Randić 1975 ; Mihalić & Trinajstić 1992)** et l'exemple de l'indice de Balaban "***J***" fondé en 1982 **(Balaban 1982)**, qui compte la somme des distances pondérées dans une molécule **(Balaban 1982 ; Mihalić & Trinajstić 1992)**.

II.4.2.4. Descripteurs 3-D

Ce type de descripteurs dépend essentiellement de la conformation tridimensionnelle des molécules. Leurs évaluations nécessitent de connaître les positions relatives de leurs atomes dans l'espace et de décrire des caractéristiques plus complexes. Une bonne connaissance de la modélisation "empirique" ou "ab-inition" ainsi que de la géométrie ***3-D*** des molécules concernées, sont souvent les clés du calcul de ce type de descripteurs. Malgré le développement remarquable des outils informatiques, la plupart de ces descripteurs s'avèrent relativement coûteux en temps de calcul, mais ils fournissent plus d'informations par rapport aux autres classes précédentes, ils sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent principalement de la structure ***3-D*** **(Kraim 2009 ; Chtita 2017)**. Il existe plusieurs familles de descripteurs appartenant à cette classe ***3-D***, le plus souvent on trouve :

- Les descripteurs géométriques qui peuvent être obtenus expérimentalement ou par modélisation, empirique ou ab-inition. Ils sont basés sur l'arrangement spécial des atomes constituant la molécule et sont définis par les coordonnées des noyaux atomiques et la taille de la molécule représentée **(Kraim 2009 ; Chtita 2017)**. Parmi ceux qui sont les plus importants, on trouve le volume moléculaire, la surface et le volume de Van Der Waals **(Higo & Gō 1989 ; Labute 2000 ; Saunders & Platts 2004)**.
- Les descripteurs quantiques électroniques qui caractérisent la distribution des charges moléculaires c'est-à-dire la polarité des molécules, mais aussi les paramètres de la chimie quantique qui nécessitent souvent des calculs plus sophistiqués. Les approches de la chimie quantique donnent accès à des informations supplémentaires telles que les données structurales, énergétiques, électroniques et spectroscopiques des molécules **(Kraim 2009 ; Chtita 2017)**. Parmi les plus importants on trouve l'énergie totale, le moment dipolaire, les énergies des orbitales frontières qui sont des paramètres très populaires dans la chimie quantique et dans les études RQSP/RQSA **(Kikuchi 1987)**, comme l'exemple des énergies

"HOMO" et "LUMO" qui jouent un rôle majeur dans de nombreuses réactions chimiques et dans les mécanismes réactionnels **(Fukui 1970 ; Zhou & Parr 1990 ; Lewis *et al*. 1994)**.

- Les descripteurs physico-chimiques et thermodynamiques qui sont à l'origine des propriétés sont utilisés comme descripteurs introduits dans un modèle afin de prédire d'autres propriétés ou activités des systèmes étudiées **(Kraim 2009 ; Chtita 2017)**. La plupart de ces descripteurs sont obtenus expérimentalement, le plus souvent on trouve des paramètres comme le coefficient de partage octanol/eau, le parachor, la tension superficielle, la constante d'Henry, l'énergie libre de Gibbs et la chaleur de formation ou bien l'enthalpie de formation **(Sugden 1924 ; Viswanadhan *et al*. 1989 ; Bosque *et al*. 2003)**.

II.4.2.5. Descripteurs 4-D

Ils correspondent à la mesure des propriétés ***3-D*** comme le potentiel électrostatique d'une molécule en tout point de l'espace. Ils permettent d'avoir des informations sur la structure cible souvent très complexes comme les protéines. Les descripteurs ***4-D*** sont obtenus par le calcul des champs d'interactions moléculaires (CoMFA, CoMSIA) entre deux molécules **(Ravi *et al*. 2001 ; Martins *et al*. 2009 ; Andrade *et al*. 2010 ; Chtita 2017)**. On peut ainsi distinguer les descripteurs ***4-D*** qui nécessitent un alignement de la molécule guidé par l'étude des complexes ligand-cibles avant d'être calculés ou, au moins, par des contraintes visant à optimiser le recouvrement spatial des champs électriques et stériques des ligands qui compensent le manque d'information sur le vrai mode de fixation dans la cible **(Cramer *et al*. 1988 ; Navajas *et al*. 1996 ; Chtita 2017)**.

II.4.3. Généralisation des descripteurs

L'une des raisons les plus importantes pour l'augmentation du nombre de descripteurs moléculaires au cours des dernières décennies est le développement énorme du domaine mathématique-informatique, qui a conduit à l'apparition de nombreux logiciels dédiés au calcul de plusieurs types de descripteurs moléculaires **(Cherkasov 2014 ; Chtita 2017)**. Il existe plusieurs logiciels pour la généralisation des descripteurs moléculaires, tel que le logiciel "**Dragon**" le plus utilisé actuellement. Sa nouvelle version, "**Dragon 7.0**" offre 5270 descripteurs **(Mauri *et al*. 2006)**. La version gratuite en ligne, "**E-Dragon 1.0**" utilisée dans cette thèse permet de calculer environ 1666 descripteurs appartenant à différentes familles de descripteurs

(**Tetko *et al*. 2005**). Il existe d'autres logiciels et plateformes en ligne à accès libre comme la plateforme "**ChemDes**" (**Dong *et al*. 2015**) qui est une plateforme web intégrée pour le calcul des descripteurs moléculaires et des empreintes digitales. Il inclut aussi pas mal de logiciels de généralisation de descripteurs tel que "**PaDEL Descriptors**" (**Dong *et al*. 2015**).

L'avantage du calcul des descripteurs moléculaires au moyen de tous ces outils est confronté à un problème de taille qui consiste à choisir les meilleurs descripteurs.

II.5. Méthodes de sélection des descripteurs

L'élaboration des modèles RQSP/RQSA nécessite parfois le calcul d'un grand nombre de descripteurs moléculaires, car les facteurs du processus étudié ne sont pas connus à priori (**Chtita 2017**). Les descripteurs calculés sont différents les uns des autres par leur principe de calcul et leurs valeurs numériques pour une même molécule donnée. Ils n'ont donc pas tous une influence significative sur la grandeur à modéliser, et les variables ne sont pas toujours indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension de la base de données d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre des observations (molécules) de la base d'apprentissage, le modèle risque d'être sur-ajusté. Dans ce cas, elle est incapable de prédire la grandeur modélisée sur de nouvelles molécules et peut contenir des informations redondantes. Ce qui conduit à faire une réduction du nombre important de descripteurs et sélectionner les plus pertinents d'entre eux.

Il ne s'agit pas d'utiliser alors le minimum de descripteurs pour expliquer la propriété ou l'activité ciblée, et il ne faut pas perdre des informations utiles. Comme Einstein a dit : "Tout devrait être fait aussi simple que possible, mais pas plus simple.". Dans une modélisation RQSP/RQSA, ce principe d'Einstein signifie que le modèle doit avoir le moins de paramètres possibles tout en traduisant au mieux l'information contenue dans la propriété ou l'activité visée (**Chtita 2017**).

Il existe plusieurs stratégies et outils pour la sélection des descripteurs moléculaires dans un procédé de construction des modèles RQSP/RQSA. Cependant, on peut classifier les stratégies utilisées en deux principales méthodes de sélections (**Kraim 2009 ; Chtita 2017**) :

- Méthodes de sélection objective.
- Méthodes de sélection subjective.

II.5.1. Sélection objective

L'analyse objective consiste en la sélection des variables afin de réduire le nombre de descripteurs sans faire participer la variable dépendante (la réponse ; c'est-à-dire l'activité ou la propriété ciblée) **(Kraim 2009 ; Chtita 2017)**. Dans ce type de sélection on doit annuler tous les descripteurs qui n'ont aucune influence significative, et éviter toute sorte de répétition. Par exemple, un descripteur ayant un pourcentage élevé de valeurs identiques pour l'ensemble des composés doit être nécessairement éliminé dès le début de la procédure de sélection **(Kraim 2009 ; Chtita 2017 ; Roubehie Fissa *et al*. 2019)**, parce que ce type de descripteurs ne fournira aucune information distinctive utile. Un autre exemple d'une paire de descripteurs qui sont considérés comme fortement corrélés, à partir du coefficient de corrélation (Statistiquement significatif R>0,95), dans ce cas, et pour éviter la redondance des informations, les deux descripteurs ne peuvent être utilisés simultanément lors de l'analyse des modèles RQSP/RQSA **(Chtita 2017 ; Roubehie Fissa *et al*. 2019)**, parce qu'ils sont de même utilité. Cet aspect utilisé pour toutes les applications de sélection en réduisant le nombre de descripteurs sans pour autant perdre de l'information.

II.5.2. Sélection subjective

Dans la plupart des cas la sélection objective ne peut pas être une solution suffisante pour sélectionner les descripteurs utilisés dans la construction des modèles finaux de type RQSP/RQSA, dans ce cas il est indispensable de trouver une autre technique de sélection pour réduire encore plus le nombre de descripteurs. A cet effet, on fait appel à la méthode de sélection subjective **(Kraim 2009 ; Chtita 2017)**.

L'analyse subjective c'est la sélection des variables finales (descripteurs optimaux) d'un sous-ensemble de descripteurs qui est déjà préalablement réduit, afin de contribuer dans l'élaboration des modèles de type RQSP/RQSA, pourvu que la variable dépendante soit impliquée dans ce processus.

Diverses techniques d'optimisations sont utilisées pour choisir d'une manière efficace, le nombre optimal de descripteurs utilisés par un modèle ou bien ce qu'on appelle "la dimension du modèle", sans pour autant examiner toutes les combinaisons possibles. Les méthodes statistiques qui seront discutées dans la section suivante, sont également utilisées pour éliminer les

descripteurs qui n'interviennent pas dans les modèles RQSP/RQSA proposés, c'est-à-dire qui n'influencent pas la propriété ou l'activité étudiée, en particulier la sélection par la méthode Ascendante Pas à Pas, en anglais : *Forward Stepwise Selection*.

Finalement, pour que tous les modèles de type RQSP/RQSA ne soient pas statistiquement non-significatives ou en cas d'erreurs ponctuelles, il faut que le rapport composés/descripteurs soit être supérieur à "**5**" (**Topliss & Edwards 1979 ; Pratim Roy *et al*. 2009 ; Hamadache 2014**).

II.6. Méthodes d'apprentissage statistiques

Au cours de la dernière décennie, les méthodes d'apprentissage statistiques ont suscité beaucoup d'intérêt dans le milieu académique et au sein d'entreprises de diverses industries **(Hamadache 2014 ; Ammi 2016)**. Elles ont été implantées avec succès pour les tâches de la modélisation RQSP/RQSA à aspect prédictif, c'est-à-dire elles ont été employées dans l'élaboration des modèles de type RQSP/RQSA pour la prédiction des différentes propriétés ou activités ciblées, mais aussi dans les tâches de sélection des descripteurs appropriés à la modélisation RQSP/RQSA.

Dans ce dernier, un tel choix peut être fait dans une variété de méthodes, y compris la sélection par l'ascendante pas à pas (en anglais : *Forward Stepwise Selection*) en utilisant un critère approprié de perfectionnement, par exemple, "Valeur-P" pour l'inclusion et l'exclusion ; basée sur la valeur de la Probabilité statistique "Valeur-P" (en anglais : *P-Value*) **(Roy *et al*. 2015 ; Roubehie Fissa *et al*. 2019)**. Tous dépondent des choix possibles de sous-ensembles, il y a aussi d'autres méthodes de sélection comme les algorithmes génétiques, et l'analyse factorielle.

En ce qui concerne, l'utilisation de ces méthodes dans le sens prédictif de la modélisation RQSP/RQSA. Plusieurs techniques linéaires et non- linéaires ont été utilisées durant ces dernières années, particulièrement la régression linéaire multiple, les réseaux de neurones artificiels, les machines à vecteurs de supports pour la régression, etc. **(Hamadache 2014 ; Ammi 2016 ; Bitam 2018)**. Dans ce qui suit, dans cette étude, je me baserai sur une classe particulière de ces méthodes ; la régression linéaire multiple et les réseaux de neurones artificiels.

II.6.1. Régression linéaire multiple

En raison de sa simplicité la méthode de Régression Linéaire Multiple ou RLM (en anglais : *Multiple Linear Regression ; MLR*), est considérée parmi les méthodes les plus couramment utilisées dans le développement des modèles prédictifs **(Sarkhosh *et al*. 2012 ; Ammi 2016 ; Bitam 2018)**. Son principe de base repose sur l'existence d'une relation linéaire quantitative entre une variable dépendante désignée parfois par $\boldsymbol{Y}$; représentant la propriété ou bien la réponse à étudier, et un nombre de $\boldsymbol{N}$ variables indépendantes désignées par $\boldsymbol{X_i}$; parfois ce sont les descripteurs moléculaires dans le cas des modèles de type RQSP/RQSA. La définition mathématique de cette méthode est donnée par l'**Eq (II.1)** :

$$\boldsymbol{Y = A_0 + \sum_{i=1}^{N} A_i * X_i} \qquad \textbf{(II.1)}$$

Avec $\boldsymbol{Y}$ et $\boldsymbol{X_i}$: la sortie et les entrées du modèle, respectivement, $\boldsymbol{A_i}$ représente les coefficients ou bien les effets correspondant à chaque variable $\boldsymbol{X_i}$, $\boldsymbol{A_0}$ est la constante moyenne du modèle, appelée aussi l'intercepter, et $\boldsymbol{N}$ est le nombre des entrées du modèle.

II.6.2. Réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNAs) représentent probablement la technologie la plus performante des deux dernières décennies qui a été largement utilisée dans une grande variété d'applications dans divers domaines **(Gurney 1997 ; Suzuki 2011)**. Un réseau de neurones artificiels, souvent simplement appelé réseau de neurones, est un modèle mathématique computationnel inspiré de la structure et de la fonction des réseaux neuronaux biologiques dans le cerveau **(Krose & Smagt 1996 ; Suzuki 2011)**. Un réseau de neurones artificiels est constitué d'un certain nombre de neurones artificiels (à savoir, des unités de traitement non linéaires) qui sont connectés les uns aux autres via des poids synaptiques (ou simplement des poids). Le réseau de neurones artificiels peut commencer la tâche d'apprentissage en ajustant ces poids **(Hagan *et al*. 1996 ; Suzuki 2011)**.

II.6.2.1. Du neurone biologique au neurone artificiel formel

Le neurone artificiel est un élément de base de tous les réseaux de neurones artificiels. Sa conception et ses fonctionnalités sont dérivées de l'observation d'un neurone biologique qui est la pierre angulaire des systèmes neuronaux biologiques qui comprennent le cerveau, la moelle épinière et les ganglions périphériques **(Krose & Smagt 1996 ; Gurney 1997)**. Des similitudes

dans la conception et les fonctionnalités peuvent être aperçu dans la **Figure II.3**. Où le côté gauche de la figure représente un neurone biologique avec son soma (corps d'une cellule neuronale), ses dendrites et son axone, où la partie droite de la figure représente un neurone artificiel formel avec ses entrées, ses poids, sa fonction de transfert, son biais et ses sorties **(Suzuki 2011 ; Khaouane 2013)**.

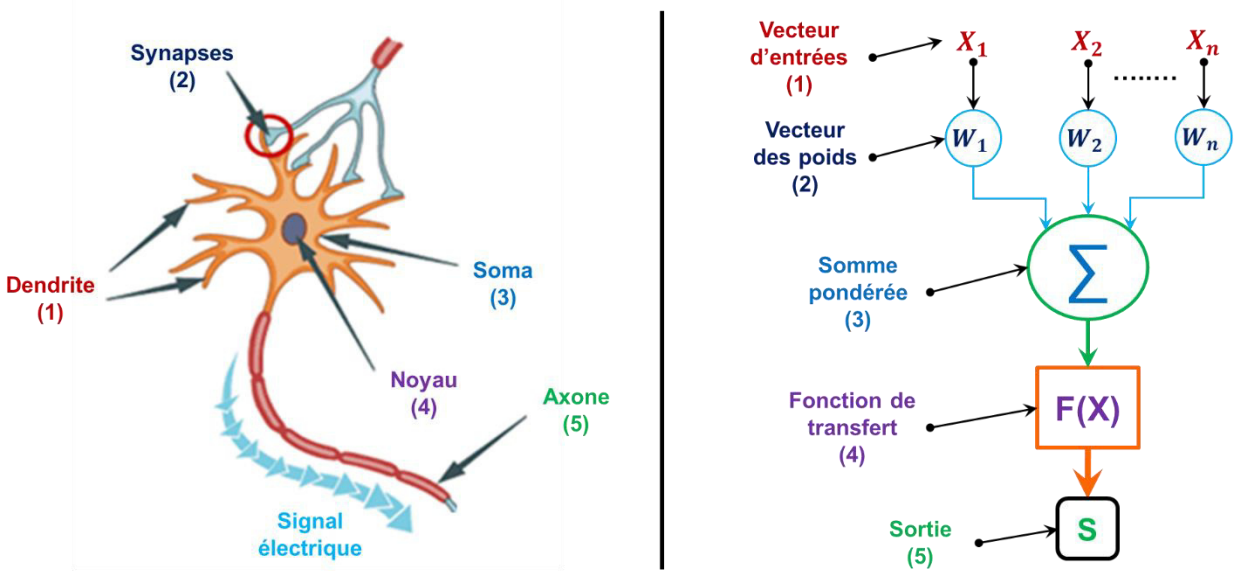


Figure II.3. Conception générale de neurones biologiques et artificiels.

Par comparaison avec le neurone biologique, chaque neurone artificiel est considéré comme un processeur élémentaire, il reçoit un nombre variable d'entrées en amont des neurones, chacune de ces entrées est associé avec un poids $\boldsymbol{w}$ (en anglais *weight*) ; qui représente l'élément clé de la connexion. Chaque neurone artificiel est accordé par une sortie unique qui se ramifie ensuite pour alimenter un nombre variable de neurones en aval comme illustrées dans la **Figure II.4**. **(Hagan *et al*. 1996 ; Haykin 1998 ; Khaouane 2013)**.

La représentation mathématique de ce neurone artificiel formel est aussi bien expliquée dans l'**Eq (II.2) (Bitam 2018)** :

$$\boldsymbol{Y_i = F(a_i) = F(b^O + \sum_{i=1}^{N} w_i * X_i)} \qquad \textbf{(II.2)}$$

Avec $\boldsymbol{b^O}$ est un poids effectué à une entrée constante appelée "biais", $\boldsymbol{w_i}$ est le poids de connexion de la $\boldsymbol{i^{ème}}$ entrée, $\boldsymbol{X_i}$ et $\boldsymbol{Y_i}$ sont les valeurs de $\boldsymbol{i^{ème}}$ entrée et sortie, respectivement. Une fonction $\boldsymbol{F}$, appelée fonction d'activation ou bien fonction de transfert, appliquée sur l'unité $\boldsymbol{a_i}$ appelé le potentiel du neurone, cette unité calcule la somme pondérée de toutes les entrées.

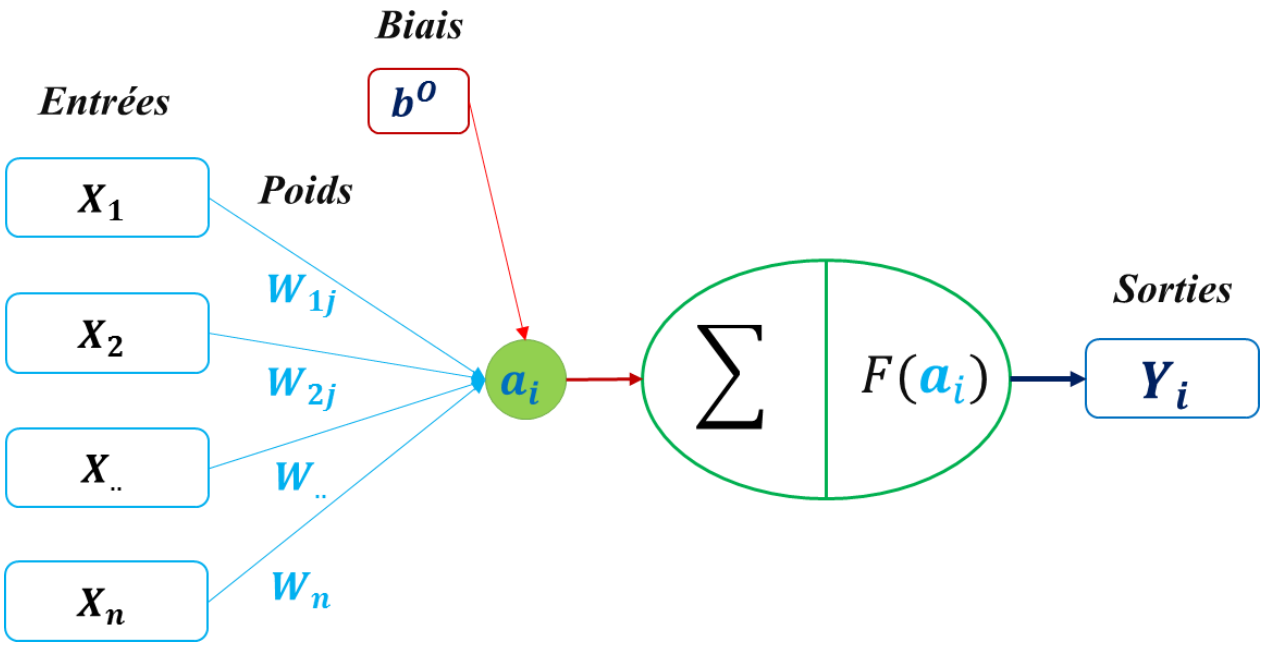


Figure II.4. Schématisation distinctive d'un neurone artificiel formel.

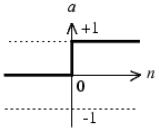
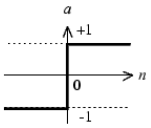
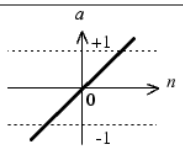
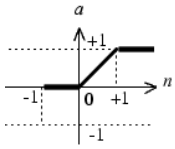
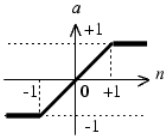
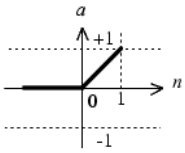
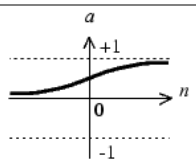
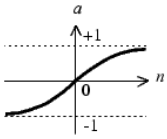
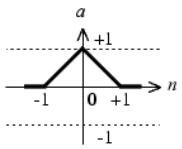
Comme on le voit à partir du modèle de neurone artificiel et de son **Eq (II.2)**, la principale variable inconnue de notre modèle est sa fonction de transfert. La fonction de transfert est un élément constitutif important pour les réseaux de neurone artificiels et pour n'importe quel modèle mathématique, car elle détermine ses essentielles propriétés. Nous l'avons choisi sur la base du problème neurone artificiel (réseau de neurones artificiels) qui doit résoudre dans la plupart des cas **(Suzuki 2011)**. Contrairement à des neurones biologiques qui sont dans un état binaire, la plupart des fonctions de transfert sont continus, proposant une infinité de valeurs possibles comprises dans l'intervalle de $[0, +1]$ ou $[-1, +1]$ **(Si-Moussa 2010)**. Il existe de nombreuses formes possibles pour la fonction de transfert. Les **Tableaux II.3** et **II.4**, donnent quelques fonctions de transfert pour les réseaux de neurone artificiels, utilisées respectivement dans le logiciel *Statistica* **(Cybenko 1989 ; Bishop 1995 ; Statsoft 2020a ; Statsoft 2020b)** et dans le programme de calcul techniques *Matlab* **(Hagan *et al*. 1996 ; Si-Moussa 2010)**.

Tableau. II.3. Fonctions de transfert (d'activation) usuelles dans logiciel *Statistica* (**Bishop 1995 ; Statsoft 2020a ; Statsoft 2020b**).

Nom du fonction	**Définition mathématique**	**Intervalle**	**Description**	**Désignation correspondante dans *Statistica***
Linéaire	$a = n$	$(-\infty, +\infty)$	L'activation du neurone est transmise directement en tant que sortie	***Identity***
Logarithmique Sigmoïde (Logistique)	$a = \frac{1}{1 + e^{-n}}$	$(0, +1)$	Une courbe en forme de S	***Logistic***
Tangente hyperbolique	$a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$	$(-1, +1)$	Une courbe sigmoïde (tangente hyperbolique) similaire à la fonction logistique. Fonctionne souvent mieux que la fonction logistique en raison de sa symétrie. Idéal pour les perceptrons multicouches, en particulier les couches cachées	***Tanh***
Exponentielle	$a = e^{-n}$	$(0, +\infty)$	La fonction exponentielle négative	***Exponential***
Sinus	$a = \sin(n)$	$[0, +1]$	Peut-être utile si la reconnaissance des données distribuées radialement	***Sine***
Exponentielle Normalisée	$a = \frac{e^{n_i}}{\sum e^{n_i}}$	$[0, +1]$	Principalement utilisé pour (mais sans s'y limiter) les tâches de classification. Utile pour construire des réseaux de neurones avec des sorties multiples normalisées, ce qui le rend particulièrement adapté à la création de classificateurs de réseaux de neurones avec des sorties probabilistes	***Softmax***
Gaussienne	$a = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	/	Ce type de fonction d'activation gaussienne isotrope est uniquement utilisé par les unités cachées d'un réseau neuronal RBF, également appelées fonctions de base radiales. L'emplacement (également connu sous le nom de vecteurs prototypes) et les paramètres de propagation sont équivalents aux poids de couche cachés en entrée d'un réseau neuronal MLP	***Gaussian***

*Avec : $\boldsymbol{a = f(n)}$, **(n)** C'est l'entrée nette d'un neurone (Somme pondérée des entrées du neurone pour les MLP).*
*Pour la fonction gaussienne : **(x)** Représente le vecteur des entrées de neurones, mu ($\boldsymbol{\mu}$) Vecteur des poids d'entrée (centre du RBF), sigma ($\boldsymbol{\sigma}$) Propagation (largeur) du RBF.*

Tableau. II.4. Fonctions de transfert usuelles dans le logiciel *Matlab* (**Samarasinghe 2006 ; Si-Moussa 2010**).

Nom du fonction	Relation entrée/sortie	Graphe du fonction	Désignation correspondante dans *Matlab*
Seuil	$a = 0$ Si $n < 0$ $a = 1$ Si $n \geq 0$		***hardlim***
Seuil symétrique	$a = -1$ Si $n < 0$ $a = 1$ Si $n \geq 0$		***hardlims***
Linéaire	$a = n$		***purelin***
Linéaire saturée	$a = 0$ Si $n < 0$ $a = n$ Si $0 \leq n \leq 1$ $a = 1$ Si $n > 1$		***satlin***
Linéaire saturée symétrique	$a = -1$ Si $n < -1$ $a = n$ Si $-1 \leq n \leq 1$ $a = 1$ Si $n > 1$		***satlins***
Linéaire positive	$a = 0$ Si $n < 0$ $a = n$ Si $n \geq 0$		***poslin***
Logarithmique sigmoïde (Logistique)	$a = \frac{1}{1 + e^{-n}}$		***logsig***
Tangente hyperbolique	$a = \frac{e^{+n} - e^{-n}}{e^{+n} + e^{-n}}$		***tansig***
Triangulaire de base	$a = 1 - \lvert n \rvert, -1 \leq n \leq +1$ $a = 0$, autrement		***tribas***

*Avec : $\boldsymbol{a = f(n)}$, **(n)** C'est l'entrée nette d'un neurone (Somme pondérée des entrées du neurone pour les MLP).*

II.6.2.2. Classification des réseaux de neurones selon leurs architectures

La façon dont les neurones artificiels individuels sont interconnectés est appelée topologie, architecture ou graphe d'un réseau de neurone artificiel **(Suzuki 2011)**. L'architecture d'un RNA désigne aussi toutes ces caractéristiques de fonctionnement de base telles que : le nombre d'entrées et de sorties, direction de signe d'information, positionnement des éléments de connexion (poids et biais), leur comportement et le mode d'interconnexion entre les neurones **(Dreyfus *et al*. 2002 ; Dreyfus 2005 ; Jain *et al*. 1996 ; Suzuki 2011)**. Le fait que l'interconnexion peut se faire de nombreuses façons dans les résultats de nombreuses topologies possibles, qui sont divisées en deux grandes familles de base, indiqué dans la **Figure II.5**, où :

- Le côté droit de la figure représente les réseaux de neurones bouclés (en anglais : *Recurrent/Feed-back Networks*), qui ont une topologie récurrente simple (graphe semi-cyclique), où une partie des informations circulent non seulement dans la direction de l'entrée vers la sortie, mais également dans le sens opposé **(Jain *et al*. 1996 ; Suzuki 2011)**.
- Le côté gauche de la figure représente les réseaux de neurones non-bouclés (en anglais : *Feed-forward Networks*), qui ont une topologie à action directe simple (graphe acyclique), dans lequel les flux d'informations circulant à partir des entrées vers les sorties dans une seule direction sans retour en arrière **(Dreyfus *et al*. 2002 ; Dreyfus 2005 ; Chamekh 2007)**.

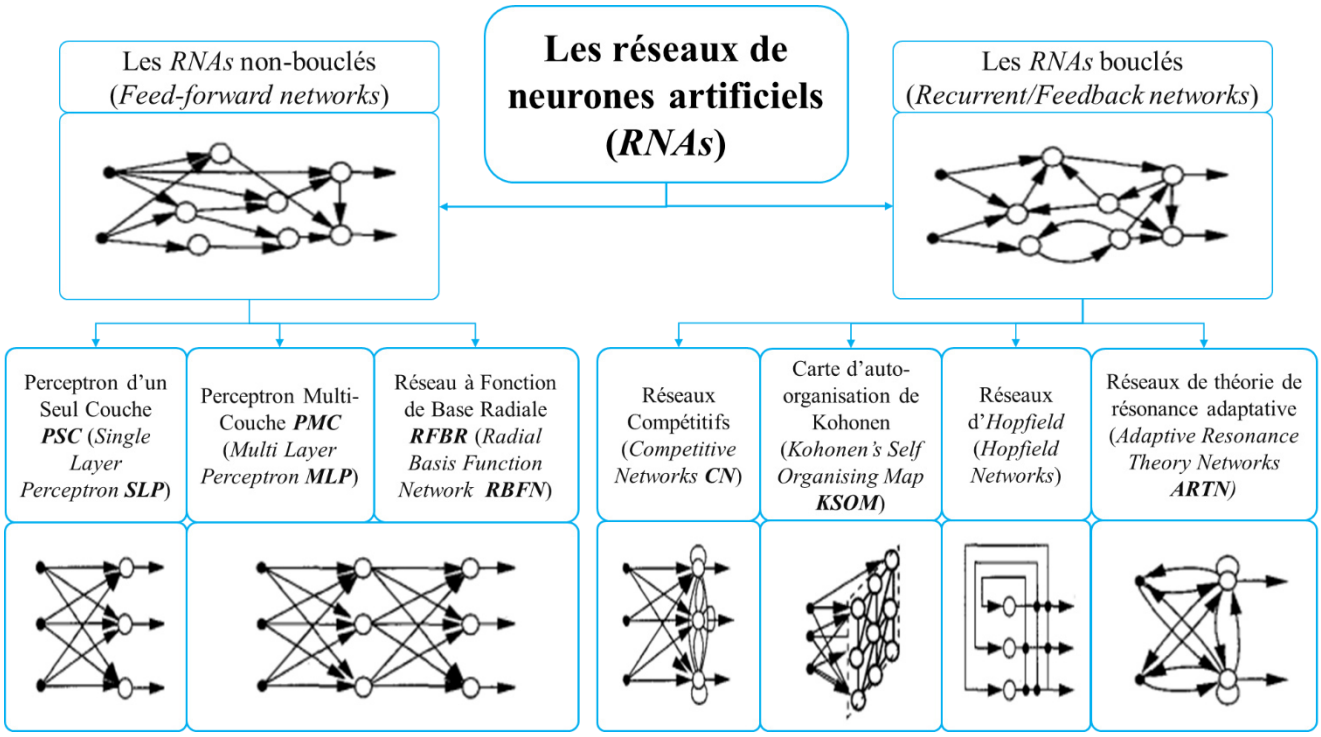


Figure II.5. Classification des réseaux de neurones artificiels selon les architectures d'interconnexions entre neurones **(Jain *et al*. 1996 ; Basheer & Hajmeer 2000 ; Bitam 2018)**.

Dans cette thèse, nous avons adopté un type de réseau de neurones non-bouclé, un Perceptron Multi Couches (***PMC***) (en anglais : *Multi Layer Perceptron* ***MLP***), qui sera expliqué ultérieurement dans la partie **II.6.2.6** de ce chapitre.

II.6.2.3. Classification des réseaux de neurones selon leurs modes d'apprentissage

L'apprentissage est probablement la propriété la plus intéressante des réseaux de neurones, c'est une phase du développement de RNA au cours de laquelle les poids des neurones sont calculés de sorte que les sorties du réseau soient aussi proches que possible aux sorties désirées. La règle d'apprentissage définit comment les poids du réseau doivent être exactement ajustés (mis à jour) entre les cycles d'entraînement successifs (époques) **(Basheer & Hajmeer 2000)**. L'apprentissage des réseaux de neurones artificiels se fait grâce à des algorithmes d'apprentissage. Dans la majorité des algorithmes actuels, l'apprentissage consiste à modifier les poids de connexions de telle sorte que la réponse du réseau soit en accord avec les données de l'expérience **(Ammar 2007 ; Chtita 2017)**.

Il existe trois grands paradigmes d'apprentissage ; apprentissage supervisé, apprentissage non-supervisé et apprentissage imposé ou forcé. Habituellement, ils peuvent être employés par

n'importe quel type d'apprentissage de réseau de neurone artificiel. Chaque paradigme possède de nombreux algorithmes d'apprentissage **(Suzuki 2011)**.

- L'apprentissage supervisé est une technique d'apprentissage automatique qui définit les paramètres d'un réseau de neurone artificiel à partir des données d'entraînement. La tâche d'apprentissage d'un RNA consiste à définir la valeur de ses paramètres pour toute valeur d'entrée valide après avoir vu la valeur de sortie. Les données d'apprentissage sont constituées de paires de valeurs d'entrée et de sortie souhaitées qui sont généralement représentées dans des vecteurs de données **(Basheer & Hajmeer 2000 ; Berkane 2010 ; Suzuki 2011 ; Laidi 2013 ; Hamadache 2014)**.
- L'apprentissage renforcé (par renforcement) est une technique similaire à l'apprentissage supervisé à la différence qu'au lieu de fournir des résultats désirés au réseau, on lui accorde plutôt un grade (ou score) qui est une mesure du degré de performance du réseau après quelques itérations **(Ammar 2007)**. En d'autres termes le RNA est fourni avec une critique sur l'exactitude des résultats plutôt que sur la bonne réponse elle-même **(Basheer & Hajmeer 2000)**, dans cet apprentissage on indique seulement au réseau si la sortie est bonne ou mauvaise en induisant ainsi une variation sélective des poids de connexions **(Laidi 2009)**. Les algorithmes utilisant la procédure d'apprentissage renforcé sont surtout utilisés dans le domaine des systèmes de contrôle **(White 1992 ; Ammar 2007)**.
- L'apprentissage non-supervisé ou bien compétitif est un type d'apprentissage qui se fait seulement sur la base d'informations locales existant à l'entrée **(Samarasinghe 2006 ; Laidi 2009)**. Il représente la capacité des réseaux de neurones à apprendre sans connaître la sortie désirée, les informations présentées à l'entrée provoquent une autoadaptation du réseau afin de produire des valeurs de sortie qui soient proches des réponses à des valeurs d'entrée similaires **(Hassoun 1995 ; Laidi 2009 ; Hamadache 2014)**.

II.6.2.4. Phénomènes de sous-apprentissage et sur-apprentissage

Le point crucial du développement d'un réseau de neurones est son apprentissage. Il s'agit d'une procédure adaptative par laquelle les connexions des neurones sont ajustées face à une source d'information **(Rumelhart *et al*. 1985 ; Ammar 2007)**. Malheureusement, la tâche d'apprentissage n'est pas toujours réussie en raison des obstacles que nous pouvons rencontrer souvent. Les plus célèbres d'entre eux sont deux phénomènes ; le sous-apprentissage et le sur-apprentissage **(Bishop 1995 ; Haykin 2010)**.

- Le phénomène de sous-apprentissage (en anglais : *under-learning* où parfois *under-fitting*) est définie comme l'insuffisance d'apprentissage dans un RNA (peu d'apprentissage) à cause de plusieurs raisons et pouvant être dû au très faible nombre d'itérations ou aussi le trop faible nombre de neurones de connexion des couches cachées **(Si-Moussa 2010)**. Ce qui affecte sur la performance du réseau obtenu, de sorte qu'on trouve une erreur très élevée dans les deux sous base d'apprentissage et de test **(Bishop 1995 ; Haykin 2010)**.
- Par contre le phénomène de sur-apprentissage (en anglais : *over-learning* où des fois *over-fitting*) est définie comme un apprentissage exagéré d'un RNA (apprentissage par cœur) **(Baughman & Liu 1995 ; Hawkins 2004)**. Cela se produit généralement lorsqu'il y a un trop grand nombre de neurones cachés, qui conduisent à un déséquilibre entre les deux sous-base d'apprentissage et de test en termes de leurs valeurs d'erreurs trouvées, alors qu'on espère les trouver très faibles dans la base d'apprentissage (résultat souhaitable) et trop élevés dans la base de test (résultat non-souhaitable) **(Si-Moussa 2010 ; Hamadache 2014)**.

La **Figure II.6** montre bien les différences entre les trois phénomènes d'apprentissage ; le sous-apprentissage, le sur-apprentissage et enfin l'apprentissage optimale ou bien la généralisation optimale. Ce dernier est l'état désiré pour n'importe modèle de réseau de neurone artificiel pendant leur processus d'apprentissage.

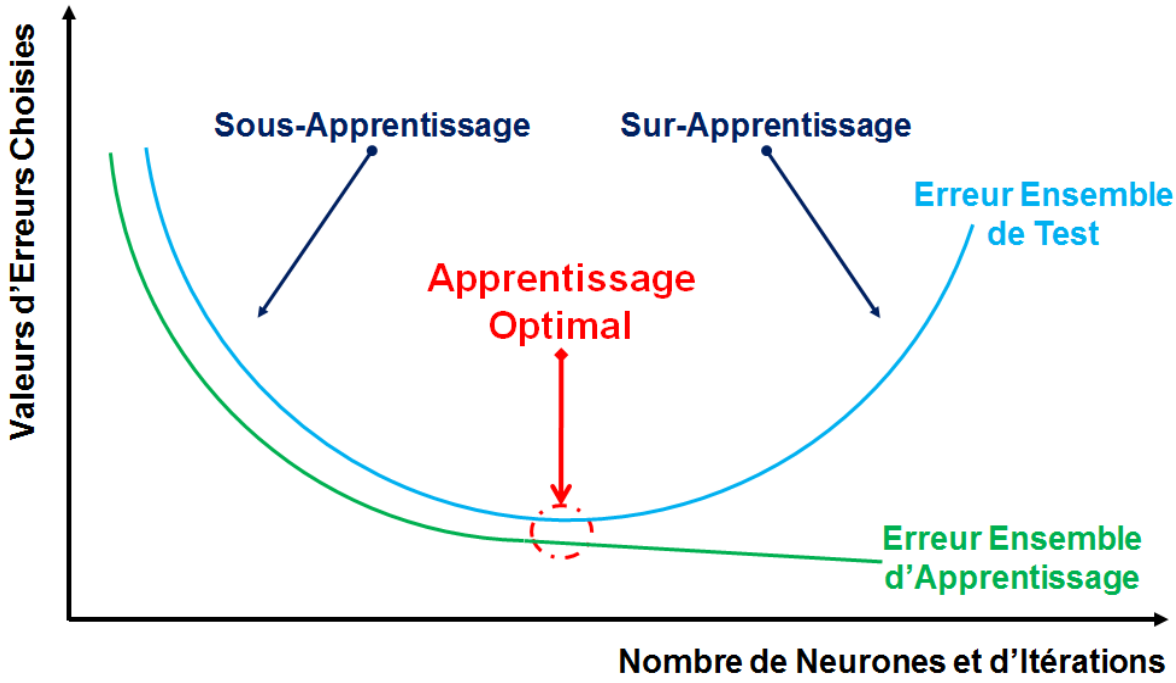


Figure II.6. Phénomènes d'apprentissage optimale, sous-apprentissage et sur-apprentissage.

D'une manière générale, plus on augmente le nombre de neurones dans la couche cachée, meilleur sera l'apprentissage, mais choir rapidement le modèle en se basant seulement sur ces performances de l'ensemble d'apprentissage peu poser un problème, de sorte qu'il y a un grand risque de retenir une structure soumise au phénomène de sur-apprentissage. Afin d'éviter ce problème. Il faut toujours contrôler les qualités de performance des deux sous-ensembles en même temps, de manière à ce qu'ils soient de même ordre de grandeur et aussi faibles que possible **(Hamadache 2014)**. Dans la plupart des situations, il n'y a pas une façon pour déterminer le nombre adéquat de neurones cachés sans faire l'apprentissage de plusieurs réseaux et d'évaluer ensuite l'erreur de généralisation de chaque réseau **(Si-Moussa 2010)**. Très peu de neurones cachés donnent des erreurs d'apprentissage et de généralisation trop élevées en raison du sous apprentissage. Par contre un nombre trop élevé de neurones cachés, donne une erreur d'apprentissage très faible mais une erreur de généralisation trop élevée, en raison de la phase d'apprentissage qui devient alors trop longue ce qui fournit beaucoup de paramètres réglables dans le système, et les performances du réseau, en général, deviennent médiocres **(Ammar 2007 ; Si-Moussa 2010)**.

Afin d'éviter tous les problèmes de sous-apprentissage et de sur-apprentissage dans cette thèse, j'ai adopté une stratégie basée sur l'équilibre entre le nombre de neurones de connexion dans les trois couches ; entrée, cachée et sortie, et puis la taille de la base des données utilisées, qui sera expliquée dans le chapitre IV (Matériels et méthodes).

II.6.2.5. Algorithme d'apprentissage

L'apprentissage dans le domaine des réseaux de neurones artificiels est la façon ou la procédure qui désigne l'architecture et les particularités principales d'un réseau. En effet, l'une de ces propriétés fondamentales est sa capacité à s'adapter et améliorer sa performance en ajustant les connexions des neurones face à une source d'informations par la procédure d'apprentissage, qui se fait grâce à des algorithmes d'apprentissage **(Ammar 2007 ; Chtita 2017)**. La majorité des algorithmes actuels consistent à modifier les poids de connexions pour que la réponse du réseau s'accorde aux exemples de l'expérience, après une initialisation aléatoire des poids, les algorithmes d'apprentissage changent progressivement ces poids, afin de réduire l'écart entre les sorties calculées par le réseau et les sorties expérimentales **(Ammar 2007 ; Si-Moussa 2010 ; Chtita 2017)**.

Plusieurs algorithmes itératifs de rétropropagation peuvent être mis en œuvre, parmi lesquels on trouve : les algorithmes Descente du gradient, l'algorithme de Levenberg–Marquardt, l'algorithme de Régularisation Bayésienne, l'algorithme de BFGS et ainsi de suite **(Si-Moussa 2010)**. Mais dans cette partie je m'intéresse juste à l'algorithme de BFGS qui sera utilisé comme un algorithme d'apprentissage dans cette étude.

Cette abréviation BFGS fait référence aux noms de ses fondateurs : **B**royden, **F**letcher, **G**oldfarb et **S**hanno **(Minoux 1983 ; Oussar 1998)**. Cet algorithme fait partie des méthodes d'optimisation dites "Quasi-Newtoniennes", ces méthodes sont une généralisation de la méthode de Newton **(Dreyfus *et al*. 2008), (Urbani 1995)**, **(Oussar 1998)**, **(Lucea 2006)**, **(Ammar 2007)**, **(SIGMA 2020)** (le site officiel du laboratoire **SIG**naux, **M**odèles, **A**pprentissage statistique).

L'algorithme de BFGS consiste à modifier les paramètres, à l'itération $\boldsymbol{i}$ de l'algorithme, par l'application de la règle présentée dans l'**Eq (II.3) (Dreyfus *et al*. 2008)** :

$$\boldsymbol{\theta_{(i)} = \theta_{(i-1)} - \mu_i M_i \nabla J\left(\theta_{(i-1)}\right)} \qquad \textbf{(II.3)}$$

Où $\boldsymbol{\mu_i}$ est une constante positive, et $\boldsymbol{M_i}$ est une approximation calculée itérativement à partir de l'inverse de la matrice Hessienne. Elle est évaluée à chaque itération par la relation montrée dans l'**Eq (II.4) (Dreyfus *et al*. 2008)** :

$$\boldsymbol{M_i = M_{i-1} + \left[1 + \frac{\gamma_{i-1}^T M_i \gamma_{i-1}}{\delta_{i-1}^T \gamma_{i-1}}\right] \frac{\delta_{i-1}^T \delta_{i-1}}{\delta_{i-1}^T \gamma_{i-1}} - \frac{\delta_{i-1} \gamma_{i-1}^T M_{i-1} + M_{i-1} \gamma_{i-1} \delta_{i-1}^T}{\delta_{i-1}^T \gamma_{i-1}}} \quad \textbf{(II.4)}$$

Où $\boldsymbol{\gamma_{i-1} = \nabla J(\theta_{(i)}) - \nabla J(\theta_{(i-1)})}$ et $\boldsymbol{\delta_{i-1} = \theta_{(i)} - \theta_{(i-1)}}$

On prend pour valeur initiale $\boldsymbol{M_0}$ la matrice identité. Si, lors d'une itération, la matrice calculée n'est pas définie positivement, elle est réinitialisée à la matrice identité **(Dreyfus *et al.* 2008)**.

Généralement, l'approximation par la méthode "Quasi-Newtonienne" n'est efficace que si elle est appliquée au voisinage d'un minimum. Il est donc recommandé d'utiliser une autre méthode comme la méthode du gradient simple ou du gradient stochastique, au début de l'apprentissage, vu qu'ils sont efficaces lorsqu'on est loin du minimum et sa convergence ralentit considérablement lorsqu'on s'approche du minimum ; de mettre, ensuite en œuvre la méthode de BFGS lorsqu'on estime être suffisamment proche d'un minimum **(Ammar 2007 ; Dreyfus *et al.* 2008 ; Si-Moussa 2010)**.

II.6.2.6. Réseaux de neurones perceptron multicouche

Comme mentionné plus tôt dans cette thèse, j'ai adopté un Perceptron Multi-Couches (***PMC***) (en anglais : *Multi Layer Perceptron* ***MLP***), un type lequel appartient à la famille des réseaux de neurones non-bouclés (en anglais : *Feed-forward Networks*). Généralement, suite à un apprentissage supervisé, ces neurones sont organisés dans des couches consécutives où les informations circulent dans un seul sens, de la couche d'entrée vers la couche de sortie, de sorte que chaque neurone dans une couche est connecté à tous les neurones de la couche précédente et de la couche suivante. Néanmoins les neurones d'une même couche ne sont pas interconnectés **(Khaouane 2013 ; Ammi 2016 ; Bitam 2018)**.

Le réseau de ce type est caractérisé principalement par trois couches différentes :

- La couche d'entrée : Elle contient des informations relatives toujours à l'entrée d'un réseau de neurones à développer, les qualités de ces informations sont différentes ; pouvant être des informations structurelles, des caractéristiques qualitatives ou quantitatives, ou bien des facteurs liés principalement aux sorties désirées.

- Les couches cachées : Elles peuvent être une ou plusieurs couches qui contiennent plusieurs neurones reliés aux informations présentes dans la couche d'entrée avec la ou les réponses désirées de la couche sortie.
- La couche de sortie : Elle est constituée d'un ou plusieurs neurones selon le nombre des réponses souhaitées, de sorte qu'ils donneront les valeurs des sorties de la combinaison linéaire des fonctions de transfert des couches cachées.

La **Figure II.7**, illustre l'architecture d'un réseau de neurone perceptron multicouche non-bouclé avec une couche d'entrée, une couche cachée et une seule couche de sortie.

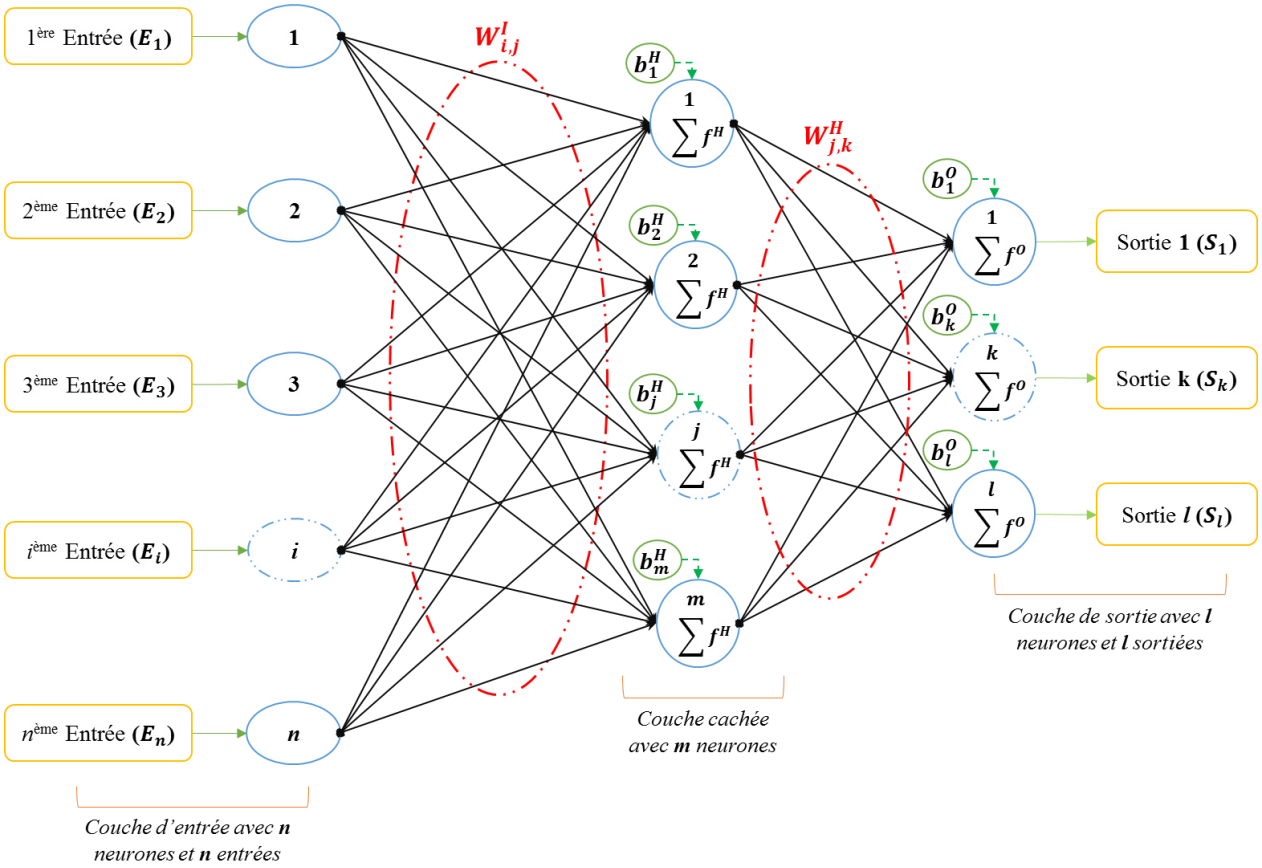


Figure II.7. Structure d'un perceptron multi couches non-bouclé.

Les neurones sont reliés entre eux par des connexions pondérées ; ce sont les poids de ces connexions qui dirigent le fonctionnement du réseau et programment une application de l'espace des entrées vers l'espace des sorties à l'aide de la transformation non linéaire (**Khaouane 2013 ; Hamadache 2014 ; Ammi 2016**). Le modèle mathématique neuronal final de l'exemple précédent est donné par l'**Eq (II.5)** :

$$S_k = f^O\left(\sum_{j=1}^{m} w_{kj}^H * f^H\left(\sum_{i=1}^{n} w_{ij}^I E_i + b_j^H\right) + b_k^O\right) \quad \textbf{(II.5)}$$

Où S_k la sortie du $k^{ème}$ neurone ($k = 1, 2, \dots, l$) ; l est le nombre de neurones dans la couche sortie, alors que E_i est l'entrée du $i^{ème}$ neurone ($i = 1, 2, \dots, n$) ; avec n est le nombre de neurones dans la couche entrée, w_{kj}^H est le poids de la couche cachée associé au $j^{ème}$ et au $k^{ème}$ neurone, w_{ij}^I est le poids de la couche entrée associé au $i^{ème}$ et au $j^{ème}$ neurone ($j = 1, 2, \dots, m$); avec m le nombre de neurones dans la couche cachée, b_j^H biais de la couche cachée connectée avec le $j^{ème}$ neurone de la même couche cachée, b_k^O biais de la couche sortie connecté avec le $k^{ème}$ neurone de la même couche sortie, f^H fonction d'activation (ou de transfert) de la couche cachée, f^O fonction d'activation (ou de transfert) de la couche sortie.

De ce fait, l'architecture du modèle neuronal final dépend aussi du type des fonctions d'activation choisies et du nombre de neurones utilisé dans les trois différentes couches d'entrée, cachées et de sortie. Dans les deux derniers chapitres je vais expliquer, en détails toutes les expressions mathématiques relatives aux cinq paramètres étudiés dans cette thèse.

II.6.3. Autres méthodes d'apprentissage

Le nombre de méthodes d'apprentissage a considérablement augmenté au cours de ces dernières années. Il existe beaucoup de méthodes modernes qui ont donné de bons résultats dans de nombreux domaines dans lesquels elles ont été utilisées, tel que le réseau à fonction de base radiale (***RFBR***) (en anglais : *Radial Basis Function Network* ***RBFN***) et la régression à vecteur de support (***RVS***) (en anglais : *Support Vector Regression* ***SVR***). Dans ce qui suit, je vais présenter brièvement quelques notions sur ces deux techniques, autour desquelles s'est axé un article que j'ai présenté sous forme de communications (au nombre de deux) lors du séminaire "*2nd Conference on **I**nformatics and **A**pplied **M**athematics, (**IAM'2019**)*".

Réseau à fonction de base radiale : Comme illustré dans la **Figure II.5**, c'est un type de réseau de neurone artificiel non-bouclé caractérisé par sa fonction d'activation de la couche cachée qui est une fonction "Gaussienne" avec l'absence de biais b_j^H dans cette couche, alors que le modèle mathématique de ce type de réseaux ressemble au modèle de *PMC* (**Yao *et al*. 2002 ; Varamesh *et al*. 2017**). L'**Eq (II.6)** montre un modèle de réseau à fonction de base radiale avec une sortie S et une seule couche cachée.

$$S = f^{O}\left[\sum_{j=1}^{m} w_{1j}^{H} * \left(f^{H}\left[\sum_{i=1}^{n} \frac{||E_i - C_j||}{2*\sigma_j}\right]\right) + b_1^{O}\right] \quad \textbf{(II.6)}$$

Où $f^{H}(x) = e^{-x^2}$ est la fonction d'activation de la couche cachée "Gaussienne", $||E_i - C_j||$ la distance Euclidienne entre les données d'entrée (E_i) et le centre de la fonction radiale (C_j) ; σ_j représente le paramètre de largeur (toujours $\sigma_j > 0$). Les paramètres (C_j, σ_j) doivent être définis séparément pour chaque unité de fonction de base radiale (FBR) (**Yao *et al*. 2002 ; Varamesh *et al*. 2017**), les autres paramètres restant de ce modèle ont les mêmes définitions que le modèle *PMC* présenté dans l'**Eq (II.5)**.

Régression à vecteurs de support : C'est une technique d'apprentissage automatique, fondée sur la base d'autres techniques semblables appelées les machines à vecteurs de supports (**MVS**) plus connues sous le nom séparateur à vaste marge (**SVM**) (en anglais : *Support Vector* ***SVM***), appartient à une classe d'algorithmes d'apprentissage initialement définis pour la discrimination ; c'est-à-dire la prévision d'une variable qualitative initialement binaire (**Bitam 2018**).

La définition mathématique du modèle de régression à vecteur de support avec une sortie S est présentée dans l'**Eq (II.7) ;** Où w^T est le vecteur transposé de la couche de sortie, $\varphi(x)$ est la fonction du noyau (en anglais : *kernel function*) qui est une fonction radiale montrée dans l'**Eq (II.8)** et b la fonction de biais (**Yao *et al*. 2002 ; Halali *et al*. 2016**). Pour calculer les paramètres w et b, nous utilisons la fonction de coût (en anglais : *Cost Function* ***Fcost***) définie par l'**Eq (II.9)**.

$$S = w^T(x)\,\varphi(x) + b \quad \textbf{(II.7)}$$

$$\varphi(x_i, x_j) = \sum_{j=1}^{k}\sum_{i=1}^{n} exp\left(\frac{-\|x_i - x_j\|^2}{(2*\gamma_j)^2}\right) \quad \textbf{(II.8)}$$

$$Fcost = \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad \textbf{(II.9)}$$

Qui est à minimiser sous les conditions suivantes :

$$w^T\varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^*$$
$$y_i - w^T\varphi(x_i) - b \leq \varepsilon + \xi_i$$
$$\xi_i, \xi_i^* \geq 0,\ i = 1, \ldots, n$$

Où x_i est l'entrée des $i^{ème}$ données et y_i fait référence à sa sortie, ε est la précision des approximations de la fonction ou bien la taille de tube (en anglais : *the tube size*), ξ_i et ξ_i^* sont les variables de relâchement (en anglais : *slack variables*), qui sont la marge d'erreur acceptable, C la constante de capacité, γ_j représente la largeur du fonction de noyau qui est un paramètre réglable de fonction de base radiale (FBR) **(Yao *et al*. 2002 ; Halali *et al*. 2016)**. Donc à la fin, l'objectif du modèle de régression à vecteur de support est de trouver les valeurs optimales des paramètres C et ε à une valeur de γ_j initialement choisis. La technique de régression à vecteur de support fournit un grand nombre de paramètres de choix comme l'utilisation d'autres types de fonction de noyau, sans oublier les autres types de RVS comme le type nu-RVS **(Bhatt *et al*. 2012 ; Lafdani *et al*. 2013)**.

II.7. Analyses statistiques, validation et comparaison

La sélection et la validation des modèles sont des étapes cruciales dans la conception des modèles par apprentissage **(Dreyfus *et al*. 2008)**. Il existe plusieurs manières ou méthodes pour analyser et valider un modèle. Je décrirai ici certaines méthodes les plus fréquemment utilisées.

Dans le cas du modèle neuronale, la validation consiste à tester le réseau après ou au cours de l'apprentissage avec des données proches des données utilisées lors de cette phase. Les résultats de ce test permettront au modèle d'éviter le phénomène de sous-apprentissage ou sur-apprentissage. En revanche la généralisation s'effectue avec des données tout à fait nouvelles et différentes de la base de données de départ. Elle permet de savoir si le réseau a développé une capacité de généralisation assez performante ou non.

L'une des méthodes d'appréciation de la performance d'un modèle est le calcul du vecteur du meilleur ajustement linéaire qui est une évaluation visuelle de la distribution des sorties visée par rapport à la droite de l'ajustement linéaire idéal. Une autre méthode est la détermination des valeurs des coefficients et d'autres paramètres statistiques des modèles développés, ensuite la vérification s'ils sont significatifs ou non. Cette vérification est basée sur des normes de références ou des domaines confirmés, déterminées préalablement par des statisticiens ou par des fondateurs de ces paramètres **(Roubehie Fissa *et al*. 2019)**. Parmi les paramètres les plus utilisés, on trouve le coefficient de régression ou de corrélation (R) et le coefficient de détermination (R^2), dont les valeurs sont toujours souhaitées les plus proches de "**1**". A titre d'exemple, sur les paramètres statistiques les plus connus dans l'étape d'analyse et de validation, le test de valeur de probabilité (***P-Value***) est généralement considéré comme

significatif lorsqu'il est supérieur à 0,05. Il y a aussi la méthode de calcul des différents types d'erreurs pour juger la performance du modèle. Généralement, on distingue deux classes d'erreurs ; les erreurs standards qui sont exprimées en unité de la sortie elle-même comme la racine d'erreur quadratique moyenne (***RMSE***), et les erreurs relatives qui sont exprimées par le pourcentage de sortie désiré, comme l'erreur relative absolu moyenne (***ERAM***) **(Chai Draxler 2014)**. Les définitions mathématiques et l'éclaircissement des paramètres utilisés lors de la validation des modèles de type RQSP-RNA développés dans cette étude seront montrés plus tard dans les deux derniers chapitres de cette thèse.

En fait, il existe d'autres manières de validation spécialisée pour certain type de modèle tel que l'analyse de sensibilité, je cite, exclusivement la méthode d'analyse de sensibilité au poids qui a été utilisés spécifiquement pour les modèles neuronaux, car elle analyse la contribution de leurs entrées sur le modèle à partir des poids de connexions **(Goh 1995 ; Gevrey et al. 2003)**. La détermination du domaine d'applicabilité est également considérée comme une méthode spécifiquement conçue pour les modèles de type RQSP/RQSA, en effet cette méthode est devenue l'un des piliers lors du développement de modèles de ce type, puisqu'elle donne un aperçu sur l'applicabilité des modèles développés, afin de permettre à l'utilisateur de choisir plus facilement les meilleurs modèles adéquats **(Sahigara *et al.* 2012)**. Ces deux dernières méthodes seront utilisées ultérieurement comme des outils de validation et d'analyse pour les modèles de type RQSP-RNA développée dans cette thèse. Leurs méthodologies d'élaboration et de calculs seront discutées plus tard dans le chapitre IV (Matériels et méthodes).

La comparaison des modèles est également considérée comme un outil très puissant pour valider et analyser les performances des nouveaux modèles par rapport aux autres modèles existant dans la littérature, mais reste incomplète lorsqu'elle est utilisée seule, compte tenu des spécifications de développement de chaque modèle **(Roubehie Fissa *et al.* 2019)**. A la fin de cette thèse je vais faire des comparaisons entre chaque modèle de type RQSP-RNA développé et d'autres modèles semblables dans la littérature.

II.8. Conclusion

Ce chapitre a été consacré aux les méthodes RQSP/RQSA, leurs approches de base, méthodologie de développement. Aussi l'analyse et la validation des modèles de ce type ont été détaillées. La manière de quantification des informations moléculaires a été montrée grâce aux différents descripteurs moléculaires. Les méthodes d'apprentissage statistique nécessaires ont été

expliquées notamment les réseaux de neurone artificiels, en se basant surtout sur les perceptrons multicouches, la technique principale d'apprentissage utilisée dans la plupart des travaux de cette thèse. J'ai discuté presque toutes les notions associées à ce type de développement, tel que l'architecture d'un réseau et ses paramètres de base comme les fonctions d'activation, le processus d'apprentissage et tous les phénomènes relatifs.

Chapitre III

Etat de l'art et revue critique de la littérature sur la détermination des propriétés des hydrocarbures

III.1. Introduction sur l'état statistique des références bibliographiques

Les propriétés physicochimiques et thermodynamiques des hydrocarbures purs et des fractions pétrolières ont été largement étudiées par plusieurs chercheurs ces dernières années (**Riazi 2005 ; Merdrignac & Espinat 2007 ; Barbosa *et al*. 2017 ; Nascimento *et al*. 2018 ; Vieira *et al*. 2019 ; Santos *et al*. 2020**), puisqu'elles occupent une grande importance dans le traitement et la production des produits pétroliers. La connaissance de ces propriétés aide les chercheures et les industriels à agir de manière correcte et efficace lors de la production, du traitement et du stockage des matériaux énergétiques. Les méthodes expérimentales de caractérisation des propriétés est le moyen le plus efficace et direct. Cependant, ces avantages sont souvent entravés par plusieurs contraintes ; telles que le temps de manipulation, le coût et le risque des expériences et parfois la technologie ou la méthode utilisée dans la détermination de ces propriétés.

En conséquence, plusieurs chercheurs ont pris l'initiative pour faire des études afin de développer des méthodes permettant de déterminer différentes propriétés physicochimiques et thermodynamiques pertinentes pour les hydrocarbures purs et les fractions pétrolières. Cela s'est traduit par l'émergence de nombreuses méthodes de modélisation, telles que : **(1)** les méthodes de la modélisation empiriques qui dépendent principalement des expériences, **(2)** les corrélations physiques ou chimiques qui ne peuvent pas être réalisées sans connaître les valeurs des autres propriétés, et **(3)** les méthodes de contribution des groupes qui se basent sur la compréhension des groupements fonctionnels des composés ou des molécules chimiques. Les résultats de ces dernières sont satisfaisants par rapport aux méthodes précédentes, mais elles contiennent quelques inconvénients qui les empêchent de donner les meilleurs résultats, comme les exigences de leur programmation qui sont parfois longues et difficiles, la diversité des choix dans la subdivision des groupements fonctionnels, en particulier dans les composés compliqués, et l'incapacité de la distinction entre les différents isomères de position (*cis-trans*, *méta*, *ortho* et *para*).

Selon les dernières recherches statistiques effectuées sur les méthodes de détermination des propriétés des hydrocarbures, généralement deux types de méthodes sont les plus répandus et donnent des résultats performants par rapport aux autres. La première classe est celle des méthodes d'intelligence et d'apprentissage artificiels, la méthode des réseaux de neurone artificiels (RNAs) (*en anglais* : ANN), la deuxième classe est celle qui repose sur les informations structurelles emmagasinées dans les composés chimiques, la méthode de relation

quantitatives structures-propriété (RQSP) (*en anglais* : QSPR). La **figure III.1** illustre l'état statistique des références bibliographiques publiées dans différentes revues scientifiques affiliées à la plateforme ***ScienceDirect*** pendant ces vingt (20) dernières années, sur trois sujets de recherche qui sont : les études RQSP appliquées aux différents domaines avec 3900 publications, les études traitant la détermination des propriétés des hydrocarbures par différentes méthodes avec 122460 publications, et les études qui appliquent les techniques des RNAs avec 1238712 publications pendant la même période, entre 2000 et 2020.

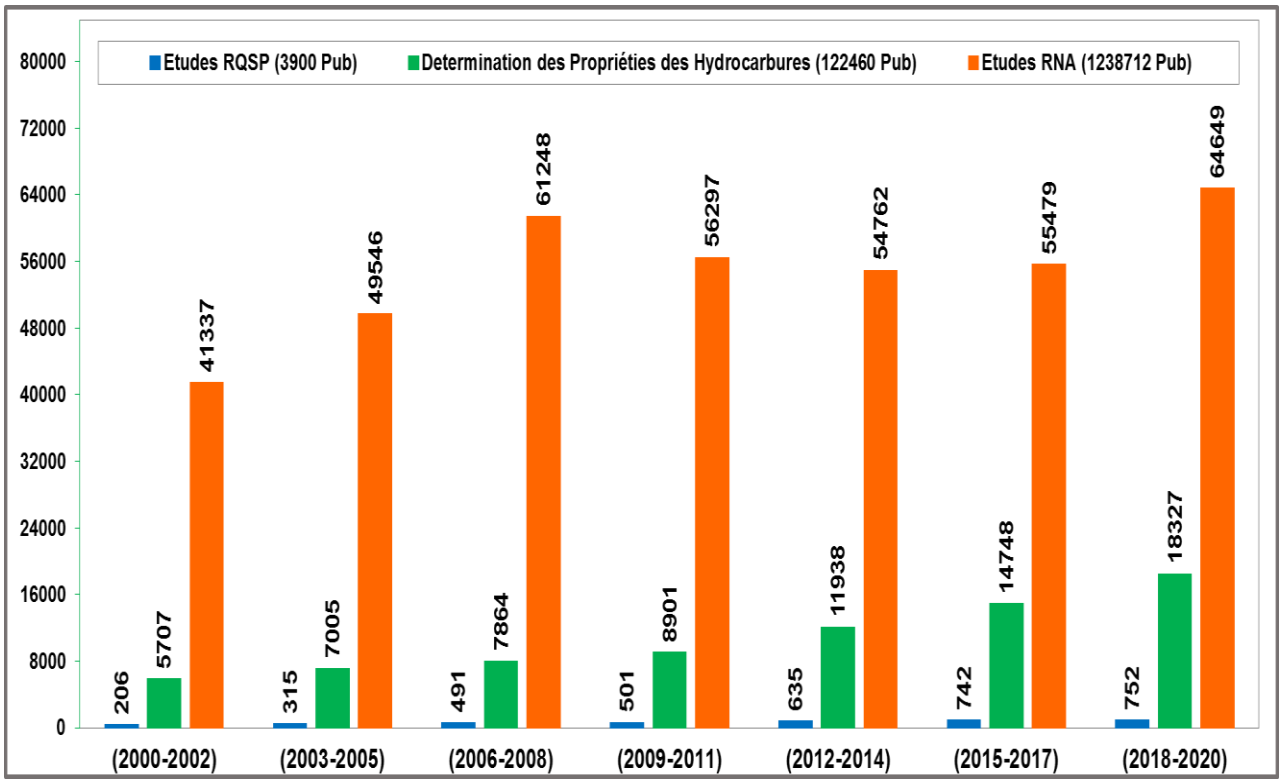


Figure III.1. Nombre de publications pour : (■) les études RQSP, (■) les études de RNA et (■) la détermination des propriétés des hydrocarbures par différentes méthodes publiées dans la plateforme *Sciencedirect* **(Sciencedirect 2020)**.

D'après ces analyses statistiques, l'augmentation du nombre de publications au cours des années indique l'importance de ces trois axes de recherche dans différents domaine tels que la chimie organique **(Yan *et al*. 2020)**, les sciences des matériaux **(Wu *et al*. 2013),** les sciences de l'environnement **(Jiao 2010)**, les sciences pharmaceutiques **(Tsakovska *et al*. 2017 ; Hassanzadeh *et al*. 2019)** et même les sciences économiques **(Taskin & Guneri 2006 ; Jahn 2020)**. La grande différence dans le nombre de publications d'un sujet à un autre est due à l'innovation et à l'ancienneté du domaine, cependant les trois sujets sont des sujets d'actualité.

En outre, dans la même période, plusieurs tentatives ont été menées pour faire progresser l'utilisation des approches relation quantitatives structures-propriété (RQSP) et/ou les réseaux de neurone artificiels (RNA) dans la détermination des propriétés des hydrocarbures. En conséquence, comme l'indique la **figure III.2**, la détermination des propriétés des hydrocarbures a été traitée dans de nombreuses publications ; par les méthodes de RNA (ANN) avec 13845 publications et par les méthodes de RQSP (QSPR) avec 620 publication. Uniquement 201 publications ont traité la détermination des différentes propriétés des hydrocarbures par la combinaison des deux techniques RQSP-RNA (QSPR-ANN), ce qui confirme encore l'importance et l'actualité de ce domaine de recherche.

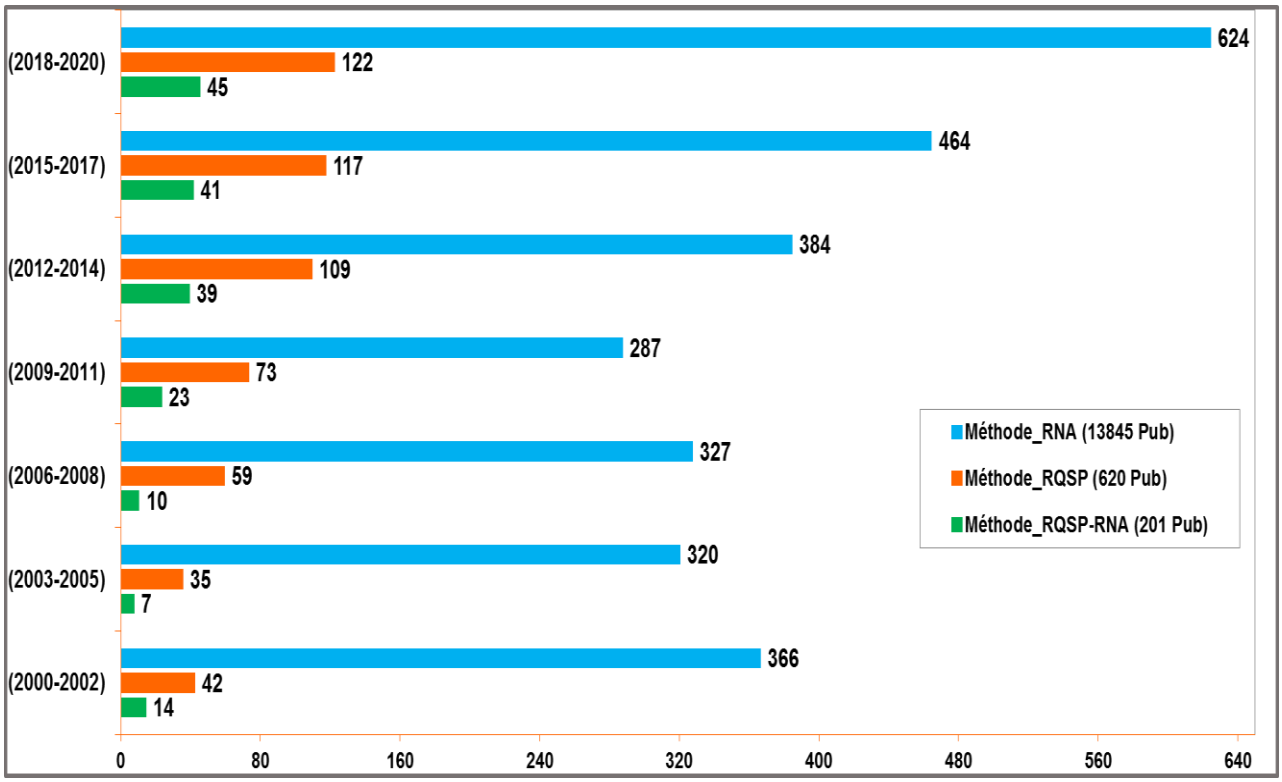


Figure III.2. Nombre de publications traitant la détermination des propriétés des hydrocarbures par les méthodes : (■) RNA, (■) RQSP et (■) RNA-RQSP publiées dans la plateforme *Sciencedirect* **(Sciencedirect 2020)**.

Le présent chapitre a été consacré à une analyse et revue bibliographique de l'état de l'art et revue critique de la littérature liée à la détermination des propriétés des hydrocarbures, divisé en deux parties principales. Dans la première partie, je parlerai d'une revue essentielle des travaux de la modélisation des hydrocarbures par la méthode RQSP en se basant sur les cinq propriétés étudiées dans notre thèse. Dans la deuxième partie je discuterai une revue des

principaux travaux traitant la prédiction des cinq paramètres d'intérêt des hydrocarbures réalisés par d'autres techniques, telles que : les méthodes computationnels, les modèles basés sur les méthodes de contribution des groupes, et les corrélations physiques-empiriques ou semi-empiriques. Une conclusion clôturera ce chapitre et sera une synthèse dans laquelle je présenterai les principales motivations pour justifier le choix du thème de cette thèse sur la modélisation des propriétés des hydrocarbures à travers les deux techniques citées précédemment.

III.2. Revue des travaux de la modélisation des hydrocarbures effectués par la méthode RQSP

Dans cette partie de la revue bibliographique, je vais parler des principales littératures scientifiques consacrées à l'utilisation des relations quantitatives structure-propriétés en vue de la prédiction des propriétés des hydrocarbures. Il existe un nombre limité de modèles RQSP, disponibles dans la littérature, pour la prédiction des cinq propriétés étudiées dans cette thèse qui sont : la température normale d'ébullition (Tb), la densité relative du liquide (d20), la température critique (Tc), la pression critique (Pc), et le volume critique (Vc). Sur la base de l'examen de la littérature, je vais relater ci-après, quelques publications intéressantes qui ont retenu mon attention.

III.2.1. Travaux réalisés sur la modélisation de la température normale d'ébullition

Il existe un bon nombre d'études traitant la modélisation de la température normale d'ébullition (Tb) par la méthode RQSP.

L'étude de Wessel et Jurs en 1995 **(Wessel & Jurs 1995)** est réalisée pour développer deux modèles de type RQSP afin de prédire la température normale d'ébullition Tb des différentes familles d'hydrocarbures purs. Le premier modèle est également basé sur la méthode de régression linéaire multiple (MLR) avec une base totale égale à 366 molécules d'hydrocarbures, tandis que le deuxième modèle est aussi basé sur la méthode de réseau de neurone concurrentiel (CNN). La base totale qui a été utilisée pour ce modèle est égale à 311 molécules d'hydrocarbures. Cette étude a été considérée parmi les premières études effectuées pour prédire la Tb des hydrocarbures par la méthode RQSP. Sa supériorité en termes de précision par rapport à la méthode de contribution des groupes a été bien démontrée par cette étude à l'époque. Les principaux points de ses résultats seront mentionnés dans la comparaison faite dans le dernier chapitre de cette thèse.

Cholakov et ses co-auteurs ont réalisé une étude en 1999, sur différents types d'hydrocarbures saturés et insaturés afin d'élaborer des modèles capables de prédire ses températures de point d'ébullition **(Cholakov *et al.* 1999)**. Deux modèles de type RQSP ont été développés, chaque modèle est basé sur huit (8) descripteurs moléculaires appartenant à deux types différents : **(1)** Descripteurs basé sur la théorie de la simulation mécanique moléculaire, tels que l'énergie totale (E_{tot}), le moment dipolaire électrostatique (DM) ou la surface totale de Van der Waals (S_t). **(2)** Descripteurs basé sur l'atome de carbone et la masse moléculaire, comme les exemples du nombre total d'atomes de carbone (C), la masse moléculaire (M) ou le nombre d'atomes de carbone dans les groupes "CH=" cycliques ($\boldsymbol{N^{c}_{DCH}}$). La base totale utilisée lors du développement de ces modèles comportait 261 composées d'hydrocarbures saturés et insaturés comme les alcanes, les alcènes, les cycloalcanes et les cycloalcènes. Les deux modèles ont aussi été basés sur la méthode de régression linéaire multiple (MLR) pas à pas. De simples paramètres statistiques ont été choisis pour valider et tester l'exactitude de la prédiction des modèles développés, on cite l'écart-type moyen des erreurs absolues, l'écart type moyen des erreurs relatives et le coefficient de détermination (R^2) qui donnent respectivement : 4,86 ±0,35 K, 1,15 ±0,07% et 0,999 pour le premier modèle, et 5,51 ±0,36 K, 1,27 ±0,08% et 0,999 pour le deuxième modèle. Ces résultats ont été considérés comme bons par comparaison aux méthodes classiques de l'époque, comme celle basées sur les contributions des groupes.

En 2001 Espinosa et ses co-auteurs **(Espinosa *et al*. 2001)**, ont effectué une étude RQSP sur 1168 composés organiques purs, y compris différentes familles d'hydrocarbures telles que les aliphatiques et les aromatiques. Afin de prédire la température d'ébullition Tb et deux autres propriétés. Dans cette étude deux modèles ont été développés sur la base d'un réseau de neurone de type carte de théorie de la résonance adaptative floue (Fuzzy ARTMAP), alors qu'un autre modèle a été basé sur un réseau de neurones à rétro propagation (bpNN). Les deux premiers modèles ont montré des résultats supérieurs mais peu convaincants à cause de leur instabilité dans les trois sous base de données, tandis que le troisième modèle a donné des résultats stables, mais pas meilleurs. Les résultats de cette étude seront aussi discutés ultérieurement dans la comparaison faite dans le dernier chapitre de cette thèse.

Ha et ses co-auteurs ont effectué une étude en 2005 **(Ha *et al.* 2005)** sur les hydrocarbures purs pour prédire la température d'ébullition (Tb) et deux autres propriétés. Trois modèles de type RQSP basés sur la méthode de régression Linéaire multiple (MLR) ont été développés sur la base de : 186 différentes molécules d'hydrocarbures purs saturés pour le

premier modèle, 200 molécules d'hydrocarbures aromatiques purs pour le second modèle, et 386 molécules d'hydrocarbures saturés et aromatiques pour le troisième modèle. Les entrées utilisées pour chaque modèle sont généralement constituées de sept (7) descripteurs. Le point critique remarqué dans cette étude, est le nombre insuffisant d'analyses et de tests statistiques réalisés pour juger ces modèles.

L'étude de Sola et ses co-auteurs a été effectuée en 2008 **(Sola *et al.* 2008)** sur les hydrocarbures purs pour prédire la température d'ébullition (Tb) et deux autres propriétés critiques. Le modèle est de type RQSP, basé sur la méthode de régression linéaire (MLR). La base utilisée dans cette étude était formée de 155 composés organiques purs, y compris des hydrocarbures. Le nombre des entrées utilisées pour développer ce modèle était égal à huit (8) descripteurs moléculaires. Certains tests et analyses statistiques intéressants, tels que la REQM, ont été effectués dans cette étude, et les résultats de ces derniers ont montré la supériorité de ce modèle en termes de précision par rapport aux autres modèles basés sur la méthode de contribution des groupes dans d'autres études. Les principaux résultats issus de ce modèle seront discutés en détails dans la comparaison du dernier chapitre de cette thèse.

Manso et ses co-auteurs ont effectué une étude en 2012 **(Manso *et al.* 2012)**, sur des hydrocarbures afin de développer des modèles linéaires simples de type RQSP, pour prédire la température normale du point d'ébullition (Tb) utilisant un nouveau descripteur topologique proposé, basé principalement sur les structures moléculaires. À cette fin, l'indice topologique *Fi* a été développé et utilisé pour corréler la topologie de différents types d'hydrocarbures tels que : les alcanes, les alcènes, les alcynes et les cycloalcanes possédant des chaînes normales ou ramifiées avec leur point d'ébullition normale (Tb). Cet indice associe une paire ordonnée (*i*, *j*) à chaque sommet du graphe structurel et introduit un nouveau concept dans la construction des indices topologiques. La robustesse du modèle RQSP avec l'indice *Fi* lors de la prédiction de Tb des hydrocarbures a été évaluée par comparaison avec d'autres modèles RQSP de même type les plus cités dans la littérature, comme ceux basés sur les indices topologiques de Wiener (***W***), Hosoya (***Z***) et Randić (χ). Dans un premier temps, le modèle développé a été testé pour prédire la température d'ébullition (Tb) des alcanes. Après, il a été appliqué avec succès pour prédire la Tb des autres composés restants (alcènes, alcynes et cycloalcanes). L'indice topologique *Fi* s'est avéré plutôt efficace et a produit de petits écarts, par rapport aux autres indices topologiques utilisés pour la comparaison. Sur la base de l'indice topologique *Fi*, d'autres propriétés d'intérêt

peuvent également être explorées et les concepts développés dans ce travail peuvent facilement être adaptés à d'autres familles de composés, principalement en phase liquide.

En 2013, Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2013)** ont mené une étude RQSP pour prédire la température d'ébullition (Tb) de plusieurs familles de composés organiques, dans laquelle la taille de la base des données utilisée pour développer ce modèle est égale à 17768 composés, y compris les hydrocarbures. Le modèle est également basé sur un réseau de neurones artificiels à propagation direct en avant (FF-ANN), et un grand nombre de descripteurs moléculaires égal à quarante-quatre (44) descripteurs ont été utilisés comme entrée du modèle. Un algorithme d'apprentissage de type Levenberg–Marquardt (LM) a été utilisé pour développer et optimiser le réseau de neurone artificiel. Deux paramètres de test et d'analyse statistiques intéressants ont été réalisés dans cette étude : le premier est la REQM, et le second est l'ERAM. Les résultats de ce modèle ne sont pas très performants compte tenu du nombre de familles et de composés traités, mais qui sont globalement dans les normes. D'autres résultats de ce modèle seront discutés plus tard, dans le dernier chapitre de cette thèse.

Dans la même année, Dai et ses co-auteurs **(Dai *et al.* 2013)** ont développé deux modèles de type RQSP afin de prédire la température d'ébullition (Tb) des hydrocarbures purs : l'un pour les alcanes et l'autre pour les hydrocarbures saturés. Dans cette étude, un troisième modèle a été développé pour prédire le même paramètre (Tb), mais sur la base d'une famille des alcools qui n'a pas d'importance dans notre thèse. Le nombre de composés utilisés dans le premier modèle est de 80 molécules, tandis que le second modèle a été développé sur la base de 65 composés. Les deux modèles RQSP sont également basés sur la méthode de régression linéaire multiple (MLR). Deux descripteurs de type topologiques (Y_C et W_C) et un troisième descripteur lié au nombre de chemins (P_3) ont été utilisés comme entrée pour les deux modèles développés dans cette étude. Certains paramètres d'analyse et de test statistiques ont été calculés pour valider le modèle de cette étude, mais le seul qui correspondait à nos paramètres est le coefficient de détermination (R^2). Les discussions relatives à ce modèle seront présentées dans le cinquième chapitre de cette thèse.

Saaidpour et Ghaderi ont réalisé, en 2016 une étude sur différents hydrocarbures issus du pétrole brut **(Saaidpour & Ghaderi 2016)**, 80 molécules ont été utilisées pour développer un modèle de type RQSP afin de prédire leur température d'ébullition (Tb). Deux autres modèles ont été développés pour prédire la constante d'Henry et la solubilité dans l'eau. Les auteurs ont suivis (6) six étapes pour développer leur modèle, qui est également basé sur la méthode de

régression linéaire multiple (MLR). Deux descripteurs uniques de type physico chimique ont été utilisés comme entrée du modèle ; le premier est la masse moléculaire (MW), et le second est le logarithme de la pression de vapeur (logVP). Parmi les paramètres statistiques d'analyse effectués pour tester le modèle ; la REQM, pour lequel les résultats étaient instables pour les deux sous-ensembles de la base des données. Les résultats issus de cette étude seront plus amplement discutés dans le dernier chapitre de cette thèse.

III.2.2. Travaux réalisés sur la modélisation de la densité relative du liquide

A l'inverse du paramètre Tb, il n'existe pas beaucoup d'études qui traitent la modélisation de la densité relative du liquide (d20) par la méthode RQSP. Dans ce qui suit, je vais citer quelques travaux intéressants sur cette dernière dans les paragraphes ci-dessous :

Wakeham et ses co-auteurs ont réalisé une étude en 2002, sur la modélisation de quatre propriétés des hydrocarbures purs **(Wakeham *et al.* 2002)**, l'un des objectifs de cette étude est le développement des deux modèles de type RQSP pour prédire la densité relative du liquide (d20), de sorte que les entrées de ces derniers sont des paramètres basés principalement sur les structures chimiques des composés. De ce fait, huit (8) descripteurs ont été utilisés comme entrée de ces deux modèles, la base des données utilisées dans ce travail contient environ 237 composés purs de différentes familles d'hydrocarbures. Les deux modèles ont également été fondés sur la base d'une méthode de régression linéaire (MLR). Plusieurs paramètres d'analyse statistique ont été réalisé afin de tester et de valider ces deux modèles. Cependant, parmi ceux qui ont en commun avec le modèle RQSP de type RQSP-RNA proposé dans cette thèse pour d20 sont : le coefficient de détermination (R^2) et l'ERAM. D'après les résultats d'analyse et de test, les deux modèles ont montré une bonne précision lors de la prédiction de paramètre d20 des hydrocarbures purs. Une comparaison entre ces deux modèles et le modèle développé dans ce travail, sera présentée dans le dernier chapitre de cette thèse.

Comme mentionné dans la partie précédente, Ha et ses co-auteurs ont réalisé, en 2005 une étude **(Ha *et al.* 2005)** sur les hydrocarbures purs pour prédire deux autres propriétés dont la densité relative du liquide (d20). Trois modèles de type RQSP basés sur la méthode de régression Linéaire multiple (MLR) ont été développées sur la base de : 186 molécules différentes d'hydrocarbures purs saturés pour le premier modèle, 200 molécules d'hydrocarbures aromatiques purs pour le deuxième modèle, et 386 molécules d'hydrocarbures saturés et aromatiques pour le troisième modèle. Sept (7) descripteurs moléculaires pertinents tels que

l'indice de Balaban (*J*), la masse moléculaire relative (relative MW) et la charge partielle minimale pour un atome H (minimum PC on H), ont été utilisés comme entrée pour ce modèle. Dans cette étude, peu de paramètres d'analyse et de tests statistiques ont été effectuées.

III.2.3. Travaux réalisés sur la modélisation des trois propriétés critiques

Vu l'importance énorme des propriétés critiques, notamment la température critique, la pression critique et le volume critique dans le domaine pétrolier, on retrouve récemment plusieurs études qui traitent la modélisation de ces propriétés par la méthode RQSP. Dans cette thèse je vais tenter de résumer certains de ces travaux jugés intéressants pour mon travail :

En 1998, Turner et ses co-auteurs ont réalisé une étude pour prédire la température critique (Tc) et le volume critique (Vc) des composés organiques purs, y compris des hydrocarbures (**Turner *et al.* 1998**). Quatre modèles de type RQSP ont été développés sur la base de 164 composés organiques, dont environ 64% sont des hydrocarbures purs. Deux modèles sont basés sur la méthode de régression linéaire multiple (MLR) et les deux autres sont basés sur un réseau de neurone concurrentiel (CNN). Vu que les quatre modèles sont de type RQSP, huit (8) descripteurs ont été utilisés comme entrée pour chaque modèle. Les architectures des modèles CNN développés dans cette étude est 8-4-1 pour Tc, et est 8-5-1 pour Pc. Parmi les paramètres statistiques utilisés pour tester les performances de ces modèles on trouve la REQM. Les principaux résultats issus de ces modèles sont discutés et comparés dans le dernier chapitre de cette thèse.

Comme déjà mentionné, dans la même étude traitant le paramètre Tb menée en 2001 (**Espinosa *et al*. 2001**), Espinosa et ses co-auteurs ont également développé six modèles de type RQSP pour prédire deux propriétés critiques : la température critique (Tc) et la pression critique (Pc) des composés organiques purs, pour différentes familles d'hydrocarbures telles que les aliphatiques et les aromatiques. Les modèles se sont également basés sur : la méthode de réseau de neurone de type carte de théorie de la résonance adaptative floue (Fuzzy ARTMAP) pour développer quatre modèles RQSP ; deux pour chaque propriété (Tc et Pc), et la méthode de réseau de neurone à rétro propagation (bpNN) pour développer deux autres modèles RQSP ; un modèle pour chaque propriété (Tc et Pc). 530 et 463 composés ont respectivement été utilisés, lors du développement des modèles des propriétés Tc et Pc. Sept ou huit descripteurs moléculaires (la somme des nombres atomiques, cinq indices de connectivité de valence, et l'indice de forme kappa du second ordre, sans ou avec le moment dipolaire) ont été utilisés

comme entrée des modèles RQSP développés. Les modèles qui sont basés sur la méthode (Fuzzy ARTMAP) ont donné des résultats meilleurs mais instables au vu des différences entre les résultats de ses sous-ensembles de données. Les résultats complets des modèles développés dans cette étude seront discutés plus tard dans le dernier chapitre de cette thèse.

Dans la même étude effectuée en 2002 sur la densité relative du liquide (d20), Wakeham et ses co-auteurs ont également développé cinq modèles RQSP, basés principalement sur la méthode de régression linéaire (MLR) (**Wakeham *et al.* 2002**). Afin d'estimer les trois propriétés critiques des hydrocarbures, 222 composés ont été utilisés pour établir un modèle pour prédire la température critique (Tc), 219 composés ont été utilisés pour développer deux modèles pour prédire la pression critique (Pc), et 237 composés pour développer deux autres modèles afin de prédire le volume critique (Vc). Huit (8) descripteurs appartenant à différentes familles ont été adaptés comme entrée pour chaque modèle. Plusieurs paramètres d'analyse statistique ont été faits pour tester les performances de ces cinq modèles. Parmi elles, le coefficient de détermination (R^2) et l'ERAM. Une comparaison entre les résultats de ces modèles et nos modèles sera présentée dans le dernier chapitre de cette thèse.

En 2002, Yao et ses co-auteurs ont mené deux études séparées sur différents composés organiques purs, y compris des hydrocarbures, l'une pour prédire leur température critique (Tc) **(Yao *et al.* 2002a)**, et l'autre pour prédire leur pression critique (Pc) **(Yao *et al.* 2002b)**. Deux modèles RQSP ont été développés pour chaque propriété, dont un basé sur la méthode de régression linéaire (MLR), et l'autre sur la méthode de réseau de neurone à fonctions de base radiale (RBFNN). Le développement des deux modèles de Tc, ont été développés sur une base de données contenant 856 composés organiques purs, dont environ 330 composés sont des hydrocarbures purs. Tandis que les deux autres modèles de la propriété Pc ont eu une base de 173 composés de benzènes substitués, y compris les hydrocarbures. Dix (10) et neuf (09) descripteurs moléculaires appartenant à différentes classes, ont été utilisés comme entrée des modèles RQSP pour la propriété Tc et la propriété Pc, respectivement. Les architectures des modèles basés sur la méthode (RBFNN) étaient : 10-33-1 pour le modèle de Tc, et 09-19-1 pour le modèle de Pc. Parmi les paramètres d'analyse et de test les plus intéressants utilisés dans ces deux études ; la REQM. Les principaux résultats issus des modèles de ces deux études seront discutés dans le dernier chapitre de cette thèse.

Godavarthy et ses co-auteurs ont effectué, en 2008 une étude sur des composés organiques purs pour la prédiction des trois propriétés critiques : Tc, Pc et Vc **(Godavarthy *et***

***al.* 2008)**. Pour atteindre cet objectif, cinq études de cas distincts ont été menées séquentiellement ; Étude du cas 1 : sur le développement des modèles RQSP linéaires utilisant des descripteurs issus de trois paquets RQSP différents. Étude du cas 2 : sur le développement des modèles RQSP linéaires utilisant une analyse de régression linéaire multiple (MLR) et des algorithmes des moindres carrés partiels (PLS). Étude du cas 3 : sur le développement des modèles RQSP linéaires comprenant des descripteurs de groupes fonctionnels structurels construits. Étude du cas 4 : sur le développement des réseaux de neurone traditionnels de type rétro-propagation avec l'utilisation de la somme des carrés (SOS) pour la propagation des erreurs. Étude du cas 5 : sur le développement des nouveaux réseaux de neurones à rétro-propagation, utilisant l'erreur quadratique moyenne pondérée (WMSE) pour la propagation des erreurs. La base des données utilisée lors du développement de ces modèles englobe 1230 composés organiques purs, y compris 73 classes différentes des hydrocarbures. Plusieurs paramètres d'analyse statistiques ont été utilisés afin de tester les performances de ces modèles dont l'ERAM qui montre bien la supériorité des modèles RQSP de type WMSE-bpNN par rapport aux autres modèles lors de la prédiction des trois propriétés critiques des composés organique et des hydrocarbures purs. Les architectures de ces modèles étaient : (7-8-1) pour la température critique Tc, (7-12-1) pour la pression critique Pc, (6-4-1) pour le volume critique Vc. Une comparaison sera effectuée entre ces résultats et ceux obtenus par les modèles développés dans cette étude dans le dernier chapitre de cette thèse.

Dans la même étude mentionnée précédemment, section traitant la température d'ébullition (Tb), Sola et ses co-auteurs ont développé, en 2008 deux autres modèles RQSP basés sur la méthode de régression linéaire multiple (MLR) **(Sola *et al.* 2008)**, traitant les composés organiques et les hydrocarbures purs. Le premier modèle destiné pour prédire la température critique (Tc), et le seconde pour prédire la pression critique (Pc). Deux bases de données comportant 153 et 139 composés ont été utilisées lors du développement de chacun des deux modèles, respectivement. Le nombre des entrées utilisées pour chaque modèle était égal à huit (8) descripteurs moléculaires. Parmi les paramètres d'analyses statistiques utilisées pour tester les deux modèles développés, on retrouve la REQM. Les principaux résultats issus par ces modèles seront discutés en détails dans le dernier chapitre de cette thèse.

Une année après, Zuas et Styarini ont réalisé une étude **(Zuas & Styarini 2009)**, afin de développer un modèle de type RQSP pour prédire le volume critique (Vc) de certains hydrocarbures purs insaturés. Le modèle a également été basé sur la méthode de régression

linéaire multiple (MLR) avec une base de données contenant environ 65 hydrocarbures aliphatiques purs insaturés tels que les alcènes et les alcynes. Les entrées utilisées dans le développement de ce modèle sont au nombre de six (6) descripteurs appartenant au type des indices de connectivités simples. Peu de paramètres d'analyse statistiques ont été utilisés pour tester la performance de ce modèle. Une comparaison des résultats issus de ce modèle et d'autres résultats obtenus par mon modèle sera discutée ultérieurement dans le dernier chapitre de cette thèse.

En 2018 Banchero et Manna ont réalisé une étude importante, pour la modélisation de la température critique (Tc) et la pression critique (Pc) des différents composés organiques et hydrocarbures purs **(Banchero et Manna 2018)**. La base de données utilisée dans cette étude englobe 306 composés organiques, dans laquelle presque 40% sont des hydrocarbures. Deux modèles RQSP ont été développés pour chaque modèle, où les premiers modèles sont basés sur la méthode de régression linéaire (MLR) et les seconds modèles sont basés sur la méthode de réseau de neurone à fonction de base radiale (RBFNN). Dix (10) descripteurs moléculaires appartenant à plusieurs classes tels que les descripteurs topologiques ont été utilisés comme entrée pour chaque modèle. Les architectures des meilleurs modèles RQSP-RBFNN trouvés sont (10-42-1) pour Tc, et (10-44-1) pour Pc. Les deux types d'erreurs, la REQM et l'ERAM, sont parmi les paramètres d'analyse et de test statistiques intéressants qui ont été utilisés dans cette étude pour mesurer la performance de prédiction de ces modèles. Les résultats de ces derniers seront discutés en détails dans le chapitre cinq de cette thèse.

III.3. Revue des travaux sur la prédiction des cinq propriétés d'intérêts des hydrocarbures réalisés par d'autres méthodologies

III.3.1. Travaux réalisés par des méthodes computationnels

Mohammadi et ses co-auteurs ont mené une étude, en 2008 sur les hydrocarbures lourds et les fractions pétrolières **(Mohammadi *et al*. 2008)**. L'objectif de cette étude était la prédiction des trois propriétés critiques : la température critique (Tc) exprimée en (K), la pression critique (Pc) exprimée en (MPa), le volume critique (Vc) exprimé en ($m^3.Kmol^{-1}$) et le facteur acentrique qui n'a pas d'importance avec notre cas d'étude. Les modèles ont été basés sur une méthode computationnelle d'intelligence artificielle qui est un réseau de neurone artificiel à alimentation en avant, dénommé en anglais : Feed-forward (ANN-FF). Ils ont utilisé une base de données comportant 67 composés et fractions pétrolières de différents nombres de carbone (C1–C45)

divisées en 44, 18 et 5 composés pour les phases d'apprentissage, de validation et de test, respectivement. Deux autres paramètres ont été utilisés comme entrée du modèle : la gravité spécifique (SG) et la température d'ébullition moyenne (Tb_{Av}). Un algorithme d'apprentissage de type "Levenberg-Marquardt (LM)" a été utilisé pour l'optimisation et l'apprentissage de réseau de neurone. De plus, les deux fonctions d'activation de type "*Exponentiel*" et "*Sigmoïde*" ont été également utilisés pour le développement de ce modèle. Nombreux paramètres d'analyse ont été utilisés pour tester la performance de ces modèles tels que l'écart relative absolu moyen (AARD) exprimé en (%). Les résultats de la prédiction étaient très appréciables pour les modèles des fractions pétrolières et des hydrocarbures lourds, dans lesquels, seulement 0,1%, 0,8% et 1,3% ont été obtenus respectivement, pour Tc, Pc et Vc pour les modèles ANN-FF des fractions pétrolières, tandis que 1%, 8% et 2,7% ont été obtenus pour les mêmes paramètres de modèles destinés aux hydrocarbures lourds, respectivement.

En 2011, Gharagheizi et ses co-auteurs ont réalisé une étude pour différentes classes de composés organiques, dont quelques familles d'hydrocarbures (**Gharagheizi *et al*. 2011**). L'étude a une large base de données contenant 1700 composées organiques divisée en 81 classes distinctes pour le développement des modèles computationnels basés sur la combinaison de deux méthodes qui sont : la méthode de contribution des groupes (GC) et le réseau de neurones artificiels de type feed-forward (FF-ANN). Six (6) modèles ont été développés afin d'estimer la température critiques (Tc), la pression critique (Pc) et le volume critique (Vc) des composés étudiés. Les entrées de chaque modèle sont des groupes fonctionnels liés à ces trois propriétés, dans lesquels : 149 entrées ont été utilisées pour les modèles de Tc et Pc, tandis que 166 entrées sont utilisées pour les modèles de Vc. L'algorithme d'apprentissage utilisé dans ces modèles est de type "Levenberg-Marquardt (LM)", la somme des carrées (SOS) a été utilisé comme fonction d'erreur. Pendant l'apprentissage de RNA, les neurones cachés varient entre 1 et 50. Les architectures des meilleurs modèles était : (149-15-1) pour le modèle de Tc, (149-15-1) pour le modèle de Pc, et (166-15-1) pour le modèle de Vc. Le coefficient de détermination (R2) et l'ERAM sont des paramètres d'analyse statistique intéressants et ont été utilisés pour tester les performances de ces modèles. Les résultats de ces derniers seront discutés dans une comparaison effectuée dans le dernier chapitre de cette thèse.

En 2017, Varamesh et ses co-auteurs ont effectué une étude (**Varamesh *et al*. 2017a)** sur la modélisation de la température d'ébullition (Tb) des composés organiques purs, y compris les hydrocarbures. Dans cette étude, la base de données utilisée pour développer les quatre (4)

modèles computationnels englobe environ 563 composés. Le premier modèle est de type "MLP-ANN", le second est de type "RBF-ANN", le troisième est de type "LSSVM", et le dernier modèle est de type "GMDH-NN". Vu qu'ils n'appartiennent pas au type des modèles RQSP, les entrées utilisées pour ces modèles sont : la masse moléculaire (MW) et la densité relative du liquide appelée dans les références anglo-saxon "Specifique Gravity" (SG). Ce sont deux propriétés très connues et qui ont des relations avec la température d'ébullition (Tb). Les auteurs ont utilisé différentes méthodes et paramètres d'analyse statistique intéressants tels que le calcul de la REQM et de l'ERAM, ainsi que la détermination du domaine d'applicabilité (DA) pour chaque modèle, qui sont des paramètres semblables à ceux utilisés dans l'analyse des modèles développés dans cette thèse. Les principaux résultats de ces modèles seront discutés dans le dernier chapitre de cette thèse.

Dans la même année, Varamesh et ses co-auteurs ont également réalisé une autre étude sur les composés organiques purs et les hydrocarbures **(Varamesh *et al*. 2017b)**, similaire à celle qui traite la température d'ébullition (Tb), pour prédire la température critique (Tc), la pression critique (Pc) et le volume critique (Vc). Quatre (4) types de modèles computationnels ont été développés pour chaque propriété, dans lesquels ils se sont basés sur quatre techniques différentes, qui sont : "MLP-ANN", "RBF-ANN", "LSSVM", et "GMDH-NN". La base de données utilisée pour développer ces modèles englobe environ 563 composés organiques, dont les hydrocarbures. Trois paramètres ont été utilisés comme entrée pour ces modèles, qui sont : la température normale d'ébullition (Tb), la masse moléculaire (MW) et la gravité spécifique (SG) dans lesquels : $Tc = f(Tb, SG)$, $Pc = f(Tb, SG)$, et $Vc = f(MW, SG)$. Différentes méthodes et paramètres d'analyse statistique intéressants ont été utilisés pour tester les performances de ces modèles, comme la REQM, l'ERAM, et la détermination du domaine d'applicabilité (DA), qui sont des paramètres semblables à ceux utilisés dans l'analyse des modèles développés dans cette thèse. Les principaux résultats de ces modèles seront aussi discutés dans le dernier chapitre de cette thèse.

III.3.2. Travaux réalisés par d'autres méthodes

En 2002, Retzekas et ses co-auteurs ont réalisé une étude sur la prédiction de quelques propriétés physiques tels que : la température du point d'ébullition normale (Tb), la température critique (Tc) et la pression critique (Pc) pour des hydrocarbures purs, des fractions pétrolières et enfin des fractions liquides de charbon qui ne sont pas en rapport avec mon cas d'étude **(Retzekas *et al*. 2002)**. La base de données utilisée englobe 293 composés d'hydrocarbures

distribués en cinq classes distinctes : 43 et 78 composés alcanes, 22 et 29 composés alcènes, 12 et 8 composés d'alcynes, 11 et 8 composés alcadiènes et enfin 22 et 60 composés aromatiques, respectivement pour les phases d'apprentissage et de validations dans chaque classe. Les modèles développés pour les hydrocarbures purs ont été basés sur la combinaison de deux termes, un terme qui comporte des paramètres empiriques, comme la densité du liquide (d20) et la masse moléculaire (MW), et un second terme basé sur des paramètres structurels d'approche de contribution des groupes, précisément sur les 15 groupements fonctionnels suivants : CH_3, –CH_2–, >CH–, >C<, =CH_2, =CH–, =C<, =C=, #CH, #C–, –CH2– (cycle), >CH– (cycle), >C< (cycle), =CH– (cycle) et =C< (cycle). L'expression générale du modèle est illustrée dans l'**Eq (III.1)** :

$$\boldsymbol{Q = a * d20^{b} * MW^{c} + d + \sum_{i} N_{i} * G_{i}} \qquad \textbf{(III.1)}$$

Où : $\boldsymbol{Q}$ représente les trois propriétés Tb, Tc et Pc, $\boldsymbol{d20}$ est la densité du liquide à 20 °C, $\boldsymbol{MW}$ est le poids moléculaire du composé ; les paramètres ($\boldsymbol{a - d}$) sont des constantes des modèles identiques pour tous les hydrocarbures mais différentes pour chaque propriété ; $\boldsymbol{N_i}$ est le nombre de fois où le groupe $\boldsymbol{i}$ apparaît dans un composé ; et $\boldsymbol{G_i}$ est la valeur de la contribution du groupe $\boldsymbol{i}$. Les valeurs des groupes utilisées lors de ce travail sont les mêmes que celle proposées par Joback.

Parmi les paramètres d'analyse utilisés pour tester la précision de prédiction de ces modèles, l'erreur relative absolue moyenne (ERAM). Les résultats sont très acceptables en comparaison avec d'autres modèles dans la littérature. La valeur de l'ERAM est de 1% pour le modèle de Tb contre 3,3% pour la méthode de Joback et 2,9% pour celle de Stein & Brown. Son principal avantage par rapport à la première méthode réside dans les composés de poids moléculaire élevé et celui par rapport à la seconde avec des composés hautement ramifiés. Pour la prédiction de Tc, le modèle proposé donne une ERAM de 1% similaire à celle des méthodes de Joback, de Riazi et de Riazi-Daubert qui nécessitent cependant la connaissance de Tb. Pour Pc, le modèle donne une ERAM de 2,7% contre 3,9% pour la méthode de Joback, 4,2% et 4,8% respectivement pour les méthodes de Riazi et de Riazi-Daubert qui nécessitent également la connaissance de Tb. La méthode proposée donne de meilleurs résultats pour ces trois propriétés par rapport à la méthode de contribution d'interaction de groupe (GIC) récemment proposée et plus difficile à utiliser de Marejon et Fontevila. Dans la deuxième partie de cette étude, les données des hydrocarbures purs ont été utilisées, afin de développer des corrélations pour la prédiction des paramètres Tc et Pc des fractions liquides du pétrole. Les prédictions des

paramètres Tc et Pc sont faites avec des erreurs de 1,2% et 5,5% similaires à celles de la méthode de Riazi-Daubert. Cependant aucune conclusion ne peut être tirée sur la fiabilité de ces méthodes en raison du petit nombre de données disponibles.

En 2006, Li et ses co-auteurs ont réalisé une étude **(Li *et al*. 2006)** sur la modélisation de la température d'ébullition normale (Tb) pour différents types d'hydrocarbures légers, lourds et complexes tels que : les alcanes, les alcènes, les alcynes et les hydrocarbures cycliques et aromatiques, y compris leurs isomères. Une méthode d'espace vectoriel de groupe modifiée (GVS) (en anglais : *a modified group vector space method*) a été appliquée avec succès dans le développement du modèle, où le principe de cette méthode est basé sur la méthode de contribution des groupes ainsi que sur les contributions topologiques, qui peuvent représenter avec précision le point d'ébullition normale des composés isomères. Les paramètres de groupe pour la méthode GVS modifiée ont été obtenus à partir de la corrélation des points d'ébullition de 1041 hydrocarbures. Des paramètres d'analyse statistique comme le coefficient de corrélation (R), la REQM exprimé en (K), et d'autre type d'erreur relative qui est l'ERAM exprimée en (%), ont été calculés pour tester la capacité de prédiction de ce modèle. Les résultats ont montré la supériorité de prédiction du point d'ébullition d'une large gamme et d'une grande variété d'hydrocarbures par ce modèle par rapport aux résultats calculés par d'autres méthodes conventionnelles basées sur la contribution de groupe qui n'utilisent pas l'approche d'indice topologique de groupe. Dans lesquels des valeurs totales acceptables de la REQM égales à 5,97K, seulement 4,06 K pour l'erreur absolue moyenne (EAM) et 0,86% pour l'erreur relative absolue moyenne (ERAM). La valeur du coefficient de corrélation (R=0,9992) proche de l'idéal a aussi montré l'efficacité de ce modèle dans la prédiction du paramètre Tb. Enfin, le modèle de type GVS montre une bonne capacité à la fois à représenter la température d'ébullition des hydrocarbures et à distinguer les différences de température entre les isomères structuraux des hydrocarbures par rapport aux méthodes classiques dans les littératures.

En 2008, Souahi et Kaabeche ont effectué une étude pour le développement des modèles de corrélation afin d'estimer les paramètres critiques des hydrocarbures et des fractions pétrolières (**Souahi & Kaabeche 2008**). Les modèles ont été basés sur la méthode des algorithmes génétiques. L'objectif principal de cette étude était de trouver une fonction mathématique reliant les propriétés critiques avec deux autres propriétés utilisées comme entrée et qui sont la température du point d'ébullition normale (Tb) et la densité relative des liquides

connus sous le nom gravité spécifique (SG), à partir de la forme de fonction générale comme montré dans l'**Eq (III.2)** :

$$\Pi_c = \left(\boldsymbol{a} * \boldsymbol{Tb}^{\boldsymbol{b}} + \boldsymbol{c} * \boldsymbol{SG}^{\boldsymbol{d}} + \boldsymbol{e} * \boldsymbol{Tb}^{\boldsymbol{f}} * \boldsymbol{SG}^{\boldsymbol{g}}\right) * \boldsymbol{exp}\left[\frac{\boldsymbol{SG}+\boldsymbol{h}}{\boldsymbol{Tb}+\boldsymbol{i}}\right] \qquad \textbf{(III.2)}$$

Où Π_c est un paramètre critique qui peut être Tc en (K), Pc en (bar) ou Vc en (cm^3/g.mol). Les paramètres $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$, $\boldsymbol{f}$, $\boldsymbol{g}$, $\boldsymbol{h}$ et $\boldsymbol{i}$ sont des constantes propres à chaque modèle, obtenus et optimisés par le bais des algorithmes génétiques (AG). Les modèles finaux obtenus pour chaque paramètre dans cette étude, sont illustrés dans : l'**Eq (III.3)** pour Tc, l'**Eq (III.4)** pour Pc, et l'**Eq (III.5)** pour Vc.

$$\boldsymbol{T_c} = \left(\boldsymbol{17,8443} * \boldsymbol{Tb}^{\boldsymbol{0,5965}} * \boldsymbol{SG}^{\boldsymbol{0,2945}}\right) * \boldsymbol{exp}\left[\frac{\boldsymbol{SG}}{\boldsymbol{Tb}-\boldsymbol{100}}\right] \qquad \textbf{(III.3)}$$

$$\boldsymbol{P_c} = \left(\boldsymbol{6,5243} * \boldsymbol{10}^{\boldsymbol{6}} * \boldsymbol{Tb}^{\boldsymbol{-1,9781}} * \boldsymbol{SG}^{\boldsymbol{1,8387}}\right) * \boldsymbol{exp}\left[\frac{\boldsymbol{SG}}{\boldsymbol{Tb}-\boldsymbol{250}}\right] \qquad \textbf{(III.4)}$$

$$\boldsymbol{V_c} = \boldsymbol{0,0038} * (\boldsymbol{Tb} - \boldsymbol{273,16})^{\boldsymbol{2}} + \boldsymbol{0,9814} * (\boldsymbol{Tb} - \boldsymbol{273,16}) * \boldsymbol{SG} + \boldsymbol{310,6783} \qquad \textbf{(III.5)}$$

Ces modèles ont été développés à partir d'une base de données qui comporte environ 109 composées appartenant aux différents hydrocarbures de C6 jusqu'à C30, répartis comme suit : 40 composés pour les alcanes, 15 composés pour les alcènes, 09 composés pour les alcynes, 20 composés aromatiques et 25 composés naphténiques. Les trois modèles ont été évalués à l'aide d'une comparaison avec les résultats obtenus par d'autres modèles en termes d'erreur absolue moyenne (AAE%). De sorte qu'elles étaient : 0,3500% pour le modèle de Tc contre 0,6230% obtenue par un autre modèle trouvé dans la littérature, 2,7225% pour le modèle de Pc contre 5,0230% obtenue par une autre corrélation, et seulement 0,8063% pour le modèle de Vc contre 2,0805% obtenue par un autre modèle dans la littérature.

Carroll et ses co-auteurs ont mène une étude en 2011, afin de développer une méthode simple pour prédire le point d'ébullition (Tb) des hydrocarbures, particulièrement des alcènes acycliques (**Carroll *et al.* 2011**). Le point d'ébullition (Tb) des alcènes acycliques sont prédits à partir de leurs nombres de points d'ébullition ($\boldsymbol{Y_{Bp}}$), comme l'indique la relation de corrélation présentée dans l'**Eq (III.6)**.

$$\boldsymbol{Tb(K)} = \boldsymbol{-16,802} * \boldsymbol{Y_{Bp}^{2/3}} + \boldsymbol{337,377} * \boldsymbol{Y_{Bp}^{1/3}} - \boldsymbol{437,883} \qquad \textbf{(III.6)}$$

Avec : $\boldsymbol{Y_{Bp}} = \boldsymbol{f}(\boldsymbol{Q_i}, \boldsymbol{C}, \boldsymbol{M_3}, \boldsymbol{M}, \boldsymbol{E_3}, \boldsymbol{E}, \boldsymbol{D}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{T}, \boldsymbol{S})$;

Où les 11 paramètres dépendent de la nature et la structure des composés d'alcènes acycliques étudiés.

Le modèle a été fondé sur une base de données de 250 composés d'alcanes acycliques de différentes tailles ou bien de différents nombres de carbone : de C3 jusqu'à C36. Plusieurs paramètres statistiques d'analyse ont été utilisés pour tester ce modèle, tels que l'écart absolu moyen (AAD) exprimé en (K) et le coefficient de détermination (R^2), qui étaient très acceptables avec seulement 1,29 K pour l'AAD et 0,999 pour R^2. Cette méthode de prédiction est non seulement très simple à utiliser, mais est également comparable en précision aux autres méthodes de réseau neurone et aux méthodes MLR nécessitant un paramètre de connectivité ou des paramètres topologiques. De plus, les valeurs ($\boldsymbol{Y_{Bp}}$) calculées avec l'**Eq (III.6)** fournissent un moyen utile de prédire le point d'ébullition (Tb) des alcènes. Des efforts pour développer des corrélations similaires pour des composés contenant d'autres groupes fonctionnels sont actuellement en cours d'étude.

En 2016, Ghasemitabar et Movagharnejad ont effectué une étude sur l'estimation de la température du point d'ébullition normale (Tb) pour de larges familles de composés organiques, y compris des hydrocarbures (**Ghasemitabar & Movagharnejad 2016**). Un grand ensemble de données expérimentales comptant 2036 composés organiques répartis sur plusieurs familles, y compris 748 composés d'hydrocarbures purs, ont été utilisées afin de développer un modèle basé sur la méthode de contribution des groupes du deuxième ordre, dont 175 composés sont des alcanes, 107 autres sont des cycloalcanes, 166 composées sont des alcènes et 44 sont des alcènes cycliques, 40 composés d'alcynes, 160 composées sont des benzène et 58 autres composés aromatiques. Les 1288 composés restants appartenant à différentes familles organiques, telle que les alcools, les ester et autres composées d'halogènes. De plus, 122 valeurs des contributions de groupes fonctionnels ont été optimisées en utilisant ces précédentes données expérimentales et l'algorithme "*fminsearch*" du logiciel MATLAB. Plusieurs paramètres d'analyse statistique ont été employés pour tester l'exactitude de ce modèle lors de la prédiction de la propriété Tb des différents composés organiques, tels que l'écart absolus moyen (AAD) et le pourcentage des écarts relatifs absolus moyens (AARD), qui était respectivement 4,35194 K et juste 1,01075% pour la totalité des composées de la base de données. Les résultats de ce modèle étaient acceptables par rapport aux valeurs de AAD des prédictions de : 4,03760 K pour 1084 isomères structurels, 2,95095 K pour 67 isomères de type *cis* avec, et seulement 2,60164 K pour 78 isomères de type *trans*. Finalement, les résultats de la prédiction par le modèle développé dans

cette étude, étaient beaucoup plus précis par rapport aux résultats des autres méthodes de même type connues dans la littérature, comme la méthode de Joback et Reid, la méthode de Constantinou et Gani, la méthode de Marrero et Pardillo, et la méthode de Marrero et Gani.

Li et ses co-auteurs ont traité une étude en 2017 pour développer des modèles mathématiques capables de prédire la température critique (Tc) de larges gammes de composés organiques **(Li *et al*. 2017)**. Une base de données de 616 composés a permis de développer 16 corrélations pour différentes classes, y compris les familles des hydrocarbures où ils étaient répartis comme suit : 161 composés pour les alcanes, 53 composés pour les oléfines, 25 composés pour les alcynes et 40 composés pour les hydrocarbures aromatiques. Les modèles ont été principalement basés sur la théorie des relations linéaires d'énergie libre (LFER) et des formules de thermodynamique. Les capacités et la précision des modèles lors de la prédiction du paramètre Tc, ont été testées par plusieurs paramètres d'analyse statistique, tels que les écarts relatifs moyens (MD) qui ont pris des valeurs de 0,01% à 2,73% où la plupart d'entre eux sont inférieurs à 2%, et les coefficients de détermination (R^2) qui vont de 0,90 à 0,98. Les résultats ont montré que les équations présentent un meilleur effet avec une forme simple d'équation et une précision de prédiction élevée par rapport à d'autres modèles du même type. Enfin, les substances étudiées dans cet article sont abondantes et appartiennent à des types larges.

En 2018, Li et ses co-auteurs ont effectué une autre étude sur la modélisation de la température d'ébullition normale (Tb) pour une large gamme de composés organiques liquides **(Li *et al*. 2018)**. Une base de données contenant environ 683 composés appartenant aux 15 classes différentes de composés organiques liquides, y compris plusieurs familles d'hydrocarbures, tels que : les alcanes avec 160 composés, les oléfines avec 58 composés, les alcynes avec 21 composés et les arènes avec 39 composés. Les corrélations proposées ont été basées sur des relations linéaires d'énergie libre (LFER) et des formules thermodynamiques, pour raccorder les Tb aux descripteurs moléculaires qui ont été obtenus en utilisant une régression linéaire multiple (MLR) pas à pas. La formule générale des modèles est indiquée par l'**Eq (III.7)**.

$$\boldsymbol{Tb(K) = c + e * E + s * S + a * A + b * B + v * V} \qquad \textbf{(III.7)}$$

Où : $\boldsymbol{E}$, $\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{B}$ et $\boldsymbol{V}$ sont les entrées ou descripteurs des modèles, $\boldsymbol{c}$, $\boldsymbol{e}$, $\boldsymbol{s}$, $\boldsymbol{a}$, $\boldsymbol{b}$ et $\boldsymbol{v}$ sont ses coefficients ou constantes associés.

Les modèles ont été testés par le biais des écarts relatifs moyens (MD) exprimés en (%). Les résultats étaient acceptables, avec une valeur de MD située entre 0,89% et 2,83%, étant pour la plupart inférieurs à 2%. Les mêmes résultats sont obtenus pour les modèles des hydrocarbures avec : 1,72% pour le modèle des alcanes, 2,6% pour le modèle des oléfines, 1,8% pour le modèle des alcynes et seulement 0,89% pour le modèle des arènes. De plus, les coefficients de corrélation de 13 équations, y compris les modèles des hydrocarbures, étaient supérieurs à 0,98. Les équations de corrélation obtenues dans cette étude sont simples, ont un sens physique clair, un large domaine d'applicabilité et fournissent une puissance prédictive et une précision de calcul assez élevées par rapport aux nombreuses méthodes classiques dans la littérature.

Tahami et ses co-auteurs ont mené une très bonne étude en 2019 sur l'estimation des trois paramètres critiques pour de larges familles de composés organiques, y compris les hydrocarbures **(Tahami *et al.* 2019)**. Une nouvelle méthode de contribution de groupes du second ordre a été utilisée pour développer des modèles mathématiques afin d'estimer les trois propriétés critiques d'intérêts : la température critique (Tc), la pression critique (Pc) et le volume critique (Vc) pour de larges gammes de composés organiques. Les groupes fonctionnels de la méthode de Joback et de Reid sont utilisés avec quelques changements en tant que groupes fonctionnels du premier ordre et certains groupes fonctionnels du second ordre sont également définis et ajoutés. Un ensemble de données expérimentales contenant 969, 715 et 539 composés, respectivement pour les paramètres Tc, Pc, et Vc ont été collectés dans cette étude. Les valeurs de contribution de 122 groupes fonctionnels ont été optimisées en utilisant les données expérimentales de composés organiques, y compris des hydrocarbures aromatiques polycycliques complexes et lourds. Les trois propriétés critiques de la plupart des isomères structuraux et des stéréo-isomères (*cis* et *trans*…) sont distinguées dans cette méthode en considérant les groupes fonctionnels du second ordre. Les hydrocarbures et dérivés d'hydrocarbures considérés dans cette recherche comprennent principalement les atomes de carbone et d'hydrogène, mais aussi des atomes d'oxygène, d'azote, de soufre, de fluor, de chlore, de brome et d'iode. Le développement des nouveaux modèles et l'optimisation des valeurs de contribution de groupes ont été réalisés par l'algorithme "*fminsearch*" du logiciel MATLAB. Les résultats de la prédiction de la nouvelle méthode étaient acceptables en comparaison avec d'autres méthodes de même type bien connues dans la littérature, tels que : la méthode de Joback et Reid, la méthode de Constantinou et Gani et la méthode de Marrero et Gani, dans laquelle, le pourcentage des écarts relatifs absolus moyens (AARD) était de 1,12% pour Tc, 2,69% pour Pc

et 2,67% pour Vc. Tous ces résultats impliquent également que la nouvelle méthode d'estimation est beaucoup plus précise que les méthodes mentionnées pour une large gamme de composés organiques.

III.4. Conclusion

L'importance des propriétés physicochimiques et thermodynamiques des hydrocarbures purs et des fractions pétrolières dans de nombreux secteurs industriels d'un côté, et les différentes contraintes liées aux méthodes expérimentales pour ses déterminations de l'autre côté, sont les principales motivations pour recourir aux méthodes de modélisation qui ont largement été utilisées par plusieurs chercheurs scientifiques, ces dernières années. Comme montré dans ce chapitre, de nombreuses méthodes ont été adaptées, chacune ayant ses avantages, ses inconvénients, ses utilisations et ses limites. Grace aux progrès technologiques dans le domaine informatique, plusieurs méthodes expérimentales ou modèles empiriques connus, ont été complétées ou remplacés par de nouvelles méthodes basées sur l'intelligence et l'apprentissage artificiel pour résoudre différents problèmes liés à ce concept.

En conséquence, dans ce chapitre j'ai essayé de donner un résumé de l'état de l'art et revue critique de la littérature sur la détermination des propriétés des hydrocarbures. Les études citées dans ce chapitre sont liées principalement aux méthodes de détermination des cinq propriétés d'intérêts (Tb, d20, Tc, Pc et Vc), qui sont traitées en se basant sur le concept RQSP en premier lieu ou par d'autres méthodes computationnelles d'intelligence artificielle telles que les différents types de réseaux de neurones artificiels. A la fin, dans la dernière partie, certaines études intéressantes développant des modèles basés sur d'autres méthodes telles que, les corrélations empiriques ou les méthodes de contribution des groupes ont été présentées.

En outre, au début de ce chapitre, j'ai fourni quelques statistiques approximatives étendues qui m'ont donné une idée sur l'évolution des différentes publications telles que : des articles de recherche qui représentent plus de 90% des documents scientifiques, des articles de revue critiques (en anglais : review articles), des livres et des chapitres, et certains documents comme des communications et résumés de conférences. Je me suis basé sur des statistiques de différentes études RQSP et ANN, et les études menées pour la détermination des propriétés des hydrocarbures par différentes méthodes et par les deux précédentes méthodes publiées dans la plateforme Sciencesdirect dans une période de 20 ans, entre les années 2000 et 2020.

Chapitre IV

Matériels et méthodes

IV.1. Introduction

Habituellement, les modèles de type RQSP doivent être construits en suivant ces trois étapes fondamentales : **(1)** Choix d'un ensemble de propriétés moléculaires, Tb, d20, Tc, Pc et Vc dans cette étude ; **(2)** Extraction des informations structurelles des molécules par le calcul et le choix des meilleurs descripteurs moléculaires pertinents ; **(3)** Lier la première et la deuxième étape à l'aide d'une méthode d'apprentissage mathématique ou numérique **(Hamadache *et al*. 2018)**. En conséquence, cette étude visait à développer des modèles de RQSP-RNA afin de prédire cinq paramètres physico-chimiques (Tb, d20, Tc, Pc et Vc) des hydrocarbures purs issus d'une fraction pétrolière légère. Entre 223 et 221 molécules d'hydrocarbures purs ont été utilisées comme base de données pour chaque paramètre. De ce fait, pour ces structures moléculaires, des centaines de descripteurs ont d'abord été calculés. Ensuite, une opération de sélection a été effectuée à l'aide d'une méthodologie statistique pour obtenir juste quelques descripteurs moléculaires pertinents. Finalement, la relation entre les structures représentées par les descripteurs moléculaires et les propriétés a été appliquée à l'aide d'un RNA-PMC. Les étapes mentionnées ci-dessus ont été respectées pour développer les modèles de RQSP-RNA. Les meilleurs modèles finaux ont été testés avec différents types d'erreur et paramètres statistiques, afin de prédire les cinq paramètres précédents (Tb, d20, Tc, Pc et Vc) des hydrocarbures purs issus des fractions pétrolières avec un grand degré de précision.

Dans ce chapitre je vais montrer tous les outils, les matériels et les méthodes utilisés dans la réalisation de cette étude.

IV.2. Outils informatiques utilisés

Plusieurs logiciels ont été utilisés pendant la réalisation de ce travail, comme le montre la **Figure IV.1** :

- J'ai utilisé le "Microsoft Office Excel 2013", surtout dans l'organisation et le prétraitement de la base des données qui se fait dans la plupart du temps manuellement ;
- Le design des structures chimiques et l'extraction des notifications SMILES ont été faite par le multi-logiciel "ChemDraw Office 12" **(Kerwin 2010 ; Cousins 2011 ; ChemOffice 2012)** ;
- Pour le calcul des descripteurs moléculaires, on a utilisé le serveur "E-Dragon en linge" ou bien "Dragon 1.0" **(VCCLAB 2005 ; Tetko *et al*. 2005)**, parce que d'une part c'est un logiciel en ligne à accès ouvert, et d'autre part à cause de son utilisation fréquente et son

succès dans plusieurs études et publications scientifiques **(Kraim *et al.* 2009 ; Hamadache *et al*. 2014 ; Hamadache *et al*. 2016b)** ;

- L'ensemble d'apprentissage des modèles RQSP-RNA et pas mal des étapes de réduction des descripteurs, ont été réalisés par le logiciel "STATISTICA Versions 07 et 08" **(Stat Soft 2007)**, en vue de son utilisation fréquente dans les publications telles que celles de Hamadache et ses co-auteurs **(Hamadache *et al*. 2016b ; Hamadache *et al*. 2014)**, ou bien celles d'Ammi et ses co-auteurs **(Ammi *et al.* 2015 ; Ammi *et al.* 2017)** ;
- Tous les graphes ont été réalisés en utilisant le logiciel "MATLAB R2013b" **(The Mathworks Inc 2013)**, avec l'utilisation de quelques programmes ou bien ce qu'on appelle les scripts, comme dans le cas de la boîte à outils du domaine d'applicabilité (DA) qui a été réalisé par un groupe de chercheurs Italiens spécialisés dans le chemo-informatique, programmé avec le langage Matlab **(Sahigara *et al.* 2012 ; Sahigara *et al.* 2014)**.

Cependant, il existe plusieurs autres logiciels gratuits et commerciaux à chaque stade du développement des modèles de RQSP/RQSA mentionnés précédemment, tels que :

- Dragon (**Mauri *et al*. 2006 ; TALETE srl 2007 ; Kode Chemoinformatics srl 2019)** et PaDel Descriptors **(Yap 2011 ; Dong *et al*. 2015)** sont principalement utilisés pour le calcul des descripteurs moléculaires et les empreintes digitales (*fingerprint*), cependant il existe plusieurs autres logiciels qui ne sont pas aussi connus, comme : PCLIENT **(VCCLAB 2005 ; Tetko *et al*. 2005)**, PreADMET **(BMDRC 2019)**, VolSurf **(Cruciani *et al*. 2000)**.
- ACD/ChemSketch **(Spessard 1998 ; Masunov 2001)**, MarvinSketch **(Csizmadia 1999)** et Gaussian (**Frisch *et al*. 2016**) utilisés essentiellement dans le dessin des différentes molécules chimiques ainsi que pour générer leurs structures 3D, mais aussi dans le calcul de quelques descripteurs moléculaires.
- Boîte à outils de réseau de neurones MATLAB (*MATLAB Neural Network Toolbox*) **(Beale *et al*. 2010 ; The Mathworks Inc 2013)**, XLSTAT **(XLSTAT 2009)**, AutoWeka **(Hall *et al*. 2009 ; Kotthoff *et al*. 2017)** pour le développement des modèles de RQSP/RQSA par différentes techniques de régression linéaire et non linéaire telles que : la RLM, le RNA, ou bien la régression du vecteur de support (RVS) (*Support Vector Regression* ***SVR***).



Figure IV.1. Capture d'écran des logiciels utilisés dans cette étude : ChemDraw Office, MATLAB, STATISTICA, Boîte à outils de DA, Microsoft Office Excel, et E-Dragon 1.0.

Il convient de noter que tous les calculs de ce travail ont été effectués à l'aide d'un ordinateur personnel TOSHIBA doté d'un processeur Intel (R) Core^{TM} i3-3120M à 2,50 GHz de CPU et de 4 Go de mémoire vive (RAM), avec Windows 7 Professional comme système d'exploitation.

IV.3. Méthodologie générale de la modélisation RQSP utilisée

Plusieurs étapes ont été utilisées dans le développement des modèles RQSP-RNA, comme décrit ci-dessus et résumé à la **Figure IV.2**.

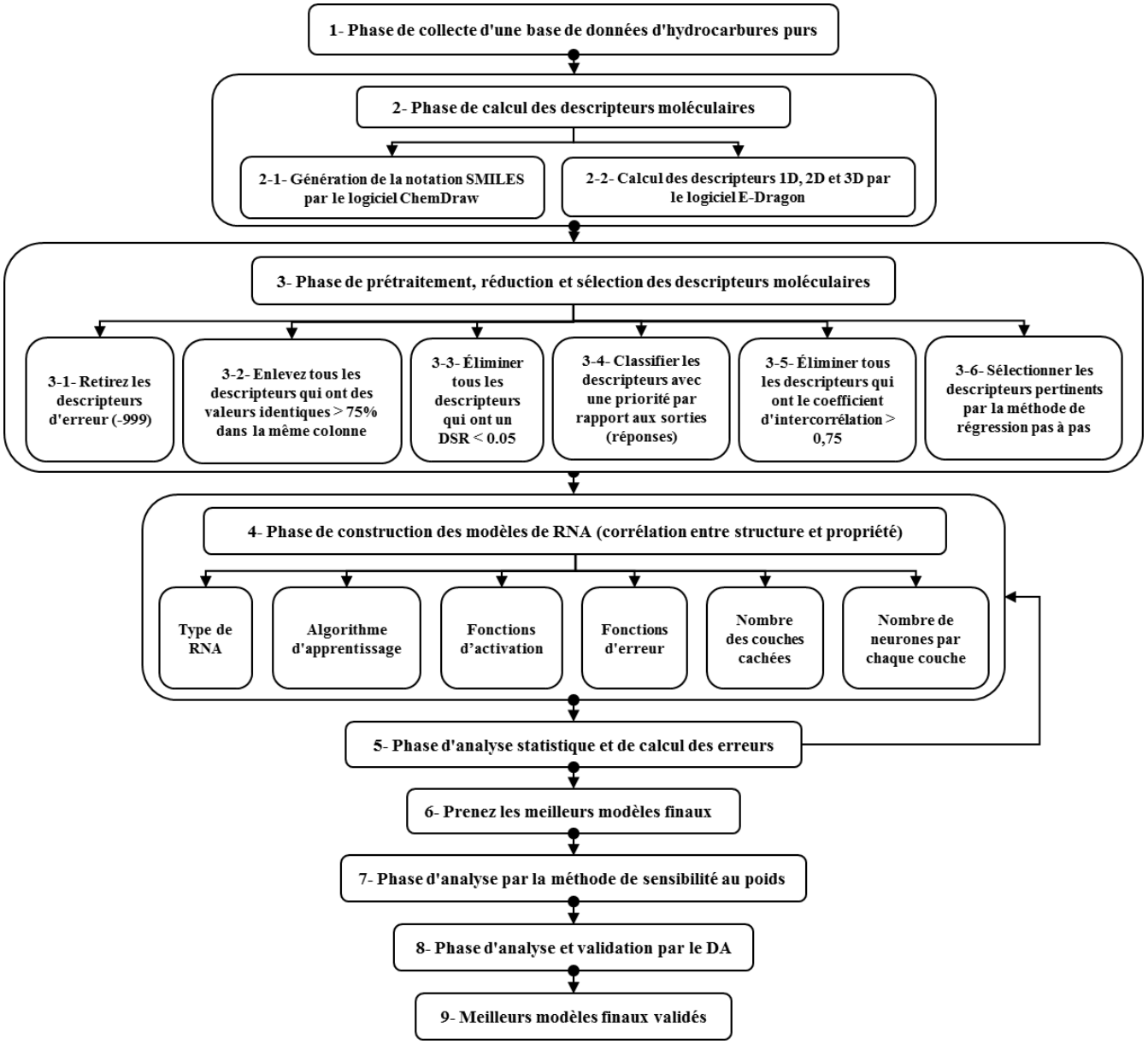


Figure IV.2. Organigramme de la méthodologie suivie dans cette étude **(Roubehie fissa *et al.* 2019)**.

IV.3.1. Collection et traitement de la base des données

On sait que la qualité de la base de données expérimentale est la pierre angulaire de tout modèle mathématique réussi **(Hamadache *et al*. 2016a ; Hamadache *et al*. 2016b)**. Cependant, collecter ces données avec la quantité et la qualité appropriées est une tâche très difficile et lourde.

La base de données utilisée dans cette étude a été collectée auprès de plusieurs sources contenant entre 223 et 221 molécules différentes des hydrocarbures purs issue d'une fraction pétrolière liéger, afin de prédire cinq paramètres physico-chimiques : le point d'ébullition normal (Tb), la densité relative du liquide (d20), la température critique (Tc), la pression critique (Pc) et le volume critique (Vc) (**Wessel & Jurs 1995 ; Lahiouel 1995b ; Wakeham *et al.* 2002 ; KDB 2017**). Ces composés comprennent diverses catégories d'hydrocarbures, saturés et insaturés représentant les principaux groupes d'importance industrielle tels que : le n-paraffine, l'iso-paraffine, la cyclo-paraffine, l'oléfine simple, l'oléfine aromatique et l'alcyne. La majorité de ces catégories ont plus d'un isomère de position tel que : *cis, trans, méta, ortho* et *para*. La totalité de la base de données utilisée dans cette étude est présentée dans le **tableau A.1 (Annexe A)**.

IV.3.2. Codage SMILES des structures chimiques des hydrocarbures

Le calcul de descripteurs moléculaires a été précédé d'une étape très intéressante consistant à générer la notation SMILES (Simplified Molecular Input Line Entry System) avec l'utilisation du logiciel "ChemDraw". Cette étape est considérée comme un passage ou bien une transformation du langage chimique vers un langage informatique, en codant tous les composés sous la forme de notation SMILES.

Les notations SMILES de toutes les molécules utilisées dans cette étude sont présentées dans le **tableau B.1 (Annexe B)**. Chaque molécule d'hydrocarbure est préparée et dessinée par le logiciel "ChemDraw" puis les notations SMILES de chaque molécule sont extraites par le même logiciel, enregistrées ensuite dans des fichiers de format ".txt" (**Figure IV.3**), afin de préparer toutes ces molécules à l'étape suivante qui est le calcul des descripteurs moléculaires.

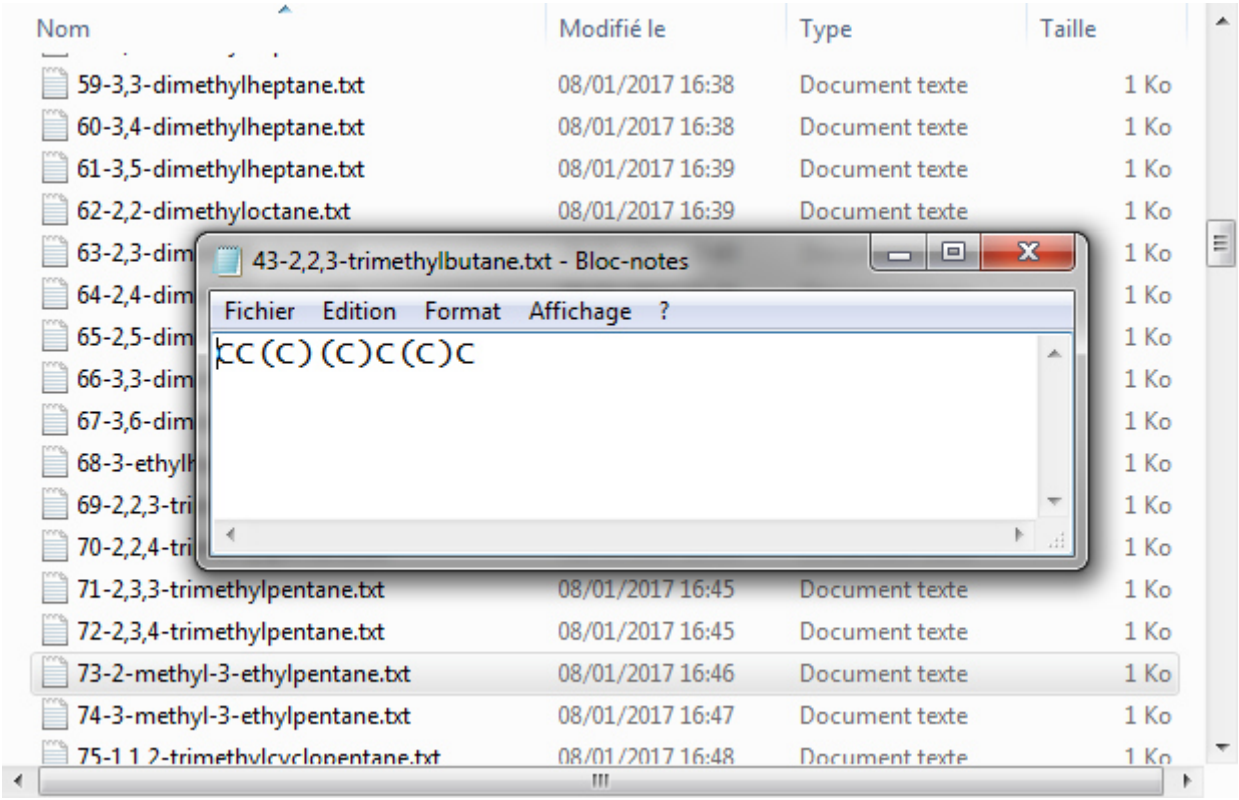


Figure IV.3. Capture d'écran des notations SMILES des molécules organisées dans des fichiers de format ".txt".

IV.3.3. Calcul des descripteurs moléculaires

L'une des étapes les plus importantes dans la construction des modèles de type RQSP/RQSA est la quantification des informations structurelles incorporées, emmagasinées dans les molécules étudiées **(Hamadache *et al*. 2016)**, qui se traduisent par différents types de descripteurs moléculaires calculés. En réalité, il existe plus de 11145 descripteurs moléculaires qui peuvent être utilisés pour résoudre plusieurs problèmes dans différents domaines tels que la chimie, la biologie et d'autres sciences connexes **(Hamadache *et al.* 2016a ; Masand & Rastija 2017)**.

Au cours de cette étude, 1666 descripteurs moléculaires ont été calculés en ligne par le logiciel E-Dragon 1.0 **(VCCLAB 2005)**, divisés en 20 classes différentes présentées dans le **tableau IV.1**.

Tableau IV.1. Nombres de descripteurs initialement calculés par chaque classe pour tous les paramètres étudiés **(VCCLAB 2005 ; Kraim 2009)**.

N° de classe	Nom des classes		Nombre des descripteurs
	"En anglais"	**"Traduisez en français"**	
(01)	*"constitutional descriptors"*	*"descripteurs constitutionnels"*	*48*
(02)	*"topological descriptors"*	*"descripteurs topologiques"*	*119*
(03)	*"walk and path counts"*	*"compte du parcours et de marche"*	*47*
(04)	*"connectivity indices"*	*"indices de connectivité"*	*33*
(05)	*"information indices"*	*"indices d'information"*	*47*
(06)	*"(2D) two dimensional autocorrelations"*	*"descripteurs 2D- d'autocorrélation"*	*96*
(07)	*"edge adjacency indices"*	*"indices de contiguïté des bords"*	*107*
(08)	*"Burden eigenvalues descriptors"*	*"descripteurs issus des valeurs propres des charges"*	*64*
(09)	*"topological charge indices"*	*"indices topologiques de charge"*	*21*
(10)	*"eigenvalue based indices"*	*"indices basés sur des valeurs propres"*	*44*
(11)	*"Randic molecular profiles"*	*"profils moléculaires de Randic"*	*41*
(12)	*"geometrical descriptors"*	*"descripteurs géométriques"*	*74*
(13)	*"RDF descriptors"*	*"descripteurs des fonctions de distribution radiale"*	*150*
(14)	*"3D MORSE descriptors"*	*"descripteurs de représentation 3D des structures moléculaires basée sur la diffraction électronique"*	*160*
(15)	*"WHIM descriptors"*	*"descripteurs de WHIM"*	*99*
(16)	*"GETAWAY descriptors"*	*"descripteurs GETAWAY"*	*197*
(17)	*"functional group counts"*	*"compte des groupes fonctionnels"*	*154*
(18)	*"atom centered fragments"*	*"fragments centrés sur l'atome"*	*120*
(19)	*"charge descriptors"*	*"descripteurs de charge"*	*14*
(20)	*"molecular properties"*	*"propriétés moléculaires"*	*31*
Nombre total des descripteurs moléculaires			***1666***

NB : *Pour plus d'informations sur ces descripteurs consulter les deux livres de Todeschini et Consonni (**Todeschini & Consonni 2008** ; **Todeschini & Consonni 2009**).*

Comme le montre le **tableau IV.1**, le nombre total de descripteurs calculés dans cette étude est de 1666, ce qui est un nombre énorme à inclure dans l'élaboration des modèles de RQSP-RNA. Il est bien connu que la majorité de ces descripteurs ne correspondent pas avec les paramètres à modéliser.

Je dois donc réduire ce nombre à un nombre acceptable en choisissant les descripteurs les plus appropriés au processus de modélisation pour chaque paramètre. Pour cette raison et dans le but d'atteindre les objectifs souhaités, j'ai adopté une procédure de réduction pour la sélection des meilleurs descripteurs pertinents pour chaque paramètre à étudier.

IV.3.4. Procédure de prétraitement pour la sélection des descripteurs pertinents

L'un des défis de cette étude est la sélection des descripteurs moléculaires pertinents pour chaque paramètre à étudier. J'ai donc utilisé une méthode statistique pour réduire et sélectionner le nombre de descripteurs, afin d'obtenir les meilleurs descripteurs pertinents pour les utiliser ensuite dans la construction des modèles RQSP-RNA. La méthode était basée sur les points suivants **(Roubehie fissa *et al.* 2019)** :

(1) Éliminer tous les descripteurs avec une valeur d'erreur ayant au moins une valeur de **"*-999*"** indiquée par le logiciel E-Dragon ;

(2) Omettre les descripteurs ayant la même valeur, autrement je recherche des valeurs minimales et maximales pour chaque descripteur ;

(3) Supprimer les descripteurs d'une même valeur répétés plus de 75%, en comparant les trois premiers quartiles ;

(4) Négliger tous les descripteurs ayant un écart type relatif (déviation standard relative ***DSR***) inférieur à 0,05. Le logiciel "STATISTICA" a été utilisé lors des deux dernières étapes **(Stat Soft 2007)** ;

(5) Supprimer les descripteurs avec un coefficient d'intercorrélation de Pearson ***(R)*** supérieur à 0,75 pour chaque paire de descripteurs ; la corrélation entre ces descripteurs et leurs réponses est aussi prise en considération. De manière que, on compare la corrélation entre la réponse et chaque descripteur calculé, puis on prend le meilleur descripteur corrélé. Ensuite, on le compare avec le reste des descripteurs, afin de ne retenir que les descripteurs pertinents. À ce stade, j'ai utilisé le logiciel "Matlab" pour calculer les coefficients de corrélations ***(R)*** **(The Mathworks Inc 2013)**. La comparaison et l'élimination de chaque descripteur indésirable a été faite manuellement en utilisant le tableur Excel 2013 ;

(6) Enfin, j'ai utilisé une méthode de régression linéaire multiple "RLM" pas à pas (*MLR forward stepwise regression*) pour supprimer les descripteurs dont la Valeur-P (valeur de probabilité) est supérieure à 0,005, en utilisant constamment le logiciel "STATISTICA" **(Stat Soft 2007)**.

IV.3.5. Sélection des paramètres de choix pour l'apprentissage des modèles de RQSP-RNA

Après le prétraitement, j'ai réduit le nombre de descripteurs autant que possible, j'ai obtenu les descripteurs pertinents pour chaque paramètre, afin de les utiliser pour la phase de construction des modèles de RQSP-RNA. Ceci est un autre défi en ce qui concerne le grand paramètre de choix dans le RNA ; pour cela, j'ai utilisé différentes options afin d'obtenir les meilleurs modèles de RQSP-RNA pour chaque paramètre **(Roubehie fissa *et al.* 2019)** :

- L'ensemble de données a été divisé au hasard en deux sous-ensembles : (80%) pour la phase d'apprentissage et (20%) pour la phase de test du modèle (**Figure IV.4**) **(Pratim Roy *et al*. 2008 ; Martin *et al.* 2012)**.
- Pour le type de RNA, j'ai utilisé un RNA-PMC.
- L'algorithme Quasi Newton à propagation en arrière (*Quasi Newton back propagation algorithm*, plus précisément : *Broyden-Fletcher-Goldfarb-Shanno algorithm* ***BFGS***) a été utilisé comme algorithme d'apprentissage **(Wessel & Jurs 1994)**.
- Les fonctions d'activation sont effectuées en choisissant plusieurs types de fonctions dans la couche cachée et sortie. Par conséquent, à la fin de l'exécution, le logiciel donne la meilleure combinaison possible entre les deux fonctions d'activation de la couche concernée (**Figure IV.5**).
- J'ai utilisé la somme des carrés (*Sum of Squares* ***SOS***) en tant que fonction d'erreur.
- Concernant le nombre de couches cachées et vu qu'il existe une couche cachée dans le logiciel STATISTICA, version 7 et 8 **(Stat Soft 2007)**, j'ai utilisé une seule couche cachée pour tous les modèles de RQSP-RNA, dans ce travail.
- Cependant, en ce qui concerne le nombre de neurones par couche cachée, j'ai généralement fait varier ce nombre de 3 à 20 neurones, afin de m'assurer que les variables des modèles de réseaux de neurones ne dépassent pas la taille de la base des données ; j'ai donc tout simplement suivi cette règle :

$$(\textbf{Nombre de neurones d'entrée} * \textbf{Nombre de neurones cachés}) + (\textbf{Nombre de neurones cachés} * \textbf{Nombre de neurones de sortie}) \leq \textbf{Taille de la base de données}.$$

- Plusieurs exécutions ont été effectuées pour chaque paramètre afin d'obtenir les meilleurs modèles de type RQSP-RNA. L'exécution comporte un nombre important d'itérations ; dépassant parfois 100 000 itérations pour une seule exécution.

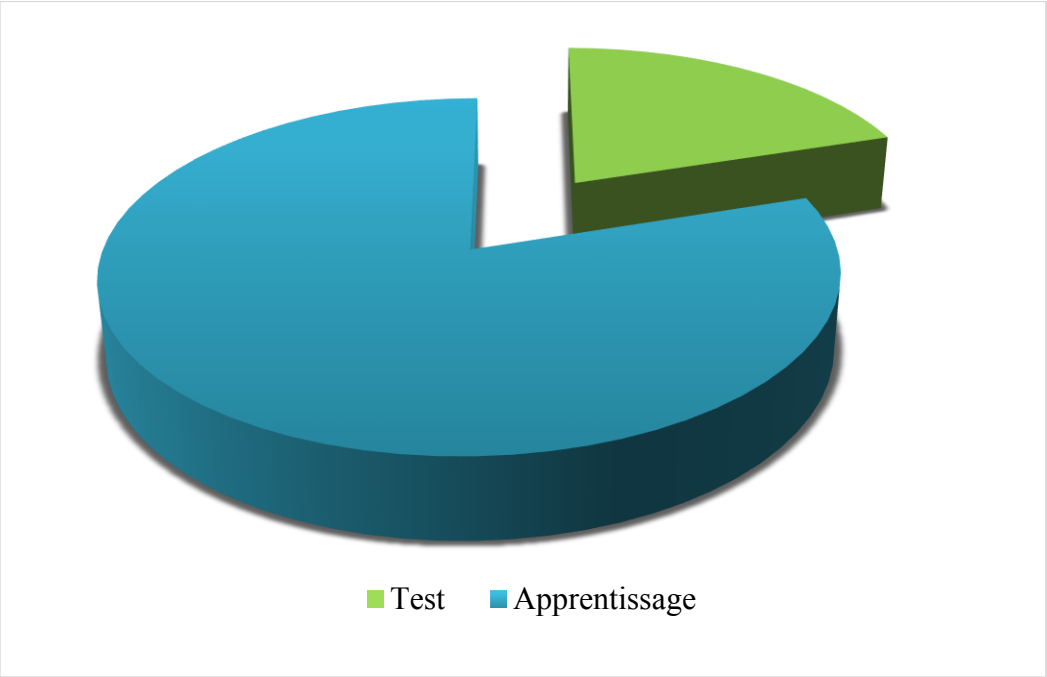


Figure IV.4. Représentation graphique des pourcentages de division de la base de données

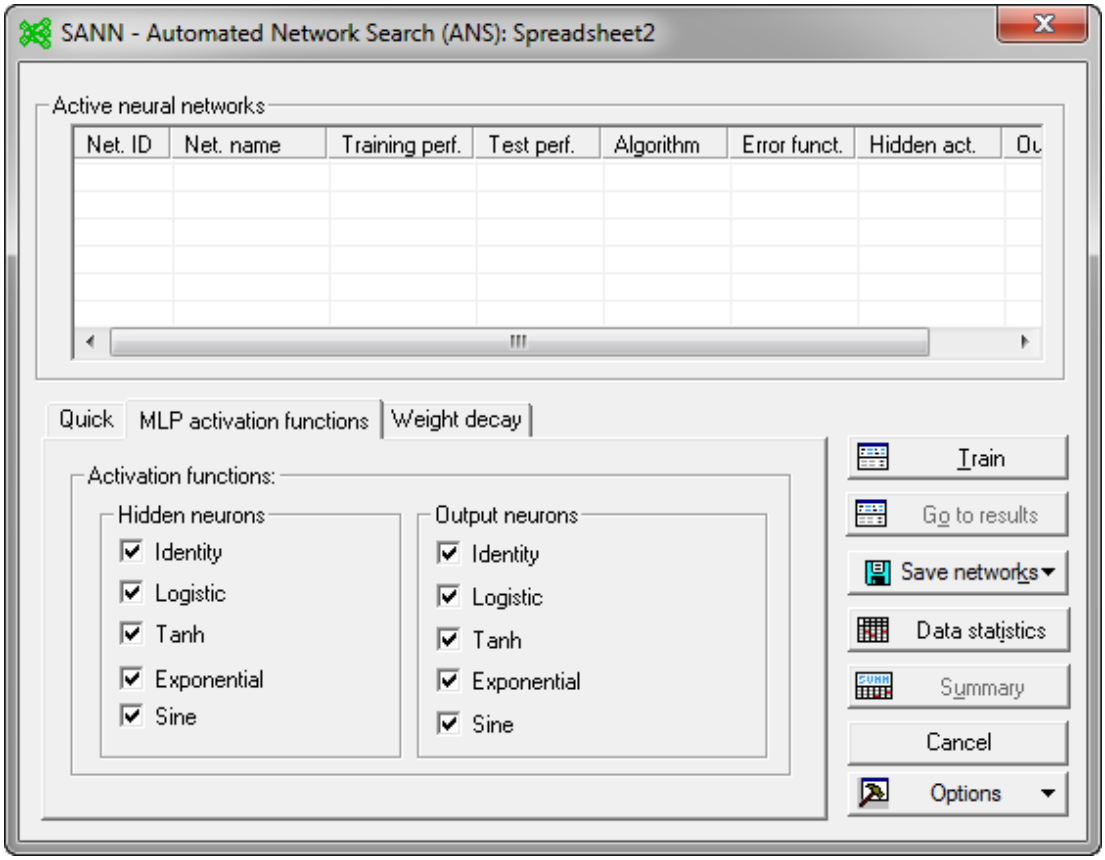


Figure IV.5. Capture d'écran pour le choix des fonctions d'activation sur STATISTICA **(Stat Soft 2007)**

IV.3.6. Estimation de la performance des meilleurs modèles de RQSP-RNA développés

IV.3.6.1. Analyses statistiques et calcul des erreurs

La performance et la précision des meilleurs modèles de RQSP-RNA obtenus dans cette étude, ont été évaluées en termes des quatre types différents d'erreurs ; la racine d'erreur quadratique moyenne (REQM), l'erreur standard de prédiction (ESP), l'erreur relative moyenne (ERM), et l'erreur relative absolue moyenne (ERAM). En plus les performances de régression ont été aussi testées à l'aide des coefficients de régression (R), de détermination (R^2), de détermination ajustée (R^2_{adj}) (comme critère dans l'étape de sélection des descripteurs) et le vecteur de régression : (les valeurs de la pente (α), et de l'interception (β)). Ceci est fait sur l'ensemble des données de test, d'apprentissage, et sur les données globales de la base.

Les définitions mathématiques de chaque type d'erreur, du coefficient de régression (R), du coefficient de détermination (R^2), et du coefficient de détermination ajusté (R^2_{adj}) sont données par les **Equations (IV.1–IV.7)**, respectivement.

$$\boldsymbol{REQM = \sqrt{\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{cal}\right)^2 / N}} \qquad \textbf{(IV.1)}$$

$$\boldsymbol{ESP = (REQM / \hat{y}^{exp}) * 100} \qquad \textbf{(IV.2)}$$

$$\boldsymbol{ERM = (100/N) * \sum_{i=1}^{N}\left(\left|\left(y_i^{exp} - y_i^{cal}\right)\right| / y_i^{cal}\right)} \qquad \textbf{(IV.3)}$$

$$\boldsymbol{ERAM = (100/N) * \sum_{i=1}^{N}\left(\left|\left(\left(y_i^{exp} - y_i^{cal}\right) / y_i^{exp}\right)\right|\right)} \qquad \textbf{(IV.4)}$$

$$\boldsymbol{R = \frac{\sum_i^N [(y_i^{exp} - \hat{y}^{exp})(y_i^{cal} - \hat{y}^{cal})]}{\sqrt{\sum_{i=1}^{N}\left(y_i^{exp} - \hat{y}^{exp}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i^{cal} - \hat{y}^{cal}\right)^2}}} \qquad \textbf{(IV.5)}$$

$$\boldsymbol{R^2 = 1 - \sum_{i=1}^{N}\left(y_i^{exp} - y_i^{cal}\right)^2 / \sum_{i=1}^{N}\left(y_i^{exp} - \hat{y}^{exp}\right)^2} \qquad \textbf{(IV.6)}$$

$$R^2_{adj} = 1 - [(1 - R^2) * (N - 1) / (N - k - 1)] \quad \textbf{(IV.7)}$$

$$\hat{y}^{cal} = \sum_{i=1}^{N} y_i^{cal} / N \quad \textbf{(IV.8)}$$

$$\hat{y}^{exp} = \sum_{i=1}^{N} y_i^{exp} / N \quad \textbf{(IV.9)}$$

Où N est le nombre des composés dans l'ensemble des données pour chaque phase (apprentissage, test, ou l'ensemble entier des données), k étant le nombre de variables du modèle (dans ce cas c'est le nombre de descripteurs sélectionnés), y^{cal} est la valeur calculée (prédite) de y ; pour l'ensemble d'apprentissage ou du test, y^{exp} est la valeur expérimentale (observée) de y, $\hat{y}^{cal}$ la moyenne des valeurs calculées de y ; (**Eq (IV.8)**), et $\hat{y}^{exp}$ la moyenne des valeurs expérimentales de y ; (**Eq (IV.9)**).

IV.3.6.2. Mesure de la contribution des descripteurs de chaque modèle de RQSP-RNA développé

L'analyse de sensibilité est un outil important pour l'évaluation systématique des modèles mathématiques, en général, et des modèles de RQSP-RNA en particulier. L'analyse de sensibilité peut être utilisée à diverses fins, notamment la validation du modèle, l'évaluation du comportement des différentes variables du modèle, et la confirmation du choix des variables du modèle à l'aide de la détermination de la contribution de chaque variable choisie sur le modèle. De ce fait, il est donc important d'utiliser cette analyse sur les modèles des études RQSP/RQSA.

Dans cette étude l'analyse de sensibilité par la méthode de "Poids" (*Weights method*) a été utilisée, afin d'évaluer l'impact de la contribution de toute variable d'entrée (descripteurs moléculaires pertinents de chaque modèle) sur les sorties (Tb, d20, Tc, Pc et Vc) de chaque modèle RQSP-RNA. Cette analyse est aussi importante pour confirmer le choix des descripteurs utilisés lors de l'élaboration des modèles de RQSP-RNA, et donc l'efficacité de la méthode de prétraitement utilisée dans la sélection des descripteurs pertinents.

La méthode utilisée dans cette étude (méthode de Poids) ; a été organisée et programmée dans un tableur "Microsoft Excel", en suivant les étapes théoriques indiquées dans l'étude de

Gevrey et ses co-auteurs **(Gevrey *et al.* 2003)**. Comme l'indique le nom de la méthode "Poids", on introduit les poids de connexion de partitionnement en : **(1)** Poids de la couche entrée $(\boldsymbol{w_{ij}^{I}})$ (connexions entre les neurones d'entrée et cachés), **(2)** Poids de la couche cachée $(\boldsymbol{w_{1j}^{H}})$ (connexions entre les neurones cachés et de sortie). Pour trouver, à la fin une signification quantifiée de chaque descripteur d'entrée sur chaque sortie du modèle, ce qui se traduit par l'importance relative **($\boldsymbol{IR}$)** de chaque descripteur.

Pour mieux expliquer cette méthode j'ai repris l'exemple d'application utilisé dans l'étude de Goh **(Goh 1995)**, et l'étude de Gevrey et ses co-auteurs **(Gevrey *et al.* 2003)**. L'exemple est donné pour un réseau de neurones d'architecture composée de : trois *(3)* neurones dans la couche entrée, quatre *(4)* neurones dans la couche cachée et un *(1)* neurone dans la couche sortie, le processus de calcul est basé sur les quatre étapes suivantes :

- **1ère étape :** Pour chaque neurone caché **($\boldsymbol{i}$)** ; multiplier la valeur absolue des poids de connexions cachée–sortie $\left|\boldsymbol{w_{1j}^{H}}\right|$ par la valeur absolue des poids de connexions cachée–entrée $\left|\boldsymbol{w_{ij}^{I}}\right|$. Faire ceci pour chaque variable d'entrée **($\boldsymbol{j}$)**, les résultats des produits **($\boldsymbol{P_{ij}}$)** sont obtenus dans la 1ère partie du **tableau IV.2**.
- **2ème étape :** Pour chaque neurone caché **($\boldsymbol{i}$)** ; diviser chaque **($\boldsymbol{P_{ij}}$)** par la somme de toutes les variables d'entrée pour obtenir **($\boldsymbol{Q_{ij}}$)**. Par exemple pour le neurone caché 01, $\boldsymbol{Q_{11} = \frac{P_{11}}{P_{11}+P_{12}+P_{13}} = 0,26445}$. Les résultats complets de tous les **($\boldsymbol{Q_{ij}}$)** sont donnés dans la 2ème partie du **tableau IV.2**.
- **3ème étape :** Pour chaque neurone d'entrée **($\boldsymbol{j}$)**, additionner les résultats des produits **($\boldsymbol{Q_{ij}}$)** obtenus par l'étape précédente, afin d'obtenir les sommes **($\boldsymbol{S_j}$)**. Par exemple : $\boldsymbol{S_1 = Q_{11} + Q_{21} + Q_{31} + Q_{41}}$. Les résultats de chaque **($\boldsymbol{S_j}$)** sont donnés dans la 3ème partie du **tableau IV.2**.
- **4ème étape :** Deviser **($\boldsymbol{S_j}$)** par la somme de toutes les variables d'entrée, exprimé en pourcentage, cela donne l'importance relative ou bien la contribution des variables d'entrée données. Par exemple : pour le neurone d'entrée 01, l'importance relative **($\boldsymbol{IR}$)** égale à : $\boldsymbol{(S_1 * 100)/(S_1 + S_2 + S_3) = 43,7\%}$, les valeurs de l'importance relative **($\boldsymbol{IR}$)** de tous les neurones d'entrée sont données dans la 4ème partie du **tableau IV.2**.

Tableau IV.2. Exemple illustrant la méthode de la sensibilité basée sur la partition des poids de connexion (méthode de Poids) **(Goh 1995 ; Gevrey *et al.* 2003)**.

Valeurs des Poids de connexions initiales : (w_{ij}^{I}) et (w_{1j}^{H})				
Neurones	Entrée 01	Entrée 02	Entrée 03	Sortie
Caché 01	–1,67624	3,29022	1,32466	4,57857
Caché 02	–0,51874	–0,22921	–0,25526	–0,48815
Caché 03	–4,01764	2,12486	–0,08168	–5,73901
Caché 04	–1,75691	–1,44702	–0,58286	–2,65221
1ère étape : Valeurs des Produits (P_{ij})				
Neurones	Entrée 01	Entrée 02	Entrée 03	
Caché 01	P_{11}= 1,67624*4,57857	P_{12}= 3,29022*4,57857	P_{13}= 1,32466*4,57857	
Caché 02	P_{21}= 0,51874*0,48815	P_{22}= 0,22921*0,48815	P_{23}= 0,25526*0,48815	
Caché 03	P_{31}= 4,01764*5,73901	P_{32}= 2,12486*5,73901	P_{33}= 0,08168*5,73901	
Caché 04	P_{41}= 1,75691*2,65221	P_{42}= 1,44702*2,65221	P_{43}= 0,58286*2,65221	
2ème étape : Valeurs des divisions (Q_{ij})				
Neurones	Entrée 01	Entrée 02	Entrée 03	
Caché 01	Q_{11}= 0,266445	Q_{12}= 0,522994	Q_{13}= 0,210560	
Caché 02	Q_{21}= 0,517081	Q_{22}= 0,228478	Q_{23}= 0,254441	
Caché 03	Q_{31}= 0,645489	Q_{32}= 0,341388	Q_{33}= 0,013123	
Caché 04	Q_{41}= 0,463958	Q_{42}= 0,382123	Q_{43}= 0,153919	
3ème étape : Valeurs des Sommes (S_j)				
Neurones	Entrée 01	Entrée 02	Entrée 03	
La somme	S_1= 1,892973	S_2= 1,474983	S_3= 0,632044	
4ème étape : Valeurs des Importance Relatives (IR)				
Neurones	Entrée 01	Entrée 02	Entrée 03	
IR (%)	47,3	36,9	15,8	

La procédure de calcul de cette méthode, est aussi schématisée sur l'organigramme dans la **figure IV.6 (Khaouane 2013)**.

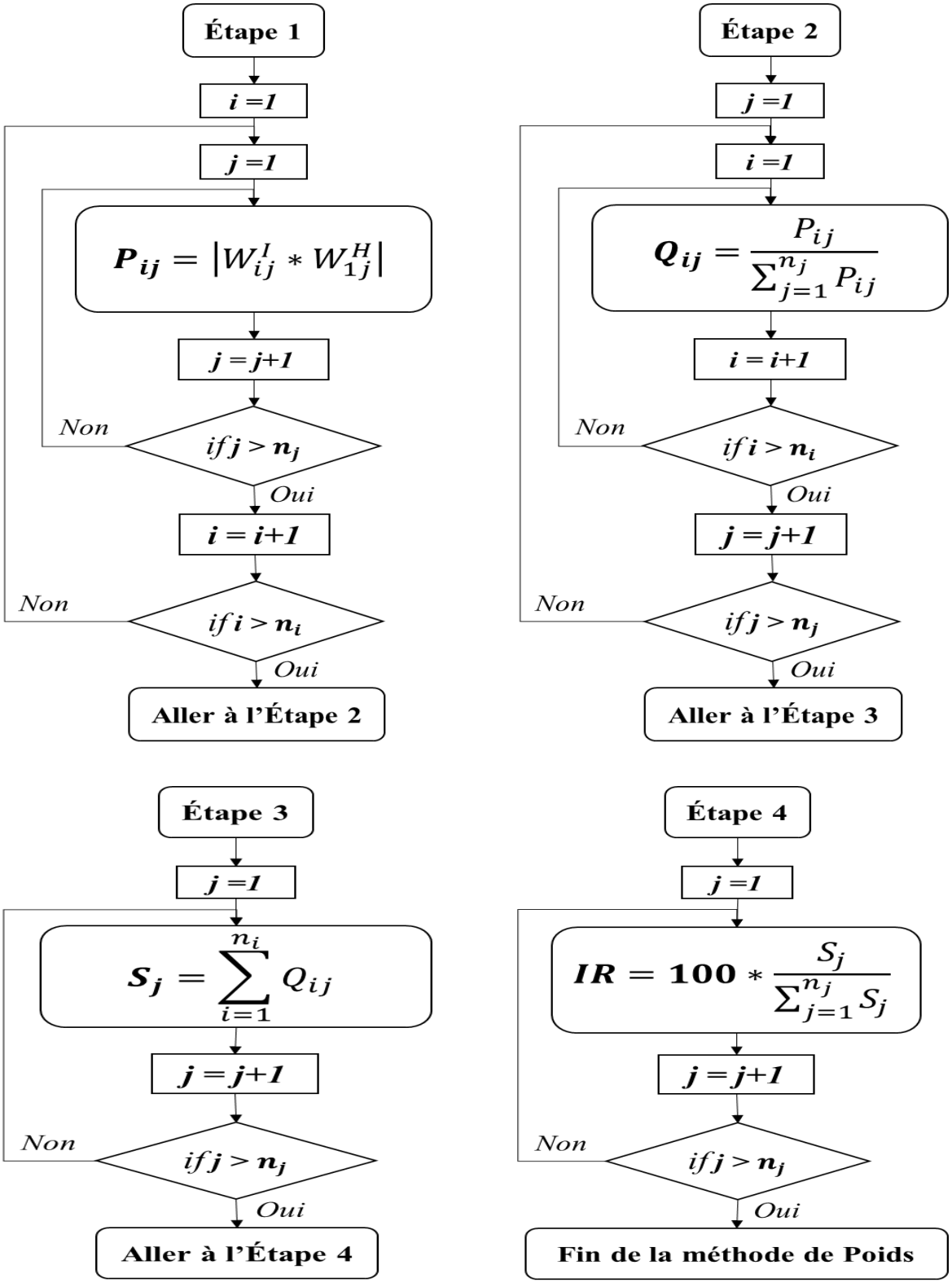


Figure IV.6. Organigramme illustrant l'analyse de sensibilité par la méthode de la partition des poids (méthode de poids) **(Khaouane 2013)**

IV.3.6.3. Détermination du domaine d'applicabilité des modèles de RQSP-RNA

Le domaine d'applicabilité (DA) est décrit comme une confirmation de la pertinence du modèle dans son champ d'application en disposant d'un ensemble approprié de précision dans l'application du modèle proposé **(Hamadache *et al*. 2016b)**. Généralement, la fiabilité des prévisions RQSP/RQSA est limitée aux produits chimiques structurellement similaires à ceux utilisés dans la construction des modèles proposés **(Weaver et Gleeson 2008 ; Hamadache *et al*. 2016a ; Hemmati-Sarapardeh *et al*. 2018)**.

Nous avons choisi dans cette étude une méthode basée sur la distance pour la détermination du domaine d'applicabilité (DA) de chaque modèle de RQSP-RNA développé. Le principe de cette méthode se base sur la détermination des valeurs de levier (***h***) (*leverage*) de chaque composé, ce qui est alors comparé à la valeur de levier standard (***h****) (*standard leverage*) **(Kraim *et al*. 2009 ; Zhou *et al*. 2017)**. L'avantage de cette méthode est qu'il est possible de quantifier l'applicabilité du modèle en le présentant sur un graphique visuel appelé le diagramme de Williams (*Williams plot*) **(Zhou *et al*. 2017 ; Roubehie fissa *et al*. 2019)**. Généralement, le domaine d'applicabilité (DA) d'un tel modèle de type RQSP/RQSA est compris entre sa valeur de levier standard (***h****) et ses valeurs de résidus standardisées (d_i) (*standardized residuals*). Il est parfois connu sous le nom de résidus normalisés, de sorte que la plupart des valeurs de résidus standardisées dans la littérature sont limitées à ± 3 **(Jaworska *et al*. 2005)**, mais parfois, dans quelques études elles sont prises égales à la valeur ± 2,5 comme l'étude de Papa **(Papa *et al*. 2005)**, indiqué aussi dans la mini-revue de Gramatica **(Gramatica 2007)**.

Comme décrit dans la partie IV.2 (Outils informatiques utilisés) de ce chapitre ; la boîte à outils (*toolbox*) développée par le groupe de chercheurs Italiens "*Milano Chemometrics and QSAR Research Group*" a été utilisée (**Figure IV.1**) **(Sahigara *et al.* 2012 ; Sahigara *et al.* 2014)**. Afin de déterminer le domaine d'applicabilité (DA) des modèles de RQSP-RNA développés dans cette étude, puis les présenter sous forme de graphes exprimés par les diagrammes de Williams pour chaque modèle.

Les expressions mathématiques de chaque paramètre utilisé lors de la détermination du domaine d'applicabilité, sont données par les **Eqs (IV.10–IV.12)**, respectivement pour les valeurs de levier (h_i), la valeur levier standard (***h****), et les valeurs de résidus standardisées (d_i).

$$h_i = X_i(X^T X)^{-1} X_i^T \qquad \textbf{(IV.10)}$$

$$h^* = 3 * (k+1)/m \quad \textbf{(IV.11)}$$

$$d_i = \frac{(y_i^{exp} - y_i^{cal})}{SD_i} \quad \textbf{(IV.12)}$$

$$DS_i = \frac{\sqrt{\sum_{i=1}^{n}(y_i^{exp} - y_i^{cal})^2}}{\sqrt{n-k-1}} \quad \textbf{(IV.13)}$$

Où X_i est le vecteur ligne des descripteurs du composé (i) concerné ; X_i^T est son transposé, X est la matrice des valeurs des descripteurs de la phase d'apprentissage ; X^T est sa matrice transposée, k est le nombre des variables du modèle ; dans ce cas s'agissant du nombre de descripteurs pertinents dans chaque modèle, m est le nombre de composés dans la phase d'apprentissage, y_i^{exp} et y_i^{cal}, respectivement les valeurs expérimentale et calculée du paramètre à modéliser (Tb, d20, Tc, Pc ou Vc) du composé (i) concerné dans la phase d'apprentissage ou de test, DS_i est la déviation standard (l'écart type) du composé (i) concerné (**Eq (IV.13)**), et n est le nombre de composés dans la phase d'apprentissage ou de test selon le cas du (d_i) des composés qu'ils veulent calculer.

Si la valeur h d'un tel composé examiné est supérieure à la valeur de h^* ; la valeur prédite du composé pourrait être considérée comme hors de la portée de domaine d'applicabilité du modèle. En conséquence, le résultat prévu peut ne pas être fiable (**Eriksson *et al*. 2003 ; Gramatica 2007 ; Sahigara *et al*. 2014 ; Zhou *et al*. 2017**).

IV.4. Conclusion

Dans ce chapitre j'ai présenté les matériels et les méthodes utilisés lors du développement des modèles de RQSP-RNA en décrivant les différentes étapes. J'ai cité tous les logiciels utilisés dans chacune des étapes, en plus d'un aperçu sur quelques logiciels et outils informatiques utiles pour le développement des modèles de RQSP-RNA qui n'ont pas été utilisés dans cette étude.

La méthodologie globale de développement des modèles RQSP-RNA a été résumé et clarifié par le bais de l'organigramme de la **figure VI.1**. Ensuite, j'ai expliqué en détails chaque méthode utilisée pour le développement des modèles de type RQSP-RNA des hydrocarbures : que ce soit les étapes utilisées lors du prétraitement pour la réduction et la sélection des

descripteurs pertinents, ou bien précédemment dans la procédure de collection et traitement de la base des données, tout aussi pour le codage SMILES des structures chimique des hydrocarbures qui est considéré comme un passage du langage chimique vers le langage informatique, ou alors la génération des descripteurs, puis la sélection des descripteurs pertinentes ; qui sont considérés comme des variables d'entrée utilisées ultérieurement dans l'élaboration des modèles de RQSP-RNA. J'ai également montré, en détails la procédure de sélection des différents paramètres de choix pour le développement des modèles de RQSP-RNA. Et enfin j'ai proposé une analyse complète pour estimer la performance des meilleurs modèles de type RQSP-RNA développés dans cette étude, en se basant sur une analyse statistique par l'utilisation de quatre types d'erreur différents afin de tester mes modèles. J'ai ensuite analysé l'effet de chaque descripteur d'entrée sur chaque paramètre de sortie, par l'utilisation de la méthode d'analyse de sensibilité de poids pour prouver l'efficacité de la méthode de sélection suivie. Et en dernier lieu, j'ai utilisé une méthode de validation nécessaire dans tous les types d'études RQSP/RQSA et qui est la méthode du domaine d'applicabilité, précisément la méthode basée sur la distance qui est exprimé graphiquement par les diagrammes de Williams.

En vertu de ce chapitre, on peut tirer quelques conclusions et recommandations :

- La méthodologie suivie dans cette étude répond à tous les principes de modélisation moderne de l'approche RQSP/RQSA.
- Il existe beaucoup de stratégies pour développer des modèles RQSP/RQSA mais la méthodologie générale est presque la même dans toutes les études.
- Compte tenu des spécifications de chaque étude RQSP/RQSA, il n'existe pas un résultat ni une valeur ou bien un objectif de référence apparente à atteindre, ceci est dû à la diversification de l'application des RQSP/RQSA dans plusieurs disciplines. En conséquence, chaque domaine a son temps d'or dans lequel il est au sommet de la science.
- Pour faire face à la croissance énorme des études de RQSP/RQSA dans plusieurs disciplines, le nombre et le type de logiciels ainsi que les outils informatiques dans ce domaine se sont multipliés aussi ces dernières années, afin de faciliter la tâche de développement des modèles de RQSP/RQSA.
- Comme tout objectif principal de modélisation, les modèles de type RQSP/RQSA sont conçus pour prédire différents paramètres dans différents domaines afin de réduire le temps, l'argent et les efforts utilisés à des fins expérimentales.

Chapitre V

Résultats et discussion

V.1. Introduction

Dans ce chapitre, je vais aborder les résultats obtenus lors de la modélisation des modèles de type RQSP-RNA des cinq propriétés physico-chimiques : Tb, d20, Tc, Pc et Vc, pour des hydrocarbures purs issus des fractions pétrolières légères, contenant les principaux groupes industriels importants des hydrocarbures.

Je commencerai par l'étude de l'évolution du nombre des descripteurs sélectionné au cours de la procédure de prétraitement par classes pour chaque paramètre étudié, ensuite je montrerai les caractéristiques de chaque descripteur pertinent sélectionné pour les cinq paramètres étudiés, puis je discuterai les architectures des meilleurs modèles de type RQSP-RNA obtenus pour chaque paramètre, puis j'expliquerai l'application des expressions mathématiques de chaque modèle obtenu pour la prédiction des cinq propriétés physico-chimiques des hydrocarbures. Après, je traiterai l'appréciation de la performance des résultats d'estimation par les modèles de type RQSP-RNA développés ; je calculerai les erreurs et j'analyserai statistiquement les modèles de RQSP-RNA obtenus. Ensuite, je vais comparer les valeurs calculées avec les valeurs expérimentales, que je ferai suivre par une analyse de sensibilité des descripteurs d'entrée de chaque modèle par la méthode de poids, afin d'interpréter et attribuer les contributions de chaque descripteur unique sur les modèles obtenus, puis les contributions de chaque classe de descripteurs sur les modèles obtenus selon la valeur de contribution moyenne (VCM) sur chaque modèle. Je déterminerai aussi le domaine d'applicabilité afin de mieux valider les modèles de type RQSP-RNA obtenus. Je ferai aussi des comparaisons de performance des modèles de type RQSP-RNA obtenus dans cette thèse et issus d'autres modèles de RNA et de RQSP existant dans la littérature. Enfin, je conclurai ce chapitre en analysant et interprétant globalement les différents résultats obtenus dans cette étude.

V.2. Évolution du nombre de descripteurs au cours de la procédure de sélection

Comme décrit dans le chapitre précédent, la procédure de sélection utilisée dans cette étude repose sur une méthodologie statistique contenant plusieurs étapes différentes, chaque étape ayant ses propres spécifications, tenant compte des mécanismes utilisés, des critères de choix, du type de paramètre à étudier, et de la nature des descripteurs à traiter. En conséquence, chaque étape a son propre effet sur le résultat du prétraitement et a donc un impact sur le nombre

de descripteurs à choisir. L'évolution du nombre de descripteurs lors du prétraitement par classe type pour chaque paramètre étudié est donnée dans le **tableau V.1**.

1666 descripteurs moléculaires issus de 20 classes différentes ont été calculés par le logiciel en ligne E-Dragon pour chaque molécule de la base des données et utilisés dans cette étude. La première étape de prétraitement consiste à éliminer chaque valeur d'erreur indiquée dans le logiciel par le signe (-999), dans cette étape le nombre de descripteurs éliminés n'a pas été important ; seulement 29 sur 1666 descripteurs ont été supprimés. Les deux étapes suivantes ont été consacrées, respectivement à l'élimination des valeurs de descripteurs qui répètent à 100% puis ; à plus de 75% pour un même type de descripteur. On a comparé tout d'abord les valeurs minimale (min) et maximale (max) pour chaque type de descripteur pour confirmer que la répétition des valeurs soit à 100%, donc il faut que le minimum soit égal au maximum des valeurs du même descripteur, dans ce cas on est obligé d'éliminer ce dernier. Ensuite, on compare les trois premiers quartiles (à 25%, à 50% et à 75% des valeurs de la base des données) si ces derniers sont égaux donc on peut dire que les valeurs dans le même descripteur sont répétées plus de 75%, et on doit alors encore éliminer ce type de descripteurs et ainsi de suite. Ces deux étapes sont manipulées pour tous les descripteurs restant de l'étape précédente ; descripteur par descripteur jusqu'au dernier descripteur de la base des données. Les descripteurs éliminés dans ces deux étapes peuvent être importants à la première étape. Environ 391 sur 1637 pour les descripteurs répétés à 100% et 141 sur 1246 pour les descripteurs répétés à plus de 75%. La prochaine étape de sélection est basée sur le calcul des valeurs de *DSR* de chaque descripteur, de sorte que toutes celles qui ont une valeur inférieure à 0,05 sont négligées. Dans cette étape 104 descripteurs sur 1105 sont éliminés. Toutes les étapes de prétraitement qui ont précédées dépendent de la méthode ou de l'analyse *objective*, parce que la sélection des descripteurs a été faite sans aucune intervention de la part des variables dépendantes, c'est-à-dire sans faire participer les cinq paramètres à modéliser. L'étape suivante repose sur l'inter-corrélation des descripteurs d'entrée entre eux, et l'inter-corrélation entre ces derniers et chaque variable dépendant de sortie à part, de sorte que juste les descripteurs qui sont faiblement corrélés les uns aux autres et fortement corrélés avec chaque sortie. Cette étape est considérée comme une étape hybride repose à la fois sur les deux types de sélection : *objective* et *subjective*, parce qu'on fait participer les deux paramètres : les descripteurs d'entrée et les variables dépendantes ; en même temps dans l'étape de sélection. En conséquence, chaque paire de descripteurs avec une valeur R supérieure à 0,75 a été supprimée. Cependant, la relation entre ces derniers et leurs sorties sont prises en considération. Il est donc toujours préférable de choisir les descripteurs ayant la

meilleure corrélation avec leurs sorties, afin de ne retenir que les descripteurs pertinents. Cette étape peut être considérée comme plus forte que les autres, étant donné le grand nombre de descripteurs éliminés. Les descripteurs sélectionnés dans cette étape sont différents quantitativement et qualitativement pour chaque paramètre, parce qu'ils appartiennent à des classes différentes et différents nombres regroupés comme illustré dans les **tableaux C.1 à C.5 (*Annexe C*)** et dans le **tableau V.1**. Au cours de cette étape, plus de 900 descripteurs parmi 1001 ont été éliminés pour chaque paramètre, exactement 912, 908, 911, 910 et 906 descripteurs ont été éliminés, respectivement pour Tb, d20, Tc, Pc et Vc. La dernière étape de prétraitement est considérée comme une étape d'analyse subjective, puisque le critère de sélection ou d'élimination des descripteurs est lié aux variables indépendantes (les sorties). Dans cette étape nous avons utilisé une méthode de régression linéaire multiple "RLM" pas à pas. Nous avons retiré les descripteurs qui ont des valeurs-P (valeur de probabilité) supérieures à 0,005, où le critère significatif de régression a été strictement posé (Valeur-P < 0,005 au lieu de Valeur-P <0,05). Cette dernière étape est considérée comme l'étape la moins importante, compte tenu de l'aspect quantitatif, à cause du faible nombre de descripteurs supprimés pour chaque paramètre : 73 descripteurs sur 89 pour Tb, 77 descripteurs sur 93 pour d20, 80 descripteurs sur 90 pour Tc, 74 descripteurs sur 91 pour Pc, et 83 descripteurs sur 95 pour Vc. Toutefois, elle est également la plus importante étape lorsqu'on tient compte du côté qualitatif des descripteurs sélectionnés pour chaque paramètre, comme le montrent leurs matrices de corrélations présentées, respectivement dans les **tableaux D.1 à D.5 (*Annexe D*)**. Donc comme le montre le **tableau V.1**, à la fin de cette procédure de sélection il ne reste que : 16, 16, 10, 17 et 12 descripteurs pertinents pour les paramètres : Tb, d20, Tc, Pc et Vc, respectivement.

Tableau V.1. Évolution du choix du nombre de descripteurs selon la procédure de prétraitement par classe pour chaque paramètre étudié.

Nom des classes de descripteurs	*(0)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*					*(6)*				
						Pour *Tb*	**Pour *d20***	**Pour *Tc***	**Pour *Pc***	**Pour *Vc***	**For *Tb***	**For *d20***	**Pour *Tc***	**Pour *Pc***	**Pour *Vc***
(1) constitutional descriptors;*	48	48	33	22	20	0	2	1	1	1	0	0	0	0	0
(2) topological descriptors;*	119	119	80	74	72	6	6	7	7	6	2	2	1	1	0
(3) walk and path counts;*	47	47	47	41	41	1	1	2	1	1	0	1	2	0	0
(4) connectivity indices;*	33	33	33	33	33	7	8	8	7	7	0	1	0	1	1
(5) information indices;*	47	37	37	37	33	7	6	7	6	7	4	5	1	2	1
(6) two dimensional (2D) autocorrelations;*	96	96	60	48	48	1	2	2	0	0	0	0	0	1	0
(7) edge adjacency indices;*	107	107	106	88	88	6	6	4	7	6	1	0	0	1	2
(8) Burden eigenvalue descriptors;*	64	64	64	64	64	1	1	2	1	1	0	0	0	0	0
(9) topological charge indices;*	21	21	21	15	7	5	5	5	4	4	2	2	0	1	1
(10) Eigenvalue based indices;*	44	44	39	39	39	1	0	0	0	0	0	0	0	0	0
(11) Randic molecular profiles;*	41	41	41	41	41	0	0	1	0	0	0	0	0	0	0
(12) geometrical descriptors;*	74	74	38	34	31	4	5	5	5	5	0	1	1	1	1
(13) RDF descriptors;*	150	150	150	95	95	9	8	10	13	9	0	2	0	3	1
(14) 3D-MORSE descriptors;*	160	160	160	160	159	19	22	16	16	23	2	0	2	0	3
(15) WHIM descriptors;*	99	99	99	99	90	8	9	7	8	9	0	1	0	2	0
(16) GETAWAY descriptors;*	197	197	197	195	121	13	10	12	14	15	4	1	2	4	2
(17) functional group counts;*	154	154	14	3	3	1	1	1	1	1	1	0	1	0	0
(18) atom-centered fragments;*	120	115	11	5	5	0	1	0	0	0	0	0	0	0	0
(19) charge descriptors;*	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(20) molecular properties*	31	31	16	12	11	0	0	0	0	0	0	0	0	0	0
Nombre total des descripteurs	**1666**	**1637**	**1246**	**1105**	**1001**	**89**	**93**	**90**	**91**	**95**	**16**	**16**	**10**	**17**	**12**

(0): Après calcul par E-Dragon; ***(1)***: Après omission de la valeur des erreurs (-999); ***(2)*** Après élimination des valeurs max et min égales pour le même descripteur; ***(3)*** Après élimination des valeur répétées plus de (> 75%) pour le même descripteur; ***(4)*** Après suppression de la valeur absolue de la *DSR* inférieure à (< 0,05) ; ***(5)*** Après inter-corrélation *R* ; ***(6)*** Après application de la méthode RLM pas à pas en avant.

V.3. Caractéristiques des descripteurs pertinents sélectionnés

A la fin de la dernière étape de prétraitement, on obtient seulement les meilleurs descripteurs pertinents pour chaque paramètre, de sorte que, le nombre de descripteurs obtenus après la sélection effectuée était : 16, 16, 10, 17 et 12, respectivement pour les paramètres : Tb, d20, Tc, Pc et Vc. Etant donné les critères rapportés au paragraphe II.7.1 du chapitre II, il est nécessaire et important de réduire le nombre de descripteurs jusqu'à ce que le rapport "nombre de molécules des hydrocarbures dans la base des données sur le nombre des descripteurs sélectionné" soit > 5 et qu'ils ne soient pas corrélés entre eux **(Tropsha *et al*. 2003)**. Ces derniers seraient utilisés ensuite dans l'élaboration des modèles RQSP-RNA. Les caractéristiques des descripteurs pertinents sélectionnés pour chaque paramètre sont données dans les **tableaux C.1 à C.5 (*Annexe C*)**, respectivement pour les paramètres : Tb, d20, Tc, Pc et Vc. Les tableaux montrent les valeurs ***P*** (Valeurs des probabilités), les valeurs ***t*** (valeur de test de Student) et les définitions des descripteurs sélectionnés, ainsi qu'un résumé de la régression RLM aux bas des tableaux pour chaque paramètre étudié.

V.4. Architectures des meilleurs modèles RQSP-RNA obtenus

Après ces étapes de collecte, de réduction, de construction et de calcul, seuls les meilleurs modèles ont été sélectionnés pour tester leurs performances, en calculant diverses erreurs. Les architectures des meilleurs modèles RQSP-RNA obtenus pour chaque paramètre avec certains paramètres statistiques primaires sont représentées dans le **tableau V.2**. Les modèles RQSP-RNA possèdent des valeurs de *SDC* très petites ; dans une gamme de 10^{-6} pour le modèle de Tb et de 10^{-5} pour les modèles de : d20, Pc et Vc, ceci pour les ensembles de test et d'apprentissage. Le modèle de Tc prend aussi une valeur d'ordre 10^{-5} pour l'ensemble d'apprentissage, et une valeur d'ordre 10^{-4} pour l'ensemble de test. Les valeurs de R également montrent une performance élevée des modèles RQSP-RNA développés dans cette étude.

Comme indiqué dans le **tableau V.2**, les architectures des meilleurs modèles de type RQSP-RNA sont (16-12-1) pour le modèle de Tb, (16-10-1) pour le modèle de d20, (10-16-1) pour le modèle de Tc, (17-11-1) pour le modèle de Pc, et (12-17-1) pour le modèle de Vc. Les réseaux comportent donc 16, 16, 10, 17 et 12 entrées pour Tb, d20, Tc, Pc et Vc, respectivement. Avec une sortie $(\boldsymbol{S})$ pour chaque modèle. En ce qui concerne les couches cachées, 12, 10, 16, 11 et 17 neurones sont apparus respectivement dans les modèles de : Tb, d20, Tc, Pc et Vc, respectivement.

Pour les fonctions d'activation, nous avons trouvé en premier lieu, la fonction d'activation *tangente hyperbolique "Tanh"* quatre fois : dans les couches cachées des modèles de Tc et de Pc, aussi dans les couches sorties des modèles de Tb et de Tc. Ensuite, nous avons trouvé également la fonction *exponentielle "Exponential"* trois fois, uniquement dans les couches cachées des modèles de Tb, d20 et de Vc. En troisième position, en termes d'importance on a trouvé deux fois la fonction d'activation *sinus "Sine"* seulement dans les deux couches cachées des modèles de d20 et de Vc. Et en dernier lieu on a trouvé une fois la fonction d'activation *Sigmoïde Logarithmique "Logistic"* pour la couche de sortie du modèle de Pc.

Tableau V.2. Architectures des meilleurs modèles RQSP-RNA obtenus pour chaque paramètre.

Paramètres	**Couches**	**Nombre de neurones**	**Fonctions d'activation**	**Coefficient de corrélation *R* et valeur de *SDC***	
Tb	Couche d'entrée	16	***(Aucune)***	*R* d'apprentissage	0,9997
	Couche cachée	12	*Exponential*		
	Couche de sortie	1	*Tanh*	*R* de test	0,9999
	Algorithme d'apprentissage	BFGS 65		*SDC* d'apprentissage	$8,0\ 10^{-6}$
	Fonction d'erreur	Somme Des Carrés (*SDC*)		*SDC* de test	$3,0\ 10^{-6}$
d20	Couche d'entrée	16	***(Aucune)***	*R* d'apprentissage	0,9965
	Couche cachée	10	*Exponential*		
	Couche de sortie	1	*Sine*	*R* de test	0,9993
	Algorithme d'apprentissage	BFGS 47		*SDC* d'apprentissage	$7,4\ 10^{-5}$
	Fonction d'erreur	Somme Des Carrés (*SDC*)		*SDC* de test	$2,9\ 10^{-5}$
Tc	Couche d'entrée	10	***(Aucune)***	*R* d'apprentissage	0,9976
	Couche cachée	16	*Tanh*		
	Couche de sortie	1	*Tanh*	*R* de test	0,9973
	Algorithme d'apprentissage	BFGS 81		*SDC* d'apprentissage	$8,2\ 10^{-5}$
	Fonction d'erreur	Somme Des Carrés (*SDC*)		*SDC* de test	$1,53\ 10^{-4}$
Pc	Couche d'entrée	17	***(Aucune)***	*R* d'apprentissage	0,9988
	Couche cachée	11	*Tanh*		
	Couche de sortie	1	*Logistic*	*R* de test	0,9977
	Algorithme d'apprentissage	BFGS 165		*SDC* d'apprentissage	$3,6\ 10^{-5}$
	Fonction d'erreur	Somme Des Carrés (*SDC*)		*SDC* de test	$8,6\ 10^{-5}$
Vc	Couche d'entrée	12	***(Aucune)***	*R* d'apprentissage	0,9995
	Couche cachée	17	*Exponential*		
	Couche de sortie	1	*Sine*	*R* de test	0,9994
	Algorithme d'apprentissage	BFGS 68		*SDC* d'apprentissage	$1\ 10^{-5}$
	Fonction d'erreur	Somme Des Carrés (*SDC*)		*SDC* de test	$1,7\ 10^{-5}$

Les définitions mathématiques des fonctions d'activations apparues dans cette étude sont données dans le **tableau V.3** (Voir les **Eqs (V.1)**, **(V.2)**, **(V.3)** et **(V.4)**) **(StatSoft 2007 ; Beale *et al.* 2012)**.

Tableau V.3. Fonctions d'activation apparues dans cette étude **(StatSoft 2007 ; Beale *et al.* 2012)**.

	Fonctions d'activation			
	Exponentielle	Tangent Hyperbolique	Sinus	Sigmoïde Logarithmique
Notations dans STATISTICA	***Exponential***	***Tanh***	***Sine***	***Logistic***
Notations dans Matlab	***exp***	***tanh***	***sin***	***logsig***
Formules Mathématiques	$a = e^{-n}$ **(V.1)**	$a = \frac{e^{+n}-e^{-n}}{e^{+n}+e^{-n}}$ **(V.2)**	$a = sin(n)$ **(V.3)**	$a = \frac{1}{1+e^{-n}}$ **(V.4)**

(e): La fonction exponentielle; ***(n)***: Les variables des modèles de RNA (Voir les **Eqs (V.5)** à **(V.14)**).

V.5. Application des expressions mathématiques des modèles RQSP-RNA obtenus

Pour utiliser les modèles de type RQSP-RNA développés dans cette étude afin de prédire les cinq propriétés des hydrocarbures. Il faut tout d'abord comprendre le contenu des expressions mathématiques de chaque modèle de RQSP-RNA obtenu.

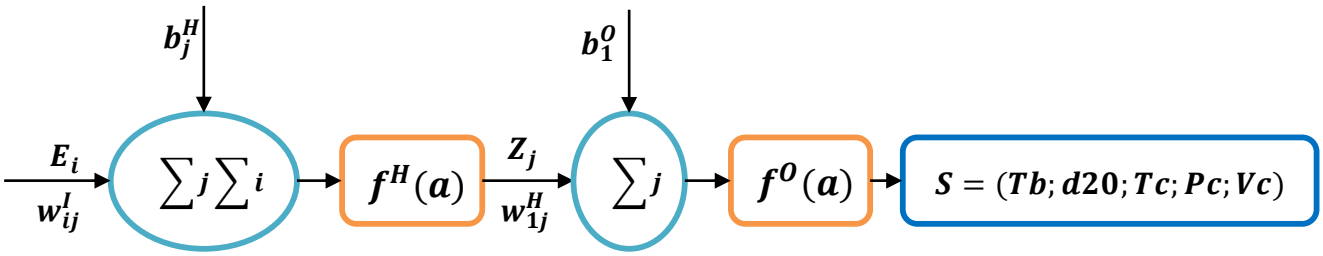


Figure V.1. Représentation schématique d'un réseau du modèle de type RQSP-RNA.

En plus des explications données sur les architectures des meilleurs modèles de RQSP-RNA, dans la partie précédente de ce chapitre, l'architecture générale d'un RNA est aussi résumée schématiquement sur la **Figure V.1**. Tous les modèles sont semblables à ceux représentés dans cette figure, la seule différence réside dans le nombre des neurones d'entrées ; dans ce cas les descripteurs moléculaires de chaque modèle, et celui des nombres de neurones dans la couche cachée de chaque modèle. Les expressions mathématiques des sorties (Z_j) de la

couche cachée et ($\boldsymbol{S}$) de la couche de sortie de chaque paramètre sont données par les équations ci-après (**V.05–V.14**) :

$$\boldsymbol{Z_j = f^H\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right] = e^{-\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right]}} \quad \textbf{(V.5)}$$

$$\boldsymbol{S = f^O\left[\sum_{j=1}^{12} w_{1j}^H Z_j + b_1^O\right] = \frac{e^{+\left[\sum_{j=1}^{12} w_{1j}^H Z_j + b_1^O\right]} - e^{-\left[\sum_{j=1}^{12} w_{1j}^H Z_j + b_1^O\right]}}{e^{+\left[\sum_{j=1}^{12} w_{1j}^H Z_j + b_1^O\right]} + e^{-\left[\sum_{j=1}^{12} w_{1j}^H Z_j + b_1^O\right]}}} \quad \textbf{(V.6)}$$

$$\boldsymbol{Z_j = f^H\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right] = e^{-\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right]}} \quad \textbf{(V.7)}$$

$$\boldsymbol{S = f^O\left[\sum_{j=1}^{10} w_{1j}^H Z_j + b_1^O\right] = \sin\left[\sum_{j=1}^{10} w_{1j}^H Z_j + b_1^O\right]} \quad \textbf{(V.8)}$$

$$\boldsymbol{Z_j = f^H\left[\sum_{i=1}^{10} w_{ij}^I E_i + b_j^H\right] = \frac{e^{+\left[\sum_{i=1}^{10} w_{ij}^I E_i + b_j^H\right]} - e^{-\left[\sum_{i=1}^{10} w_{ij}^I E_i + b_j^H\right]}}{e^{+\left[\sum_{i=1}^{10} w_{ij}^I E_i + b_j^H\right]} + e^{-\left[\sum_{i=1}^{10} w_{ij}^I E_i + b_j^H\right]}}} \quad \textbf{(V.9)}$$

$$\boldsymbol{S = f^O\left[\sum_{j=1}^{16} w_{1j}^H Z_j + b_1^O\right] = \frac{e^{+\left[\sum_{j=1}^{16} w_{1j}^H Z_j + b_1^O\right]} - e^{-\left[\sum_{j=1}^{16} w_{1j}^H Z_j + b_1^O\right]}}{e^{+\left[\sum_{j=1}^{16} w_{1j}^H Z_j + b_1^O\right]} + e^{-\left[\sum_{j=1}^{16} w_{1j}^H Z_j + b_1^O\right]}}} \quad \textbf{(V.10)}$$

$$\boldsymbol{Z_j = f^H\left[\sum_{i=1}^{17} w_{ij}^I E_i + b_j^H\right] = \frac{e^{+\left[\sum_{i=1}^{17} w_{ij}^I E_i + b_j^H\right]} - e^{-\left[\sum_{i=1}^{17} w_{ij}^I E_i + b_j^H\right]}}{e^{+\left[\sum_{i=1}^{17} w_{ij}^I E_i + b_j^H\right]} + e^{-\left[\sum_{i=1}^{17} w_{ij}^I E_i + b_j^H\right]}}} \quad \textbf{(V.11)}$$

$$\boldsymbol{S = f^O\left[\sum_{j=1}^{11} w_{1j}^H Z_j + b_1^O\right] = \frac{1}{1 + e^{-\left[\sum_{j=1}^{11} w_{1j}^H Z_j + b_1^O\right]}}} \quad \textbf{(V.12)}$$

$$\boldsymbol{Z_j = f^H\left[\sum_{i=1}^{12} w_{ij}^I E_i + b_j^H\right] = e^{-\left[\sum_{i=1}^{12} w_{ij}^I E_i + b_j^H\right]}} \quad \textbf{(V.13)}$$

$$\boldsymbol{S = f^{O}\left[\sum_{j=1}^{17} w_{1j}^{H} Z_j + b_1^{O}\right] = \sin\left[\sum_{j=1}^{17} w_{1j}^{H} Z_j + b_1^{O}\right]} \quad \textbf{(V.14)}$$

Chaque neurone caché calcule la fonction d'activation, en envoyant des résultats ($\boldsymbol{Z_j}$; $j =$ 1 à 10− 17 pour chaque modèle) au neurone de la couche de sortie, qui donne finalement la réponse du réseau ($\boldsymbol{S}$). Le signal de sortie de chaque neurone caché $\left(\boldsymbol{Z_j}\right)$ est calculé par les équations : **Eq (V.5)**, **Eq (V.7)**, **Eq (V.9)**, **Eq (V.11)**, et **Eq (V.13),** respectivement pour les modèles de : Tb, d20, Tc, Pc et Vc. Les sorties de chaque réseau sont données par les équations : **Eq (V.6)** pour le modèle de Tb, **Eq (V.8)** pour le modèle de d20, **Eq (V.10)** pour le modèle de Tc, **Eq (V.12)** pour le modèle de Pc, et **Eq (V.14)** pour le modèle de Vc.

Avec $\boldsymbol{f^H}$ la fonction d'activation de la couche cachée (reliant la couche d'entrée avec la couche cachée), $\boldsymbol{f^O}$ la fonction d'activation de la couche de sortie (reliant la couche cachée à la couche de sortie), $\boldsymbol{S}$ la sortie du réseau (paramètres de sortie : Tb, d20, Tc, Pc et Vc), $\boldsymbol{Z}$ le signal de sortie des neurones cachés, $\boldsymbol{w_{ij}^I}$ les poids de la couche d'entrée (connexions entre les neurones d'entrée et cachés), $\boldsymbol{w_{1j}^H}$ Les poids de la couche cachée (connexions entre les neurones cachés et de sortie), $\boldsymbol{E_i}$ les neurones d'entrée (dans le cas actuel, les descripteurs moléculaires pertinents), $\boldsymbol{b_j^H}$ sont les biais de la couche cachée (biais des neurones cachés), $\boldsymbol{b_1^O}$ est le biais de la couche de sortie (biais du neurone de sortie), $\boldsymbol{i}$ le nombre d'entrées pour chaque modèle ($i = 1$ à 16 ; pour les modèles de Tb et de d20, $i = 1$ à 10 ; pour le modèle de Tc, $i = 1$ à 17 ; pour le modèle de Pc, et $i = 1$ à 12 ; pour le modèle de Vc), et $\boldsymbol{j}$ est le nombre de neurones cachés ($j = 1$ à 12 ; pour le modèle de Tb, $j = 1$ à 10 ; pour le modèle de d20, $j = 1$ à 16 ; pour le modèle de Tc, $j = 1$ à 11 ; pour le modèle de Pc, et $j = 1$ à 17 ; pour le modèle de Vc).

Les relations mathématiques des modèles RQSP-RNA finaux pour la prédiction des cinq paramètres physicochimiques (Tb, d20, Tc, Pc et Vc) des hydrocarbures purs, sont obtenues en utilisant la méthode de RNA-PMC et sont donnés par les **Eq (V.6)** pour Tb, **Eq (V.8)** pour d20, **Eq (V.10)** pour Tc, **Eq (V.12)** pour Pc, et **Eq (V.14)** pour Vc.

Les valeurs des poids ($\boldsymbol{w_{ij}^I}$, $\boldsymbol{w_{1j}^H}$) et des biais ($\boldsymbol{b_j^H}$, $\boldsymbol{b_1^O}$) des cinq meilleurs modèles RQSP-RNA pour chaque couche cachée et sortie, sont données respectivement dans : le **tableau V.4** pour le modèle de Tb, dans le **tableau V.5** pour le modèle de d20, dans le **tableau V.6** pour le modèle de Tc, dans le **tableau V.7** pour le modèle de Pc, et dans le **tableau V.8** pour le modèle de Vc.

Tableau V.4. Poids et biais du meilleur modèle RQSP-RNA pour *Tb* (16-12-1).

Connexion des couches : Entrée – Cachée																	**Connexion des couches : Cachée – Sortie**	
Poids																Biais	Poids	Biais
w_{j1}^I	w_{j2}^I	w_{j3}^I	w_{j4}^I	w_{j5}^I	w_{j6}^I	w_{j7}^I	w_{j8}^I	w_{j9}^I	w_{j10}^I	w_{j11}^I	w_{j12}^I	w_{j13}^I	w_{j14}^I	w_{j15}^I	w_{j16}^I	b_j^H	w_{1j}^H	b_1^0
0,009864	0,131273	-0,159378	0,024488	-0,091216	0,143698	-0,039865	0,003232	0,137718	0,199995	-0,022475	0,094949	0,265254	0,072699	-0,034833	0,124340	0,025964	-0,019756	
0,069998	0,075598	-0,021941	0,001980	0,006373	-0,009344	0,112673	0,129439	0,019633	-0,087861	-0,056111	-0,010917	0,397759	-0,186341	0,223640	0,197385	-0,033483	0,241887	
0,116355	0,011479	0,199730	0,087473	0,051825	0,028454	0,030022	0,050706	0,005365	0,029440	-0,080623	0,014642	0,413759	-0,090053	0,038812	0,107484	0,025180	-0,300627	
-0,070792	-0,025558	0,079572	0,103414	0,031218	-0,007215	0,093104	-0,028382	-0,034935	-0,040122	-0,029904	0,143385	-0,004062	-0,074794	0,141121	-0,010990	-0,134288	0,182875	
0,024399	0,069101	-0,034797	-0,030407	-0,080989	0,089859	0,138529	-0,204162	0,075355	-0,049025	-0,060384	0,022731	0,134390	-0,077829	0,094782	-0,023460	0,039860	-0,078214	
0,071626	-0,019226	0,189392	0,122801	0,115880	0,065268	0,007156	-0,047033	-0,056113	-0,012374	0,062477	-0,096835	0,216472	0,225703	0,063699	0,208738	-0,065939	0,358785	-0,238553
0,135111	0,130306	-0,020245	-0,036751	0,056104	0,076463	0,045008	0,228372	0,060847	-0,102220	-0,083894	-0,033674	0,277753	0,122682	0,114253	-0,275343	0,059628	-0,262796	
0,034835	-0,113186	-0,061448	0,049020	-0,042338	-0,055638	0,005716	0,042338	0,067990	-0,078587	-0,071901	0,047684	0,013703	0,029865	-0,111052	0,054865	0,216495	-0,543612	
-0,061365	0,145887	0,102048	0,166568	-0,023350	0,042352	-0,016121	0,103132	-0,002081	-0,063945	-0,092095	0,041444	0,128004	-0,011819	-0,122846	0,195958	-0,226731	0,441421	
0,082497	0,013939	0,074779	-0,095042	-0,074363	0,020530	-0,030233	0,196304	0,085556	-0,052631	0,007707	-0,202520	0,140653	0,026646	0,467387	0,176520	-0,021818	0,574267	
0,124394	0,019825	0,064614	0,005369	0,070898	0,088739	0,146040	0,019060	-0,083061	0,090296	0,068613	-0,043028	-0,011746	0,261647	0,290865	0,071172	0,103426	-0,514280	
0,060392	0,007065	-0,060827	0,042004	-0,008327	0,113876	-0,061116	-0,008284	0,159160	-0,211767	0,078398	-0,091167	0,117979	0,093913	-0,070512	-0,034277	-0,031219	0,430837	

Tableau V.5. Poids et biais du meilleur modèle RQSP-RNA pour *d20* (16-10-1).

Connexion des couches : Entrée – Cachée																	**Connexion des couches : Cachée – Sortie**	
Poids																Biais	Poids	Biais
w_{j1}^{I}	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	w_{j8}^{I}	w_{j9}^{I}	w_{j10}^{I}	w_{j11}^{I}	w_{j12}^{I}	w_{j13}^{I}	w_{j14}^{I}	w_{j15}^{I}	w_{j16}^{I}	b_{j}^{H}	w_{1j}^{H}	b_{1}^{0}
0,062821	0,005073	0,115269	-0,238803	-0,358561	0,256714	-0,049285	0,431838	0,237497	-0,100146	-0,011440	-0,143733	0,333071	-0,079278	-0,022923	-0,163323	-0,036450	0,088829	
-0,012124	-0,086282	0,239915	-0,558316	-0,118554	-0,061053	-0,029464	-0,065819	0,004777	0,317418	0,095945	0,041557	0,136082	0,020095	-0,135007	-0,215363	0,026637	0,409268	
0,030587	0,053311	0,080411	0,236126	0,227635	-0,019908	-0,111884	0,013157	0,237525	-0,008611	-0,214334	-0,213450	-0,206603	0,056443	0,034694	-0,075190	0,024590	0,421947	
-0,031150	0,024171	-0,192073	0,187201	-0,107494	0,167835	-0,178944	0,129681	-0,154646	0,239620	0,088722	0,090139	0,127146	-0,113437	-0,056483	-0,022605	-0,022309	0,223981	
-0,112589	-0,010526	-0,354779	0,281519	0,206561	0,007354	-0,269775	-0,261835	-0,012277	-0,121527	-0,050556	-0,209818	-0,184914	0,092562	0,253307	-0,213246	0,082509	-0,478869	
0,038157	0,006942	0,293059	-0,004016	0,041515	-0,138464	0,100690	-0,113383	0,160129	-0,107120	-0,075575	-0,209873	-0,058162	-0,087198	-0,335935	-0,113587	-0,048136	-0,177317	0,048832
-0,072574	0,261802	-0,239721	-0,185052	-0,145032	-0,233951	-0,247444	-0,110889	-0,227703	0,563301	0,050328	0,003641	0,009245	0,068542	0,129292	0,047183	-0,075357	-0,117345	
-0,071034	-0,095698	-0,032301	0,318407	-0,145645	0,143420	-0,070729	-0,499925	0,095738	-0,117533	0,205199	-0,271359	-0,200542	0,038140	0,088674	-0,103414	0,059010	-0,220648	
-0,046272	0,509844	-0,283572	0,035221	0,099706	0,041304	-0,026065	0,149718	0,218163	0,429139	0,072789	0,337555	-0,136832	-0,035953	-0,296564	-0,054112	-0,077512	-0,112767	
-0,055639	-0,080644	-0,162031	-0,288892	-0,025937	-0,079025	-0,370050	-0,164888	0,054800	-0,274146	0,163328	-0,079330	-0,022314	-0,039497	0,145382	-0,012521	-0,058363	0,318762	

Tableau V.6. Poids et biais du meilleur modèle RQSP-RNA pour *Tc* (10-16-1).

Connexion des couches : Entrée – Cachée											**Connexion des couches : Cachée – Sortie**	
Poids										Biais	Poids	Biais
w_{j1}^{I}	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	w_{j8}^{I}	w_{j9}^{I}	w_{j10}^{I}	b_j^H	w_{1j}^{H}	b_1^0
-0,15062	-0,04835	-0,39689	0,47320	-0,02841	-0,18534	0,03817	0,31585	0,02332	-0,02752	0,05874	-0,19936	
0,02147	0,31160	0,51162	0,75510	-0,19245	-0,09777	-0,04994	0,03833	-0,04666	0,02934	0,59810	-0,80999	
-0,26009	-0,15153	-0,14701	-0,15974	-0,12369	-1,45394	0,01441	-0,09436	-0,11168	-0,08741	0,12517	-0,40962	
0,03480	-0,24006	0,69171	0,06338	0,00820	0,21320	-0,01170	0,33738	0,01877	0,03036	0,21771	-1,17357	
-0,00143	-0,00983	0,13454	-0,08957	0,01141	0,15888	0,07253	0,00732	-0,15834	-0,07532	-0,14396	0,40673	
-0,13810	0,52523	-0,06821	-0,33050	-0,01588	-0,33601	0,02006	-0,10717	0,04716	-0,05385	-0,66358	-1,04174	
-0,06398	-0,03321	0,33320	-0,10352	-0,04236	-0,15479	0,12958	0,07088	-0,04862	-0,12308	-0,09802	-0,48533	
0,08240	-0,13626	0,16346	0,15009	-0,01181	0,51227	-0,09353	-0,02031	-0,10027	-0,20365	-0,09695	-0,12975	0,57809
-0,02064	0,40932	0,31528	-0,25765	-0,05788	-0,03737	0,02749	0,27117	-0,04734	-0,17881	0,38197	0,82515	
0,09122	0,30628	-0,02535	-0,66550	-0,05819	-0,31422	0,01583	0,09327	0,11375	0,18785	-0,04937	0,06644	
-0,32229	0,29216	0,02286	-0,06072	-0,02517	0,02120	0,02339	0,29242	-0,01784	-0,11614	-0,10043	0,02422	
-0,48615	-0,01221	0,15035	0,20302	-0,10620	0,01099	-0,12251	-0,02330	0,09969	-0,07143	-0,30831	0,71381	
-0,17348	0,15408	-0,50977	-0,18178	-0,04437	0,12093	0,03535	0,04526	-0,01735	-0,06485	-0,20947	0,69193	
0,38507	0,07945	-0,29320	-0,01402	-0,02873	-0,06547	-0,06479	0,28507	-0,12290	0,21096	0,09950	-0,09009	
-0,04653	-0,18394	0,18712	-0,29016	0,00425	-0,01994	0,33729	0,09003	-0,19145	-0,06275	-0,03536	-0,20525	
-0,24276	-0,26084	-0,39020	-0,12907	-0,00438	-0,00520	-0,08621	-0,13019	0,07776	0,01058	0,21820	-0,31756	

Tableau V.7. Poids et biais du meilleur modèle RQSP-RNA pour *Pc* (17-11-1).

Connexion des couches : Entrée – Cachée																		**Connexion des couches : Cachée – Sortie**	
Poids																	Biais	Poids	Biais
w_{j1}^{I}	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	w_{j8}^{I}	w_{j9}^{I}	w_{j10}^{I}	w_{j11}^{I}	w_{j12}^{I}	w_{j13}^{I}	w_{j14}^{I}	w_{j15}^{I}	w_{j16}^{I}	w_{j17}^{I}	b_j^H	w_{1j}^{H}	b_1^0
-0,90792	-0,22003	-0,00954	-0,34444	0,02462	-0,57670	-0,70405	0,10657	-0,51347	-1,08236	-0,10852	-0,37277	0,11367	0,16433	0,08528	-0,06245	-0,81650	-1,50529	-0,75957	
0,41018	0,68653	-0,20604	-1,61980	0,41291	0,10640	0,07300	-0,22791	0,44062	-0,90942	0,01210	0,02703	-0,41490	0,09198	0,09004	0,15543	-1,04943	-0,33861	-0,59869	
-0,21451	-0,34483	0,21899	0,00165	0,39137	-0,02731	-0,14133	-0,28855	0,17829	0,78403	-0,06959	-0,68309	-0,15980	0,10601	-0,01428	0,07381	-1,44260	-0,17015	0,84999	
1,22365	0,96440	-0,12170	0,44511	-0,01503	0,02229	0,14465	-0,17394	1,01817	-0,06836	-0,10885	-1,23867	0,46180	-0,18219	0,07354	0,02520	0,18154	-0,49356	-0,76528	
0,22352	1,10510	0,01350	-1,09463	2,18883	-0,41918	0,00821	0,24657	0,41799	0,05810	-0,00216	1,09163	0,19260	0,21218	0,05002	0,07197	0,59640	0,35193	0,36120	
0,80189	-0,55745	-0,05684	-0,25003	-0,75685	-0,18805	-0,71995	0,34778	0,90862	0,02587	0,02944	0,09100	-0,65031	-0,35764	0,00516	-0,54038	-0,16166	0,49512	0,94145	-0,02354
-0,63243	0,43793	-0,01040	0,15244	-0,95846	0,42459	-0,01157	-0,14116	-0,37777	-0,27059	0,05612	-0,14703	-0,07882	-0,26139	0,12785	-0,85852	-0,01394	0,00502	0,13537	
-0,23489	-0,37447	-0,22595	-1,37216	0,88447	-0,07541	0,15675	-0,03916	-0,40008	-0,07344	-0,03614	-0,07898	1,16325	-0,35033	-0,00884	1,20368	-0,03051	0,52198	-0,51932	
0,26179	0,00299	-0,21818	0,25917	0,53151	-0,77885	-0,18291	-0,82141	0,53795	-0,05184	0,06932	-1,27748	0,50146	0,10783	-0,00578	0,03554	-0,77477	-0,25549	-0,69075	
0,40964	1,00952	-0,34931	-0,13622	0,01051	-0,17363	-0,21052	0,49576	1,08031	0,02066	-0,29514	-0,57984	-0,05708	-0,08452	0,12533	0,36459	-0,81896	0,32257	-0,03509	
0,53078	0,20722	0,15332	0,00269	0,76564	0,05183	-0,54833	0,34894	-0,05986	-0,00063	0,61791	0,49373	-0,00464	0,06473	0,12127	-0,76142	0,38783	0,03646	1,00925	

Tableau V.8. Poids et biais du meilleur modèle RQSP-RNA pour *Vc* (12-17-1).

Connexion des couches : Entrée – Cachée													**Connexion des couches : Cachée – Sortie**	
Poids												Biais	Poids	Biais
w_{j1}^{I}	w_{j2}^{I}	w_{j3}^{I}	w_{j4}^{I}	w_{j5}^{I}	w_{j6}^{I}	w_{j7}^{I}	w_{j8}^{I}	w_{j9}^{I}	w_{j10}^{I}	w_{j11}^{I}	w_{j12}^{I}	b_{j}^{H}	w_{1j}^{H}	b_{1}^{0}
-0,019192	-0,009954	0,102814	-0,081152	0,021129	0,124626	-0,087825	0,020905	-0,028094	0,099030	0,064316	0,094526	-0,019362	-0,090290	
0,053516	-0,016830	0,068774	0,003939	0,030533	-0,081736	0,053104	-0,094718	0,022712	-0,038003	-0,159530	0,174095	-0,010153	0,010047	
0,034318	-0,001769	0,068658	0,112124	0,019075	0,024490	0,031615	-0,044952	0,017280	-0,042863	0,136649	-0,060210	0,053389	-0,280587	
0,005852	-0,001892	0,047046	-0,032749	-0,008543	-0,049510	0,055946	0,016557	0,045455	-0,098576	0,228449	0,043592	-0,018649	0,061338	
0,043349	-0,009129	-0,057707	0,031008	0,062704	0,222432	-0,002401	0,018043	0,006697	-0,110573	0,044811	0,112874	0,008181	0,119186	
-0,091205	0,005078	-0,021898	0,000567	-0,092605	-0,100685	-0,032419	-0,047019	0,062331	-0,065643	0,048835	-0,004751	0,026222	0,143248	
-0,032447	-0,004211	0,005296	0,037646	-0,105302	0,186104	-0,035894	-0,163460	0,007649	-0,080595	-0,035858	-0,010201	0,402856	-0,540056	
-0,004368	0,000609	0,071910	0,008020	-0,144679	-0,142321	0,046405	0,130658	-0,030291	-0,079205	0,077153	-0,023808	0,029742	0,327527	
0,036954	0,008175	-0,024138	-0,005256	-0,018880	-0,073353	0,074619	-0,053634	0,006384	0,217617	0,171967	-0,028116	0,111933	0,178914	0,039784
-0,047398	0,017424	0,047938	0,033362	-0,585031	0,348534	-0,015454	-0,130088	-0,063006	-0,074179	0,002498	-0,005594	0,117314	-0,173635	
0,016052	-0,208801	-0,041478	-0,031148	-0,071252	0,005973	-0,015997	-0,032541	0,069676	-0,003876	0,075475	0,031376	0,111393	-0,222701	
0,048765	-0,080095	0,005579	0,032015	0,033550	0,000307	-0,194220	0,010789	-0,141495	0,146250	0,193637	-0,011977	0,062348	0,015424	
0,010344	-0,206820	0,090406	-0,056812	-0,104579	-0,071977	0,005189	-0,060394	-0,006610	0,083297	0,124854	-0,000713	0,066370	0,201630	
-0,003928	0,100549	0,032386	-0,000507	0,181316	0,088248	0,012313	0,001286	0,157319	-0,065095	0,015456	0,009603	0,080348	0,253551	
0,005344	0,025547	0,024488	0,083328	0,170256	-0,111506	0,017928	-0,028096	-0,046069	-0,135708	0,126529	0,001651	0,043580	0,338514	
-0,004605	0,076006	-0,030082	-0,028652	-0,094883	-0,052962	-0,022344	0,045863	-0,006963	0,149357	-0,167949	-0,013309	-0,022993	-0,247600	
-0,010914	-0,145947	0,029908	0,021413	0,001561	0,148881	0,050038	-0,016441	0,041777	0,061694	-0,188571	0,003361	-0,007641	0,069069	

V.6. Prédiction des propriétés physicochimiques des hydrocarbures par les modèles RQSP-RNA

Afin d'estimer les propriétés physicochimiques des hydrocarbures : Tb, d20, Tc, Pc et Vc, j'utilise, respectivement les équations (**Eq (V.6)**, **Eq (V.8)**, **Eq (V.10)**, **Eq (V.12)**, et **Eq (V.14)**) en introduisant les valeurs des entrées $\boldsymbol{E_i}$ qui sont les valeurs des descripteurs pertinents choisis pour chaque paramètre, données dans les **tableaux E.1–E.5 (*Annexe E*)**, classés du premier au dernier descripteur :

- Pour le modèle de Tb (*STN, TIE, IVDM, Yindex, IC3, CIC3, EEig07x, GGI4, JGI1, Mor10v, Mor16e, ISH, H5e, R3e, R4e* et *nCt*) (comme ils sont rangés dans le **tableau C.1 (*Annexe C*)**).
- Également pour le modèle de d20 (*GNar, J, piPC03, X3Av, IVDE, HDcpx, Yindex, IC3, CIC3, GGI5, JGI1, SPH, RDF030m, RDF055m, L3p* et *HATS3u*) (comme ils sont rangés dans le **tableau C.2 (*Annexe C*)**).
- Pour le modèle de Tc (*piPC07, H5e, MWC10, Xindex, SPH, D/D, Mor11v, Mor15u, nCt* et *HATS4u*) (comme ils sont rangés dans le **tableau C.3 (*Annexe C*)**).
- Pour le modèle de Pc (*CIC0, P2u, JGT, RDF035u, PJI3, RDF030v, R2m, BAC, RDF055m, EEig06r, Vindex, R1e, H0e, R5e, L3u, X4Av* et *GGI5*) (comme ils sont rangés dans le **tableau C.4 (*Annexe C*)**).
- Pour le modèle de Vc (*ADDD, BIC1, RDF045u, Mor07u, Mor30v, EEig09x, JGI1, HATS3p, ESpm09d, X2A, H1e* et *Mor08p*) (comme ils sont rangés dans le **tableau C.4 (*Annexe C*)**).

En plus de l'utilisation des "poids" et des "biais" des connexions des deux couches, cachée et sortie, données : dans le **tableau V.4** pour Tb, dans le **tableau V.5** pour d20, dans le **tableau V.6** pour Tc, dans le **tableau V.7** pour Pc, et dans le **tableau V.8** pour Vc.

V.7. Estimation de la performance des résultats d'estimation issus des modèles RQSP-RNA développés

IV.7.1. Comparaison graphique entre les valeurs expérimentales et les valeurs calculées à partir des modèles RQSP-RNA

Les comparaisons visuelles entre les valeurs expérimentales et calculées obtenues en utilisant les modèles de RQSP-RNA développés sont illustrées sur les figures : **Figure V.2** pour

le modèle de Tb, **Figure V.3** pour le modèle de d20, **Figure V.4** pour le modèle de Tc, **Figure V.5** pour le modèle de Pc, et sur la **Figure V.6** pour le modèle de Vc. Dans chaque figure, les trois ensembles sont respectivement désignés par les lettres : **(a)** pour l'ensemble entier des données, **(b)** pour l'ensemble des données d'apprentissage et **(c)** pour l'ensemble des données du test.

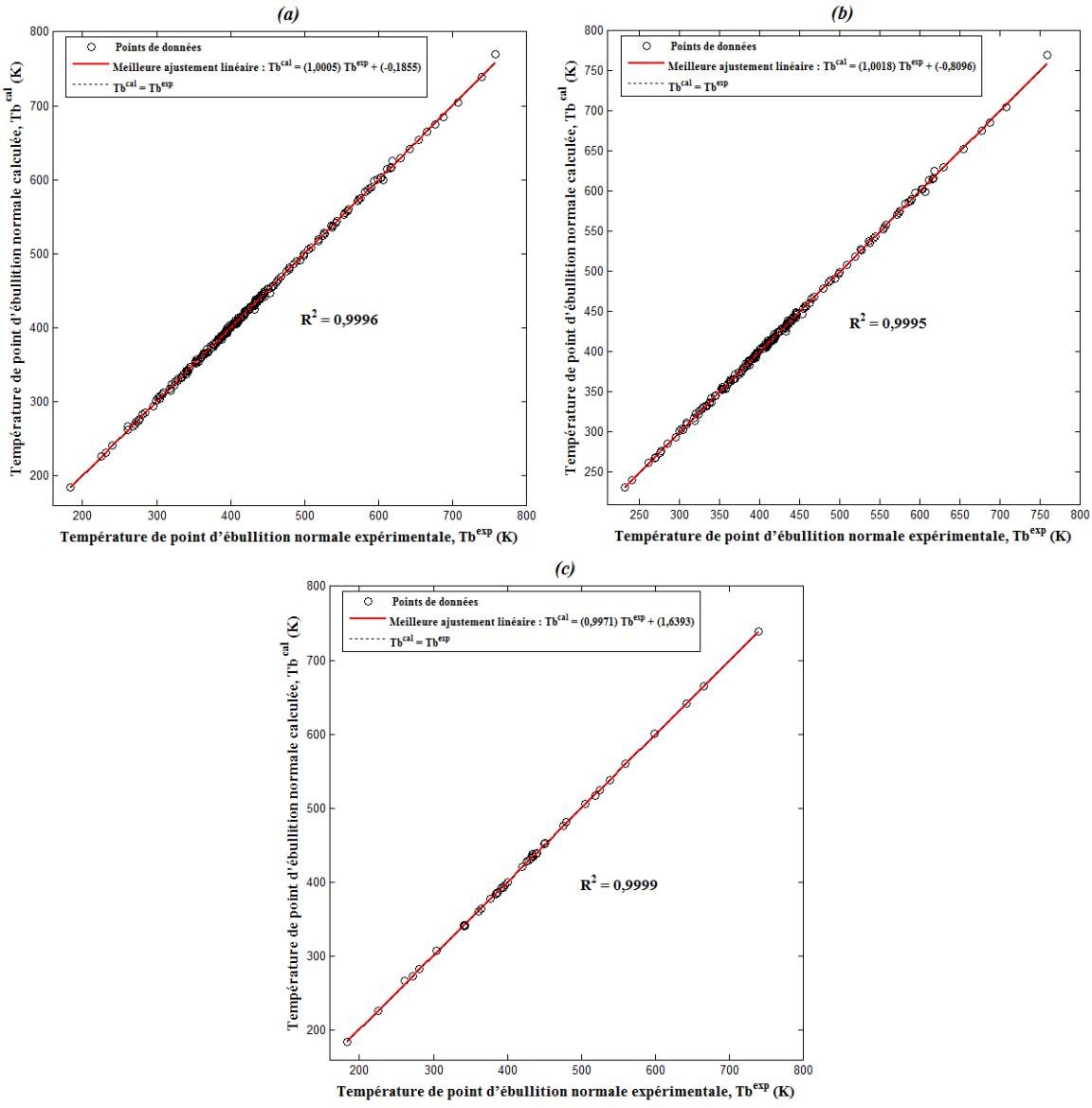


Figure V.2. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la température du point d'ébullition normal Tb pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test.

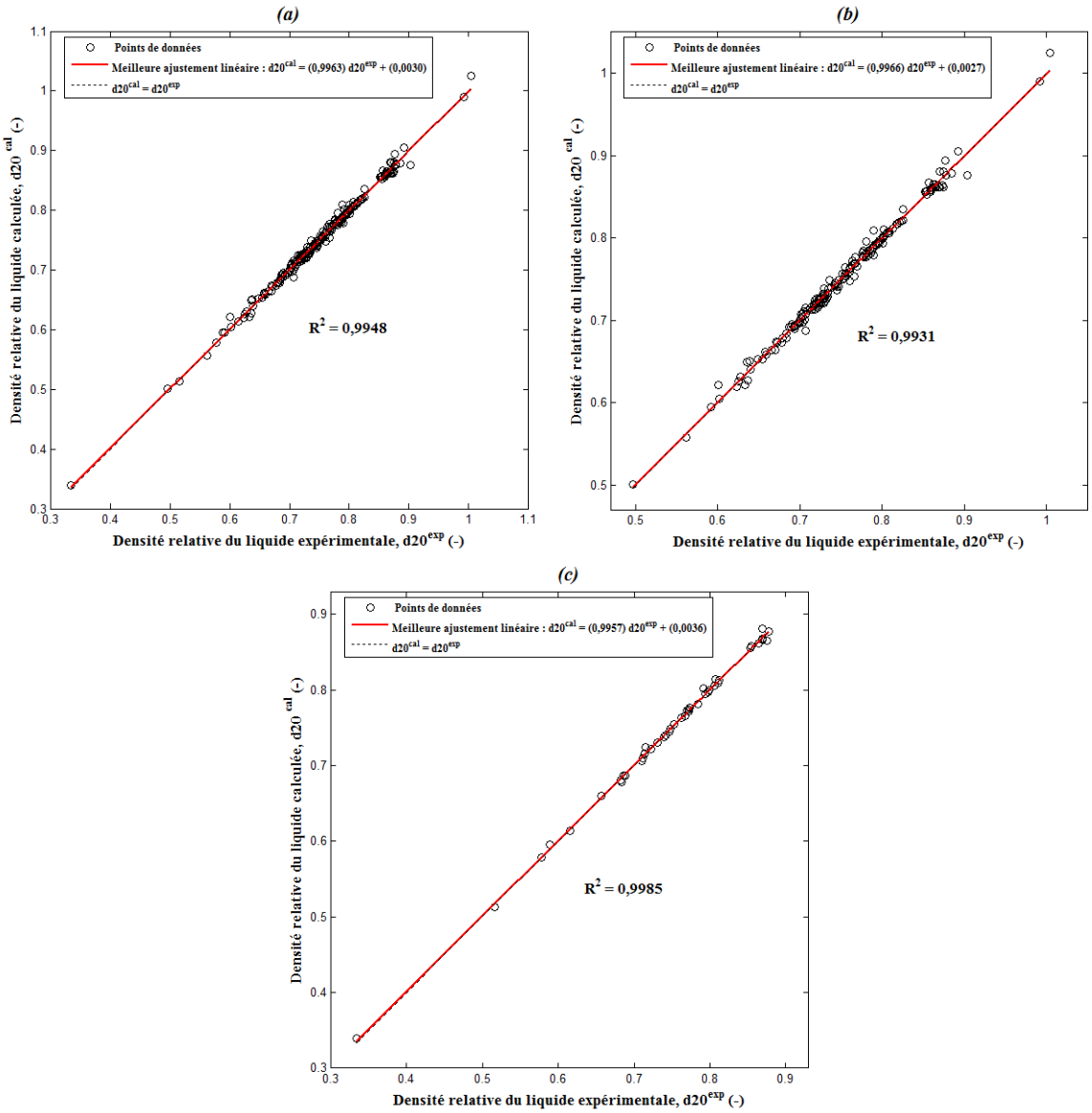


Figure V.3. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la densité relative liquide d20 pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test.

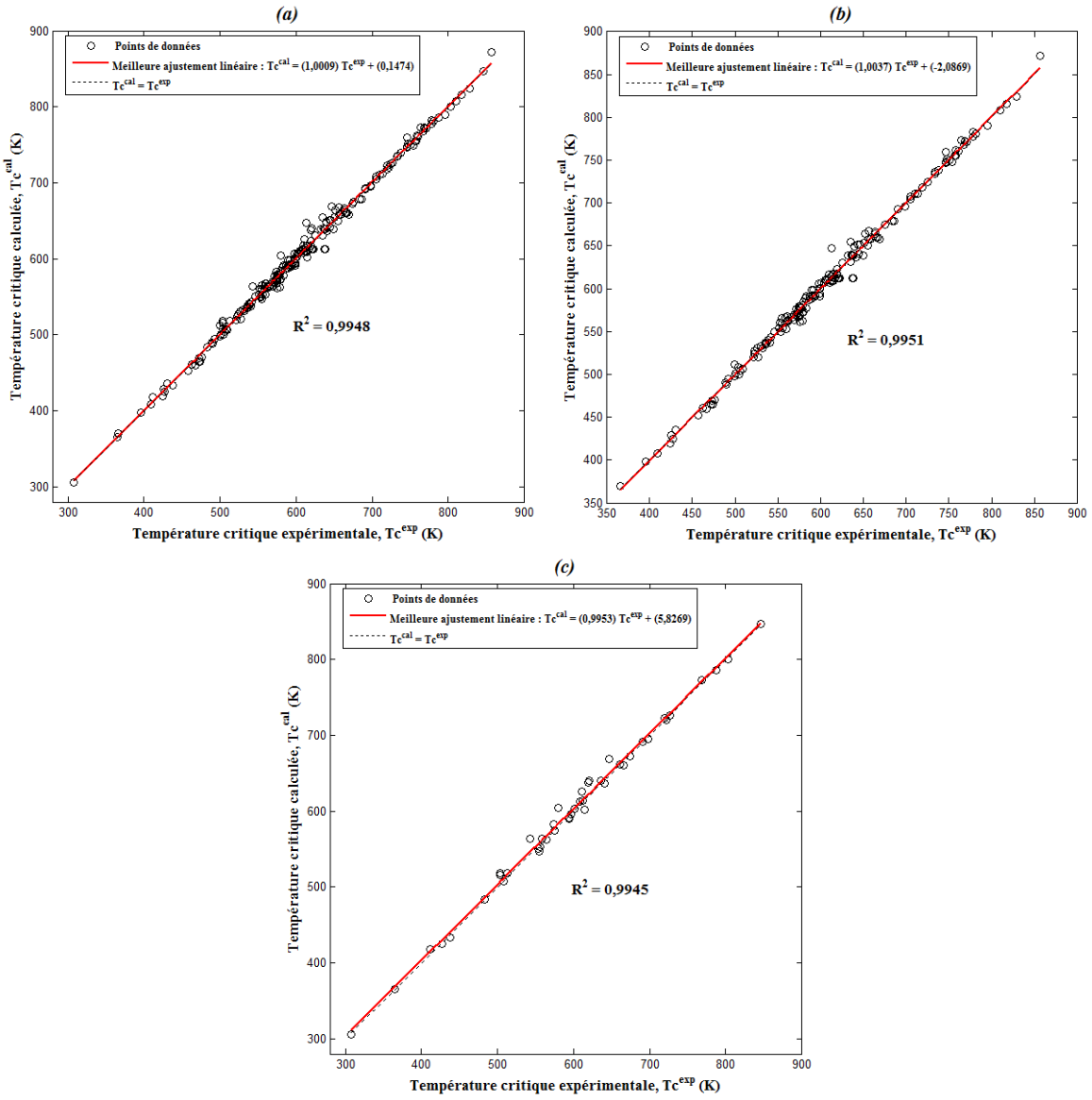


Figure V.4. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la température critique Tc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test.

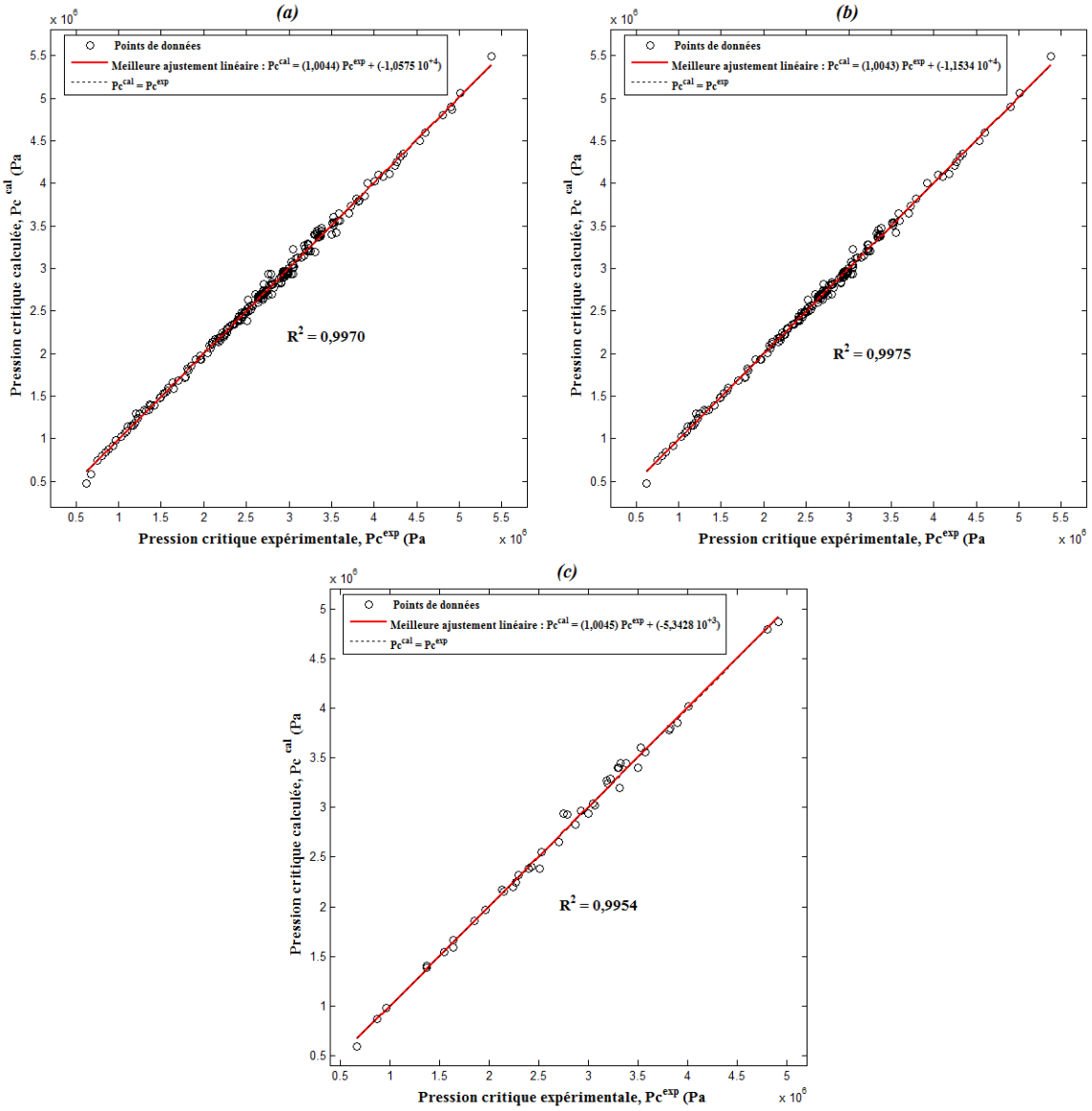


Figure V.5. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de la pression critique Pc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test.

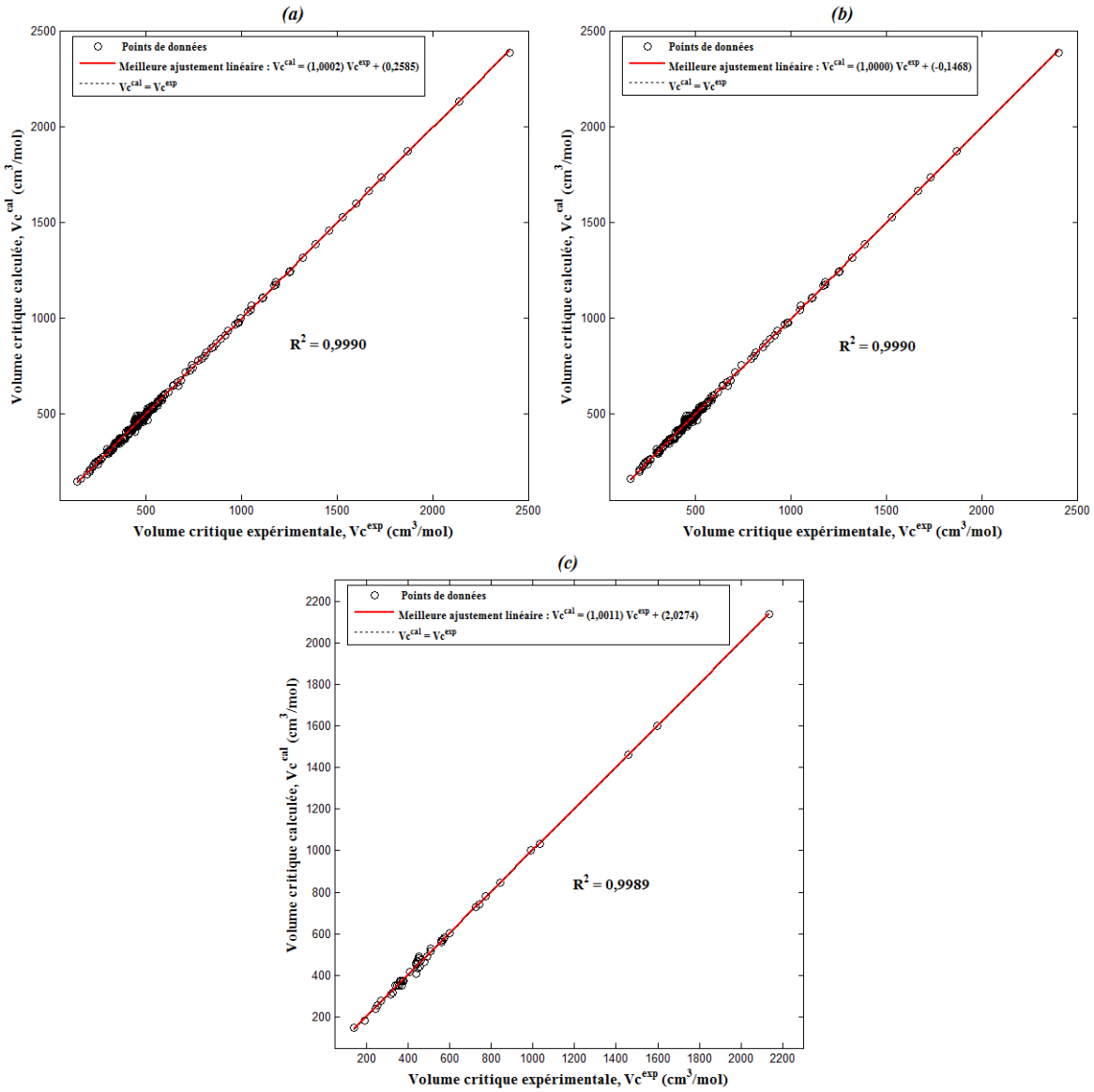


Figure V.6. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA du volume critique Vc pour l'ensemble de base des données : ***(a)*** entier, ***(b)*** d'apprentissage et ***(c)*** de test.

Une concentration étroite de points de données autour de la ligne de pente unitaire (indiquée par la ligne en pointillé) révèle que les modèles de RQSP-RNA peuvent prédire avec précision les valeurs de Tb, d20, Tc, Pc et de Vc. Il a été confirmé que les modèles proposés donnent des résultats satisfaisants, en accord de précision avec les valeurs vectorielles de corrélation proches de l'idéal (i.e. $\alpha = 1$ (pente), $\beta = 0$ (ordonnée à l'origine), $R = 1$), dans les graphiques appropriés des profils de Tb, d20, Tc, Pc et Vc, pour l'ensemble entier de données, d'apprentissage et de test (voir les équations de régression en haut de chaque figure). Sur la base des figures précédentes, les modèles de RQSP-RNA donnent de bons résultats avec des valeurs de R^2 élevées (voir également le **tableau V.9**). Les valeurs expérimentales et calculées pour les

cinq modèles (RQSP-RNA de : Tb, d20, Tc, Pc et Vc) ainsi que la manière de division sont présentées dans le **tableau A.1 (*Annexe A*)**.

V.7.2. Calculs des paramètres de performance et analyse statistique des modèles de type RQSP-RNA obtenus

Les performances des meilleurs modèles de RQSP-RNA sont résumées dans le **tableau V.9**, en termes de la racine d'erreur quadratique moyenne (*REQM*), l'erreur standard de prédiction (*ESP*), l'erreur relative moyenne (*ERM*), et l'erreur relative absolue moyenne (*ERAM*), correspondant à chaque modèle dans l'ensemble de base des données : entier, d'apprentissage et du test. En plus des valeurs de R et R^2, les valeurs de la pente (α), de l'interception (β) et le nombre de points de données (N) dans chaque phase. Les définitions mathématiques des types d'erreurs précédentes, du coefficient de régression (R), et du coefficient de détermination (R^2), sont respectivement, données par les **Eqs (IV.1–IV.6)** dans la partie IV.3.6.1 du chapitre précédent (chapitre IV : Matériels et méthodes).

Tableau V.9. Paramètres statistiques des meilleurs modèles RQSP-RNA pour les cinq propriétés.

Paramètres Statistiques	*R*	R^2	*(REQM)*	*(ESP)*	*(ERM)*	*(ERAM)*	*α*	*β*	*N*
Phase d'apprentissage									
Tb^{App}	0,9997	0,9995	2,1458	0,5009	0,3536	0,3539	1,0018	-0,8096	179
$d20^{App}$	0,9965	0,9931	0,0064	0,8429	0,5804	0,5797	0,9966	0,0027	178
Tc^{App}	0,9976	0,9951	6,4445	1,0602	0,7425	0,7424	1,0037	-2,0869	179
Pc^{App}	0,9988	0,9975	$4{,}279\ 10^{+4}$	1,6054	1,2389	1,2781	1,0043	$-1{,}1534\ 10^{+4}$	177
Vc^{App}	0,9995	0,9990	10,0040	1,7680	1,6405	1,6409	1,0000	-0,1468	179
Phase de test									
Tb^{Test}	0,9999	0,9999	1,3718	0,3255	0,2613	0,2600	0,9971	1,6393	44
$d20^{Test}$	0,9993	0,9985	0,0040	0,5381	0,3975	0,3960	0,9957	0,0036	44
Tc^{Test}	0,9972	0,9945	8,7785	1,4626	1,0152	0,9974	0,9953	5,8269	44
Pc^{Test}	0,9977	0,9954	$6{,}5715\ 10^{+4}$	2,4024	1,9958	2,0212	1,0045	$-5{,}3428\ 10^{+3}$	44
Vc^{Test}	0,9994	0,9989	12,8870	2,2855	2,2683	2,2497	1,0011	2,0274	44
Phase entière (apprentissage + test)									
Tb^{Ent}	0,9998	0,9996	2,0168	0,4723	0,3354	0,3353	1,0005	-0,1855	223
$d20^{Ent}$	0,9974	0,9948	0,0060	0,7930	0,5441	0,5433	0,9963	0,0030	222
Tc^{Ent}	0,9974	0,9948	6,9672	1,1491	0,7963	0,7927	1,0009	0,1474	223
Pc^{Ent}	0,9985	0,9970	$4{,}8231\ 10^{+4}$	1,8001	1,3896	1,4261	1,0044	$-1{,}0575\ 10^{+4}$	221
Vc^{Ent}	0,9995	0,9990	10,6349	1,8808	1,7644	1,7610	1,0002	0,2585	223

Selon les résultats de différents coefficients statistiques et types d'erreur présentés dans le **tableau V.9**, les modèles développés étaient très intéressants, comme le montrent les valeurs de R^2 supérieures à **0,9930** dans les cinq modèles de : Tb, d20, Tc, Pc et Vc ; pour la phase de base des données entière, d'apprentissage et de test.

Les types d'erreur ont confirmé que la valeur de REQM du modèle Tb allait d'une valeur maximale de 2,1458 K pour la phase d'apprentissage à une valeur minimale de 1,3718 K pour la phase de test. Néanmoins, la valeur de REQM de la base des données entière ; ce qui signifie la valeur réelle du modèle de Tb développé, comprise dans la même gamme, qui est exactement 2,0168 K, confirmant la stabilité du modèle. La même remarque en ce qui concerne les modèles de d20, Tc, Pc et Vc, avec une valeur de REQM comprise entre la valeur extrême de 0,0064 et la plus petite valeur de 0,0040, respectivement pour la phase d'apprentissage et de test du modèle de d20. Une valeur maximale de 8,7785 K et une valeur minimale de 6,4445 K pour le modèle de Tc, la plus grande valeur de 6,5715 10^{+4} Pa et la plus petite valeur de 4,279 10^{+4} Pa pour le modèle de Pc, l'extrême valeur de 12,887 cm^3/mol et le minimum valeur de 10,0040 cm^3/mol pour le modèle, respectivement, pour les phases de test et d'apprentissage de chaque modèle. Les valeurs de REQM des ensembles de base des données totales appartenant aux restes des modèles précédentes développées sont presque semblable aux deux autres ensembles de base des données de chaque modèle, à savoir, égal à : 0,0060; 6,9672 K; 4,8231 10^{+4} Pa; et de 10,6349 cm^3/mol, respectivement pour les modèles de d20, Tc, Pc et de Vc, prouvant également l'équilibre entre le sous-apprentissage (sous-ajustement) et le sur-apprentissage (sur-ajustement) des modèles de RQSP-RNA actuels, ce qui a confirmé encore leurs stabilités **(Bishop 1995 ; Pigram & MacDonald 2001 ; Hawkins 2004)**.

On sait que l'unité des valeurs de REQM dépend de la propriété concernée comptée (c'est-à-dire que la valeur du REQM de Tb par exemple est de la même unité que l'unité de Tb, K), il semble donc que la REQM ne soit pas un indicateur exhaustif pour juger un modèle, car il existe de grandes différences entre les valeurs si vous utilisez différentes unités. De plus, dans la même unité, il est difficile d'identifier quelles sont les valeurs de référence de REQM ; sont-elles vraiment de petites ou de grandes valeurs **(Chai & Draxler 2014)**. En conséquence, j'essaye de rechercher d'autres types d'erreurs sur la base d'une unité plus représentative, telles que les erreurs relatives exprimées en pourcentage (%).

Compte tenu de la définition mathématique du REQM et de l'ESP calculées par les équations **Eq (IV.3)** et **Eq (VI.4)**, la valeur de l'ESP est en fait de même type que celle de la

REQM mais exprimée en pourcentage (%). Les valeurs des ESP approuvant les résultats précédents des REQM sont inférieures à 1% dans les modèles de Tb et de d20, et ne dépassent pas les 2,4 % pour le reste des modèles de Tc, Pc et Vc. Le meilleur résultat est associé au modèle de Tb (inférieure à 0,51%), ceci étant pour les trois ensembles de base des données. Les valeurs de l'ERM et de l'ERAM confirment que nos modèles de RQSP-RNA sont plus précis lors de la prédiction des paramètres : Tb, d20, Tc, Pc et Vc. Les valeurs des ERAM comprises entre 0,26 et 0,36 % pour le modèle de Tb, entre 0,40 et 0,58 % pour le modèle de d20, entre 0,75 et 1 % pour le modèle de Tc, entre 1,28 et 2,022 % pour le modèle de Pc, et enfin comprise entre 1,65 et 2,25 % pour le modèle de Vc.

V.7.3. Analyse de sensibilité des descripteurs dans chaque modèle RQSP-RNA par la méthode de poids

Comme je l'ai mentionné précédemment dans le paragraphe VI.3.6.2 (chapitre VI), je vais utiliser la méthode d'analyse de sensibilité au poids, pour apprécier la contribution des descripteurs d'entrée sur chaque modèle de RQSP-RNA, et déterminant l'impact de chaque descripteur d'entrée individuel sur la sortie de chaque modèle de RQSP-RNA pour tout paramètre.

Les contributions des descripteurs pertinents sur les modèles de RQSP-RNA ont été établies et illustrées sur les figures : **Figure V.7**, **Figure V.8**, **Figure V.9**, **Figure V.10** et la **Figure V.11**, respectivement pour les modèles de Tb, d20, Tc, Pc et Vc.

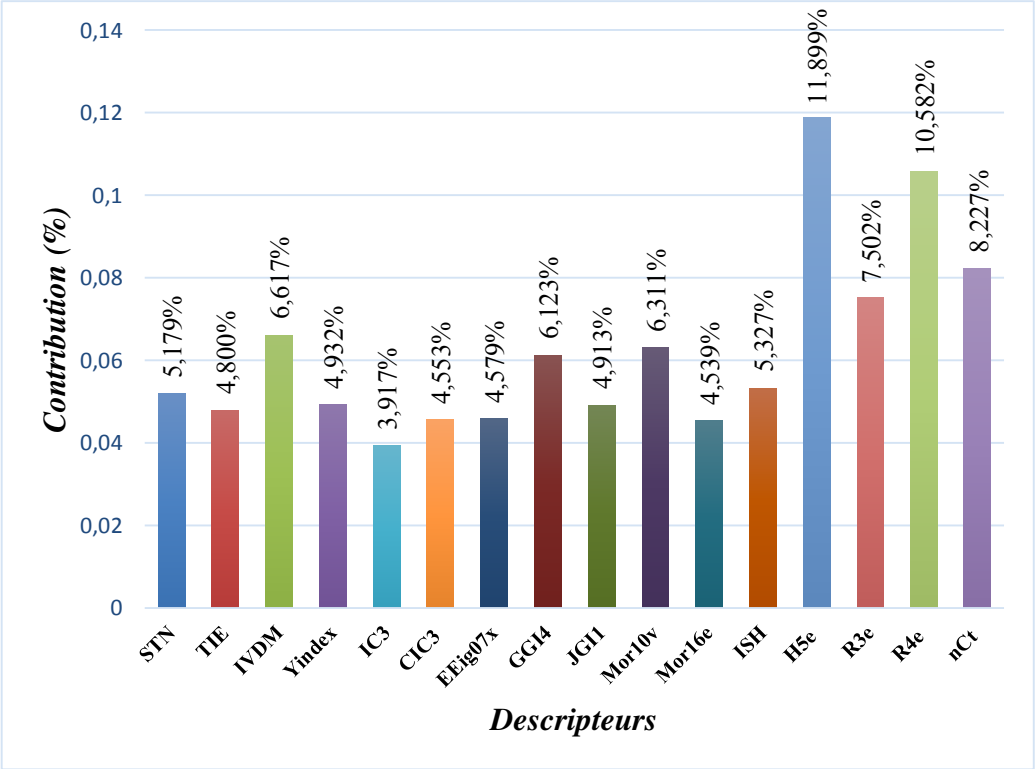


Figure V.7. Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la température du point d'ébullition normal Tb.

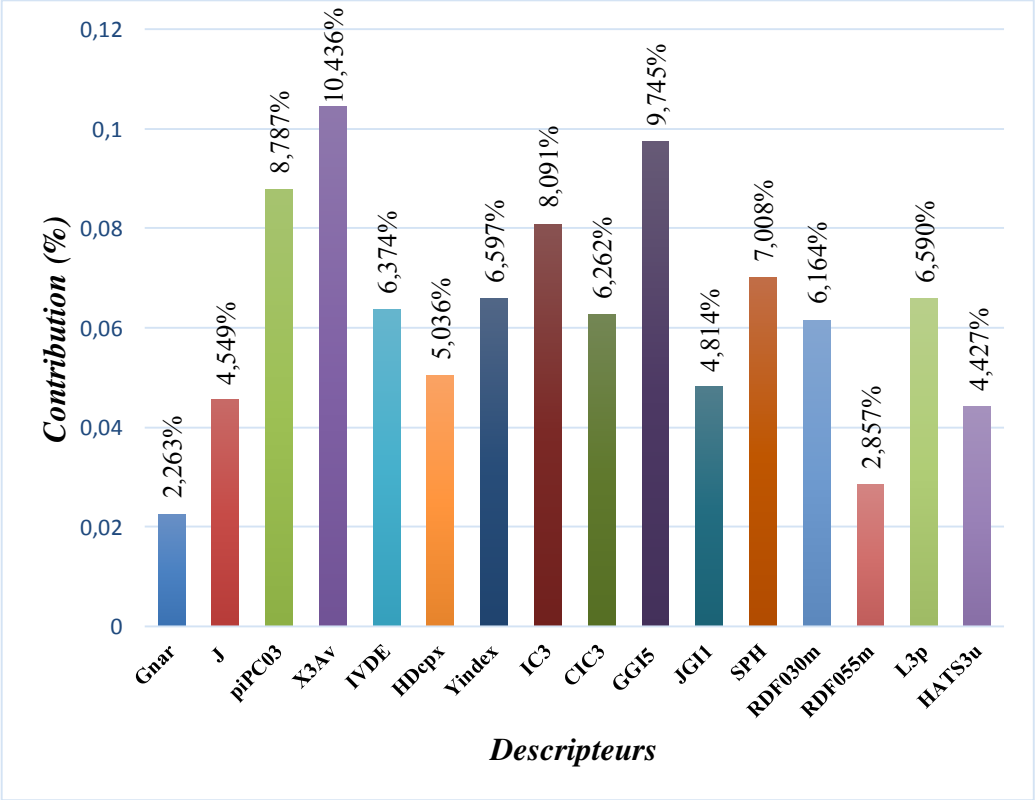


Figure V.8. Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la densité relative du liquide d20.

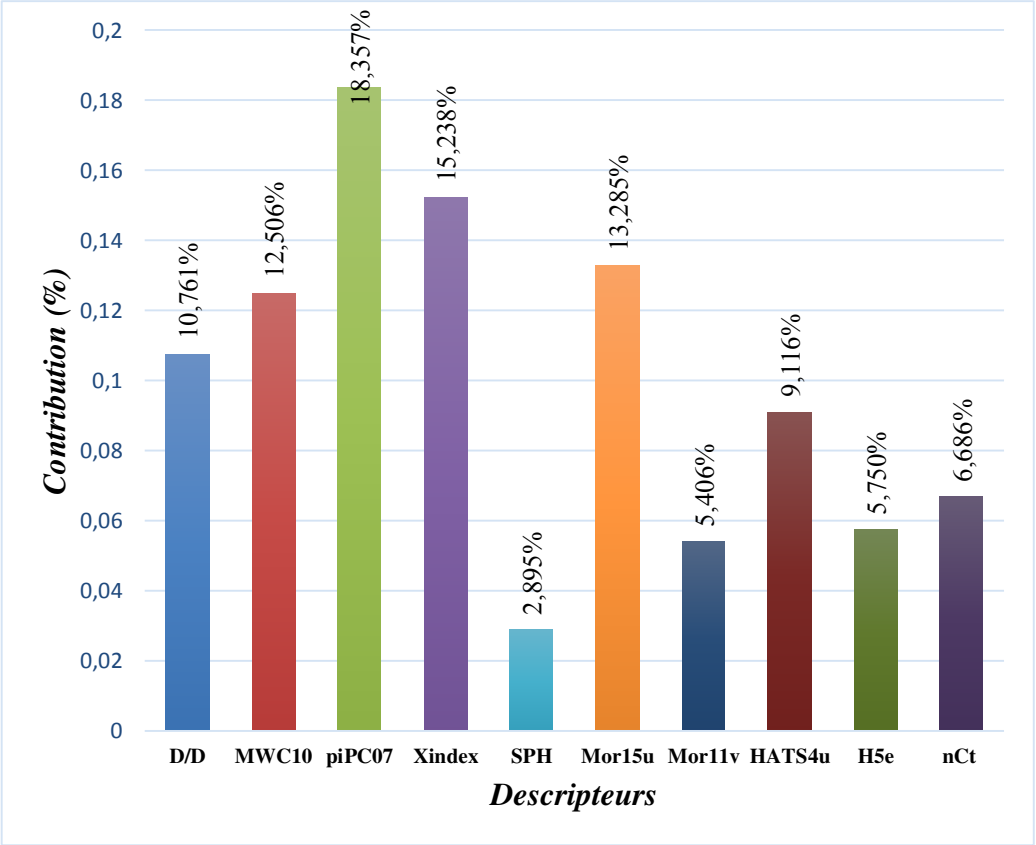


Figure V.9. Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la température critique Tc.

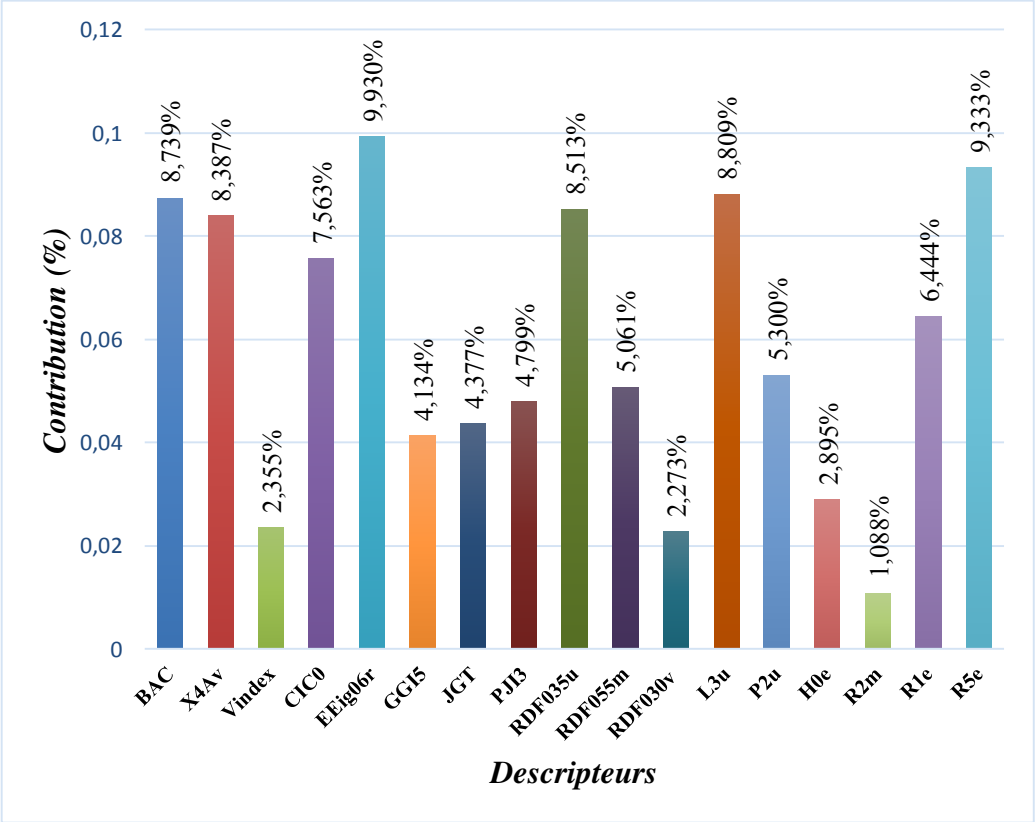


Figure V.10. Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA de la pression critique Pc.

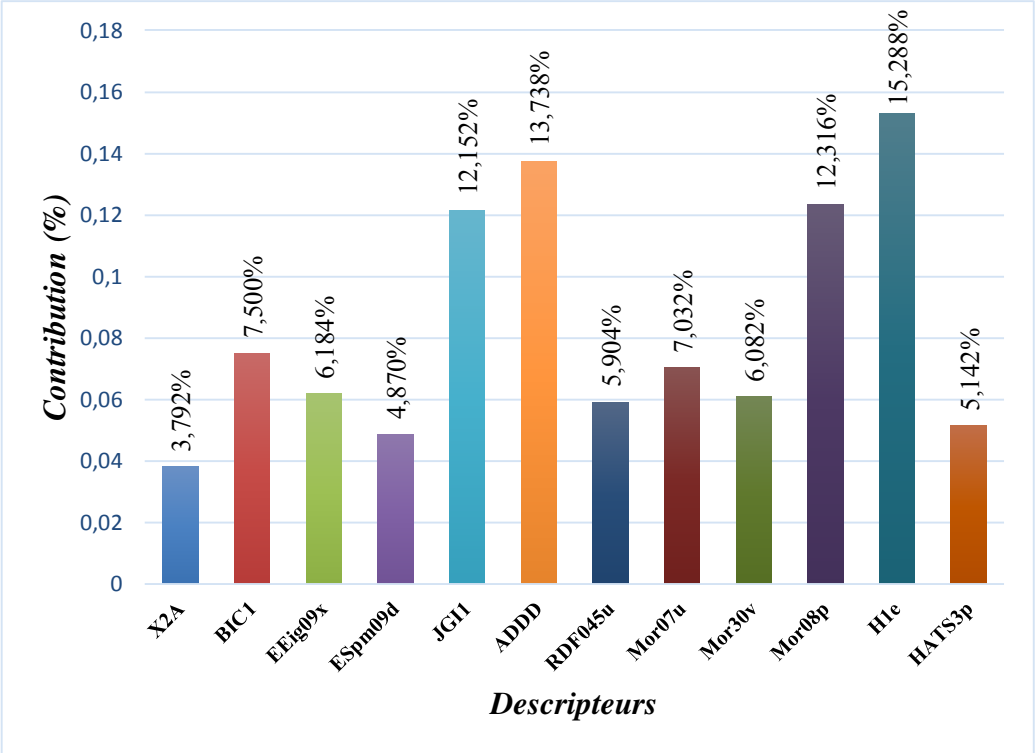


Figure V.11. Contributions relatives des descripteurs pertinents d'entrée pour le modèle de RQSP-RNA du volume critique Vc.

IV.7.3.1. Interprétation et classification des contributions selon la (VCM) et selon chaque groupe

Selon les résultats de l'analyse de sensibilité et les valeurs des contributions moyennes (VCM) qui sont calculées par la division de la contribution totale de chaque modèle (100 %) sur le nombre des descripteurs dans le même modèle concerné. Les VCM calculées sont de 6,25% pour les modèles de Tb et de d20 ; 10% pour le modèle de Tc ; 5,88% pour le modèle de Pc ; et 8,33% pour le modèle de Vc. Il en découle le classement suivant de la contribution des descripteurs :

- En trois groupes pour le modèle de Tb : **(1)** contribution élevée, allant de 7 à 12 % $\{H5e(11{,}899\%) > R4e(10{,}582\%) > nCt(8{,}227\%) > R3e(7{,}502\%)\}$, **(2)** contribution moyenne, allant de 6 à 7% $\{IVDM(6{,}617\%) > Mor10v(6{,}311\%) > GGI4(6{,}123\%)\}$ et **(3)** faible contribution, allant de 4 à 6% $\{ISH(5{,}327\%) > STN(5{,}179\%) > Yindex(4{,}932\%) > JGI1(4{,}913\%) > TIE(4{,}800\%) > EEig07x(4{,}579\%) > CIC3(4{,}553\%) > Mor16e(4{,}539\%) > IC3(3{,}917\%)\}$.
- En quatre groupes pour le modèle d20 : **(1)** contribution élevée, allant de 8 à 11% $\{X3Av(10{,}436\%) > GGI5(9{,}745\%) > piPC03(8{,}787\%) > IC3(8{,}091\%)\}$, **(2)** contribution moyenne, allant de 6 à 7% $\{SPH(7{,}008\%) > Yindex(6{,}597\%) >$

$L3p(6{,}590\%) > IVDE(6{,}374\%) > CIC3(6{,}262\%) > RDF030m(6{,}164\%)\}$, **(3)** faible contribution, allant de 4 à 5% $\{HDcpx(5{,}036\%) > JGI1(4{,}814\%) > J(4{,}549\%) > HATS3u(4{,}427\%)\}$, et **(4)** très faible contribution, moins de 3% $\{RDF055m(2{,}857\%) > GNar(2{,}263\%)\}$.

- En cinq groupes pour le modèle de Tc : **(1)** contribution très élevée, plus de 14 % $\{piPC07(18{,}357\%) > Xindex(15{,}238\%)\}$, **(2)** contribution élevée, allant de 12 à 14% $\{Mor15u(13{,}285\%) > MWC10(12{,}506\%)\}$, **(3)** contribution moyenne, allant de 8 à 12% $\{D/D(10{,}761\%) > HATS4u(9{,}116\%)\}$, **(4)** faible contribution, allant de 3 à 8% $\{nCt(6{,}686\%) > H5e(5{,}750\%) > Mor11v(5{,}406\%)\}$, et **(5)** très faible contribution, moins de 3% $\{SPH(2{,}895\%)\}$.
- En quatre groupes pour le modèle de Pc : **(1)** contribution très élevée, plus de 9 % $\{EEig06r(9{,}930\%) > R5e(9{,}333\%)\}$, **(2)** contribution élevée, allant de 7 à 9 % $\{L3u(8{,}809\%) > BAC(8{,}739\%) > RDF035u(8{,}513\%) > X4Av(8{,}387\%) > CIC0(7{,}563\%)\}$, **(3)** contribution moyenne, allant de 4 à 7% $\{R1e(6{,}444\%) > P2u(5{,}300\%) > RDF055m(5{,}061\%) > PJI3(4{,}799\%) > JGT(4{,}377\%) > GGI5(4{,}134\%)\}$, **(4)** faible contribution, moins de 4% $\{H0e(2{,}895\%) > Vindex(2{,}355\%) > RDF030v(2{,}273\%) > R2m(1{,}088\%)\}$.
- Et enfin en quatre groupes pour le modèle de Vc: **(1)** contribution très élevée, plus de 13 % $\{H1e(15{,}288\%) > ADDD(13{,}738\%)\}$, **(2)** contribution élevée, allant de 10 à 13% $\{Mor08p(12{,}316\%) > JGI1(12{,}152\%)\}$, **(3)** contribution moyenne, allant de 5 à 10% $\{BIC1(7{,}500\%) > Mor07u(7{,}032\%) > EEig09x(6{,}184\%) > Mor30v(6{,}082\%) > RDF045u(5{,}904\%) > HATS3p(5{,}142\%)\}$, **(4)** faible contribution, moins de 5% $\{ESpm09d(4{,}870\%) > X2A(3{,}792\%)\}$.

A l'exception de quelques descripteurs qui appartiennent aux groupes qui ont de faibles à de très faibles contributions, la grande majorité des descripteurs dans les cinq modèles appartiennent aux groupes qui possèdent au moins une moyenne contribution. Il convient de noter qu'il n'y a pas une différence significative entre les descripteurs, en particulier ceux du même groupe de classification sus-cités. Ce qui signifie que tous les descripteurs sélectionnés étaient importants à des degrés divers dans le développement des modèles de RQSP-RNA selon la classification du groupe auquel ils appartiennent.

Selon l'analyse de sensibilité, il est évident que les classes de descripteurs les plus significatives qui ont influencé le modèle de Tb sont les suivantes (Voir la **Figure V.12**) : la

classe "*GETAWAY descriptors*" avec $RI = 35{,}310\%$ sur 4 descripteurs, la classe "*information indices*" avec $RI = 20{,}019\%$ également sur 4 descripteurs, ensuite les classe de : "*topological charge indices*", "*3D-MoRSE descriptors*" et "*topological descriptors*" avec $RI = 11{,}036\%;\ 10{,}850\%$ et $9{,}979\%$; respectivement, avec 2 descripteurs pour chaque classe, et dans la gamme finale avec seulement 1 descripteur pour chaque classe, on trouve les classes de : "*functional group counts*" et "*edge adjacency indices*" avec $RI = 8{,}227\%$ et $4{,}579\%$, respectivement.

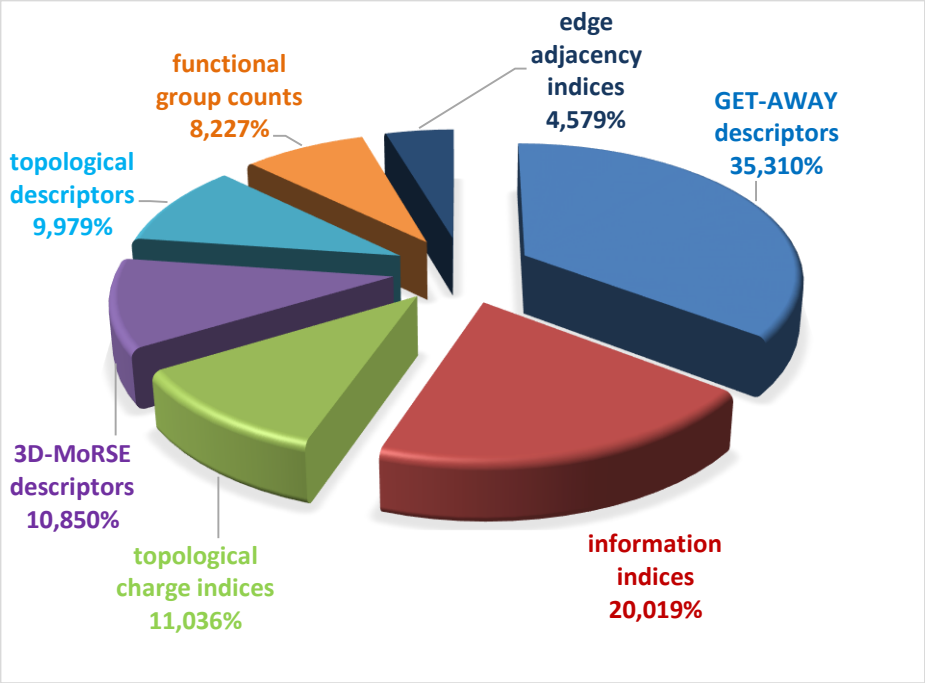


Figure V.12. Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la température du point d'ébullition normal Tb.

En ce qui concerne le modèle de d20, les classes de descripteurs significatives les plus influencées sont les suivantes (Voir la **Figure V.13**) : la classe "*information indices*" avec $RI = 32{,}360\%$ sur 5 descripteurs ; la classe "*topological charge indices*" avec $RI = 14{,}559\%$ sur 2 descripteurs ; la classe "*connectivity indices*" avec $RI = 10{,}436\%$ uniquement sur 1 descripteur ; la classe "*RDF descriptors*" avec $RI = 9{,}021\%$ sur 2 descripteurs ; les classes de : "*walk and path counts*" et "*geometrical descriptors*" avec $RI = 8{,}787\%$ et $RI = 7{,}008\%$ respectivement, sur 1 descripteur pour chaque classe ; la classe "*topological descriptors*" avec $RI = 6{,}812\%$ sur 2 descripteurs ; la classe "*WHIM descriptors*" avec $RI = 6{,}590\%$ sur 1 descripteur ; et enfin la classe "*GETAWAY descriptors*" avec $RI = 4{,}427\%$ sur 1 descripteur également.

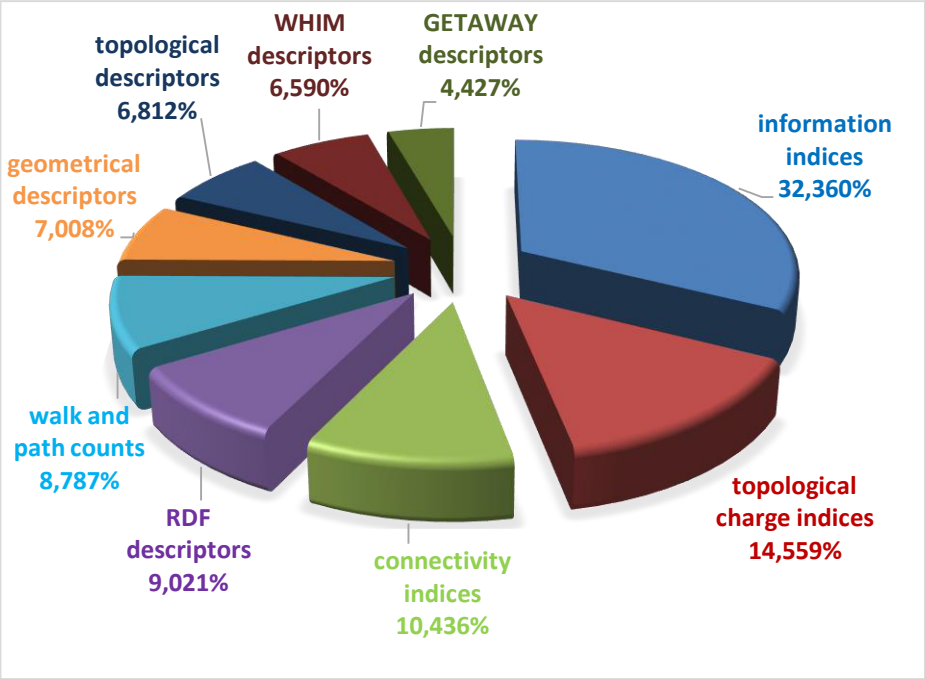


Figure V.13. Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la densité relative du liquide d20.

Concernant le modèle de Tc, les classes de descripteurs significatives les plus influencées sont les suivantes (Voir la **Figure V.14**) : la classe "*walk and path counts*" avec $RI = 30{,}863\%$ sur 2 descripteurs ; la classe "*3D-MoRSE descriptors*" avec $RI = 18{,}691\%$ aussi sur 2 descripteurs ; la classe "*information indices*" avec $RI = 15{,}238\%$ seulement sur 1 descripteur ; la classe de "*GETAWAY descriptors*" avec $RI = 14{,}866\%$ sur 2 descripteurs ; et enfin les classes de : "*topological descriptors*", "*functional group counts*" et "*geometrical descriptors*" avec $RI = 10{,}761\%, RI = 6{,}686\%$ et $RI = 2{,}895\%$ respectivement, sur seulement 1 descripteur pour chaque classe.

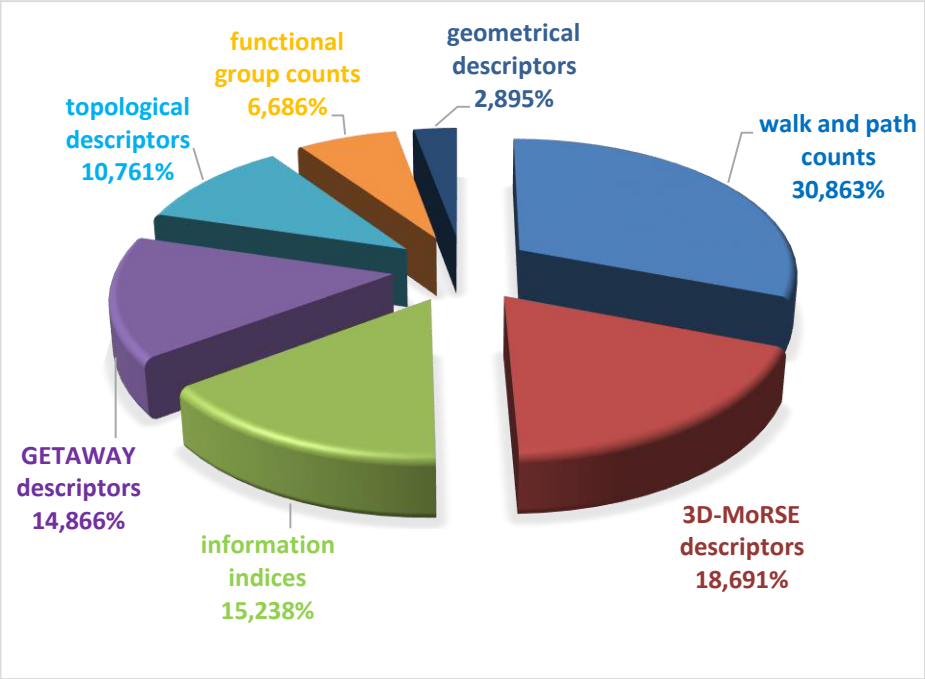


Figure V.14. Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la température critique Tc.

Pour le modèle de Pc, les classes de descripteurs significatifs les plus influentes sont les suivantes (Voir la **Figure V.15**) : la classe "*GETAWAY descriptors*" avec $RI = 19{,}760\%$ sur 4 descripteurs ; la classe "*RDF descriptors*" avec $RI = 15{,}847\%$ sur 3 descripteurs ; la classe "*WHIM descriptors*" avec $RI = 14{,}109\%$ sur 2 descripteurs ; la classe "*edge adjacency indices*" avec $RI = 9{,}930\%$ seulement sur un 1 descripteur ; la classe "*information indices*" avec $RI = 9{,}918\%$ sur 2 descripteurs ; et enfin les classes de : "*topological descriptors*", "*connectivity indices*", "*geometrical descriptors*", "*2D autocorrelations*", et "*topological charge indices*" respectivement, avec $RI = 8{,}739\%$; $RI = 8{,}387\%$; $RI = 4{,}799\%$; $RI = 4{,}377\%$ et $RI = 4{,}134\%$, seulement sur 1 descripteur pour chaque classe.

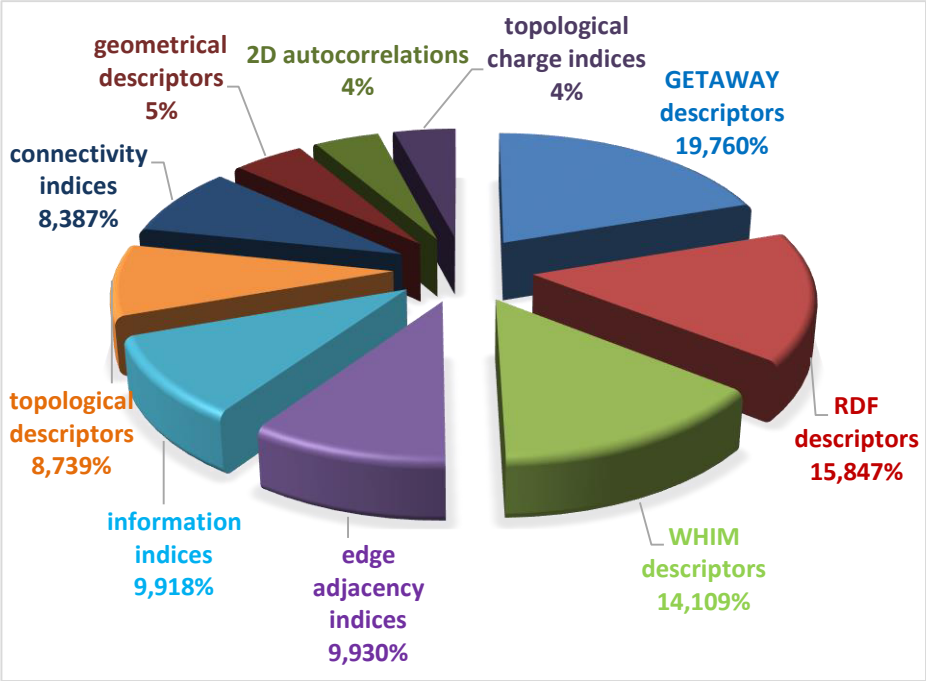


Figure V.15. Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA de la pression critique Pc.

Et enfin, concernant le modèle de Vc, les classes de descripteurs significatifs les plus influentes sont les suivantes (Voir la **Figure V.16**) : la classe "*3D-MoRSE descriptors*" avec $RI = 25{,}430\%$ sur 3 descripteurs ; la classe "*GETAWAY descriptors*" avec $RI = 20{,}430\%$ sur 2 descripteurs ; la classe "*geometrical descriptors*" avec $RI = 13{,}738\%$ seulement sur 1 descripteur ; la classe "*topological charge indices*" avec $RI = 12{,}152\%$ également sur 1 descripteur ; la classe "*edge adjacency indices*" avec $RI = 11{,}054\%$ sur 2 descripteurs ; et enfin les classes : "*information indices*", "*RDF descriptors*", et "*connectivity indices*" respectivement, avec $RI = 7{,}500\%$; $RI = 5{,}904\%$; et $RI = 3{,}792\%$, seulement sur 1 descripteur pour chaque classe.

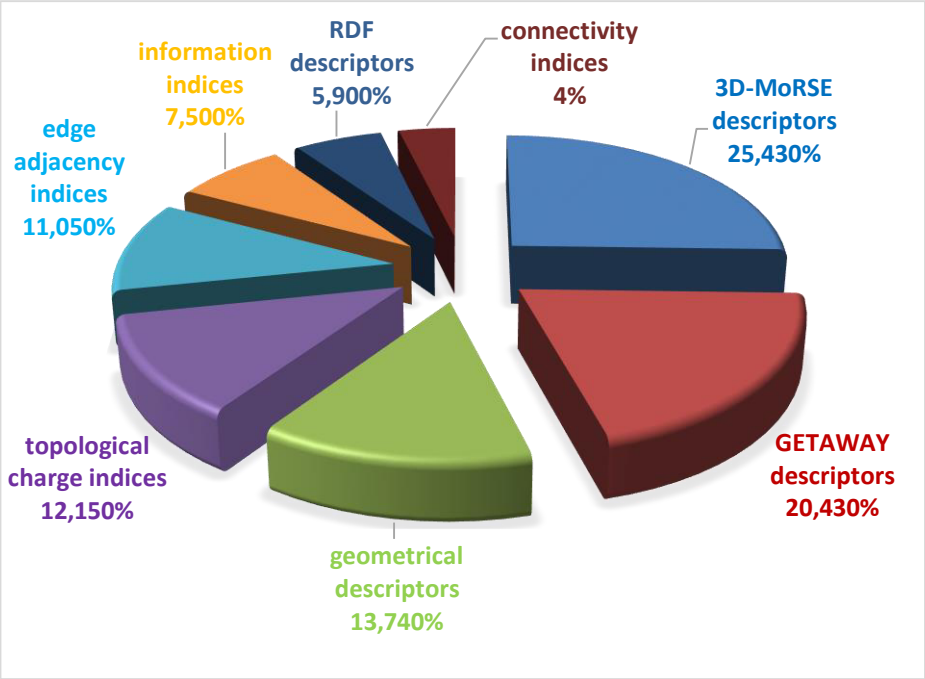


Figure V.16. Contributions relatives des classes de descripteurs pertinents selon leurs importances pour le modèle RQSP-RNA du volume critique Vc.

V.7.4. Domaine d'applicabilité pour la validation des modèles RQSP-RNA obtenus

Les domaines d'applicabilité (DA) de chaque modèle RQSP-RNA développé, ont été montrées dans le diagramme de Williams ; sur les figures : **Figure V.17**, **Figure V.18**, **Figure V.19**, **Figure V.20** et la **Figure V.21**, respectivement pour les modèles de Tb, d20, Tc, Pc et Vc. Les domaines d'applicabilité (DA) de chaque modèle sont établis dans une zone carrée comprise entre les valeurs de levier standard (***h****) de chaque modèle RQSP-RNA et une valeur de résidus standardisée égale à ± 2. Elles sont considérées comme assez strictes, par rapport au reste des calculs dans les autres études de littératures, de sorte que la plupart des valeurs de résidus standardisées sont limitées à ± 3 **(Jaworska *et al*. 2005)**.

La valeur de ***h**** du modèle RQSP-RNA de Tb est égale à 0,2374 (**Figure V.17**), ce qui signifie qu'il n'y a que 17 composés sur 223 en dehors du domaine d'applicabilité. En d'autres termes, 92% des composés au sein du DA, c'est-à-dire seulement 12 parmi les 179 composés d'ensemble de base d'apprentissage, et 5 parmi les 44 composées d'ensemble de base de test sont hors le domaine d'applicabilité. Comme il apparaît dans la **Figure V.17**, et selon leur ordre de division dans le **tableau A.1** (***Annexe A***) ; les points aberrants du domaine d'applicabilité sont : les 1er, 2ème, 23ème, 64ème, 109ème, 111ème, 112ème, 130ème, 131ème, 132ème, 164ème et la 175ème pour les composés dans l'ensemble de base de données d'apprentissage (●), et les 1er, 7ème, 8ème, 17ème et la 20ème pour les composés de l'ensemble de base de données de test (◊). Par conséquent,

plus de 93% et 88% des composés se situent dans la zone du DA pour l'ensemble de données de d'apprentissage et de test, respectivement.

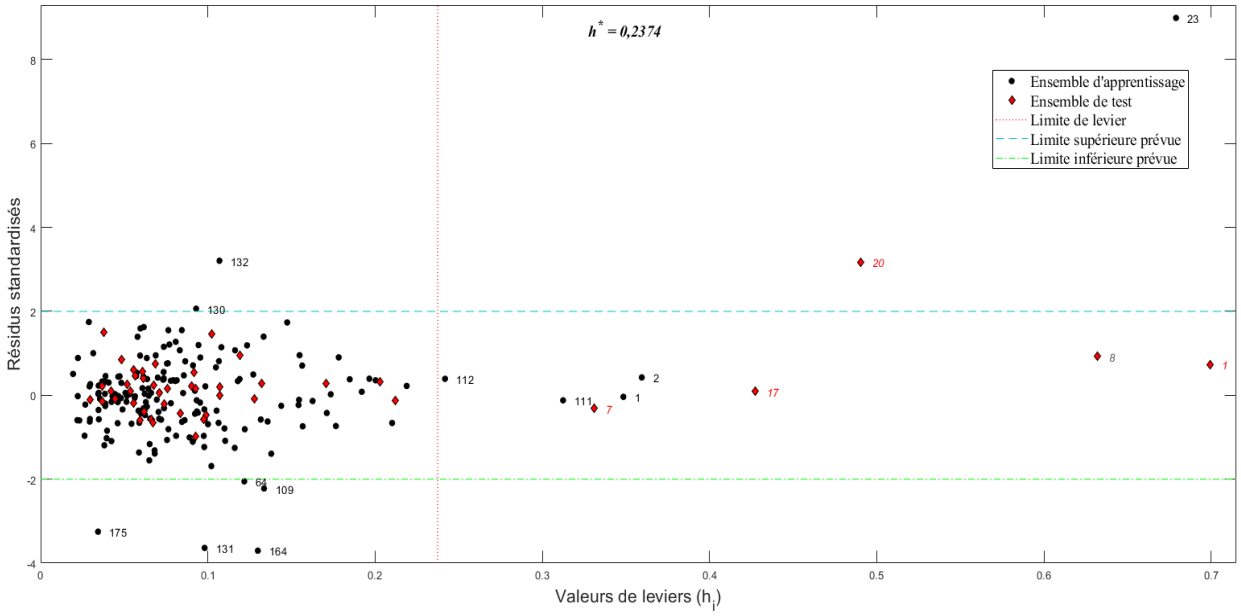


Figure V.17. Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la température du point d'ébullition normal Tb.

Pour le modèle de RQSP-RNA du paramètre d20 (**Figure V.18**), la valeur de ***h**** est égale à 0,2388, ceci signifie que seulement 22 composés sur 222 sont considérés comme hors du domaine d'applicabilité. Donc environ 90% des composés sont dans le domaine d'applicabilité ; en d'autres termes 17 composés parmi les 178 de l'ensemble de base des données d'apprentissage, uniquement 5 composés parmi les 44 de l'ensemble de la base des données sont hors du domaine d'applicabilité. Conformément à l'ordre de division indiqué dans le **tableau A.1** (***Annexe A***) et comme il apparait dans la **Figure V.18**, les points aberrants du domaine d'applicabilité sont : les 1er, 2ème, 26ème, 64ème, 98ème, 99ème, 111ème, 112ème, 120ème, 136ème, 142ème, 157ème, 162ème, 169ème, 171ème, 173ème et la 177ème pour l'ensemble de la base de données d'apprentissage (●) et les 1er, 2ème, 8ème, 17ème et 20ème pour les composés de l'ensemble de la base de données de test (■). En conséquence, environ 93% et plus de 88% des composés sont considérés dans le DA, respectivement pour l'ensemble de la base de données d'apprentissage et de test.

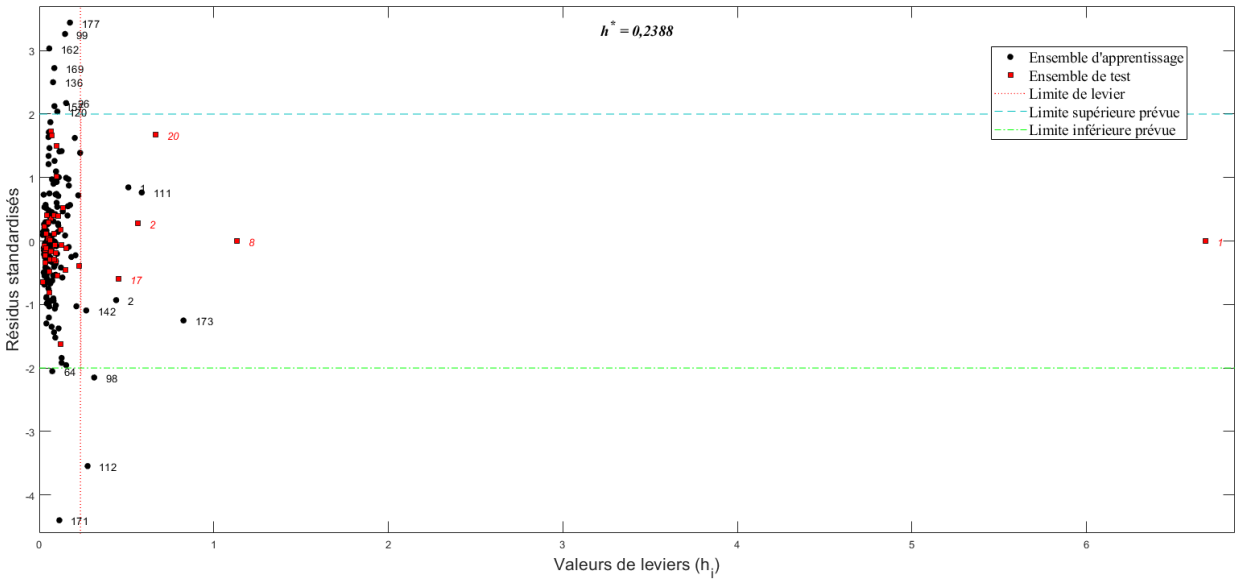


Figure V.18. Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la densité relative de liquide d20.

Concernant le modèle de RQSP-RNA du paramètre Tc (**Figure V.19**), la valeur de ***h**** est égale à 0,1536 ce qui désigne que seulement 26 composés sur 223 sont observés comme hors du domaine d'applicabilité. De ce fait, presque 88,34% des composés dans la base de données globale sont dans le domaine d'applicabilité ; d'autre part 14 composés parmi les 179 de l'ensemble de la base de données d'apprentissage, et 12 composés parmi les 44 de l'ensemble de la base de données sont hors du domaine d'applicabilité. Conformément à l'ordre de division des composés désignés dans le **tableau A.1** (***Annexe A***) et ses apparences dans la **Figure V.19**, les points aberrants du domaine d'applicabilité sont : les 1er, 2ème, 23ème, 39ème, 62ème, 69ème, 99ème, 111ème, 115ème, 133ème, 161ème, 164ème, 174ème et la 178ème pour l'ensemble de la base des données d'apprentissage (●) et les 1er, 7ème, 8ème, 17ème, 20ème, 22ème, 24ème, 26ème, 27ème, 28ème, 29ème et la 42ème pour les composés de l'ensemble de la base des données de test (▲). Par conséquence, presque 92% et plus de 72% des composés sont considérés dans le DA, respectivement pour l'ensemble de la base des données d'apprentissage et de test.

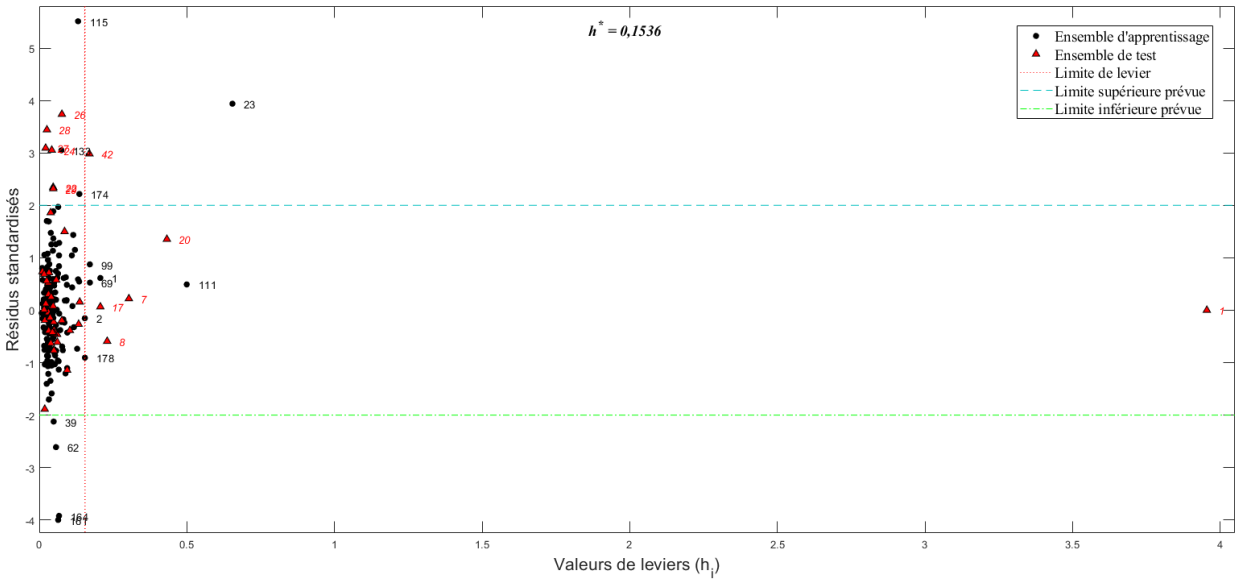


Figure V.19. Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la température critique Tc.

Concernant le modèle de RQSP-RNA du paramètre Pc (**Figure V.20**), la valeur de ***h**** est égale à 0,2542 ce qui désigne qu'uniquement 28 composés sur 221 sont observés comme hors du domaine d'applicabilité. De ce fait, 87,33% des composés dans la base de données globale sont dans le domaine d'applicabilité ; d'autre part 16 composés parmi les 177 de l'ensemble de la base de données d'apprentissage, et 12 composés parmi les 44 de l'ensemble de la base des données sont hors du domaine d'applicabilité. Conformément à l'ordre de division des composés désignés dans le **tableau A.1** (***Annexe A***) et ses apparences dans la **Figure V.20**, les points aberrants du domaine d'applicabilité sont : les 23ème, 54ème, 97ème, 105ème, 110ème, 115ème, 136ème, 141ème, 150ème, 158ème, 160ème, 163ème, 166ème, 168ème, 173ème et la 176ème pour l'ensemble de la base de données d'apprentissage (●) et les 1er, 7ème, 8ème, 22ème, 24ème, 26ème, 27ème, 32ème, 33ème, 34ème, 42ème et la 43ème pour les composés de l'ensemble de la base de données de test (▼). Par conséquence, plus de 90% et environ 72% des composées sont considérés dans le DA, respectivement pour l'ensemble de la base de données d'apprentissage et de test.

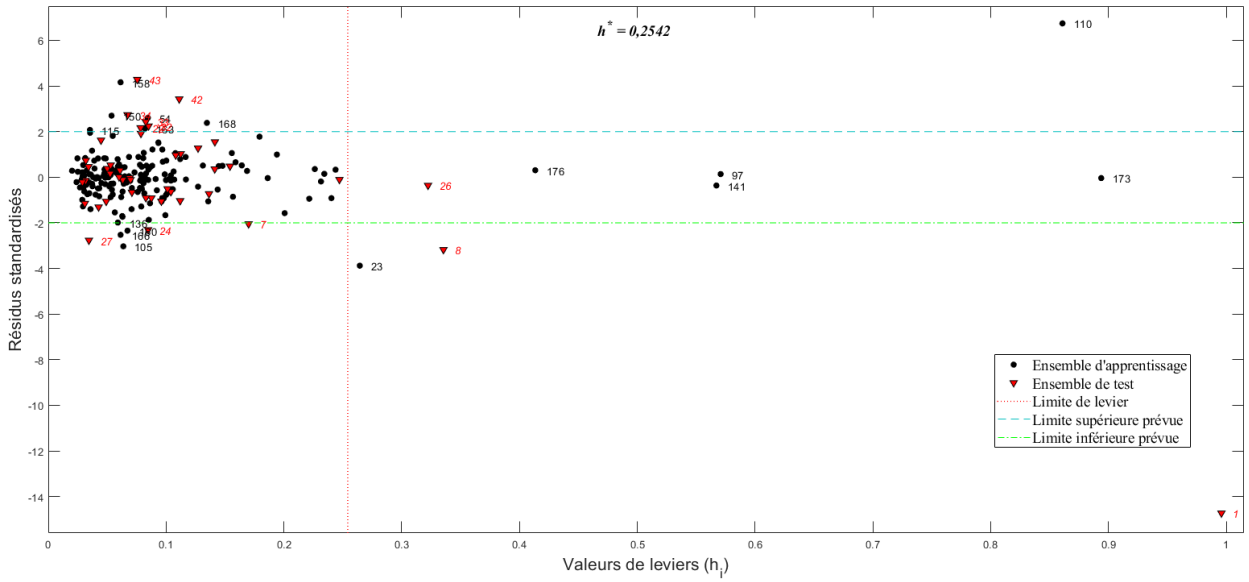


Figure V.20. Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA de la pression critique Pc.

Concernant le modèle de RQSP-RNA du paramètre Pc (**Figure V.21**), la valeur de ***h**** est égale à 0,1816, ce qui signifie qu'uniquement 26 composés sur 223 sont observés comme hors le domaine d'applicabilité. De ce fait, plus de 88% des composés dans la base de données globale sont dans le domaine d'applicabilité ; d'autre part 16 composés parmi les 179 de l'ensemble de la base de données d'apprentissage, et 10 composés parmi les 44 de l'ensemble de la base de données sont hors du domaine d'applicabilité. Conformément à l'ordre de division des composés désignés dans le **tableau A.1** (***Annexe A***) et ses apparences dans la **Figure V.21**, les points aberrants du domaine d'applicabilité sont : les 1^{ére}, 9ème, 23ème, 106ème, 115ème, 116ème, 139ème, 140ème, 164ème, 166ème, 167ème, 172ème, 173ème, 174ème, 177ème et la 178ème pour l'ensemble de la base de données d'apprentissage (●) et les 1er, 7ème, 8ème, 17ème, 28ème, 29ème, 35ème, 40ème, 41ème et la 44ème pour les composés de l'ensemble de la base de données de test (◀). Par conséquence, environ 91% et plus de 77% des composés sont considérés dans le DA, respectivement pour l'ensemble de la base de données d'apprentissage et de test.

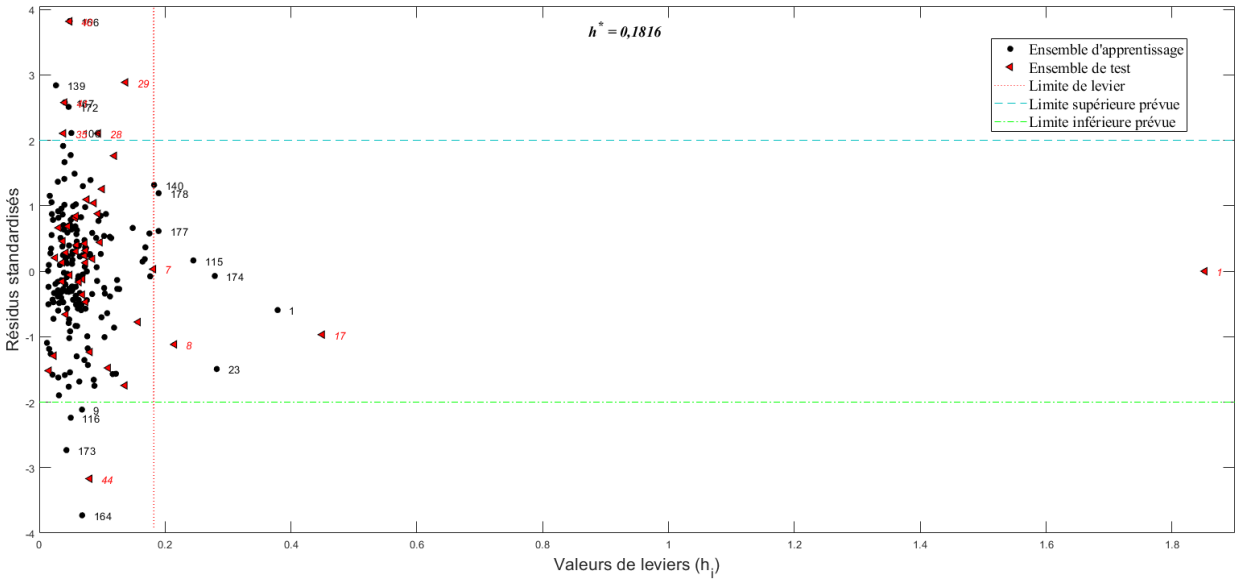


Figure V.21. Diagramme de Williams décrivant le Domaine d'Applicabilité (DA) pour le modèle RQSP-RNA du volume critique Vc.

Ceci étant quelle que soit la valeur résiduelle standard (± 2) prise, ce qui confirme une nouvelle fois l'efficacité des modèles RQSP-RNA développés dans cette étude. Les noms des composés indiqués par ordre de classement dans les paragraphes ci-dessus sont montrés en détails dans le **tableau A.1** (***Annexe A***).

V.7.5. Comparaisons des performances des modèles de type RQSP-RNA obtenus avec d'autres modèles dans la littérature

Dans le paragraphe V.7.1 de ce chapitre, j'ai abordé seulement les résultats obtenus lors de la modélisation des modèles RQSP-RNA pour les cinq propriétés physico-chimiques (Tb, d20, Tc, Pc et Vc) des hydrocarbures, néanmoins je n'ai pas fait une comparaison de la performance entre ces modèles et d'autres modèles dans la littérature. De ce fait dans cette partie, j'ai pensé à faire une comparaison entre ces cinq modèles de type RQSP-RNA et d'autres modèles dans la littérature.

Pour plus d'appréciation de mes modèles de type RQSP-RNA développés dans cette thèse, et afin de connaître l'importance de leurs résultats obtenus, j'ai essayé de trouver des études analogues présentant une grande similarité pour faire une comparaison systématique. Les résultats des comparaisons de chaque modèle de type RQSP-RNA sont montrés en détails dans le **tableau V.10** pour le modèle du paramètre Tb, dans le **tableau V.11** pour le modèle du paramètre d20, dans le **tableau V.12** pour le modèle de paramètre Tc, dans le **tableau V.13** pour

le modèle de paramètre Pc et dans le **tableau V.14** pour le modèle du paramètre Vc. J'ai choisi des études presque différentes qui ont des méthodes communes avec la présente étude (RQSP, RNA, etc.). Malgré cela, il est bien connu que l'évaluation de leurs forces et leurs faiblesses est assez difficile, compte tenu de la spécificité de chaque étude, par exemple la qualité et la quantité de l'ensemble de la base de données ainsi que les descripteurs calculés, les approches de la modélisation utilisée, et les différentes stratégies de validation adaptées, etc.

Bien que mes modèles incluent toutes les familles des hydrocarbures, ils ont surpassé la majorité des modèles mentionnés dans cette comparaison, et même dépassé les études traitant chaque famille individuellement, comme dans l'étude de Ha et ses co-auteurs **(Ha *et al.* 2005)** pour les deux paramètres de Tb et d20, dans l'étude de Zuas et Styarini **(Zuas & Styarini 2009)** pour le paramètre Vc et dans l'étude de Dai et ses co-auteurs **(Dai *et al.* 2013)** pour le paramètre Tb. Ils ont aussi dépassé ceux qui ont presque pris toutes les familles des hydrocarbures ; comme dans l'étude de Wessel et Jurs **(Wessel & Jurs 1995)** pour le paramètre Tb, et dans l'étude de Saaidpour et Ghaderi **(Saaidpour & Ghaderi 2016)** pour le même paramètre Tb, ainsi que dans l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)** pour les paramètres d20, Pc et Vc, puis à l'exception du modèle du paramètre Tc dans la même étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)**, qui a dépassé mon modèle d'une légère différence en termes d'ERAM.

Comme prévu, la précision de mes modèles du type RQSP-RNA pour les cinq paramètres étudiés est meilleure conformément aux majorités des études qui ont utilisé différentes familles organiques ; y compris les hydrocarbures, telles que l'étude de Sola et ses co-auteurs **(Sola *et al.* 2008)** pour les paramètres Tb, Tc et Pc, et aussi dans les études de Gharagheizi et ses co-auteurs **(Gharagheizi *et al*. 2013)** et de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017a)** pour le paramètre Tb. Ainsi que dans les deux études de Turner et ses co-auteurs **(Turner *et al*. 1998)** et de Banchero et Manna **(Banchero et Manna 2018)** pour les paramètres Tc et Pc, de même dans les études de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)** et de Godavarthy et ses co-auteurs **(Godavarthy *et al*. 2008)** pour les propriétés critiques Tc, Pc et Vc, également dans les deux études de Yao et ses co-auteurs **(Yao *et al*. 2002a ; Yao *et al*. 2002b)**, respectivement pour les deux paramètres Tc et Pc. Ils ont même dépassé cinq sur neuf modèles mentionnés dans l'étude de Espinosa et ses co-auteurs **(Espinosa *et al*. 2001)** pour les trois paramètres Tb, Tc et Pc, bien que les quatre modèles dans la même étude de Espinosa et ses co-auteurs **(Espinosa *et al*. 2001)** aient dépassé mes modèles, mais ils restent peu convaincants, compte tenu de la différence remarquable entre les résultats issus de l'ensemble de la base entière et la base d'apprentissage. En ce qui concerne l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al*.**

2011), mes modèles ont été surpassés. Quatre modèles sur six mentionnés dans cette étude, précisément deux modèles pour le paramètre de Tc, et un modèle pour chaque paramètre Pc et Vc, regardant les résultats des deux modèles restant pour les deux derniers paramètres de la même étude, ils ont dépassé nos modèles par un léger pourcentage en termes des valeurs des ERAM.

Comme déjà mentionné dans le paragraphe précédent, la précision de mes modèles est meilleure par rapport à la majorité des études qui ont utilisé différentes familles organiques, car lorsqu'on travaille largement ou bien sur une large gamme, la valeur d'erreur augmente, en revanche pour les familles chimiques. Travailler avec les composés les plus proches conduits à des erreurs moins importantes. Cela indique que tous les modèles de type RQSP-RNA présentés dans cette thèse dépassent la majorité des études disponibles et donnent les prédictions les plus précises, malgré la diversité des méthodes de modélisation et les différences dans les familles des composés chimiques, dans cette comparaison, mes modèles ont montré une bonne précision dans la plupart des cas.

Le **tableau V.10** révèle que le modèle de type RQSP-RNA de Tb proposé dans cette thèse, a une valeur de REQM plus de dix fois moins comparé à la méthode de "QSPR-FF-ANN" dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2013)** ainsi que les méthodes de "MLP-ANN", "LSSVM" et de "GMDH-NN" dans l'étude de Varamesh et ses co-auteurs, huit fois moins pour la méthode de "RBF-ANN" dans la même étude **(Varamesh *et al.* 2017a)**, presque les mêmes proportions avec les valeurs de l'ERAM dans les deux études **(Gharagheizi *et al.* 2013 ; Varamesh *et al.* 2017a)**. Le modèle de type RQSP-RNA de Tb proposé dans cette thèse a aussi une valeur de REQM plus de cinq fois moins par rapport à la méthode de "QSPR MLR" dans l'étude de Sola et ses co-auteurs **(Sola *et al.* 2008)** pour l'ensemble d'apprentissage et plus de quatre fois moins pour l'ensemble de test dans la même étude **(Sola *et al.* 2008)**, également plus de sept fois moins toujours en termes de REQM par rapport à la même méthode "QSPR MLR" dans l'étude de Saaidpour et Ghaderi **(Saaidpour & Ghaderi 2016)** pour l'ensemble d'apprentissage et plus de deux fois moins pour l'ensemble de test dans la même étude **(Saaidpour & Ghaderi 2016)**. Concernant les deux modèles mentionnés dans l'étude de Wessel et Jurs **(Wessel & Jurs 1995)**, les résultats de mon modèle en termes de REQM, sont plus de dix fois moins pour l'ensemble d'apprentissage et presque de trois fois moins pour l'ensemble de test par rapport à la méthode de "QSPR MLR", également plus de quatre fois moins par rapport à la méthode de "QSPR CNN" pour les deux ensembles d'apprentissage et de test dans la même étude de Wessel et Jurs **(Wessel & Jurs 1995)**. A propos de l'étude de

Espinosa et ses co-auteurs **(Espinosa *et al.* 2001)**, les résultats de notre modèle de type RQSP-RNA pour le paramètre Tb a dépassé les trois modèles, de sorte qu'il a une valeur d'ERAM plus d'une fois et demi moins par rapport au premier modèle de la méthode "QSPR Fuzzy-ARTMAP" pour l'ensemble entier des données, et approximativement trois fois moins pour le deuxième modèle de la même méthode dans la même étude **(Espinosa *et al.* 2001)** pour l'ensemble entier des données, bien que les deux derniers modèles de la méthode "QSPR Fuzzy-ARTMAP" ont des résultats assez bons, mais ils ne sont pas convaincants à cause de l'instabilité due à la différence entre les résultats des trois ensembles de données ; d'apprentissage, de test et le résultat de la base entière. Toujours en terme d'ERAM, le résultat de notre modèle est plus de vingt fois moins par rapport au modèle de la méthode "QSPR bpNN" pour l'ensemble entier des données dans la même étude **(Espinosa *et al*. 2001)**.

D'autres modèles comme les trois modèles de la méthode "QSPR MLR" dans l'étude de Ha et ses co-auteurs **(Ha *et al.* 2005)** et les deux modèles de même méthode "QSPR MLR" dans l'étude de Dai et ses co-auteurs **(Dai *et al.* 2013)** pour le paramètre de Tb, n'ont pas calculé les mêmes types d'erreur comme dans la présente thèse. J'ai donc évalué en fonction de leurs valeurs R^2, la comparaison a montré que ses valeurs ne dépassaient pas nos modèles qui ont des valeurs de 0,9995 pour l'ensemble d'apprentissage, de 0,9999 pour l'ensemble de test et de 0,9996 pour l'ensemble de base de données entière. Ce dernier indique la valeur réelle de R2 du modèle de type RQSP-RNA pour le paramètre Tb.

En ce qui concerne les comparaisons des modèles de paramètre d20, le **tableau V.11** montre que le modèle de type RQSP-RNA proposé dans cette thèse est meilleur, de sorte qu'il a une valeur proche d'une fois et demi moins en terme d'ERAM par rapport aux deux modèles de la méthode "QSPR MLR" dans l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)** pour l'ensemble de la base d'apprentissage malgré qu'il a légèrement dépassé ma valeur de R^2 dans le premier modèle de la même étude **(Wakeham *et al.* 2002)** qui vaut exactement 0,9970. Notre modèle a aussi une valeur approximativement trois fois moins par rapport au premier modèle de la méthode "QSPR MLR" et plus de deux fois et demi moins par rapport au deuxième modèle de la même méthode "QSPR MLR" dans l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)** pour l'ensemble de la base de test, toujours en termes d'ERAM. En utilisant les valeurs des R^2 pour effectuer une comparaison entre les trois modèles de la méthode "QSPR MLR" mentionnés dans l'étude de Ha et ses co-auteurs **(Ha *et al.* 2005)** et le modèle de type RQSP-RNA développé dans cette thèse pour le même paramètre d20, puisqu'ils n'ont pas calculé les mêmes types d'erreur que la présente thèse. Bien que ce n'est pas un vrai critère pour

évaluer un modèle mais j'ai comparé selon leurs valeurs R^2, le résultat de meilleur modèle parmi les trois modèles de l'étude précédente **(Ha *et al.* 2005)** en termes de R^2 a montré que sa valeur qui vaut 0,9910 ne dépasse pas la valeur de notre R^2 qui est exactement égale à 0,9948 pour l'ensemble de la base de données entière, ce qui signifier également que notre modèle de type RQSP-RNA est meilleur par rapport à cette étude.

Pour la comparaison effectuée pour le paramètre Tc entre le modèle de type RQSP-RNA développé dans cette thèse et d'autres modèles disponibles dans la littérature, les résultats ont été organisés et résumés dans le **tableau V.12**. Ces résultats révèlent que le modèle développé dans cette thèse a surpassé la plupart des résultats issus des modèles utilisés dans la comparaison, alors qu'il a une valeur d'ERAM plus d'une fois et demi inférieure, comparée à la méthode de "QSPR RBFNN" dans les deux études de Banchero et Manna **(Banchero et Manna 2018)** et de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)** pour l'ensemble de la base entière, presque deux fois moins comparée à des méthodes de "MLP ANN", "LSSVM" et "GMDH-NN" dans la même étude de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)**, identique pour le résultat du premier modèle de la méthode "GC FF-ANN" dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)**, toujours pour l'ensemble de la base entière. Également plus de cinq fois moins par rapport à la méthode de "QSPR MLR" mentionnée dans l'étude de Banchero et Manna **(Banchero et Manna 2018)**, et plus de sept fois moins par rapport à la méthode de "QSPR bpNN" citée dans l'étude de Espinosa et ses co-auteurs **(Espinosa *et al.* 2001)**. Notre modèle à légèrement dépassé le deuxième modèle de la méthode "GC FF-ANN" dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)** et le modèle de la méthode "QSPR-WMSE-bpNN" mentionnée dans l'étude de Godavarthy et ses co-auteurs **(Godavarthy *et al.* 2008)**, de sorte que la différence a été estimée de l'ordre de 0,1 %, toujours en terme d'ERAM pour l'ensemble de la base entière. Concernant les études qui n'ont pas calculé l'ERAM, on a comparé en utilisant les valeurs de la REQM, et le résultat obtenu par mon modèle a une valeur plus de deux fois moins comparée à la méthode de "QSPR RBFNN" citée dans l'étude de Yao et ses co-auteurs **(Yao *et al.* 2002a)** pour l'ensemble de la base entière, plus d'une fois et demi moins par rapport à la méthode de "QSPR MLR" dans la même étude **(Yao *et al.* 2002a)** pour les deux ensembles de la base d'apprentissage et de test en termes de la REQM. Pour le modèle proposé dans cette étude, le résultat est aussi meilleur par rapport à la méthode de "QSPR MLR" dans l'étude de Sola et ses co-auteurs **(Sola *et al.* 2008)**, il a une valeur presque deux fois moins pour le résultat de l'ensemble de la base d'apprentissage et une légère différence de 0,88 K pour le résultat de l'ensemble de la base de test. De même les résultats issus du modèle de type RQSP-

RNA présenté dans cette thèse, ont surpassé les résultats de la même méthode "QSPR MLR" mentionnée dans l'étude de Turner et ses co-auteurs **(Turner *et al.* 1998)**. Le modèle a donc une valeur presque une fois et demi moins par rapport au résultat de l'ensemble d'apprentissage équivalent à 2,72 K de différence, et une légère différence de 0,35 K pour l'ensemble de la base de test en termes de la REQM. Dans la même étude **(Turner *et al.* 1998)**, mais cette fois pour le modèle de la méthode "QSPR CNN", donc il n'y pas une grande différence entre ce modèle et mon modèle, car le résultat de mon modèle a légèrement dépassé par une différence de 0,885 K pour l'ensemble de la base d'apprentissage, par contre le modèle de Turner et ses co-auteurs **(Turner *et al.* 1998)** a légèrement surpassé de 0,438 K par rapport à mon modèle pour l'ensemble de la base de test.

Les seuls modèles qui ont dépassé le modèle proposé dans l'estimation du paramètre Tc sont ceux de la méthode de "QSPR MLR" dans l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)**, avec une différence de 0,132 % et de 0,287 %, respectivement pour les ensembles d'apprentissage et de test, en termes d'ERAM. Aussi pour les deux modèles de la méthode "QSPR Fuzzy-ARTMAP" citée dans l'étude de Espinosa et ses co-auteurs **(Espinosa *et al.* 2001)**, la valeur plus de trois fois moins pour le premier modèle et plus de deux fois moins pour le deuxième modèle par rapport à mon résultat d'ensemble de la base entière, toujours en termes d'ERAM. Comme nous l'avons dit plus tôt dans les premiers paragraphes de cette partie, bien qu'il ait dépassé notre modèle mais il est encore incomplet en raison de la grande différence entre les résultats des trois ensembles, ou ce qu'on appelle techniquement dans ce domaine, le sur-apprentissage et le sous-apprentissage.

Pour le paramètre Pc, le **tableau V.13** montre les résultats de la comparaison entre le modèle de type RQSP-RNA développé dans cette thèse et d'autres modèles trouvés dans la littérature. La comparaison révèle que mon résultat est meilleur, puisqu'il a surpassé la plupart des modèles présentés dans cette comparaison. Mon modèle a une valeur de plus de cinq fois moins comparée au modèle de la méthode "QSPR bpNN" présentée dans l'étude de Espinosa et ses co-auteurs **(Espinosa *et al.* 2001)**, et plus de quatre fois et demi par rapport aux méthodes de "QSPR MLR" et de "GMDH-NN" indiquées dans l'étude de Banchero et Manna **(Banchero et Manna 2018)** et dans l'étude de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)**, respectivement, en termes d'ERAM pour l'ensemble de la base entière. Plus de quatre fois moins par rapport aux deux modèles de la méthode "MLP-ANN" et la méthode "LSSVM", et plus de trois fois et demi moins pour le modèle de la méthode "RBF-ANN" mentionnée dans la même étude de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)** toujours pour les résultats de

l'ensemble de la base entière en termes d'ERAM. Le modèle proposé a aussi une valeur plus d'une fois et demie moins comparée à la méthode de "QSPR-RBFNN" dans l'étude de Banchero et Manna **(Banchero et Manna 2018)**. Même proportion pour le premier modèle de la méthode "GC FF-ANN" présentée dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)** en termes d'ERAM pour les résultats d'ensemble de la base entière. Le résultat de mon modèle a aussi dépassé légèrement le résultat du modèle de la méthode "QSPR-WMSE-bpNN" mentionnée dans l'étude de Godavarthy et ses co-auteurs **(Godavarthy *et al.* 2008)**, presque 0,0739 % de différence en termes d'ERAM pour l'ensemble de la base entière. Le modèle de type RQSP-RNA développé dans cette thèse a également dépassé les deux modèles de la méthode "QSPR MLR" présentée dans l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)**, il a une valeur presque deux fois inférieure aux résultats pour l'ensemble de la base de test dans les deux modèles, plus d'une fois et demi moins et plus de deux fois et demi moins, respectivement pour le premier et le deuxième modèle de la même étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)** pour l'ensemble de la base d'apprentissage en termes d'ERAM.

Les deux modèles de la méthode "QSPR Fuzzuy-ARTMAP" présentés dans l'étude de Espinosa et ses co-auteurs **(Espinosa *et al.* 2001)** ont dépassés les résultats obtenu dans cette thèse, en termes d'ERAM par des valeurs plus de deux fois et demie inférieures pour le premier modèle et plus de deux fois inférieures pour le deuxième modèle par rapport au résultat de l'ensemble de la base entière. Bien qu'ils soient dépassés notre résultat mais ils ne sont pas convaincus et restent de mauvais modèles, toujours en raison du phénomène de sur-apprentissage ou sous-apprentissage. Le deuxième modèle de la méthode "GC FF-ANN" présenté dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)** a aussi légèrement surpassé le résultat obtenu par mon modèle, d'environ 0,33 % de différence en termes d'ERAM pour l'ensemble de la base entière.

Les autres modèles présentés dans cette étude pour lesquels l'ERAM n'a pas été calculée, sont évalués en termes du résultat de la REQM. Par comparaison, mon modèle est meilleur en termes de la REQM avec une valeur plus de trois fois plus petite que celle des deux modèles de la méthode de "QSPR MLR" et de la méthode de "QSPR RBFNN" mentionnées dans l'étude de Yao et ses co-auteurs **(Yao *et al.* 2002a)** pour l'ensemble de la base entière. Mon modèle a également surpassé le modèle de la méthode de "QSPR MLR" avec une valeur de plus cinq fois et demie inférieure dans la base de l'ensemble d'apprentissage et plus de quatre fois inférieure dans la base de l'ensemble de test en termes de la REQM, mentionnée dans l'étude de Sola et ses co-auteurs **(Sola *et al.* 2008)**. Presque cinq fois inférieure et plus de quatre fois inférieure

respectivement pour l'ensemble de la base d'apprentissage et de test pour le modèle de la méthode de "QSPR MLR" présentée dans l'étude de Turner et ses co-auteurs **(Turner *et al.* 1998)**, et plus de trois fois et demie moins pour les deux ensembles d'apprentissage et de test du modèle de la méthode "QSPR CNN" dans la même étude de Turner et ses co-auteurs **(Turner *et al.* 1998)**, toujours en termes des valeurs de la REQM.

Le **tableau V.14** montre les résultats de la comparaison entre le modèle de type RQSP-RNA développé dans cette thèse et d'autres modèles trouvés dans la littérature pour le paramètre Vc. La comparaison révèle que mon résultat est supérieur à la majorité des modèles mentionnés dans cette comparaison. Mon modèle a une valeur plus d'une fois et demie comparée aux résultats des modèles des méthodes "MLP-ANN", "RBF-ANN" et "LSSVM" cités dans l'étude de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)** en termes d'ERAM pour l'ensemble de la base entière, et presque même proportion comparée aux résultats de la REQM pour les trois méthodes dans la même étude de Varamesh **(Varamesh *et al.* 2017b)**. En termes d'ERAM, mon modèle a une valeur plus de deux fois inférieure par rapport au modèle de la méthode "GMDH-NN", et plus de trois fois inférieure en termes de la REQM dans l'étude de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)** pour l'ensemble de la base entière. À propos de l'étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)**, mon modèle a également dépassé les résultats des deux modèles de la méthode "QSPR MLR" mentionnés dans cette étude, il a une valeur d'ERAM plus de deux fois inférieure aux résultats de l'ensemble de la base de test des deux modèles, même proportion comparée aux résultats de l'ensemble d'apprentissage du deuxième modèle et plus d'une fois et demie par rapport au premier modèle du même ensemble de données, dans la même étude de Wakeham et ses co-auteurs **(Wakeham *et al.* 2002)**.

Le modèle de la méthode "QSPR MLR" présenté dans l'étude de Zuas et Styarini **(Zuas & Styarini 2009)** pour le paramètre Vc, est considéré comme le seul modèle qui n'a pas calculé le même type d'erreur. De ce fait, j'ai évalué ce modèle selon la valeur de R^2 qui est égale à 0,997 pour l'ensemble de la base entière, légèrement supérieure à celle obtenue pour mon modèle, égale à 0,999 pour le même ensemble de base de données entière.

Le modèle de type RQSP-RNA de Vc proposé dans cette thèse, a aussi légèrement dépassé le premier modèle de la méthode "GC FF-ANN" mentionné dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)**. La différence en termes d'ERAM, égale à 0,44 % pour les résultats de l'ensemble de la base entière. Les résultats du modèle de la méthode "QSPR-WMSE-bpNN" mentionné dans l'étude de Godavarthy et ses co-auteurs **(Godavarthy *et***

***al*. 2008)** sont presque identiques, avec une légère différence estimée à 0,061 % entre ce modèle et le modèle de type RQSP-RNA développé dans cette thèse pour le paramètre Vc, par rapport à l'ensemble de la base entière en termes d'ERAM.

Le seul modèle qui a légèrement dépassé notre résultat est le deuxième modèle de la méthode "GC FF-ANN" mentionné dans l'étude de Gharagheizi et ses co-auteurs (**Gharagheizi *et al.* 2011**). Il y a une différence de 0,361 % inférieure, en termes d'ERAM, par rapport au résultat issu du modèle de type RQSP-RNA développé dans cette thèse, toujours pour l'ensemble de la base entière.

Enfin, sur la base de cette comparaison détaillée entre mes modèles pour chaque paramètre et d'autres modèles dans la littérature, on peut dire que tous les indicateurs confirment encore la supériorité, l'efficacité et la précision de mes modèles de type RQSP-RNA développés dans cette thèse, pour l'estimation des cinq paramètres Tb, d20, Tc, Pc et Vc pour les hydrocarbures purs.

Tableau V.10. Comparaison entre le modèle de RQSP-RNA du paramètre *Tb* développé dans cette étude et d'autres modèles présents dans la littérature.

Modèles	Méthodes	Nature des hydrocarbures	Ensemble d'apprentissage				Ensemble de test				Ensemble entière			
			N	R^2 (-)	*REQM* (K)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (K)	*ERAM* (%)	*N*	R^2 (-)	*RMSE* (K)	*ERAM* (%)
Proposé	**RQSP-RNA**	**223 différents hydrocarbures purs**	**179**	**0,9995**	**2,1458**	**0,3539**	**44**	**0,9999**	**1,3718**	**0,2600**	**223**	**0,9996**	**2,0168**	**0,3353**
(Varamesh *et al.* 2017a)	**MLP-ANN**	563 composés organiques purs, y compris des hydrocarbures	450	(-)	21,22	3,24	113	(-)	24,70	4,07	563	(-)	21,95	3,41
	RBF-**ANN**		450	(-)	17,92	2,60	113	(-)	21,87	2,82	563	(-)	18,78	2,65
	LSSVM		450	(-)	21,52	3,09	113	(-)	18,54	3,16	563	(-)	20,95	3,11
	GMDH-**NN**		450	(-)	25,59	3,79	113	(-)	23,98	3,73	563	(-)	25,27	3,78
(Saaidpour & Ghaderi 2016)	**QSPR** MLR	80 hydrocarbures du pétrole brut	64	0,9938	5,8971	(-)	16	(-)	10,8742	(-)	80	(-)	(-)	(-)
(Dai *et al.* 2013)	**QSPR** MLR	80 alcanes	60	0,9993	(-)	(-)	20	(-)	(-)	(-)	(-)	(-)	(-)	(-)
	QSPR MLR	65 hydrocarbures insaturés	50	0,9910	(-)	(-)	15	(-)	(-)	(-)	(-)	(-)	(-)	(-)
(Ghargheizi *et al.* 2013)	**QSPR** FF-**ANN**	Un grand ensemble de données de 17768 composés chimiques purs	**14216**	0,9430	22	3,19	**1776**	0,9470	21	3,05	**17768**	0,9430	22	3,16
(Sola *et al.* 2008)	**QSPR** MLR	155 composés organiques purs, y compris des hydrocarbures	135	(-)	9,10	(-)	20	(-)	7,33	(-)	155	0,9864	(-)	(-)
(Ha *et al.* 2005)	**QSPR** MLR	186 hydrocarbures saturés purs	152	(-)	(-)	(-)	34	(-)	(-)	(-)	186	0,9979	(-)	(-)
	QSPR MLR	200 hydrocarbures aromatiques purs	139	(-)	(-)	(-)	61	(-)	(-)	(-)	200	0,9960	(-)	(-)
	QSPR MLR	386 hydrocarbures saturés et aromatiques purs	291	(-)	(-)	(-)	95	(-)	(-)	(-)	386	0,9947	(-)	(-)
(Espinosa *et al.* 2001)	**QSPR** Fuzzy-ARTMAP	1168 composés organiques purs, y compris des hydrocarbures (aliphatique, aromatique …)	1015	(-)	(-)	**0,06**	153	(-)	(-)	3,27	1168	(-)	(-)	0,49
	QSPR Fuzzy-ARTMAP		1015	(-)	(-)	0,40	153	(-)	(-)	5,03	1168	(-)	(-)	1,00
	QSPR bpNN		1015	(-)	(-)	7,20	153	(-)	(-)	5,40	1168	(-)	(-)	7,02
(Wessel & Jurs 1995)	**QSPR** MLR	366 différents hydrocarbures purs	296	(-)	6,3	(-)	55	(-)	14,4	(-)	366	(-)	(-)	(-)
	QSPR CNN	311 différents hydrocarbures purs	267	(-)	5,7	(-)	29	(-)	5,7	(-)	311	(-)	(-)	(-)

***NB_ Tableau V.10* :** Les résultats du sous ensemble de validation dans les deux études : de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2013)** et celui de Wessel et Jurs **(Wessel & Jurs 1995)** n'est pas présenté dans cette comparaison.

Tableau V.11. Comparaison entre le modèle de RQSP-RNA du paramètre *d20* développé dans cette étude et d'autres modèles présents dans la littérature.

Modèles	Méthodes	Nature des hydrocarbures	Ensemble d'apprentissage				Ensemble de test				Ensemble entière			
			N	R^2 (-)	*REQM* (-)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (-)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (-)	*ERAM* (%)
Proposé	**RQSP-RNA**	**222 différents hydrocarbures purs**	**178**	**0,9931**	**0,0064**	**0,5797**	**44**	**0,9985**	**0,0040**	**0,3960**	**222**	**0,9948**	**0,0060**	**0,5433**
(Ha *et al.* 2005)	**QSPR** MLR	186 hydrocarbures saturés purs	152	(-)	(-)	(-)	34	(-)	(-)	(-)	186	0,9910	(-)	(-)
	QSPR MLR	200 hydrocarbures aromatiques purs	164	(-)	(-)	(-)	36	(-)	(-)	(-)	200	0,9881	(-)	(-)
	QSPR MLR	**386** hydrocarbures saturés et aromatiques purs	**316**	(-)	(-)	(-)	**70**	(-)	(-)	(-)	**386**	0,9805	(-)	(-)
(Wakeham *et al.* 2002)	**QSPR** MLR	237 différents hydrocarbures purs	219	**0,9970**	(-)	0,82	18	(-)	(-)	1,17	237	(-)	(-)	(-)
	QSPR MLR		219	(-)	(-)	0,84	18	(-)	(-)	1,09	237	(-)	(-)	(-)

Tableau V.12. Comparaison entre le modèle de RQSP-RNA du paramètre *Tc* développé dans cette étude et d'autres modèles présents dans la littérature.

Modèles	Méthodes	Nature des hydrocarbures	Ensemble d'apprentissage				Ensemble de test				Ensemble entière			
			N	R^2 (-)	*REQM* (K)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (K)	*ERAM* (%)	*N*	R^2 (-)	*RMSE* (K)	*ERAM* (%)
Proposé	**RQSP-RNA**	**223 différents hydrocarbures purs**	**179**	**0,9951**	**6,4445**	**0,7424**	**44**	**0,9945**	**8,7785**	**0,9974**	**223**	**0,9948**	**6,9672**	**0,7927**
(Banchero et Manna 2018)	**QSPR** MLR	306 composés organiques purs, y compris des hydrocarbures (environ 40%)	215	(-)	22,0	3,2	91	(-)	37,4	6,2	306	(-)	27,5	4,1
	QSPR RBF**NN**		215	(-)	7,2	0,92	91	(-)	11,9	1,7	306	(-)	8,8	1,2
(Varamesh *et al.* 2017b)	**MLP-ANN**	563 composés organiques purs, y compris des hydrocarbures	450	(-)	21,45	1,48	113	(-)	18,29	1,60	(-)	563	20,85	1,51
	RBF-**ANN**		450	(-)	19,90	1,38	113	(-)	19,94	1,47	(-)	563	19,91	1,40
	LSSVM		450	(-)	20,43	1,43	113	(-)	20,20	1,57	(-)	563	20,38	1,45
	GMDH-**NN**		450	(-)	20,72	1,56	113	(-)	23,13	1,48	(-)	563	21,23	1,54
(Ghargheizi *et al.* 2011)	GC FF-**ANN-1**	Environ 1700 composés chimiques purs de différentes classes (81 classes, y compris des hydrocarbures)	**1373**	0,983	(-)	1,5	**171**	0,984	(-)	1,8	**1715**	0,983	(-)	1,5
	GC FF-**ANN-2**		1359	0,994	(-)	0,9	169	0,992	(-)	1,0	1697	0,994	(-)	0,9
(Sola *et al.* 2008)	**QSPR** MLR	153 composés organiques purs, y compris des hydrocarbures	133	0,9856	12,6	(-)	20	(-)	9,66	(-)	153	0,9861	(-)	(-)
(Godavarthy *et al.* 2008)	**QSPR**-WMSE-**bpNN**	1230 composés organiques purs, y compris des hydrocarbures (73 classes différentes des hydrocarbures)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	1230	0,995	(-)	0,9
(Yao *et al.* 2002a)	**QSPR** MLR	856 composés organiques purs, y compris des hydrocarbures (environ 330 composés)	733	(-)	16,29	(-)	123	(-)	16,44	(-)	856	(-)	(-)	(-)
	QSPR RBF**NN**		733	(-)	12,32	(-)	123	(-)	14,23	(-)	856	(-)	13,97	(-)
(Wakeham *et al.* 2002)	**QSPR** MLR	222 différents hydrocarbures purs	204	**0,999**	(-)	**0,61**	18	(-)	(-)	**0,71**	(-)	(-)	(-)	(-)
(Espinosa *et al.* 2001)	**QSPR** Fuzzy-ARTMAP	530 composés organiques purs, y compris des hydrocarbures (aliphatiques, aromatiques, …)	461	(-)	(-)	**0,05**	69	(-)	(-)	1,45	530	(-)	(-)	**0,24**
	QSPR Fuzzy-ARTMAP		461	(-)	(-)	**0,06**	69	(-)	(-)	2,30	530	(-)	(-)	**0,35**
	QSPR bpNN		463	(-)	(-)	5,80	67	(-)	(-)	3,90	530	(-)	(-)	5,60
(Turner *et al.* 1998)	**QSPR** MLR	164 composés organiques purs, y compris des hydrocarbures (environ 46%)	147	(-)	9,16	(-)	17	(-)	9,13	(-)	164	(-)	(-)	(-)
	QSPR C**NN**		132	(-)	7,33	(-)	17	(-)	**8,34**	(-)	164	(-)	(-)	(-)

NB_ Tableau V.12 **:** Les résultats du sous ensemble de validation dans les deux études : de Ghargheizi et ses co-auteurs **(Ghargheizi *et al.* 2011)** et celui de Turner et ses co-auteurs **(Turner *et al.* 1998)** n'est pas présenté dans cette comparaison.

Tableau V.13. Comparaison entre le modèle de RQSP-RNA du paramètre *Pc* développé dans cette étude et d'autres modèles présents dans la littérature.

Modèles	Méthodes	Nature des hydrocarbures	Ensemble d'apprentissage				Ensemble de test				Ensemble entière			
			N	R^2 (-)	*REQM* (MPa)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (MPa)	*ERAM* (%)	*N*	R^2 (-)	*RMSE* (MPa)	*ERAM* (%)
Proposé	**RQSP-RNA**	**221 différents hydrocarbures purs**	**177**	**0,9975**	**0,04279**	**1,2781**	**44**	**0,9954**	**0,065715**	**2,0212**	**221**	**0,9970**	**0,048231**	**1,4261**
(Banchero & Manna 2018)	**QSPR** MLR	306 composés organiques purs, y compris des hydrocarbures (environ 40%)	215	(-)	0,40	6,1	91	(-)	0,47	8,5	306	(-)	0,42	6,8
	QSPR RBF**NN**		215	(-)	0,11	1,9	91	(-)	0,18	3,5	306	(-)	0,13	2,3
(Varamesh *et al.* 2017b)	**MLP-ANN**	563 composés organiques purs, y compris des hydrocarbures	450	(-)	0,395	6,04	113	(-)	0,437	7,65	(-)	563	0,404	6,36
	RBF-**ANN**		450	(-)	0,328	5,13	113	(-)	0,414	5,84	(-)	563	0,347	5,27
	LSSVM		450	(-)	0,365	5,85	113	(-)	0,309	6,19	(-)	563	0,355	5,92
	GMDH-**NN**		450	(-)	0,412	6,57	113	(-)	0,531	7,77	(-)	563	0,439	6,81
(Ghargheizi *et al.* 2011)	GC FF-**ANN-1**	Environ 1700 composés chimiques purs de différentes classes (81 classes, y compris des hydrocarbures)	**1373**	0,987	(-)	2,5	**171**	**0,999**	(-)	2,1	**1715**	0,986	(-)	2,5
	GC FF-**ANN-2**		1358	**0,999**	(-)	**1,0**	169	**0,999**	(-)	**1,0**	1696	**0,999**	(-)	**1,1**
(Godavarthy *et al.* 2008)	**QSPR**-WMSE-**bpNN**	1230 composés organiques purs, y compris des hydrocarbures (73 classes différentes d'hydrocarbures)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	1230	0,992	(-)	1,5
(Sola *et al.* 2008)	**QSPR** MLR	139 composés organiques purs, y compris des hydrocarbures	121	0,9209	0,25	(-)	18	(-)	0,28	(-)	139	0,9293	(-)	(-)
(Yao *et al.* 2002b)	**QSPR** MLR	173 composés benzènes substitués, y compris des hydrocarbures.	148	(-)	0,176	(-)	25	(-)	0,190	(-)	173	(-)	0,1767	(-)
	QSPR RBF**NN**		148	(-)	0,1385	(-)	25	(-)	0,2372	(-)	173	(-)	0,1547	(-)
(Espinosa *et al.* 2001)	**QSPR** Fuzzy-ARTMAP	463 composés organiques purs, y compris des hydrocarbures (aliphatiques, aromatiques, …)	409	(-)	(-)	**0,01**	54	(-)	(-)	4,31	463	(-)	(-)	**0,52**
	QSPR Fuzzy-ARTMAP		409	(-)	(-)	**0,08**	54	(-)	(-)	4,90	463	(-)	(-)	**0,65**
	QSPR bpNN		409	(-)	(-)	7,96	54	(-)	(-)	6,00	463	(-)	(-)	7,71
(Wakeham *et al.* 2002)	**QSPR** MLR	219 différents hydrocarbures purs	201	**0,999**	(-)	2,33	18	(-)	(-)	4,25	219	(-)	(-)	(-)
	QSPR MLR		201	**0,999**	(-)	3,37	18	(-)	(-)	3,85	219	(-)	(-)	(-)
(Turner *et al.* 1998)	**QSPR** MLR	164 composés organiques purs, y compris des hydrocarbures (environ 46%)	147	(-)	0,20974	(-)	17	(-)	0,28371	(-)	164	(-)	(-)	(-)
	QSPR CNN		132	(-)	0,153	(-)	17	(-)	0,24216	(-)	164	(-)	(-)	(-)

NB_ Tableau V.13 **:** Dans ce travail les valeurs de REQM de Pc ont été converties en ***MPa*** sachant qu'elles étaient principalement exprimées en ***Pa*** ; ***(2)*** Les valeurs de REQM de Pc dans les études de Varamesh et ses co-auteurs **(Varamesh *et al.* 2017b)**, de Yao et ses co-auteurs **(Yao *et al.* 2002b)**, et de Turner et ses co-auteurs **(Turner *et al.* 1998)** ont été également converties en ***MPa*** sachant que les valeurs obtenues dans les deux premières études sont exprimées en ***bar*** et dans la dernière étude en ***atm***, ***(3)*** Les résultats du sous ensemble de validation dans les deux études : de Ghargheizi et ses co-auteurs **(Ghargheizi *et al.* 2011)** et de Turner et ses co-auteurs **(Turner *et al.* 1998)** ne sont pas présentés dans cette comparaison.

Tableau V.14. Comparaison entre le modèle de RQSP-RNA du paramètre *Vc* proposé dans cette étude et d'autres modèles présents dans la littérature.

Modèles	Méthodes	Nature des hydrocarbures	Ensemble d'apprentissage				Ensemble de test				Ensemble entière			
			N	R^2 (-)	*REQM* (cm^3/mol)	*ERAM* (%)	*N*	R^2 (-)	*REQM* (cm^3/mol)	*ERAM* (%)	*N*	R^2 (-)	*RMSE* (cm^3/mol)	*ERAM* (%)
Proposé	**RQSP-RNA**	**221 différents hydrocarbures purs**	**179**	**0,9990**	**10,0040**	**1,6409**	**44**	**0,9989**	**12,8870**	**2,2497**	**223**	**0,9990**	**10,6349**	**1,7610**
Zuas & Styarini 2009)	**QSPR** MLR	65 hydrocarbures aliphatiques insaturés purs	35	0,997	(-)	(-)	30	(-)	(-)	(-)	(-)	(-)	(-)	(-)
(Varamesh *et al.* 2017b)	**MLP-ANN**	563 composés organiques purs, y compris des hydrocarbures	450	(-)	20,90	3,09	113	(-)	16,41	3,13	(-)	563	20,08	3,10
	RBF-**ANN**		450	(-)	19,88	2,91	113	(-)	15,99	2,72	(-)	563	19,38	2,93
	LSSVM		450	(-)	20,89	2,94	113	(-)	16,37	3,13	(-)	563	20,07	2,98
	GMDH-**NN**		450	(-)	36,24	4,07	113	(-)	34,68	3,66	(-)	563	35,91	3,99
(Godavarthy *et al.* 2008)	**QSPR-**WMSE-**bpNN**	1230 composés organiques purs, y compris des hydrocarbures (73 classes différentes des hydrocarbures)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	1230	0,999	(-)	1,7
(Wakeham *et al.* 2002)	**QSPR** MLR	208 différents hydrocarbures purs	190	**0,999**	(-)	2,95	18	(-)	(-)	4,53	208	(-)	(-)	(-)
	QSPR MLR		190	(-)	(-)	3,63	18	(-)	(-)	4,97	208	(-)	(-)	(-)
(Gharagheizi *et al.* 2011)	GC FF-**ANN-1**	Environ 1700 composés chimiques purs de différentes classes (81 classes, y compris des hydrocarbures)	**1375**	0,998	(-)	2,2	**171**	0,998	(-)	2,3	**1717**	0,998	(-)	2,2
	GC FF-**ANN-2**		1366	**0,999**	(-)	**1,4**	170	**0,999**	(-)	**1,2**	1706	0,999	(-)	**1,4**

***NB_ Tableau V.14* :** Les résultats du sous ensemble de validation dans l'étude de Gharagheizi et ses co-auteurs **(Gharagheizi *et al.* 2011)** ne sont pas présentés dans cette comparaison.

V.8. Conclusion

Dans ce chapitre, on a discuté l'ensemble des résultats obtenus par les modèles de type RQSP-RNA développés dans cette thèse, afin de prédire cinq propriétés physico-chimiques importantes pour des hydrocarbures purs : la température normale d'ébullition (Tb), la densité relative de liquide (d20), la température critique (Tc), la pression critique (Pc) et le volume critique (Vc). Le développent de ces modèles a induit plusieurs étapes fondamentales, telles que la préparation de la base des données pour le calcul des descripteurs moléculaires, le prétraitement et la réduction de ces descripteurs, le choix des paramètres de chaque RNA, l'apprentissage et le test des modèles de type RQSP-RNA, l'estimation des erreurs et le calcul des coefficients statistiques pour chaque modèle, puis la validation des modèles finaux de type RQSP-RNA.

Dans ce qui précède, j'ai commencé par l'étude de l'évolution des descripteurs quantitativement au cours de la procédure de prétraitement, puis on a étudié chaque descripteur qualitativement, de sorte à montrer les caractéristiques de chaque descripteur pertinent sélectionné pour les cinq paramètres étudiés, ensuite j'ai discuté les architectures des meilleurs modèles de type RQSP-RNA obtenus pour chaque paramètre. Puis j'ai expliqué l'application ou bien l'utilisation des expressions mathématiques de chaque modèle obtenu pour la prédiction des cinq propriétés physicochimiques des hydrocarbures. Ensuite, j'ai traité l'appréciation de la performance des résultats d'estimation issus des modèles de type RQSP-RNA développés dans cette thèse, par le calcul des erreurs et l'analyse statistique de chaque modèle obtenu. J'ai aussi comparé graphiquement les résultats des valeurs expérimentales et les résultats des valeurs calculées à partir de chaque modèle de type RQSP-RNA pour les cinq paramètres. J'ai également fait une analyse de sensibilité des descripteurs d'entrée de chaque modèle par la méthode de poids, afin d'interpréter et attribuer les contributions de chaque descripteur individuellement sur le modèle auquel il appartient. Puis, selon la valeur de la contribution moyenne (VCM) de chaque modèle, j'ai aussi mesuré les contributions de chaque classe des descripteurs sur chaque modèle obtenu. J'ai déterminé le domaine d'applicabilité afin de mieux valider les modèles de type RQSP-RNA obtenus. J'ai fait aussi des comparaisons systématiques de performance de mes modèles de type RQSP-RNA, développés dans cette thèse et d'autres modèles similaires existant dans la littérature afin de mieux confirmer l'efficacité et la précision de mes modèles dans l'estimation des cinq paramètres Tb, d20, Tc, Pc et Vc des hydrocarbures purs. Finalement, j'ai clôturé ce chapitre par l'analyse et l'interprétation globales des différents résultats obtenus dans cette thèse.

Conclusion générale et perspectives

Les hydrocarbures est la plus large famille de pétrole brut. Elle comporte plusieurs constituants saturés et insaturés, qui sont considérés comme des composantes principales pour différents types de mélanges pétroliers. Les propriétés physicochimiques et thermodynamiques des hydrocarbures sont des paramètres importants pour plusieurs ingénieries et industries, comme l'industrie pétrochimique et d'autres industries liées, car elles sont parmi les clés qui aident à la fois les scientifiques et les industriels pour comprendre de nombreux phénomènes et opérations unitaires liées aux traitements, productions, transports et stockages de ces matériaux d'intérêts. Cependant, le calcul des propriétés de mélanges pétroliers dépend des propriétés de ses composants purs. Certaines propriétés comme les propriétés critiques ne peuvent pas être facilement mesurées, cela est dû principalement aux tests de laboratoire coûteux et lents.

Dans ce travail de thèse, j'ai abordé une problématique qui traite un sujet très important dans le domaine scientifique de recherche lié avec le milieu industriel. Ma contribution a porté sur la modélisation moléculaire, précisément sur le développement des modèles de type RQSP-RNA afin de prédire cinq (5) propriétés physicochimiques des hydrocarbures purs : la température du point d'ébullition normale (Tb), la densité relative du liquide (d20), la température critique (Tc), la pression critique (Pc), et le volume critique (Vc). Seulement, à partir des informations structurelles incorporées dans les molécules des hydrocarbures étudiés, utilisant les réseaux de neurones artificiels perceptron multicouches (RNA-PMC) qui font partie des méthodes computationnelles d'intelligences artificielles.

La méthode de développement adoptée dans cette thèse comporte plusieurs étapes basées sur les précédentes études dans la littérature. L'ensemble des données expérimentales a été sélectionné à partir de plusieurs sources, et contient environ 220 molécules d'hydrocarbures purs, appartenant aux principales catégories d'importance industrielles où les majorités ont plus d'un isomère de position. La base a été divisée en deux sous-ensembles : un pour l'apprentissage et l'autre pour le test. Un nombre important de descripteurs moléculaires ont été calculées en ligne à l'aide du logiciel E-Dragon. Plusieurs étapes basées sur des méthodes statistiques, y compris la méthode de RLM ont été suivies afin de réduire ce nombre et de ne conserver que les descripteurs pertinents. Les meilleurs modèles de type RQSP-RNA ont été obtenus par l'utilisation de la stratégie d'essais et d'erreurs avec un algorithme d'apprentissage de type "BFGS", les architectures des meilleurs modèles obtenus sont : "16-12-1" pour Tb, "16-10-1" pour d20, "10-16-1" pour Tc, "17-11-1" pour Pc et "12-17-1" pour Vc. Ultérieurement, chaque modèle a été analysé et testé par des méthodes et des paramètres statistiques tels que : le coefficient de détermination (R^2), la Racine d'Erreur Quadratique Moyenne (REQM) et l'Erreur

Relative Absolu Moyenne (ERAM), en plus de l'analyse de sensibilité par la méthode de poids et la détermination du domaine d'applicabilité.

Les résultats des modèles de type RQSP-RNA proposés dans cette thèse ont montré une bonne précision en fonction des valeurs de $\boldsymbol{R^2}$, qui varient d'une valeur maximale égale à 0,9999 pour la $\boldsymbol{Tb}^{Test}$ et une valeur minimale égale à 0,9931 pour la $\boldsymbol{d20}^{App}$. En outre, la robustesse et la capacité prédictive des modèles de type RQSP-RNA ont aussi été confirmées par divers types d'erreurs (***REQM***, ***ESP***, ***ERM*** et ***ERAM***) ; les résultats étant très encourageants : 2,0168 K, 0,0060, 6,9672 K, 0,048231 MPa, et 10,6349 cm^3/mol pour les REQM des paramètres Tb, d20, Tc, Pc et Vc, respectivement. Concernant les résultats des trois types d'erreurs relatives restantes (l'ESP, l'ERM et l'ERAM) ; elles ne dépassent pas les 2,5% pour les cinq paramètres dans les trois sous-ensembles de données. Précisément, elles sont comprises entre une valeur minimale égale à 0,2600% pour l'**ERAM** de $\boldsymbol{Tb}^{Test}$, et une valeur maximale égale à 2,4024% pour l'**ESP** de $\boldsymbol{Pc}^{Test}$. Les résultats des différentes comparaisons faites entre les modèles de type RQSP-RNA développés dans cette thèse et d'autres modèles similaires dans la littérature ont confirmé que nos modèles sont supérieurs lors de la prédiction des cinq paramètres : Tb, d20, Tc, Pc et Vc, car ils fournissent les meilleures performances, et donc ils donnent de meilleurs résultats.

Les modèles de type RQSP-RNA développés dans cette thèse ne sont pas seulement destinés aux familles particulières ou spécifiques telles que les alcanes ou les alcènes, mais ils englobent également toutes les familles des hydrocarbures dans le même modèle. De plus, ils sont capables de distinguer, même entre les isomères de position (à savoir ; *cis*, *trans*, *méta*, *ortho* et *para*) avec un haut degré de précision. Ils peuvent être utilisés dans le domaine de l'industrie et de l'ingénierie pétrolière, ainsi que dans d'autres domaines liés aux hydrocarbures purs, afin d'estimer leurs Tb, d20, Tc, Pc et Vc, uniquement à partir de leur structure moléculaire et en utilisant des descripteurs moléculaires simples et faciles à calculer.

L'ensemble des travaux présentés dans cette thèse a permis de répondre, en grande partie à l'objectif principal, qui était de développer des modèles basés sur des méthodes computationnelles d'intelligence artificielle et les informations structurelles des molécules d'hydrocarbures pour prédire quelques propriétés physicochimiques des hydrocarbures. Comme je l'ai montré précédemment dans la liste des travaux et activités de recherches, la grande partie de nos travaux présentés dans cette thèse a fait l'objet de plusieurs communications dans des congrès scientifiques nationaux et internationaux. En plus d'un article scientifique publié dans

une revue indexée dans les deux bases de données internationales connues : **Scopus** et **WOS**. Les articles et/ou les résumes complets de ces travaux ont été présentés à la fin de cette thèse.

En vertu de ce travail de thèse, je peux entrevoir quelques perspectives qui peuvent être des informations prometteuses et utiles pour les prochaines recherches dans ce domaine :

- L'enrichissement de la base des données quantitativement par l'utilisation d'une base plus large qui contiendrait plus de molécules, et de manière qualitative par le traitement des autres types d'hydrocarbures, permettrait d'améliorer la qualité des prochains travaux dans ce domaine de recherche. Par conséquent, ce changement serait accompagné de quelques modifications dans la méthodologie du développement, telles que l'ajout d'un troisième sous-ensemble qui est la base de validation, et d'essayer différentes partitions possibles.
- Le travail par la même méthodologie sur les mélanges pétroliers serait toujours une excellente perspective, pourvu que les données nécessaires, les moyens logistiques et surtout la coopération les services concernés soient possible.
- Compte tenu de la spécification des études RQSP/RQSA, il est possible d'appliquer la méthodologie suivie dans ce travail à d'autres propriétés ou activités dans divers domaines, qui sont liées aux informations structurelles des molécules, telles que les paramètres pharmacocinétiques (ADME) dans le domaine de recherche et de développent des médicaments, ou les propriétés liées aux différents types de toxicité (EC_{50}, IC_{50}…) dans le domaine de l'environnement et des autres domaines en relation.
- La combinaison de différentes méthodologies et stratégies de modélisation basées sur l'approche de RQSP/RQSA est l'une des solutions proposées qui pourrait améliorer les qualités des futurs travaux. A titre d'exemple l'utilisation des différentes méthodes statistiques, que ce soit d'autres types de RNAs telles que le réseau de neurone de type RBF, ou bien d'autres techniques récentes comme les Machines de Vecteur de Support (SVM ou SVR). De plus, l'utilisation des méthodes d'apprentissage approfondi (*en anglais* : Deep learning) pourrait être adaptée pour développées des modèles traitant des systèmes compliqués.
- Une autre perspective serait de tester différents algorithmes d'apprentissage, par exemple l'algorithme d'apprentissage de type "Levenberg-Marquardt : *trainlm*". En outre, l'utilisation des méthodes de normalisation des entrées est une autre solution suggérée, car elle permet de réduire l'écart entre les données dans la même base de données, ce qui facilite les tâches d'apprentissage des RNAs.

- Utiliser d'autres logiciels, plateformes ou bien des outils informatiques car il existe bon nombre de logiciels commerciaux ou open sources conçues pour faciliter la tâche du développement des modèles de type RQSP-RNA, tels que le logiciel commercial le plus connu dans le calcul des descripteurs moléculaires "Dragon 7", qui permet à la fois de calculer et de traiter environs 5270 descripteurs distincts, ou bien le logiciel Matlab qui est un outil très puissant utilisé par plusieurs ingénieurs dans divers domaines, notamment dans le développement des modèles neuronales et d'autres modèles computationnels d'intelligence artificielle. Il fournit l'emploi des interfaces intégrées standards ou bien encore permet de programmer un script personnel, qui donne plus de liberté, de choix, et de créativité pour le modélisateur, ce qui fournit des programmes plus stables et des résultats plus précis et reproductibles.

Références bibliographiques

Références biobibliographies

Abdel-Aal, H. K., & Alsahlawi, M. A. (Eds.). (2013). *Petroleum economics and engineering*. CRC Press.

Abdul Jameel, A. G., Van Oudenhoven, V., Emwas, A. H., & Sarathy, S. M. (2018). Predicting octane number using nuclear magnetic resonance spectroscopy and artificial neural networks. *Energy & fuels*, *32*(5), 6309-6329.

Albahri, T. A. (2003). Structural group contribution method for predicting the octane number of pure hydrocarbon liquids. *Industrial & engineering chemistry research*, *42*(3), 657-662.

Albahri, T. A., Riazi, M. R., & Alqattan, A. A. (2002). Octane number and aniline point of petroleum fuels. *Fuel Chemistry Division Preprints*, *47*(2), 710.

Alqaheem, S. S., & Riazi, M. R. (2017). Flash points of hydrocarbons and petroleum products: Prediction and evaluation of methods. *Energy & Fuels*, *31*(4), 3578-3584.

Ammar, M. Y. (2007). "*Mise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition batch/continu*". Thèse de doctorat en Génie des Procédés et de l'Environnement, *Ecole doctorale : Transfert, Dynamique des Fluides, Energétique & Procédés*, Institut national polytechnique de Toulouse, France. Disponible en ligne : http://ethesis.inp-toulouse.fr/archive/00000607/.

Ammi, Y. (2016). "*Modélisation des phénomènes de transport lors d'osmose inverse*". Thèse de Doctorat en génie des procédés, Option : Modélisation des phénomènes de transport en sciences de l'eau et d'environnement, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Ammi, Y., Khaouane, L., & Hanini, S. (2015). Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. Korean Journal of Chemical Engineering, 32(11), 2300-2310.

Ammi, Y., Khaouane, L., & Hanini, S. (2018). A Model Based on Bootstrapped Neural Networks for Modeling the Removal of Organic Compounds by Nanofiltration and Reverse Osmosis Membranes. Arabian Journal for Science and Engineering, 43(11), 6271-6284.

Amri, F. (2017). The relationship amongst energy consumption (renewable and non-renewable), and GDP in Algeria. *Renewable and Sustainable Energy Reviews*, *76*, 62-71.

Andrade, C. H., Pasqualoto, K. F., Ferreira, E. I., & Hopfinger, A. J. (2010). 4D-QSAR: perspectives in drug design. *Molecules*, *15*(5), 3281-3294.

Apergis, N., & Payne, J. E. (2012). Renewable and non-renewable energy consumption-growth nexus: Evidence from a panel error correction model. *Energy economics*, *34*(3), 733-738.

Arab, M., Bracene, R., Roure, F., Zazoun, R. S., Mahdjoub, Y., & Badji, R. (2015). Source rocks and related petroleum systems of the Chelif Basin,(western Tellian domain, north Algeria). *Marine and Petroleum Geology*, *64*, 363-385.

Atlas, R. M., & Bartha, R. (1992). Hydrocarbon biodegradation and oil spill bioremediation. In *Advances in microbial ecology*(pp. 287-338). Springer, Boston, MA.

Bagheri, M., Borhani, T. N. G., & Zahedi, G. (2012). Estimation of flash point and autoignition temperature of organic sulfur chemicals. *Energy Conversion and Management*, *58*, 185-196.

Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, *89*(5), 399-404.

Balaban, A. T. (1985). Applications of graph theory in chemistry. *Journal of chemical information and computer sciences*, *25*(3), 334-343.

Banchero, M., & Manna, L. (2018). Comparison between Multi-Linear-and Radial-Basis-Function-Neural-Network-Based QSPR Models for The Prediction of The Critical Temperature, Critical Pressure and Acentric Factor of Organic Compounds. *Molecules*, *23*(6), 1379.

Barbosa, L. L., Montes, L. F., Kock, F. V., Morgan, V. G., Souza, A., Song, Y. Q., & Castro, E. R. (2017). Relative hydrogen index as a fast method for the simultaneous determination of physicochemical properties of petroleum fractions. *Fuel*, *210*, 41-48.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, *43*(1), 3-31.

Baughman, D. R., & Liu, Y. A. (1995). *Neural networks in bioprocessing and chemical engineering*. Academic press.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (2010). Neural network toolbox. *User's Guide, MathWorks*, *2*, 77-81.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (2012). Neural network toolbox™ user's guide. In *R2012a, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098, www. mathworks. com*.

Belaid, F., & Youssef, M. (2017). Environmental degradation, renewable and non-renewable electricity consumption, and economic growth: Assessing the evidence from Algeria. *Energy Policy*, *102*, 277-287.

Belghit, C. (2019). "*Nouvelles corrélations pour la détermination des propriétés physiques et thermodynamiques des hydrocarbures et des fractions pétrolières*", Thèse de doctorat en Génie des Procédés, Option : Chimie Industrielle, Université 08 Mai 1945 Guelma, Département de Génie des Procédés, Laboratoire des Silicates, Polymère et des Nanocomposites (LSPN), Guelma.

Belghit, C., Lahiouel, Y., & Albahri, T. A. (2018). New empirical correlation for estimation of vaporization enthalpy of Algerian saharan blend petroleum fractions. *Petroleum Science and Technology*, *36*(15), 1181-1186.

Berkane, M. (2010). "*Estimation et analyse de mouvement par approches neuronales*". Thèse de doctorat en Images et Systèmes, *Ecole doctorale : Electronique, Electrotechnique, Automatique Transfert*. Laboratoire CREATIS-LRMN, CNRS UMR 5220, Inserm U6. Institut National des Sciences Appliquées de Lyon, France. Disponible en ligne : http://theses.insa-lyon.fr/publication/2010ISAL0069/these.pdf.

Bhatt, D., Aggarwal, P., Bhattacharya, P., & Devabhaktuni, V. (2012). An enhanced mems error modeling approach based on nu-support vector regression. *Sensors*, *12*(7), 9448-9466.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press, New York, USA.

Bitam, S. (2018). "*Modélisation de l'activité thérapeutique à partir de la structure moléculaire*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

BMDRC (2019). Bioinformatics and Molecular Design Research Center. Page d'accueil : *http://bmdrc.org/website/*. PreADMET Software version 2.0. Disponible en ligne : *https://preadmet.bmdrc.kr/preadmet-pc-version-2-0/*. (Consulté le 08/09/2019).

Bonachera, F. (2011). "*Les triplets pharmacophoriques ous Développement et applications*". Thèse de doctorat en Chimie théorique, physique et analytique. Ecole doctorale biologie santé, Université des sciences et sechnologies de Lille 1, France. Disponible en ligne : https://pdfs.semanticscholar.org/1da9/5b49c738012b8b229ec0765cdc84410481d5.pdf.

Bonchev, D. (1991). *Chemical graph theory: introduction and fundamentals* (Vol. 1). CRC Press.

Borhani, T. N. G., Afzali, A., & Bagheri, M. (2016). QSPR estimation of the auto-ignition temperature for pure hydrocarbons. *Process Safety and Environmental Protection*, *103*, 115-125.

Bosque, R., Sales, J., Bosch, E., Roses, M., García-Alvarez-Coque, M. C., & Torres-Lapasió, J. R. (2003). A QSPR study of the p solute polarity parameter to estimate retention in HPLC. *Journal of chemical information and computer sciences*, *43*(4), 1240-1247.

BP, B. P. (2012). Statistical Review of World Energy, June 2010. *British Petroleum*, *201*(6).

Brown, A. C., & Fraser, T. R. (1868). V.—On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *25*(1), 151-203.

Brown, A. C., & Fraser, T. R. (1869). XX.—On the Connection between Chemical Constitution and Physiological Action. Part II. —On the Physiological Action of the Ammonium Bases derived from Atropia and Conia. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *25*(2), 693-739.

Carroll, F. A., Godinho, J. M., & Quina, F. H. (2011). Development of a simple method to predict boiling points and flash points of acyclic alkenes. *Industrial & engineering chemistry research*, *50*(24), 14221-14225.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.

Chamekh, A. (2007). "*Optimisation des procédés de mise en forme par les réseaux de neurones artificiels*". Thèse de doctorat en génie mécaniques, Cotutelle entre *Laboratoire de Génie Mécanique*, Ecole Nationale d'Ingénieurs de Monastir, Tunisie ; et *Laboratoire en Sûreté de fonctionnement, Qualité et Organisation*, Institut des Sciences et Techniques de l'Ingénieur d'Angers, France. Disponible en ligne : https://tel.archives-ouvertes.fr/tel-00445341.

Charton, M. (2002). *Advances in Quantative Structure-Property Relationships*. Elsevier.

ChemOffice (2012). PerkinElmer Informatics. CambridgeSoft. 100 CambridgePark Drive, Cambridge, MA 02140. *www.cambridgesoft.com*.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Consonni, V. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, *57*(12), 4977-5010.

Chitour, C. E. (1999). *Raffinage: les propriétés physiques des hydrocarbures et des fractions pétrolières*. Office des publications universitaires.

Cholakov, G. S., Wakeham, W. A., & Stateva, R. P. (1999). Estimation of normal boiling points of hydrocarbons from descriptors of molecular structure. *Fluid Phase Equilibria*, *163*(1), 21-42.

Chtita, S. (2017). *Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments*. Thèse de doctorat en Chimie. Faculté des sciences, Université Moulay Ismaïl. Meknès, Maroc.

Constantinou, L., & Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, *40*(10), 1697-1710.

Cousins, K. R. (2011). Computer review of ChemDraw Ultra 12.0. *Journal of the American Chemical Society*, *133*(21), 8388-8388.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, *110*(18), 5959-5967.

Cros, A. F. A. (1863). Action de l'alcool amylique sur l'organisme, *Thèse de Doctorat, Faculté de Médecine, Université Strasbourg, Strasbourg*.

Cruciani, G., Pastor, M., & Guba, W. (2000). VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *European Journal of Pharmaceutical Sciences*, *11*, S29-S39.

Csizmadia, P. (1999). MarvinSketch and MarvinView: molecule applets for the World Wide Web. In *Proceedings of ECSOC-3, the third international electronic conference on synthetic organic chemistry, September 1ą30* (p. 367ą369).

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

Dai, Y. M., Zhu, Z. P., Cao, Z., Zhang, Y. F., Zeng, J. L., & Li, X. (2013). Prediction of boiling points of organic compounds by QSPR tools. *Journal of Molecular Graphics and Modelling*, *44*, 113-119.

Daubert, T. E., & Danner, R. P. (1997). API Technical Data Book Petroleum Refining: American Petroleum Institute (API): Washington, DC.

Dearden, J. C., Cronin, M. T., & Kaiser, K. L. (2009). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, *20*(3-4), 241-266.

Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., ... & Chen, A. F. (2015). ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of cheminformatics*, *7*(1), 60.

Dreyfus, G. (2005). Neural networks: methodology and applications. *Springer Science & Business Media*.

Dreyfus, G., Martinez, J. M., Samuelides, M., Gordon, M. B., Badran, F., Thiria, S., & Hérault, L. (2002). Réseaux de neurones-Méthodologie et applications. *(No. BOOK). Eyrolles*.

Dreyfus, G., Martinez, J. M., Samuelides, M., Gordon, M. B., Badran, F., & Thiria, S. (2008). Apprentissage statistique : Réseaux de neurones : Cartes topologiques Machines à vecteurs supports, eyrolles.

Duch, W., Swaminathan, K., & Meller, J. (2007). Artificial intelligence approaches for rational drug design and discovery. *Current pharmaceutical design*, *13*(14), 1497-1508.

Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening*, *9*(3), 213-228.

Dudley, B. (2018). BP statistical review of world energy. *BP Statistical Review, London, UK, accessed Aug*, *6*, 2018.

Dupuy, A. (2015). "Formation professionnalisante PRO/RCP1 Groupe 01, Ingénieurs procédés en raffinage et pétrochimie hydrocarbures et produits pétrolières", Centre IAP, Hassi Messaoud.

Elsharkawy, A. M., & Elkamel, A. (2000). Compressibility factor for sour gas reservoirs. In *SPE Asia Pacific Oil and Gas Conference and Exhibition*. Society of Petroleum Engineers.

Elsharkawy, A. M., Hashem, Y. S. K. S., & Alikhan, A. A. (2001). Compressibility factor for gas condensates. *Energy & fuels*, *15*(4), 807-816.

Engel, T., & Gasteiger, J. (Eds.). (2018). *Applied chemoinformatics: achievements and future opportunities*. John Wiley & Sons.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives*, *111*(10), 1361-1375.

Espinosa, G., Yaffe, D., Arenas, A., Cohen, Y., & Giralt, F. (2001). A fuzzy ARTMAP-based quantitative structure− property relationship (QSPR) for predicting physical properties of organic compounds. *Industrial & engineering chemistry research*, *40*(12), 2757-2766.

Eyssautier, J. (2012). *Caractérisation et Modélisation des Asphaltènes en Conditions Réactionnelles d'Hydrotraitement*. Thèse de doctorat. Palaiseau, Ecole polytechnique. France.

Fahim, M. A., Al-Sahhaf, T. A., & Elkilani, A. (2009). *Fundamentals of petroleum refining*. Elsevier.

Farzi, R., & Esmaeilzadeh, F. (2016). Prediction of surface tension of pure hydrocarbons using Esmaeilzadeh-Roshanfekr equation of state and group contribution method. *Fluid Phase Equilibria*, *427*, 353-361.

Free, S. M., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, *7*(4), 395-399.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., ... & Li, X. (2016). Gaussian 16. *Revision A*, *3*.

Fukui, K. (1970). Theory of orientation and stereoselection. In *Orientation and Stereoselection* (pp. 1-85). Springer, Berlin, Heidelberg.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 160(3), 249-264.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452-459.

Gharagheizi, F., Eslamimanesh, A., Mohammadi, A. H., & Richon, D. (2011). Determination of critical properties and acentric factors of pure compounds using the artificial neural network group contribution algorithm. *Journal of Chemical & Engineering Data*, *56*(5), 2460-2476.

Gharagheizi, F., Mirkhani, S. A., Ilani-Kashkouli, P., Mohammadi, A. H., Ramjugernath, D., & Richon, D. (2013). Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy: application to a very large dataset. *Fluid Phase Equilibria*, *354*, 250-258.

Ghasemitabar, H., & Movagharnejad, K. (2016). Estimation of the normal boiling point of organic compounds via a new group contribution method. *Fluid Phase Equilibria*, *411*, 13-23.

Godavarthy, S. S., Robinson Jr, R. L., & Gasem, K. A. (2008). Improved structure–property relationship models for prediction of critical properties. *Fluid Phase Equilibria*, *264*(1-2), 122-136.

Goh, A. T. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3), 143-151.

Goulon-Sigwalt-Abram, A. (2008). "*Une nouvelle méthode d'apprentissage de données structurées: applications à l'aide à la découverte de médicaments*" Thèse de doctorat en Electronique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://pastel.archives-ouvertes.fr/pastel-00004103/.

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & combinatorial science*, *26*(5), 694-701.

Groenzin, H., & Mullins, O. C. (2000). Molecular size and structure of asphaltenes from various sources. *Energy & Fuels*, *14*(3), 677-684.

Gurbanov, T. (2007). *Le pétrole de la Caspienne et la politique extérieure de l'Azerbaïdjan: Tome 1-Questions économiques et juridiques*. Editions L'Harmattan.

Gurney, K. (1997). An Introduction to Neural Networks, Routledge, *ISBN 1-85728-673-1 London*.

Guthrie, V. B., & Guthrie, V. B. (1960). *Petroleum products handbook* (Vol. 33). New York/Toronto/London: McGraw-Hill.

Ha, Z., Ring, Z., & Liu, S. (2005). Quantitative Structure− Property Relationship (QSPR) Models for Boiling Points, Specific Gravities, and Refraction Indices of Hydrocarbons. *Energy & fuels*, *19*(1), 152-163.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). Neural network design, PWS Pub. *Co., Boston*, *3632*.

Halali, M. A., Azari, V., Arabloo, M., Mohammadi, A. H., & Bahadori, A. (2016). Application of a radial basis function neural network to estimate pressure gradient in water–oil pipelines. *Journal of the Taiwan Institute of Chemical Engineers*, *58*, 189-202.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Hamadache, M. (2014). "*Modélisation en vue de la prédiction ou de la corrélation de l'activité toxicologique à partir de la structure moléculaire*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Hamadache, M., Amrane, A., Hanini, S., & Benkortbi, O. (2018). Multilayer perceptron model for predicting acute toxicity of fungicides on rats. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, *3*(1), 100-118.

Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L., & Moussa, C. S. (2016a). A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction. Journal of hazardous materials, 303, 28-40.

Hamadache, M., Hanini, S., Benkortbi, O., Amrane, A., Khaouane, L., & Moussa, C. S. (2016b). Artificial neural network-based equation to predict the toxicity of herbicides on rats. Chemometrics and Intelligent Laboratory Systems, 154, 7-15.

Hamadache, M., Khaouane, L., Benkortbi, O., Si Moussa, C., Hanini, S., & Amrane, A. (2014). Prediction of acute herbicide toxicity in rats from quantitative structure–activity relationship modeling. Environmental Engineering Science, 31(5), 243-252.

Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616-1626.

Hansen, P. J., & Jurs, P. C. (1988). Chemical applications of graph theory. Part I. Fundamentals and topological indices. *Journal of Chemical Education*, *65*(7), 574.

Hassanzadeh, P., Atyabi, F., & Dinarvand, R. (2019). The significance of artificial intelligence in drug delivery system design. *Advanced Drug Delivery Reviews*, *151*, 169-190.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. MIT press.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1-12.

Haykin, S. (1998). Neural Networks: A Comprehensive Foundation, Prentice Hall PTR. *Upper Saddle River, NJ, USA*.

Haykin, S. (2010). *Neural Networks and Learning Machines, 3/E*. Pearson Education India.

Hemmati-Sarapardeh, A., Varamesh, A., Husein, M. M., & Karan, K. (2018). On the evaluation of the viscosity of nanofluid systems: Modeling and data assessment. *Renewable and Sustainable Energy Reviews*, *81*, 313-329.

Henry, J. T. (1873). *The early and later history of petroleum, with authentic facts in regard to its development in western Pennsylvania* (Vol. 108). Burt Franklin.

Higo, J. I., & Gō, N. (1989). Algorithm for rapid calculation of excluded volume of large molecules. *Journal of computational chemistry*, *10*(3), 376-379.

Jahn, M. (2020). Artificial neural network regression models in a panel setting: Predicting economic growth. *Economic Modelling*, *91*, 148-154.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31-44.

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Alternatives to laboratory animals*, *33*(5), 445-459.

Jensen, F. (2007). Electron correlation methods. *Jensen F. Introduction to computational chemistry. 2nd ed. New Jersey: John Wiley & Sons*, 133-144.

Jiao, L. (2010). QSPR studies on soot–water partition coefficients of persistent organic pollutants by using artificial neural network. *Chemosphere*, *80*(6), 671-675.

Karelson, M. (2000). *Molecular descriptors in QSAR/QSPR* (Vol. 230). New York: Wiley-Interscience.

Katritzky, A. R., Petrukhin, R., Jain, R., & Karelson, M. (2001). QSPR analysis of flash points. *Journal of chemical information and computer sciences*, *41*(6), 1521-1530.

KDB, (2017). Korean Thermophysical Properties Data Bank–Cheric (KDB). Disponible en ligne : http://www.cheric.org/research/kdb/, (Consulté le 01/01/2017).

Kerwin, S. M. (2010). ChemBioOffice Ultra 2010 suite. *Journal of the American Chemical Society*, *132*(7), 2466-2467.

Khaouane, L. (2013). "*Etude et Modélisation de la Biosynthèse des Antibiotiques à partir de Différentes Souches Productrices – Cas de Pleuromutiline*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Kikuchi, O. (1987). Systematic QSAR procedures with quantum chemical descriptors. *Quantitative Structure-Activity Relationships*, *6*(4), 179-184.

Kode Chemoinformatics srl (2019). DRAGON Professional 7 program for Windows and Linux. Registered office : Via Nino Pisano, 14 - 56122 Pisa (PI) – Italy VAT : 02040400505 Main office : Lungarno Antonio Pacinotti, 50 - 56126 Pisa (PI) – Italy Contact us at the email address: *chm@kode-solutions.net*.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *The Journal of Machine Learning Research*, *18*(1), 826-830.

Kraim, K. (2009). *"Etudes de QSAR sur des activités biologiques utilisant des produits d'origines naturels"*. Thèse de doctorat en génie des procédés, Option : Chimie industrielle, Département de génie des procédés, Université 08 mai 1945 Guelma. Guelma, Algérie.

Kraim, K., Khatmi, D., Saihi, Y., Ferkous, F., & Brahimi, M. (2009). Quantitative structure activity relationship for the computational prediction of α-glucosidase inhibitory. Chemometrics and Intelligent Laboratory Systems, 97(2), 118-126.

Krose, B., & Smagt, P. V. (1996). An Introduction to Neural Network. The University of Amsterdam. *Amesterdam, Netherland*.

Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, *18*(4-5), 464-477.

Lafdani, E. K., Nia, A. M., & Ahmadi, A. (2013). Daily suspended sediment load prediction using artificial neural networks and support vector machines. *Journal of Hydrology*, *478*, 50-62.

Lahiouel, Y. (1995a). "*Estimation de la température d'ébullition et de facteur acentrique des composés organiques par les méthodes de contribution de groupes et les indices topologiques*". Mémoire de Magister en Chimie Industrielle, Option : Conception Assistée par Ordinateur, Ancien Institut de Chimie Industrielle, Ancien Centre Universitaire de Guelma. Guelma, Algérie.

Lahiouel, Y. (1995b). Industrial Experimental Dataset in Petroleum Fractions and Pure Hydrocarbons, Refinery of Skikda, Skikda, Algeria.

Laidi, M. (2009). "*Approche Neuronale Pour l'Estimation du coefficient de Diffusion Moléculaire Apparent dans un Milieu Poreux Anisotrope (Bois)*". Mémoire de Magister en Génie mécanique, Option : Energétique, Ancien Faculté des Sciences et de la Technologie, Université Yahia Fares de Médéa. Médéa, Algérie.

Laidi, M. (2013). "*Etude et Modélisation d'un Réfrigérateur Solaire*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Lewis, D. F. V., Ioannides, C., & Parke, D. V. (1994). Interaction of a series of nitriles with the alcohol-inducible isoform of P450: computer analysis of structure—activity relationships. *Xenobiotica*, *24*(5), 401-408.

Li, H., Higashi, H., & Tamura, K. (2006). Estimation of boiling and melting points of light, heavy and complex hydrocarbons by means of a modified group vector space method. *Fluid phase equilibria*, *239*(2), 213-222.

Li, Z., Wu, W., & Chen, L. (2018). A new method for calculating normal boiling point of liquids. *Russian Chemical Bulletin*, *67*(10), 1823-1830.

Li, Z., Zuo, L., Wu, W., & Chen, L. (2017). The new method for correlation and prediction of thermophysical properties of fluids. critical temperature. *Journal of Chemical & Engineering Data*, *62*(11), 3723-3731.

Lucea, M. (2006). "*Modélisation dynamique par réseaux de neurones et machines à vecteurs supports : contribution à la maîtrise des émissions polluantes de véhicules automobiles*". Thèse de doctorat en Electronique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://pastel.archives-ouvertes.fr/pastel-00001943.

Manso, F. C., Júnior, H. S., Bruns, R. E., Rubira, A. F., & Muniz, E. C. (2012). Development of a new topological index for the prediction of normal boiling point temperatures of hydrocarbons: The Fi index. *Journal of Molecular Liquids*, *165*, 125-132.

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., & Tropsha, A. (2012). Does rational selection of training and test sets improve the outcome of QSAR modeling?. Journal of chemical information and modeling, 52(10), 2570-2578.

Martins, J. P. A., Barbosa, E. G., Pasqualoto, K. F., & Ferreira, M. M. (2009). LQTA-QSAR: a new 4D-QSAR methodology. *Journal of chemical information and modeling*, *49*(6), 1428-1436.

Masand, V. H., & Rastija, V. (2017). PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors. Chemometrics and Intelligent Laboratory Systems, 169, 12-18.

Masseron, J. (1982). L'economie des hydrocarbures. *Science et technique du pétrole*.

Masseron, J. (1990). *Petroleum economics*. Editions Technip.

Masunov, A. (2001). ACD/I-Lab 4.5: an internet service review. *Journal of chemical information and computer sciences*, *41*(4), 1093-1095.

Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular descriptors. *Handbook of Computational Chemistry*, 2065-2093.

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, *56*(2), 237-248.

Merdrignac, I., & Espinat, D. (2007). Physicochemical characterization of petroleum fractions: the state of the art. *Oil & Gas Science and Technology-Revue de l'IFP*, *62*(1), 7-32.

Meyer, H. (1899). Zur theorie der alkoholnarkose. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *42*(2), 109-118.

Mihalić, Z., & Trinajstić, N. (1992). A graph-theoretical approach to structure-property relationships. *Journal of chemical education*, *69*(9), 701-712.

Minoux, M. (1983). Programmation mathematique. Theorie et algorithmes dunod. *Minoux–Paris : Bordas et CNETENST*.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8-30.

Mohammadi, A. H., Afzal, W., & Richon, D. (2008). Determination of critical properties and acentric factors of petroleum fractions using artificial neural networks. *Industrial & engineering chemistry research*, *47*(9), 3225-3232.

Nascimento, M. H., Oliveira, B. P., Rainha, K. P., Castro, E. V., Silva, S. R., & Filgueiras, P. R. (2018). Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics. *Fuel*, *234*, 643-649.

Navajas, C., Poso, A., Tuppurainen, K., & Gynther, J. (1996). Comparative Molecular Field Analysis (CoMFA) of MX Compounds using different Semi-empirical Methods: LUMO Field and its Correlation with Mutagenic Activity. *Quantitative Structure-Activity Relationships*, *15*(3), 189-193.

Nieto-Draghi, C., Fayet, G., Creton, B., Rozanska, X., Rotureau, P., de Hemptinne, J. C., ... & Adamo, C. (2015). A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes. *Chemical reviews*, *115*(24), 13093-13164.

Oussar, Y. (1998). "*Réseaux d'ondelettes et réseaux de neurones pour la modélisation statique et dynamique de processus*". Thèse de doctorat en Robotique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://hal.archives-ouvertes.fr/pastel-00000677.

Overton, C. E. (1901). *Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie*. Fischer.

Pan, Y., Jiang, J., & Wang, Z. (2007). Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network. *Journal of hazardous materials*, *147*(1-2), 424-430.

Pan, Y., Jiang, J., Wang, R., Cao, H., & Cui, Y. (2009). Predicting the auto-ignition temperatures of organic compounds from molecular structure using support vector machine. *Journal of hazardous materials*, *164*(2-3), 1242-1249.

Pao, H. T., & Fu, H. C. (2013). Renewable energy, non-renewable energy and economic growth in Brazil. *Renewable and Sustainable Energy Reviews*, *25*, 381-392.

Papa, E., Villa, F., & Gramatica, P. (2005). Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales p romelas (Fathead Minnow). *Journal of chemical information and modeling*, *45*(5), 1256-1266.

Perrodon, A. (1985). *Histoire des grandes découvertes pétrolières: un certain art de l'exploration* (Vol. 10). Editions OPHRYS.

Pigram, G. M., & MacDonald, T. R. (2001). Use of neural network models to predict industrial bioreactor effluent quality. *Environmental science & technology*, *35*(1), 157-162.

Pratim Roy, P., Leonard, J. T., & Roy, K. (2008). Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems, 90(1), 31-42.

Pratim Roy, P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. *Molecules*, *14*(5), 1660-1701.

Randić, M. (1975). Characterization of molecular branching. *Journal of the American Chemical Society*, *97*(23), 6609-6615.

Randić, M. (1991). Generalized molecular descriptors. *Journal of Mathematical Chemistry*, *7*(1), 155-168.

Ravi, M., Hopfinger, A. J., Hormann, R. E., & Dinan, L. (2001). 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *Journal of chemical information and computer sciences*, *41*(6), 1587-1604.

Rebas, O. (2006). "*Etablissement de nouvelles corrélations de contribution de groupes avec interactions pour l'estimation des propriétés physiques des hydrocarbures et de leurs mélanges*", Mémoire de Magister en Génie des Procédés, Option : Génie Chimique, École Nationale Polytechnique (ENP), Département Génie Chimique, Alger.

Rebas, O. (2016). "*Évaluation des propriétés des pétroles bruts et des fractions pétrolières par les méthodes de contribution de groupes avec interactions moléculaires*", Thèse de doctorat en Génie des Procédés, Option : Génie Chimique, Ecole Nationale Polytechnique (ENP), Département Génie Chimique, Laboratoire de Valorisation des Energies Fossiles, Alger.

Retzekas, E., Voutsas, E., Magoulas, K., & Tassios, D. (2002). Prediction of physical properties of hydrocarbons, petroleum, and coal liquid fractions. *Industrial & engineering chemistry research*, *41*(6), 1695-1702.

Riazi, M. R. (2005). *Characterization and properties of petroleum fractions* (Vol. 50). ASTM international.

Riazi, M. R., & Roomi, Y. A. (2001). Use of the refractive index in the estimation of thermophysical properties of hydrocarbons and petroleum mixtures. *Industrial & engineering chemistry research*, *40*(8), 1975-1984.

Richet, M. C. (1893). Note sur le rapport entre la toxicité et les propriétés physiques des corps. *Compt Rend Soc Biol (Paris)*, *45*, 775-6.

Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120.

Roussel J.C et Boulet R. (1994). "Pétrole bruit, produits pétroliers, schémas de fabrication, Cha01 : Composition des pétroles bruts et des produits pétrolières", EDITIONS TECHNIP, Paris-P02

Roy, K., Kar, S., & Das, R. N. (2015). *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Saaidpour, S., & Ghaderi, F. (2016). Quantitative Modeling of Physical Properties of Crude Oil Hydrocarbons Using Volsurf+ Molecular Descriptors. *structure*, *22*, 26.

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2014). Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective. Current computer-aided drug design, 10(2), 137-147.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791-4810.

Samarasinghe, S. (2006). Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. *Auerbach publications, Taylor & Francis, Group. Boca Raton New York, USA*. ISBN : 978-0-8493-3375-0.

Santos, S. M., Nascimento, D. C., Costa, M. C., Neto, A. M., & Fregolente, L. V. (2020). Flash point prediction: Reviewing empirical models for hydrocarbons, petroleum fraction, biodiesel, and blends. *Fuel*, *263*, 116375.

Sarkhosh, M., Ghasemi, J. B., & Ayati, M. (2012). A quantitative structure-property relationship of gas chromatographic/mass spectrometric retention data of 85 volatile organic compounds as air pollutant materials by multivariate methods. *Chemistry Central Journal*, *6*(2), S4.

Saunders, R. A., & Platts, J. A. (2004). Scaled polar surface area descriptors: development and application to three sets of partition coefficients. *New Journal of Chemistry*, *28*(1), 166-172.

Sciencedirect (2020). Plateforme des publications scientifiques, lien : consulté le (15/10/2020).

SIGMA. (2020). Laboratoire SIGnaux, Modèles, Apprentissage statistique. *10, Rue Vauquelin 75231 PARIS Cedex 05*, France. Site Officiel : https://www.neurones.espci.fr/. (Consulté le 10/02/2020).

Si-Moussa, C. (2010). "*Modélisation des équilibres de phases à pressions élevées par les réseaux de neurones artificiels*". Thèse de doctorat. École nationale supérieure polytechnique. Alger. Algérie.

Slaoui, K. H. (1986). "*Application des techniques mathématiques à la gestion des mélanges : histosplines et optimisation*", Thèse de doctorat en Mathématiques Appliquées, Institut National Polytechnique de Grenoble (Grenoble INPG), France. Disponible en ligne : https://hal.archives-ouvertes.fr/tel-00320471.

Sola, D., Ferri, A., Banchero, M., Manna, L., & Sicardi, S. (2008). QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method. *Fluid Phase Equilibria*, *263*(1), 33-42.

Souahi, F., & Kaabeche, H. (2008). Developing correlations for prediction of petroleum fraction properties using genetic algorithms. *Oil & Gas Science and Technology-Revue de l'IFP*, *63*(2), 229-237.

Speight, J. G. (2015). *Handbook of petroleum product analysis*. John Wiley & Sons.

Spessard, G. O. (1998). ACD Labs/LogP dB 3.5 and ChemSketch 3.5. *Journal of chemical information and computer sciences*, *38*(6), 1250-1253.

StatSoft (2007). STATISTICA (data analysis software system) vs 8.0. StatSoft Inc, Tulsa, OK.

Statsoft. (2020a). Statsoft.documentation *STATISTICA Help SANN Overviews*. Disponible en ligne :*https://documentation.statsoft.com/STATISTICAHelp.aspx?path=SANN/Overview/SANNOverviewsActivationFunctions*. (Consulté le 14/01/2020).

Statsoft. (2020b). Concepts Fondamentaux en Statistique- Réseaux de Neurones- Fonctions d'Activation. Disponible en ligne : *http://www.statsoft.fr/concepts-statistiques/reseaux-de-neurones-automatises/reseaux-de-neurones-automatises.htm#fonctions*. (Consulté le 14/01/2020).

Sugden, S. (1924). VI.—The variation of surface tension with temperature and some related functions. *Journal of the Chemical Society, Transactions*, *125*, 32-41.

Sutton, R. P. (1985). Compressibility factors for high-molecular-weight reservoir gases. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Suzuki, K. (Ed.). **(2011)**. Artificial neural networks: methodological advances and biomedical applications. *BoD–Books on Demand*.

Tahami, S., Movagharnejad, K., & Ghasemitabar, H. (2019). Estimation of the critical constants of organic compounds via a new group contribution method. *Fluid Phase Equilibria*, *494*, 45-60.

TALETE srl (2007). DRAGON Professional 5.4 program for Windows and Linux. Talete srl, Milano, Italy. *URL: http://www.talete.mi.it/products/dragon_description.htm*.

Taskin, A., & Guneri, A. F. (2006). Economic analysis of risky projects by ANNs. *Applied mathematics and computation*, *175*(1), 171-181.

Tauler, R., Walczak, B., & Brown, S. D. (2009). *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier.

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., ... & Tanchuk, V. Y. (2005). Virtual computational chemistry laboratory–design and description. *Journal of computer-aided molecular design*, *19*(6), 453-463.

The MathWorks, Inc. USA (2013), Matlab software, Version 2013b.

Todeschini, R., & Consonni, V. (2008). *Handbook of molecular descriptors* (Vol. 11). John Wiley & Sons.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons.

Topliss, J. G., & Edwards, R. P. (1979). Chance factors in studies of quantitative structure-activity relationships. Journal of Medicinal Chemistry, 22(10), 1238-1244.

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, *22*(1), 69-77.

Tsakovska, I., Pajeva, I., Al Sharif, M., Alov, P., Fioravanzo, E., Kovarich, S., ... & Cronin, M. T. (2017). Quantitative structure-skin permeability relationships. *Toxicology*, *387*, 27-42.

Tugcu, C. T., Ozturk, I., & Aslan, A. (2012). Renewable and non-renewable energy consumption and economic growth relationship revisited: evidence from G7 countries. *Energy economics*, *34*(6), 1942-1950.

Turner, B. E., Costello, C. L., & Jurs, P. C. (1998). Prediction of critical temperatures and pressures of industrially important organic compounds from molecular structure. *Journal of chemical information and computer sciences*, *38*(4), 639-645.

Urbani, D. (1995). "*Méthodes statistiques de sélection d'architectures neuronales : application à la conception de modèles de processus dynamiques*". Thèse de doctorat en Microélectronique et Microinformatique, Université Pierre et Marie Curie-Paris VI, France. Disponible en ligne : https://www.neurones.espci.fr/Theses_PS/Urbani_D/these.pdf.

Varamesh, A., Hemmati-Sarapardeh, A., Dabir, B., & Mohammadi, A. H. (2017a). Development of robust generalized models for estimating the normal boiling points of pure chemical compounds. *Journal of Molecular Liquids*, *242*, 59-69.

Varamesh, A., Hemmati-Sarapardeh, A., Moraveji, M. K., & Mohammadi, A. H. (2017b). Generalized models for predicting the critical properties of pure chemical compounds. *Journal of Molecular Liquids*, *240*, 777-793.

Vatani, A., Mehrpooya, M., & Gharagheizi, F. (2007). Prediction of standard enthalpy of formation by a QSPR model. *International Journal of Molecular Sciences*, *8*(5), 407-432.

VCCLAB, (2005). Virtual Computational Chemistry Laboratory. Disponible en ligne : *http://www.vcclab.org*, (Consulté le 15/02/2018).

Vidal, J. (1997). *Thermodynamique : application au génie chimique et à l'industrie pétrolière*. Editions Technip.

Vieira, A. P., Portela, N. A., Neto, A. C., Lacerda Jr, V., Romao, W., Castro, E. V., & Filgueiras, P. R. (2019). Determination of physicochemical properties of petroleum using 1H NMR spectroscopy combined with multivariate calibration. *Fuel*, *253*, 320-326.

Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of chemical information and computer sciences*, *29*(3), 163-172.

Wakeham, W. A., Cholakov, G. S., & Stateva, R. P. (2002). Liquid density and critical properties of hydrocarbons estimated from molecular structure. *Journal of Chemical & Engineering Data*, *47*(3), 559-570.

Wauquier, J. P. (1998). *Raffinage du pétrole (Le). Tome 1. Pétrole brut. Produits pétroliers. Schémas de fabrication* (Vol. 1). Editions Technip.

Weaver, S., & Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, *26*(8), 1315-1326.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

Weininger, D. (1990). SMILES. 3. DEPICT. Graphical depiction of chemical structures. *Journal of chemical information and computer sciences*, *30*(3), 237-243.

Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences*, *29*(2), 97-101.

Wessel, M. D., & Jurs, P. C. (1994). Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks. Analytical Chemistry, 66(15), 2480-2487.

Wessel, M. D., & Jurs, P. C. (1995). Prediction of normal boiling points of hydrocarbons from molecular structure. *Journal of chemical information and computer sciences*, *35*(1), 68-76.

White, H. (1992). *Artificial neural networks: approximation and learning theory*. Blackwell Publishers, Inc.

Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, *69*(1), 17-20.

Wu, K., Natarajan, B., Morkowchuk, L., Krein, M., & Breneman, C. M. (2013). From drug discovery QSAR to predictive materials QSPR: the evolution of descriptors, methods, and models. In *Informatics for Materials Science and Engineering* (pp. 385-422). Butterworth-Heinemann.

XLSTAT (2009). Add-in software, XLSTAT Company. *www.xlstat.com*.

Yan, X., Lan, T., Jia, Q., Yan, F., & Wang, Q. (2020). A norm index-based QSPR model to predict the standard absolute entropy of organic compounds in three phase states. *Fluid Phase Equilibria*, *526*, 112815.

Yao, X., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2002a). Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemometrics and Intelligent Laboratory Systems*, *62*(2), 217-225.

Yao, X., Zhang, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2002b). Radial basis function neural network based QSPR for the prediction of critical pressures of substituted benzenes. *Computers & chemistry*, *26*(2), 159-169.

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, *32*(7), 1466-1474.

Zhou, L., Wang, B., Jiang, J., Pan, Y., & Wang, Q. (2017). Predicting the gas-liquid critical temperature of binary mixtures based on the quantitative structure property relationship. *Chemometrics and Intelligent Laboratory Systems*, *167*, 190-195.

Zhou, Z., & Parr, R. G. (1990). Activation hardness: new index for describing the orientation of electrophilic aromatic substitution. *Journal of the American Chemical Society*, *112*(15), 5720-5724.

Zuas, O., & Styarini, D. (2009). A quantitative structure-property relationship (QSPR) evaluation of critical volume of unsaturated hydrocarbon alkenes and alkynes using simple connectivity indices. *Reaktor*, *12*(4), 260-267.

Annexes

Annexe A.

Tableau A.1. Valeurs expérimentales et prédites par les modèles RQSP-RNA des cinq propriétés physicochimiques des hydrocarbures ainsi leurs ordres de partition dans les bases d'apprentissage et de test pour chaque paramètre.

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
1	184,6[c]	183,6967	1st Test	0,339[b]	0,333782	1st Test	305,3[b]	307,7842	1st Test	4872000[b]	4914558	1st Test	145,5[b]	140,228	1st Test
2	231,1[c]	231,168	1st Train	0,5004[b]	0,496451	1st Train	369,8[b]	366,1707	1st Train	4248000[b]	4270566	1st Train	200[b]	204,841	1st Train
3	272,65[a]	272,2369	2nd Test	0,5788[a]	0,577582	2nd Test	425,18[a]	426,4263	2nd Test	3796900[a]	3824115	2nd Test	254,9[a]	253,613	2nd Test
4	261,43[a]	260,6658	2nd Train	0,5572[a]	0,561857	2nd Train	408,14[a]	409,0721	2nd Train	3648000[a]	3711429	2nd Train	262,7[a]	263,432	2nd Train
5	309,22[a]	308,4504	3rd Train	0,62622[a]	0,625311	3rd Train	469,7[a]	472,6487	3rd Train	3370000[a]	3355878	3rd Train	313[a]	308,444	3rd Train
6	300,99[a]	299,7926	4th Train	0,61965[a]	0,62323	4th Train	460,43[a]	462,6757	4th Train	3381200[a]	3360704	4th Train	305,83[a]	313,836	4th Train
7	341,88[a]	341,3421	3rd Test	0,65935[a]	0,657016	3rd Test	507,6[a]	507,9008	3rd Test	3025000[a]	3064157	3rd Test	371[a]	366,984	3rd Test
8	371,58[a]	368,9663	5th Train	0,68374[a]	0,683784	5th Train	540,2[a]	538,9524	5th Train	2740000[a]	2725580	5th Train	428[a]	431,049	5th Train
9	398,83[a]	398,0236	6th Train	0,7025[a]	0,701609	6th Train	568,7[a]	575,4082	6th Train	2490000[a]	2486371	6th Train	486[a]	496,333	6th Train
10	423,97[a]	423,0077	7th Train	0,71761[a]	0,716901	7th Train	594,6[a]	599,071	7th Train	2290000[a]	2282991	7th Train	544[a]	561,839	7th Train
11	447,3[a]	446,3276	8th Train	0,73003[a]	0,728395	8th Train	617,7[a]	618,7557	8th Train	2110000[a]	2104180	8th Train	600[a]	597,396	8th Train
12	469,08[a]	468,071	9th Train	0,7402[a]	0,738767	9th Train	639[a]	638,7122	9th Train	1930000[a]	1957730	9th Train	645[a]	666,128	9th Train
13	489,5[c]	488,9911	10th Train	0,7488[b]	0,747086	10th Train	658[b]	658,507	10th Train	1820000[b]	1810631	10th Train	754[b]	739,007	10th Train
14	508,6[c]	508,529	11th Train	0,7561[b]	0,754888	11th Train	675[b]	675,5256	11th Train	1680000[b]	1699887	11th Train	823[b]	815,964	11th Train
15	526,7[c]	526,8834	12th Train	0,7626[b]	0,76096	12th Train	693[b]	690,4507	12th Train	1570000[b]	1567875	12th Train	894[b]	892,213	12th Train
16	543,8[c]	543,7257	13th Train	0,7684[b]	0,767006	13th Train	708[b]	704,8853	13th Train	1480000[b]	1484130	13th Train	966[b]	963,505	13th Train
17	560[c]	559,78	4th Test	0,7735[b]	0,771987	4th Test	723[b]	719,5922	4th Test	1400000[b]	1367448	4th Test	1034[b]	1032,655	4th Test
18	575,3[c]	575,2468	14th Train	0,778[b]	0,776488	14th Train	736[b]	733,58	14th Train	1340000[b]	1302771	14th Train	1103[b]	1105,739	14th Train
19	589,9[c]	589,9392	15th Train	0,7828[b]	0,780849	15th Train	747[b]	745,8739	15th Train	1290000[b]	1204041	15th Train	1189[b]	1178,711	15th Train
20	603[c]	603,7416	16th Train	0,7835[b]	0,784553	16th Train	755[b]	756,5611	16th Train	1160000[b]	1154806	16th Train	1249[b]	1252,145	16th Train
21	616,9[c]	616,9289	17th Train	0,7843[b]	0,787784	17th Train	768[b]	767,1814	17th Train	1070000[b]	1069843	17th Train	1319[b]	1321,702	17th Train
22	629,6[c]	629,5442	18th Train	0,7916[b]	0,791275	18th Train	778[b]	777,8079	18th Train	1030000[b]	1031342	18th Train	1390[b]	1388,25	18th Train

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
23	641,8[c]	642,1874	5th Test	0,7943[b]	0,794074	5th Test	786[b]	787,454	5th Test	980000[b]	963473	5th Test	1461[b]	1457,964	5th Test
24	653,3[c]	654,2211	19th Train	0,7968[b]	0,796732	19th Train	790[b]	795,572	19th Train	920000[b]	932717	19th Train	1530[b]	1527,754	19th Train
25	664,5[c]	665,7633	6th Test	0,799[b]	0,798916	6th Test	800[b]	802,961	6th Test	870000[b]	874779	6th Test	1600[b]	1597,626	6th Test
26	675[c]	676,7473	20th Train	0,8011[b]	0,801162	20th Train	808[b]	810,4277	20th Train	846000[b]	851628	20th Train	1669[b]	1666,62	20th Train
27	685,3[c]	687,6741	21st Train	0,8031[b]	0,803051	21st Train	816[b]	817,4988	21st Train	795000[b]	806390	21st Train	1737[b]	1731,978	21st Train
28	704,8[c]	707,7089	22nd Train	0,8066[b]	0,806773	22nd Train	824[b]	828,5487	22nd Train	744000[b]	750375	22nd Train	1873[b]	1868,066	22nd Train
29	738,8[c]	739,3861	7th Test	0,8123[b]	0,812709	7th Test	847[b]	845,7786	7th Test	588000[b]	671924	7th Test	2136[b]	2135,698	7th Test
30	770,1[c]	758,6414	23rd Train	0,8168[b]	0,817377	23rd Train	872[b]	856,6428	23rd Train	472000[b]	621758	23rd Train	2389[b]	2402,114	23rd Train
31	282,6[c]	281,3343	8th Test	0,6135[b]	0,615127	8th Test	433,8[b]	437,2439	8th Test	3196000[b]	3312548	8th Test	307[b]	317,256	8th Test
32	333,41[a]	333,4679	24th Train	0,65313[a]	0,654515	24th Train	497,5[a]	499,6543	24th Train	3010300[a]	3049454	24th Train	368[b]	367,016	24th Train
33	336,43[a]	336,9246	25th Train	0,66429[a]	0,670404	25th Train	504,43[a]	505,3973	25th Train	3124000[a]	3079147	25th Train	368[b]	362,332	25th Train
34	322,88[a]	320,3776	26th Train	0,64914[a]	0,635784	26th Train	488,78[a]	490,215	26th Train	3080600[a]	3028374	26th Train	358[b]	364,922	26th Train
35	331,13[a]	329,4036	27th Train	0,66162[a]	0,657801	27th Train	499,98[a]	504,3929	27th Train	3126800[a]	3147449	27th Train	361[b]	359,868	27th Train
36	363,2[a]	363,9745	28th Train	0,67857[a]	0,679191	28th Train	530,37[a]	525,7331	28th Train	2733800[a]	2722011	28th Train	421[a]	423,453	28th Train
37	365[a]	366,3321	29th Train	0,6915[a]	0,692742	29th Train	535,25[a]	535,6044	29th Train	2813800[a]	2813921	29th Train	404[a]	415,266	29th Train
38	352,34[a]	353,0202	30th Train	0,67383[a]	0,670695	30th Train	520,5[a]	526,8028	30th Train	2773300[a]	2822521	30th Train	416[a]	417,993	30th Train
39	362,93[a]	362,5847	31st Train	0,69506[a]	0,696738	31st Train	537[a]	540,7319	31st Train	2908000[a]	2946288	31st Train	393[a]	409,202	31st Train
40	353,64[a]	357,2362	32nd Train	0,67268[a]	0,67214	32nd Train	519,79[a]	521,4523	32nd Train	2736800[a]	2742636	32nd Train	418[a]	420,974	32nd Train
41	359,21[a]	360,0799	33rd Train	0,69325[a]	0,693015	33rd Train	536,4[a]	536,7952	33rd Train	2945500[a]	2926530	33rd Train	414[a]	403,991	33rd Train
42	366,62[a]	369,597	34th Train	0,69814[a]	0,699221	34th Train	540,64[a]	538,2383	34th Train	2890800[a]	2935112	34th Train	416[a]	412,164	34th Train
43	354,03[a]	352,384	35th Train	0,69009[a]	0,69357	35th Train	531,17[a]	532,1276	35th Train	2953600[a]	2942937	35th Train	398[a]	409,721	35th Train
44	390,8[a]	390,2134	36th Train	0,6979[a]	0,700632	36th Train	559,64[a]	554,6218	36th Train	2484500[a]	2497347	36th Train	488[a]	485,198	36th Train
45	392,08[a]	392,313	9th Test	0,7058[a]	0,71009	9th Test	563,67[a]	558,8083	9th Test	2546300[a]	2525960	9th Test	464[a]	479,617	9th Test
46	390,86[a]	392,1247	37th Train	0,70461[a]	0,703978	37th Train	561,74[a]	560,9967	37th Train	2542200[a]	2529159	37th Train	476[a]	472,46	37th Train
47	416,15[a]	417,3643	38th Train	0,7265[a]	0,732937	38th Train	591[a]	583,6537	38th Train	2399375,44[a]	2411795	38th Train	533,5[d]	515,59	38th Train
48	414,372[a]	415,2589	39th Train	0,7281[a]	0,729186	39th Train	561,738[a]	575,4764	39th Train	2392480,679[a]	2417060	39th Train	505[a]	499,306	39th Train
49	380[c]	379,1597	40th Train	0,6953[b]	0,690438	40th Train	549,8[b]	553,3906	40th Train	2530000[b]	2505640	40th Train	478[b]	469,05	40th Train
50	388,76[a]	386,5388	41st Train	0,71212[a]	0,708607	41st Train	563,4[a]	561,5178	41st Train	2630000[a]	2691878	41st Train	468,2[a]	457,39	41st Train

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
51	382,58[a]	384,443	42nd Train	0,70034[a]	0,706204	42nd Train	553,5[a]	551,2911	42nd Train	2560000[a]	2570240	42nd Train	472[a]	463,959	42nd Train
52	382,26[a]	384,3292	43rd Train	0,69352[a]	0,693571	43rd Train	550[a]	546,0076	43rd Train	2490000[a]	2496792	43rd Train	482[a]	470,172	43rd Train
53	385,12[a]	385,9917	10th Test	0,70998[a]	0,711173	10th Test	562,02[a]	564,5877	10th Test	2653791,97[a]	2700345	10th Test	443[a]	456,215	10th Test
54	390,9[c]	390,7776	44th Train	0,7192[b]	0,717244	44th Train	568,8[b]	571,4539	44th Train	2690000[b]	2757540	44th Train	466[b]	456,272	44th Train
55	405,97[a]	404,7978	45th Train	0,7105[a]	0,704597	45th Train	577,8[a]	583,4904	45th Train	2350000[a]	2353493	45th Train	525[a]	529,463	45th Train
56	413,6[a]	409,7288	46th Train	0,7259[a]	0,72487	46th Train	588,8[a]	582,5333	46th Train	2390000[a]	2422953	46th Train	518[a]	523,197	46th Train
57	406,3[a]	408,5533	47th Train	0,7153[a]	0,712825	47th Train	578[a]	576,6322	47th Train	2390000[a]	2403351	47th Train	518[a]	521,077	47th Train
58	408,372[a]	409,8038	48th Train	0,7089[a]	0,70358	48th Train	577,8[a]	577,2432	48th Train	2302848,84[a]	2282134	48th Train	535[a]	538,417	48th Train
59	410,17[a]	410,3972	49th Train	0,7256[a]	0,722855	49th Train	588,1[a]	588,8193	49th Train	2480000[a]	2443281	49th Train	515[a]	514,958	49th Train
60	413,761[a]	413,042	50th Train	0,7314[a]	0,730339	50th Train	591,927[a]	590,7608	50th Train	2461428,249[a]	2462773	50th Train	503[a]	509,139	50th Train
61	408,8[a]	408,6696	51st Train	0,7205[a]	0,729	51st Train	584,6[a]	579,6334	51st Train	2450000[a]	2441933	51st Train	513[a]	515,044	51st Train
62	430[a]	428,669	11th Test	0,7245[a]	0,714972	11th Test	602,3[a]	614,6985	11th Test	2170000[a]	2126493	11th Test	580[a]	575,78	11th Test
63	437,46[a]	434,1493	12th Test	0,7379[a]	0,738879	12th Test	612,6[a]	607,9351	12th Test	2200000[a]	2238907	12th Test	572[a]	569,131	12th Test
64	429,03[a]	431,4327	52nd Train	0,7264[a]	0,72015	52nd Train	599,42[a]	599,303	52nd Train	2137374,67[a]	2170603	52nd Train	566[a]	566,911	52nd Train
65	431,65[a]	431,0751	13th Test	0,7302[a]	0,731249	13th Test	603,15[a]	601,1762	13th Test	2151164,18[a]	2152362	13th Test	569[a]	562,059	13th Test
66	434,372[a]	435,7338	53rd Train	0,739[a]	0,734327	53rd Train	612,2611[a]	620,189	53rd Train	*(-)*	*(-)*	*(-)*	557[a]	562,783	53rd Train
67	434[a]	434,2564	54th Train	0,7363[a]	0,744978	54th Train	609,6[a]	614,5916	54th Train	2250000[a]	2216445	53rd Train	568[a]	562,83	54th Train
68	391,68[a]	392,6167	55th Train	0,71356[a]	0,717951	55th Train	565,48[a]	554,4392	55th Train	2813750,33[a]	2701704	54th Train	445[a]	461,568	55th Train
69	383[c]	383,8038	56th Train	0,716[b]	0,719611	56th Train	563,5[b]	569,2265	56th Train	2730000[b]	2730835	55th Train	436[b]	455,332	56th Train
70	372,4[c]	374,5567	57th Train	0,6919[b]	0,688957	57th Train	543,8[b]	541,9022	57th Train	2570000[b]	2563035	56th Train	468[b]	464,509	57th Train
71	387,9[c]	387,6668	58th Train	0,7262[b]	0,72618	58th Train	573,5[b]	580,1602	58th Train	2820000[b]	2801748	57th Train	455[b]	449,025	58th Train
72	386,6[c]	386,5522	59th Train	0,719[b]	0,723178	59th Train	566,4[b]	566,8209	59th Train	2730000[b]	2695584	58th Train	460[b]	450,65	59th Train
73	388,8[c]	386,301	60th Train	0,7193[b]	0,716607	60th Train	567,1[b]	558,9952	60th Train	2700000[b]	2681329	59th Train	442[b]	450,376	60th Train
74	391,4[c]	390,6657	61st Train	0,7274[b]	0,730901	61st Train	576,5[b]	571,1203	61st Train	2810000[b]	2771598	60th Train	455[b]	451,333	61st Train
75	386,88[a]	386,2933	14th Test	0,7725[a]	0,769237	14th Test	582,87[a]	573,3548	14th Test	2940000[a]	2944501	61st Train	417[a]	408,343	14th Test
76	385,12[a]	383,1457	62nd Train	0,74823[a]	0,760206	62nd Train	562,02[a]	578,8401	62nd Train	2826850,37[a]	2871930	14th Test	417[a]	420,345	62nd Train
77	389,91[a]	388,8706	63rd Train	0,76343[a]	0,760724	63rd Train	579,81[a]	575,9162	63rd Train	2877871,57[a]	2883411	62nd Train	416[a]	407,435	63rd Train
78	383,3[a]	387,6239	64th Train	0,7535[a]	0,766688	64th Train	572,15[a]	578,3873	64th Train	2826967,5[a]	2907509	63rd Train	416[a]	419,766	64th Train

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
79	406,8[d]	406,3322	65th Train	0,7205[a]	0,726282	65th Train	588[b]	588,5224	65th Train	2493000[b]	2514791	64th Train	498[b]	502,121	65th Train
80	399,69[a]	399,8705	15th Test	0,7155[a]	0,71363	15th Test	574,7[a]	574,6351	15th Test	2440000[a]	2416511	65th Train	514[a]	509,352	15th Test
81	397,24[d]	395,7781	66th Train	0,7072[a]	0,707258	66th Train	569,8[a]	569,6425	66th Train	2249000[b]	2270709	15th Test	519[b]	520,428	66th Train
82	404,51[a]	402,5863	67th Train	0,72189[a]	0,725198	67th Train	579,4[a]	574,5877	67th Train	2410000[a]	2407551	66th Train	511[a]	504,554	67th Train
83	409,774[a]	411,3016	68th Train	0,77876[a]	0,78276	68th Train	612,15[a]	609,119	68th Train	2695245[a]	2659196	67th Train	469[a]	482,538	68th Train
84	419,34[a]	419,8176	69th Train	0,75357[a]	0,756185	69th Train	610,05[a]	606,8724	69th Train	2675000[a]	2669187	68th Train	473[a]	488,294	69th Train
85	413,4[c]	414,707	70th Train	0,7566[b]	0,751896	70th Train	607,5[b]	604,1219	70th Train	2740000[b]	2719722	69th Train	478[b]	482,406	70th Train
86	406,177[a]	404,6525	71st Train	0,73893[a]	0,729432	71st Train	592,15[a]	593,2543	71st Train	2564849,6[a]	2566992	70th Train	490[b]	492,637	71st Train
87	395,4[c]	394,626	72nd Train	0,7194[b]	0,718861	72nd Train	574,6[b]	573,3929	72nd Train	2490000[b]	2477092	71st Train	504[b]	502,256	72nd Train
88	414,7[c]	413,8847	73rd Train	0,7547[b]	0,754926	73rd Train	607,5[b]	610,1928	73rd Train	2720000[b]	2724270	72nd Train	481[b]	486,725	73rd Train
89	416,43[a]	415,299	74th Train	0,7134[a]	0,714263	74th Train	586,75[a]	585,9276	74th Train	2290000[a]	2273206	73rd Train	541[a]	545,846	74th Train
90	417,38[a]	418,7007	75th Train	0,7205[a]	0,72396	75th Train	590,15[a]	591,1899	75th Train	2340600[a]	2358201	74th Train	529[a]	541,195	75th Train
91	415,59[a]	417,7311	76th Train	0,7202[a]	0,719237	76th Train	587,65[a]	587,9355	76th Train	2340600[a]	2338077	75th Train	523[a]	535,956	76th Train
92	440,1[c]	438,1264	77th Train	0,7264[a]	0,724666	77th Train	610,4[b]	605,0774	77th Train	2097000[b]	2064906	76th Train	596[b]	587,742	77th Train
93	441[a]	441,0442	78th Train	0,7334[a]	0,7341	78th Train	614,2[a]	612,9931	78th Train	2180000[a]	2167331	77th Train	579[a]	583,399	78th Train
94	438,261[a]	439,5144	79th Train	0,7326[a]	0,727781	79th Train	609,761[a]	605,9434	79th Train	2137374,67[a]	2100834	78th Train	573[a]	581,398	79th Train
95	461[a]	462,3894	80th Train	0,7422[a]	0,742596	80th Train	639[a]	632,0526	80th Train	1930000[a]	1974040	79th Train	645[a]	637,863	80th Train
96	439,65[a]	439,4338	16th Test	0,7399[a]	0,742185	16th Test	613,76[a]	612,1262	16th Test	2185637,97[a]	2184058	80th Train	561[a]	561,58	16th Test
97	428,8[d]	427,385	81st Train	0,7432[b]	0,741644	81st Train	609,5[b]	613,5239	81st Train	2320000[b]	2300026	16th Test	537[b]	537,507	81st Train
98	433,48[d]	434,2614	82nd Train	0,7644[b]	0,755529	82nd Train	623[b]	619,3584	82nd Train	2510000[b]	2555494	81st Train	518[b]	512,716	82nd Train
99	225,4[c]	225,2432	17th Test	0,5133[b]	0,51625	17th Test	364,9[b]	364,5176	17th Test	4600000[b]	4607432	82nd Train	184[b]	191,442	17th Test
100	266,9[c]	268,58	83rd Train	0,5945[b]	0,591603	83rd Train	419,5[b]	424,3528	83rd Train	4020000[b]	4005068	17th Test	240[b]	234,868	83rd Train
101	303,1[c]	303,9617	84th Train	0,6405[b]	0,639819	84th Train	464,8[b]	473,5585	84th Train	3560000[b]	3599746	83rd Train	298[b]	294,752	84th Train
102	336,6[c]	339,4538	85th Train	0,6732[b]	0,677526	85th Train	504[b]	505,3904	85th Train	3210000[b]	3214093	84th Train	355[b]	351,163	85th Train
103	366,8[c]	368,0092	86th Train	0,697[b]	0,703258	86th Train	537,3[b]	535,661	86th Train	2920000[b]	2948014	85th Train	409[b]	406,368	86th Train
104	394,4[c]	393,4868	87th Train	0,7149[b]	0,721227	87th Train	567[b]	573,7989	87th Train	2680000[b]	2663957	86th Train	468[b]	469,397	87th Train
105	420[c]	418,9974	88th Train	0,7292[b]	0,734091	88th Train	594[b]	598,9775	88th Train	2429000[b]	2450388	87th Train	526[b]	533,443	88th Train
106	443,8[c]	442,8066	89th Train	0,7408[b]	0,744777	89th Train	617[b]	619,2779	89th Train	2220000[b]	2234682	88th Train	584[b]	580,108	89th Train

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*d20^{exp} (-)*	*d20^{cal} (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
107	465,8[c]	465,1348	90th Train	0,7503[b]	0,753128	90th Train	638,7[b]	637,3286	90th Train	2062000[b]	2081232	89th Train	652[b]	644,071	90th Train
108	486,5[c]	485,7771	91st Train	0,7584[b]	0,760568	91st Train	658[b]	655,8173	91st Train	1930000[b]	1906437	90th Train	716[b]	707,19	91st Train
109	505,9[c]	505,4265	18th Test	0,7653[b]	0,766785	18th Test	673[b]	674,269	18th Test	1730000[b]	1786778	91st Train	781[b]	774,229	18th Test
110	524,2[c]	524,5024	19th Test	0,7713[b]	0,77222	19th Test	691,7[b]	690,8999	19th Test	1590000[b]	1641004	18th Test	846[b]	843,898	19th Test
111	541,6[c]	542,0354	92nd Train	0,7764[b]	0,777229	92nd Train	705[b]	705,1348	92nd Train	1540000[b]	1546326	19th Test	912[b]	915,491	92nd Train
112	558[c]	558,3455	93rd Train	0,7811[b]	0,781635	93rd Train	718[b]	719,0747	93rd Train	1390000[b]	1421201	92nd Train	978[b]	981,929	93rd Train
113	573,5[c]	573,6488	94th Train	0,7852[b]	0,785459	94th Train	734[b]	733,202	94th Train	1340000[b]	1349858	93rd Train	1045[b]	1044,033	94th Train
114	588[c]	588,3403	95th Train	0,7888[b]	0,789289	95th Train	748[b]	746,113	95th Train	1300000[b]	1248637	94th Train	1111[b]	1111,231	95th Train
115	602,2[c]	602,3382	96th Train	0,792[b]	0,792324	96th Train	755[b]	757,6394	96th Train	1190000[b]	1189863	95th Train	1177[b]	1177,404	96th Train
116	615,5[c]	616,0865	97th Train	0,795[b]	0,795286	97th Train	772[b]	768,1535	97th Train	1140000[b]	1103801	96th Train	1244[b]	1246,77	97th Train
117	268,7[c]	270,4061	98th Train	0,6211[b]	0,632996	98th Train	425[b]	427,0159	98th Train	4320000[b]	4316020	97th Train	221[b]	219,607	98th Train
118	276,9[c]	277,6193	99th Train	0,6210001[d]	0,600881	99th Train	435,5[b]	430,219	99th Train	4210000[b]	4242913	98th Train	233,8[b]	224,033	99th Train
119	274[c]	275,4693	100th Train	0,6051[b]	0,601664	100th Train	428,6[b]	425,543	100th Train	4100000[b]	4056254	99th Train	237,8[b]	247,677	100th Train
120	266,2[c]	261,1068	20th Test	0,5953[b]	0,588866	20th Test	417,9[b]	411,131	20th Test	4000000[b]	3927614	100th Train	238,8[b]	246,186	20th Test
121	307,2[c]	304,1067	21st Test	0,6805[b]	0,682789	21st Test	484[b]	483,0337	21st Test	3850000[b]	3893844	20th Test	276[b]	271,649	21st Test
122	328,8[c]	328,4288	101st Train	0,6779[b]	0,683998	101st Train	501[b]	500,9426	101st Train	3242000[b]	3188490	21st Test	343[b]	359,898	101st Train
123	346,3[c]	345,5059	102nd Train	0,708[b]	0,701677	102nd Train	524[b]	522,8337	102nd Train	3364000[b]	3342557	101st Train	351[b]	342,673	102nd Train
124	342[c]	342,9201	22nd Test	0,6869[b]	0,688762	22nd Test	518[b]	502,809	22nd Test	3283000[b]	3231191	102nd Train	351[b]	340,095	22nd Test
125	341[c]	342,4154	23rd Test	0,6784[b]	0,683653	23rd Test	516[b]	503,9037	23rd Test	3273000[b]	3179408	22nd Test	351[b]	349,141	23rd Test
126	327[c]	324,9348	103rd Train	0,6642[b]	0,665879	103rd Train	495[b]	492,3391	103rd Train	3290000[b]	3218690	23rd Test	353,6[b]	356,782	103rd Train
127	374,6[c]	373,8863	104th Train	0,715[b]	0,706949	104th Train	560,2[b]	551,9786	104th Train	2645000[b]	2674907	103rd Train	457[b]	454,537	104th Train
128	378,1[c]	377,5071	24th Test	0,7218[b]	0,722894	24th Test	563,2[b]	543,3768	24th Test	2655000[b]	2635283	104th Train	457[b]	439,866	24th Test
129	304,3[c]	302,2391	105th Train	0,6504[b]	0,638294	105th Train	465[b]	471,8544	105th Train	3400000[b]	3499318	24th Test	294[b]	299,126	105th Train
130	311,7[c]	308,9543	106th Train	0,658[b]	0,658521	106th Train	470[b]	476,1494	106th Train	3420000[b]	3551921	105th Train	318[b]	296,692	106th Train
131	293,2[c]	295,5025	107th Train	0,6272[b]	0,636401	107th Train	452,7[b]	457,5027	107th Train	3530000[b]	3508218	106th Train	304,9[b]	301,891	107th Train
132	341,9[c]	339,9692	108th Train	0,7267[b]	0,723356	108th Train	527,2[b]	523,2168	108th Train	3536000[b]	3514460	107th Train	333[b]	327,2	108th Train
133	314,4[c]	319,0478	109th Train	0,6529[b]	0,648665	109th Train	490,2[b]	489,0467	109th Train	3151000[b]	3173447	108th Train	340[d]	343,764	109th Train
134	340,4[c]	341,6988	25th Test	0,6863[b]	0,68555	25th Test	518[b]	513,4847	25th Test	3283000[b]	3218717	109th Train	351[b]	357,685	25th Test

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
135	332,6[c]	333,8126	110th Train	0,6923[b]	0,686359	110th Train	508[b]	503,1462	110th Train	3445400[b]	3380827	25th Test	328[b]	329,474	110th Train
136	240,4[c]	240,6259	111th Train	0,6309998[d]	0,627719	111th Train	398[b]	395,6944	111th Train	5495000[b]	5381844	110th Train	162[b]	158,549	111th Train
137	285,7[c]	284,9319	112th Train	0,6873[b]	0,707446	112th Train	460[b]	466,9643	112th Train	5056000[b]	5015933	111th Train	210[b]	204,573	112th Train
138	322,4[a]	323,7218	113th Train	0,74536[a]	0,742907	113th Train	511,76[a]	499,5508	113th Train	4502300[a]	4539849	112th Train	258,32[a]	260,929	113th Train
139	353,87[a]	355,366	114th Train	0,77853[a]	0,790024	114th Train	553,54[a]	555,371	114th Train	4074800[a]	4110220	113th Train	307,88[a]	309,185	114th Train
140	391,9[c]	391,2521	26th Test	0,8098[b]	0,810532	26th Test	604,2[b]	580,3787	26th Test	3820000[b]	3792914	114th Train	353[b]	369,8	26th Test
141	423,8[c]	423,6332	115th Train	0,8349[b]	0,825218	115th Train	647,2[b]	613,1376	115th Train	3560000[b]	3573275	26th Test	410[b]	408,518	115th Train
142	374,08[a]	376,6083	116th Train	0,76937[a]	0,765636	116th Train	572,19[a]	574,7189	116th Train	3471000[a]	3379381	115th Train	368[a]	390,593	116th Train
143	404,95[a]	407,9673	117th Train	0,7879[a]	0,788598	117th Train	609,15[a]	611,9009	117th Train	3040000[a]	3049645	116th Train	430[a]	434,958	117th Train
144	429,9[a]	432,529	118th Train	0,79358[a]	0,79649	118th Train	639,15[a]	636,5968	118th Train	2807000[a]	2796325	117th Train	477[a]	466,724	118th Train
145	427,714[a]	426,824	27th Test	0,80219[a]	0,791025	27th Test	640,15[a]	619,8681	27th Test	2837100[a]	2798578	118th Train	473[a]	462,7	27th Test
146	454,1[c]	454,458	119th Train	0,7996[b]	0,800444	119th Train	667,2[b]	656,0571	119th Train	2385000[b]	2507266	27th Test	534[b]	536,384	119th Train
147	452,45[a]	450,6064	28th Test	0,814[a]	0,807521	28th Test	669,15[a]	646,6199	28th Test	2674980[a]	2638920	119th Train	530[a]	509,246	28th Test
148	344,96[a]	344,2123	120th Train	0,74862[a]	0,735718	120th Train	532,79[a]	530,0151	120th Train	3784500[a]	3813174	28th Test	318,92[a]	318,229	120th Train
149	376,62[a]	377,8779	121st Train	0,76645[a]	0,76308	121st Train	569,52[a]	571,5249	121st Train	3397700[a]	3368704	120th Train	374,52[a]	380,477	121st Train
150	404,1[c]	404,1237	122nd Train	0,7763[b]	0,78082	122nd Train	594,7[b]	599,1957	122nd Train	2962000[b]	2999018	121st Train	426[b]	429,332	122nd Train
151	399,569[a]	396,0939	123rd Train	0,77653[a]	0,767821	123rd Train	601,15[a]	599,0153	123rd Train	2999220[a]	2993818	122nd Train	422[a]	430,407	123rd Train
152	429,75[a]	428,884	124th Train	0,7846[a]	0,787854	124th Train	612,2[a]	622,5109	124th Train	2710000[a]	2707529	123rd Train	482[a]	474,429	124th Train
153	421,122[a]	419,9006	29th Test	0,7809[a]	0,784338	29th Test	625,65[a]	610,6462	29th Test	2735775[a]	2704717	124th Train	482[a]	454,24	29th Test
154	453,8[d]	454,2093	125th Train	0,7915[b]	0,792502	125th Train	640,8[b]	644,143	125th Train	2385000[b]	2394462	29th Test	543[b]	538,253	125th Train
155	476,3[d]	475,9472	30th Test	0,7976[b]	0,798102	30th Test	660,7[b]	665,6651	30th Test	2162000[b]	2195777	125th Train	603[b]	599,857	30th Test
156	497,3[d]	498,576	126th Train	0,8016[b]	0,8019	126th Train	678,9[b]	684,2037	126th Train	1971000[b]	1963855	30th Test	665[b]	665,034	126th Train
157	516,9[d]	518,991	31st Test	0,8056[b]	0,806353	31st Test	695,5[b]	697,9783	31st Test	1806000[b]	1818178	126th Train	727[b]	728,219	31st Test
158	535,3[d]	537,9346	127th Train	0,8086[b]	0,809097	127th Train	710,7[b]	710,5328	127th Train	1663000[b]	1639821	31st Test	789[b]	789,76	127th Train
159	552,5[d]	553,8998	128th Train	0,8116[b]	0,812842	128th Train	724,8[b]	724,7631	128th Train	1537000[b]	1526310	127th Train	852[b]	853,175	128th Train
160	584,1[d]	581,5419	129th Train	0,8167[b]	0,818815	129th Train	749,7[b]	750,3582	129th Train	1326000[b]	1324257	128th Train	979[b]	984,705	129th Train
161	598,6[d]	594,1797	130th Train	0,8187[b]	0,820864	130th Train	760,8[b]	760,7157	130th Train	1238000[b]	1228895	129th Train	1043[b]	1046,457	130th Train
162	599[d]	606,7687	131st Train	0,8207[b]	0,823302	131st Train	771,1[b]	770,9305	131st Train	1158000[b]	1167275	130th Train	1107[b]	1106,398	131st Train

ID	Tb^{exp} *(K)*	Tb^{cal} *(K)*	*Partition des Tb*	$d20^{exp}$ *(-)*	$d20^{cal}$ *(-)*	*Partition des d20*	Tc^{exp} *(K)*	Tc^{cal} *(K)*	*Partition des Tc*	Pc^{exp} *(Pa)*	Pc^{cal} *(Pa)*	*Partition des Pc*	Vc^{exp} *(cm^3/mol)*	Vc^{cal} *(cm^3/mol)*	*Partition des Vc*
163	625[d]	618,1861	132nd Train	0,8217[b]	0,82508	132nd Train	780,6[b]	780,9404	132nd Train	1087000[b]	1092132	131st Train	1170[b]	1167,408	132nd Train
164	443,87[a]	441,6034	133rd Train	0,8[a]	0,801337	133rd Train	654,15[a]	634,6978	133rd Train	2599999,5[a]	2635986	132nd Train	525[a]	518,686	133rd Train
165	360,96[a]	360,97	32nd Test	0,75446[a]	0,753366	32nd Test	547[a]	554,2202	32nd Test	3400000[a]	3353909	133rd Train	373[a]	360,673	32nd Test
166	363,9[a]	364,9226	33rd Test	0,74476[a]	0,746779	33rd Test	553[a]	555,8512	33rd Test	3400000[a]	3303536	32nd Test	373[a]	374,67	33rd Test
167	364,7[a]	364,2483	34th Test	0,74878[a]	0,748034	34th Test	551[a]	553,7364	34th Test	3400000[a]	3295169	33rd Test	373[a]	377,683	34th Test
168	365,02[a]	363,281	134th Train	0,75142[a]	0,753602	134th Train	553,15[a]	559,5762	134th Train	3445100[a]	3326315	34th Test	360[a]	366,275	134th Train
169	392,7[a]	395,0026	135th Train	0,78092[a]	0,786713	135th Train	590,928[a]	598,5577	135th Train	2964745,51[a]	2986549	134th Train	450[a]	446,021	135th Train
170	402,94[a]	399,3985	136th Train	0,79625[a]	0,780205	136th Train	606,15[a]	597,9972	136th Train	2938400[a]	2962418	135th Train	460[a]	443,094	136th Train
171	396,58[a]	396,2551	35th Test	0,77599[a]	0,773345	35th Test	596,15[a]	596,9523	35th Test	2938400[a]	3025218	136th Train	460[a]	438,616	35th Test
172	393,2[c]	393,8316	137th Train	0,766[b]	0,76985	137th Train	591,1[b]	587,1106	137th Train	2938000[b]	2995652	35th Test	448,9[d]	440,571	137th Train
173	397,61[a]	396,9607	138th Train	0,7847[a]	0,776824	138th Train	598,15[a]	589,2964	138th Train	2964745,51[a]	2949935	137th Train	460[a]	446,053	138th Train
174	397,47[a]	396,6368	139th Train	0,78283[a]	0,775862	139th Train	598[b]	591,3089	139th Train	2938000[b]	2929082	138th Train	471,3[d]	442,284	139th Train
175	392,505[a]	393,7084	36th Test	0,76253[a]	0,762892	36th Test	589,816[a]	593,7389	36th Test	2964745,51[a]	2971211	139th Train	450[a]	441,819	36th Test
176	317,4[c]	317,6272	140th Train	0,772[b]	0,763878	140th Train	506,5[b]	508,907	140th Train	4800000[b]	4804013	36th Test	245[b]	232,657	140th Train
177	356,1[c]	353,6693	141st Train	0,8102[b]	0,801364	141st Train	560,4[b]	558,6707	141st Train	4350000[b]	4338708	140th Train	292[b]	307,206	141st Train
178	353,24[a]	352,7994	142nd Train	0,87889[a]	0,885143	142nd Train	562,05[a]	561,5579	142nd Train	4895000[a]	4905792	141st Train	256[a]	249,679	142nd Train
179	383,78[a]	383,9503	37th Test	0,86696[a]	0,868807	37th Test	591,75[a]	594,1734	37th Test	4108000[a]	4179116	142nd Train	316[a]	328,255	37th Test
180	409,35[a]	407,0375	143rd Train	0,867[a]	0,856347	143rd Train	617,15[a]	614,7282	143rd Train	3609000[a]	3526224	37th Test	374[a]	390,096	143rd Train
181	432,388[a]	433,8728	144th Train	0,86202[a]	0,861544	144th Train	638,4[b]	649,4853	144th Train	3200000[b]	3255430	143rd Train	440[b]	438,782	144th Train
182	425,561[a]	426,1248	145th Train	0,8617[a]	0,863982	145th Train	631,1[a]	634,6968	145th Train	3209000[a]	3236502	144th Train	427[a]	434,45	145th Train
183	456,5[c]	457,9644	146th Train	0,862[b]	0,865283	146th Train	660,5[b]	664,5234	146th Train	2890000[b]	2907664	145th Train	497[b]	506,261	146th Train
184	417,58[a]	418,4075	147th Train	0,88018[a]	0,874405	147th Train	630,37[a]	625,471	147th Train	3734200[a]	3730817	146th Train	369,17[a]	383,425	147th Train
185	412,27[a]	412,6436	148th Train	0,86415[a]	0,873838	148th Train	617,05[a]	609,8926	148th Train	3541200[a]	3537236	147th Train	375,8[a]	392,211	148th Train
186	411,51[a]	410,4914	149th Train	0,86103[a]	0,862311	149th Train	616,26[a]	613,5384	149th Train	3510800[a]	3528243	148th Train	379,11[a]	381,619	149th Train
187	478,6[c]	479,6241	150th Train	0,8645[b]	0,862771	150th Train	678,9[b]	685,8754	150th Train	2634000[b]	2636833	149th Train	552[b]	553,651	150th Train
188	499,3[c]	499,657	151st Train	0,8585[b]	0,860372	151st Train	695,8[b]	698,701	151st Train	2634000[b]	2515883	150th Train	614[b]	618,81	151st Train
189	519,2[c]	519,0738	152nd Train	0,8583[b]	0,858877	152nd Train	711,3[b]	713,7555	152nd Train	2189000[b]	2200315	151st Train	677[b]	680,27	152nd Train
190	537,5[c]	537,9634	38th Test	0,8575[b]	0,854897	38th Test	725,6[b]	726,5393	38th Test	2009000[b]	2046995	152nd Train	741[b]	742,606	38th Test

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
191	555,2[c]	555,0886	153rd Train	0,8569[b]	0,854908	153rd Train	738,8[b]	737,9581	153rd Train	1854000[b]	1853547	38th Test	805[b]	806,479	153rd Train
192	571[c]	571,9432	154th Train	0,8564[b]	0,851926	154th Train	750,9[b]	747,6044	154th Train	1716000[b]	1776723	153rd Train	870[b]	867,676	154th Train
193	586,4[c]	585,9471	155th Train	0,8561[b]	0,852694	155th Train	762,2[b]	758,5224	155th Train	1595000[b]	1582735	154th Train	935[b]	927,857	155th Train
194	600,8[c]	598,8048	39th Test	0,8558[b]	0,853339	39th Test	772,7[b]	768,987	39th Test	1485000[b]	1490630	155th Train	1001[b]	992,528	39th Test
195	614,4[c]	611,4751	156th Train	0,8556[b]	0,853866	156th Train	782,4[b]	777,9279	156th Train	1388000[b]	1374851	39th Test	1066[b]	1053,021	156th Train
196	438,6[c]	435,251	157th Train	0,9057[b]	0,892133	157th Train	654[b]	650,7054	157th Train	3405000[b]	3325855	156th Train	410[b]	427,311	157th Train
197	434,5[c]	434,3767	40th Test	0,861[b]	0,864093	40th Test	637,1[b]	641,2048	40th Test	2837000[b]	2912564	157th Train	490[b]	451,43	40th Test
198	438,3[c]	438,2253	158th Train	0,8617[b]	0,869528	158th Train	651,1[b]	644,051	158th Train	3040000[b]	3044668	40th Test	460[b]	445,693	158th Train
199	435,2[c]	434,1755	41st Test	0,8806[b]	0,869861	41st Test	640,2[b]	636,5433	41st Test	3230000[b]	3048521	158th Train	470[b]	443,85	41st Test
200	406,93[a]	409,0089	159th Train	0,7235[a]	0,730189	159th Train	580,15[a]	578,8672	159th Train	2399375,44[a]	2429754	41st Test	504[a]	508,397	159th Train
201	431,37[a]	432,0373	160th Train	0,7406[a]	0,746607	160th Train	606,76[a]	601,0378	160th Train	2227006,511[a]	2250541	159th Train	541[a]	532,748	160th Train
202	425,31[a]	426,5271	161st Train	0,8062[a]	0,808342	161st Train	612,15[a]	637,8127	161st Train	2695245[a]	2797128	160th Train	469[a]	474,109	161st Train
203	421,614[a]	418,5729	162nd Train	0,809[a]	0,789315	162nd Train	612,15[a]	617,2137	162nd Train	2695245[a]	2714858	161st Train	469[a]	471,349	162nd Train
204	424,87[a]	423,2367	163rd Train	0,794[a]	0,79451	163rd Train	612,15[a]	621,3409	163rd Train	2695245[a]	2715645	162nd Train	469[a]	462,57	163rd Train
205	424,87[a]	432,6397	164th Train	0,794[a]	0,800447	164th Train	612,5[a]	637,6273	164th Train	2695245[a]	2603174	163rd Train	469[a]	506,281	164th Train
206	425,561[a]	426,1248	165th Train	0,8617[a]	0,863982	165th Train	631,1[a]	634,6968	165th Train	3209000[a]	3236502	164th Train	427[a]	434,45	165th Train
207	434,48[a]	434,348	166th Train	0,8645[a]	0,863945	166th Train	637,15[a]	640,9224	166th Train	2837100[a]	2911356	165th Train	490[a]	451,387	166th Train
208	435,16[a]	434,1717	167th Train	0,86116[a]	0,869857	167th Train	640,15[a]	636,4819	167th Train	2938400[a]	3048541	166th Train	470[a]	443,835	167th Train
209	438,33[a]	438,3817	168th Train	0,88067[a]	0,869529	168th Train	651,15[a]	645,7689	168th Train	3039800[a]	3043662	167th Train	460[a]	440,559	168th Train
210	449,27[a]	445,6679	169th Train	0,89436[a]	0,876957	169th Train	664,53[a]	651,8823	169th Train	3454200[a]	3354466	168th Train	414[a]	423,893	169th Train
211	442,53[a]	443,0534	170th Train	0,8758[a]	0,877837	170th Train	649,13[a]	639,5358	170th Train	3232300[a]	3235729	169th Train	430[a]	431,464	170th Train
212	437,89[a]	438,1311	42nd Test	0,86516[a]	0,875332	42nd Test	637,36[a]	619,3368	42nd Test	3126900[a]	3105591	170th Train	433[a]	447,439	42nd Test
213	451,33[a]	449,6967	43rd Test	0,8766[a]	0,877796	43rd Test	662[a]	661,5067	43rd Test	2930000[a]	2784630	42nd Test	489[a]	492,583	43rd Test
214	448,23[a]	444,8901	171st Train	0,8758[a]	0,903453	171st Train	666,2[a]	664,0043	171st Train	2938545,43[a]	2753175	43rd Test	478[a]	491,06	171st Train
215	480,8[c]	479,6449	44th Test	*(-)*	*(-)*	*(-)*	720[b]	721,6516	44th Test	3648000[b]	3582724	171st Train	408[b]	439,482	44th Test
216	442,3[a]	445,674	172nd Train	0,86648[a]	0,869307	44th Test	660[a]	667,258	172nd Train	2970300[a]	2929154	44th Test	492[a]	466,61	172nd Train
217	445,94[a]	445,1724	173rd Train	0,85319[a]	0,85428	172nd Train	650,15[a]	655,132	173rd Train	3039800[a]	3035076	172nd Train	456[a]	483,689	173rd Train
218	537,4[c]	536,684	174th Train	0,8559[b]	0,859388	173rd Train	760[b]	746,3318	174th Train	2710000[d]	2710558	173rd Train	563[b]	563,638	174th Train

ID	*Tb^{exp} (K)*	*Tb^{cal} (K)*	*Partition des Tb*	*$d20^{exp}$ (-)*	*$d20^{cal}$ (-)*	*Partition des d20*	*Tc^{exp} (K)*	*Tc^{cal} (K)*	*Partition des Tc*	*Pc^{exp} (Pa)*	*Pc^{cal} (Pa)*	*Partition des Pc*	*Vc^{exp} (cm^3/mol)*	*Vc^{cal} (cm^3/mol)*	*Partition des Vc*
219	446,48[a]	453,6651	175th Train	0,86205[a]	0,868839	174th Train	664,54[a]	663,1877	175th Train	2951000[a]	2981409	174th Train	497[a]	490,101	175th Train
220	456,9[c]	457,1811	176th Train	0,862[b]	0,874026	175th Train	657,9[b]	669,0207	176th Train	2807000[b]	2769509	175th Train	480[b]	495,61	176th Train
221	528,1[c]	526,2607	177th Train	0,9900001[d]	0,991519	176th Train	773[b]	764,044	177th Train	3380000[b]	3369511	176th Train	497[b]	491,274	177th Train
222	491,1[c]	493,7525	178th Train	1,0253[b]	1,00443	177th Train	748,4[b]	753,9188	178th Train	*(-)*	*(-)*	*(-)*	407[b]	395,887	178th Train
223	379,4[c]	380,2639	179th Train	0,7242[b]	0,718085	178th Train	567,8[b]	561,26	179th Train	2908000[b]	2909611	177th Train	461[b]	447,173	179th Train

NB1_Tableau A.1 : Les valeurs de [(a)], [(b)], [(c)] et [(d)] sont respectivement prises à partir de : la base de données expérimentale de Lahiouel **(Lahiouel 1995b)**, l'étude de Wakeham et ces coauteurs **(Wakeham *et al.* 2002)**, l'étude de Wessel et Jurs **(Wessel & Jurs 1995)**, et la base de KDB (**KDB 2017**).

NB2_Tableau A.1 : Concernant les valeurs de la densité relative du liquide *d20* obtenues à partir de KDB **(KDB 2017)**, elles ont été prises en g/cm^3 qui est la densité de la molécule à 20 °C ; Ensuite, elle est divisée par la densité de l'eau égale à 1 g/cm^3 à 4 °C.

Annexe B.

Tableau B.1. Écriture *SMILES* des molécules des hydrocarbures utilisés dans cette étude.

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
1	ethane	C_2H_6	CC
2	propane	C_3H_8	CCC
3	n-butane	C_4H_{10}	CCCC
4	i-butane	C_4H_{10}	CC(C)C
5	n-pentane	C_5H_{12}	CCCCC
6	i-pentane	C_5H_{12}	CCC(C)C
7	n-hexane	C_6H_{14}	CCCCCC
8	n-heptane	C_7H_{16}	CCCCCCC
9	n-octane	C_8H_{18}	CCCCCCCC
10	n-nonane	C_9H_{20}	CCCCCCCCC
11	n-decane	$C_{10}H_{22}$	CCCCCCCCCC
12	n-undecane	$C_{11}H_{24}$	CCCCCCCCCCC
13	n-dodecane	$C_{12}H_{26}$	CCCCCCCCCCCC
14	n-tridecane	$C_{13}H_{28}$	CCCCCCCCCCCCC
15	n-tetradecane	$C_{14}H_{30}$	CCCCCCCCCCCCCC
16	n-pentadecane	$C_{15}H_{32}$	CCCCCCCCCCCCCCC
17	n-hexadecane	$C_{16}H_{34}$	CCCCCCCCCCCCCCCC
18	n-heptadecane	$C_{17}H_{36}$	CCCCCCCCCCCCCCCCC
19	n-octadecane	$C_{18}H_{38}$	CCCCCCCCCCCCCCCCCC
20	n-nonadecane	$C_{19}H_{40}$	CCCCCCCCCCCCCCCCCCC

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
21	n-eicosane	$C_{20}H_{42}$	CCCCCCCCCCCCCCCCCCCC
22	n-heneicosane	$C_{21}H_{44}$	CCCCCCCCCCCCCCCCCCCCC
23	n-docosane	$C_{22}H_{46}$	CCCCCCCCCCCCCCCCCCCCCC
24	n-tricosane	$C_{23}H_{48}$	CCCCCCCCCCCCCCCCCCCCCCC
25	n-tetracosane	$C_{24}H_{50}$	CCCCCCCCCCCCCCCCCCCCCCCC
26	n-pentacosane	$C_{25}H_{52}$	CCCCCCCCCCCCCCCCCCCCCCCCC
27	n-hexacosane	$C_{26}H_{54}$	CCCCCCCCCCCCCCCCCCCCCCCCCC
28	n-octacosane	$C_{28}H_{58}$	CCCCCCCCCCCCCCCCCCCCCCCCCCCC
29	n-dotriacontane	$C_{32}H_{66}$	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
30	n-hexatriacontane	$C_{36}H_{74}$	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
31	2,2-dimethylpropane (neopentane)	C_5H_{12}	CC(C)(C)C
32	2-methylpentane	C_6H_{14}	CC(C)CCC
33	3-methylpentane	C_6H_{14}	CCC(C)CC
34	2,2-dimethylbutane	C_6H_{14}	CC(C)(C)CC
35	2,3-dimethylbutane	C_6H_{14}	CC(C)C(C)C
36	2-methylhexane	C_7H_{16}	CC(C)CCCC
37	3-methylhexane	C_7H_{16}	CCC(C)CCC
38	2,2-dimethylpentane	C_7H_{16}	CC(C)(C)CCC
39	2,3-dimethylpentane	C_7H_{16}	CC(C)C(C)CC
40	2,4-dimethylpentane	C_7H_{16}	CC(C)CC(C)C
41	3,3-dimethylpentane	C_7H_{16}	CCC(C)(C)CC
42	3-ethylpentane	C_7H_{16}	CCC(CC)CC
43	2,2,3-trimethylbutane	C_7H_{16}	CC(C)(C)C(C)C
44	2-methylheptane	C_8H_{18}	CC(C)CCCCC
45	3-methylheptane	C_8H_{18}	CCC(C)CCCC

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
46	4-methylheptane	C_8H_{18}	CCCC(C)CCC
47	3-ethylheptane	C_9H_{20}	CCC(CC)CCCC
48	4-ethylheptane	C_9H_{20}	CCCC(CC)CCC
49	2,2-dimethylhexane	C_8H_{18}	CC(C)(C)CCCC
50	2,3-dimethylhexane	C_8H_{18}	CC(C)C(C)CCC
51	2,4-dimethylhexane	C_8H_{18}	CC(C)CC(C)CC
52	2,5-dimethylhexane	C_8H_{18}	CC(C)CCC(C)C
53	3,3-dimethylhexane	C_8H_{18}	CCC(C)(C)CCC
54	3,4-dimethylhexane	C_8H_{18}	CCC(C)C(C)CC
55	2,2-dimethylheptane	C_9H_{20}	CC(C)(C)CCCCC
56	2,3-dimethylheptane	C_9H_{20}	CC(C)C(C)CCCC
57	2,4-dimethylheptane	C_9H_{20}	CC(C)CC(C)CCC
58	2,6-dimethylheptane	C_9H_{20}	CC(C)CCCC(C)C
59	3,3-dimethylheptane	C_9H_{20}	CCC(C)(C)CCCC
60	3,4-dimethylheptane	C_9H_{20}	CCC(C)C(C)CCC
61	3,5-dimethylheptane	C_9H_{20}	CCC(C)CC(C)CC
62	2,2-dimethyloctane	$C_{10}H_{22}$	CC(C)(C)CCCCCC
63	2,3-dimethyloctane	$C_{10}H_{22}$	CC(C)C(C)CCCCC
64	2,4-dimethyloctane	$C_{10}H_{22}$	CC(C)CC(C)CCCC
65	2,5-dimethyloctane	$C_{10}H_{22}$	CC(C)CCC(C)CCC
66	3,3-dimethyloctane	$C_{10}H_{22}$	CCC(C)(C)CCCCC
67	3,6-dimethyloctane	$C_{10}H_{22}$	CCC(C)CCC(C)CC
68	3-ethylhexane	C_8H_{18}	CCC(CC)CCC
69	2,2,3-trimethylpentane	C_8H_{18}	CC(C)(C)C(C)CC
70	2,2,4-trimethylpentane	C_8H_{18}	CC(C)(C)CC(C)C

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
71	2,3,3-trimethylpentane	C_8H_{18}	CC(C)C(C)(C)CC
72	2,3,4-trimethylpentane	C_8H_{18}	CC(C)C(C)C(C)C
73	2-methyl-3-ethylpentane	C_8H_{18}	CC(C)C(CC)CC
74	3-methyl-3-ethylpentane	C_8H_{18}	CCC(CC)(C)CC
75	1,1,2-trimethylcyclopentane	C_8H_{16}	CC1C(C)(C)CCC1
76	1,1,3-trimethylcyclopentane	C_8H_{16}	CC1CC(C)(C)CC1
77	1-cis,2-trans,4-trimethylcyclopentane	C_8H_{16}	C[C@H]1C[C@H](C)[C@H](C)C1
78	1-trans,2-cis,3-trimethylcyclopentane	C_8H_{16}	C[C@H]1C(C)[C@H](C)CC1
79	2,2,3-trimethylhexane	C_9H_{20}	CC(C)(C)C(C)CCC
80	2,2,4-trimethylhexane	C_9H_{20}	CC(C)(C)CC(C)CC
81	2,2,5-trimethylhexane	C_9H_{20}	CC(C)(C)CCC(C)C
82	2,3,5-trimethylhexane	C_9H_{20}	CC(C)C(C)CC(C)C
83	1-cis-2-cis-3-cis-trimethylcyclohexane	C_9H_{18}	C[C@@H]1[C@H](C)[C@H](C)CCC1
84	3,3-diethylpentane	C_9H_{20}	CCC(CC)(CC)CC
85	2,2,3,3-tetramethylpentane	C_9H_{20}	CC(C)(C)C(C)(C)CC
86	2,2,3,4-tetramethylpentane	C_9H_{20}	CC(C)(C)C(C)C(C)C
87	2,2,4,4-tetramethylpentane	C_9H_{20}	CC(C)(C)CC(C)(C)C
88	2,3,3,4-tetramethylpentane	C_9H_{20}	CC(C)C(C)(C)C(C)C
89	2-methyloctane	C_9H_{20}	CC(C)CCCCCC
90	3-methyloctane	C_9H_{20}	CCC(C)CCCCC
91	4-methyloctane	C_9H_{20}	CCCC(C)CCCC
92	2-methylnonane	$C_{10}H_{22}$	CC(C)CCCCCC
93	3-methylnonane	$C_{10}H_{22}$	CCC(C)CCCCC
94	5-methylnonane	$C_{10}H_{22}$	CCCCC(C)CCC
95	4-methyldecane	$C_{11}H_{24}$	CCCC(C)CCCCC

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
96	3-ethyloctane	$C_{10}H_{22}$	CCC(CC)CCCCC
97	3,3,5-trimethylheptane	$C_{10}H_{22}$	CCC(C)(C)CC(C)CC
98	2,2,3,3-tetramethylhexane	$C_{10}H_{22}$	CC(C)(C)C(C)(C)CCC
99	propylene	C_3H_6	C=CC
100	1-butene	C_4H_8	C=CCC
101	1-pentene	C_5H_{10}	C=CCCC
102	1-hexene	C_6H_{12}	C=CCCCC
103	1-heptene	C_7H_{14}	C=CCCCCC
104	1-octene	C_8H_{16}	C=CCCCCCC
105	1-nonene	C_9H_{18}	C=CCCCCCCC
106	1-decene	$C_{10}H_{20}$	C=CCCCCCCCC
107	1-undecene	$C_{11}H_{22}$	C=CCCCCCCCCC
108	1-dodecene	$C_{12}H_{24}$	C=CCCCCCCCCCC
109	1-tridecene	$C_{13}H_{26}$	C=CCCCCCCCCCCC
110	1-tetradecene	$C_{14}H_{28}$	C=CCCCCCCCCCCCC
111	1-pentadecene	$C_{15}H_{30}$	C=CCCCCCCCCCCCCC
112	1-hexadecene	$C_{16}H_{32}$	C=CCCCCCCCCCCCCCC
113	1-heptadecene	$C_{17}H_{34}$	C=CCCCCCCCCCCCCCCC
114	1-octadecene	$C_{18}H_{36}$	C=CCCCCCCCCCCCCCCCC
115	1-nonadecene	$C_{19}H_{38}$	C=CCCCCCCCCCCCCCCCCC
116	1-eicosene	$C_{20}H_{40}$	C=CCCCCCCCCCCCCCCCCCC
117	1,3-butadiene	C_4H_6	C=CC=C
118	cis-2-butene	C_4H_8	C/C=C\C
119	trans-2-butene	C_4H_8	C/C=C/C
120	i-butene	C_4H_8	C=C(C)C

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
121	isoprene	C_5H_8	C=CC(C)=C
122	2,3-dimethyl-1-butene	C_6H_{12}	C=C(C)C(C)C
123	2,3-dimethyl-2-butene	C_6H_{12}	C/C(C)=C(C)/C
124	cis-2-hexene	C_6H_{12}	C/C=C\CCC
125	trans-2-hexene	C_6H_{12}	C/C=C/CCC
126	4-methyl-1-pentene	C_6H_{12}	C=CCC(C)C
127	2,4,4-trimethyl-1-pentene	C_8H_{16}	C=C(C)CC(C)(C)C
128	2,4,4-trimethyl-2-pentene	C_8H_{16}	C/C(C)=C\C(C)(C)C
129	2-methyl-1-butene	C_5H_{10}	C=C(C)CC
130	2-methyl-2-butene	C_5H_{10}	C/C(C)=C/C
131	3-methyl-1-butene	C_5H_{10}	C=CC(C)C
132	2,3-dimethyl-1,3-butadiene (2,3-dimethylbutadiene)	C_6H_{10}	C=C(C)C(C)=C
133	3,3-dimethyl-1-butene	C_6H_{12}	C=CC(C)(C)C
134	2-methyl-2-pentene	C_6H_{12}	C/C(C)=C\CC
135	1,5-hexadiene	C_6H_{10}	C=CCCC=C
136	cyclopropane	C_3H_6	C1CC1
137	cyclobutane	C_4H_8	C1CCC1
138	cyclopentane	C_5H_{10}	C1CCCC1
139	cyclohexane	C_6H_{12}	C1CCCCC1
140	cycloheptane	C_7H_{14}	C1CCCCCC1
141	cyclooctane	C_8H_{16}	C1CCCCCCC1
142	methylcyclohexane	C_7H_{14}	CC1CCCCC1
143	ethylcyclohexane	C_8H_{16}	CCC1CCCCC1
144	n-propylcyclohexane	C_9H_{18}	CCCC1CCCCC1
145	i-propylcyclohexane	C_9H_{18}	CC(C1CCCCC1)C

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
146	butylcyclohexane	$C_{10}H_{20}$	CCCCC1CCCCC1
147	sec-butylcyclohexane	$C_{10}H_{20}$	CCC(C1CCCCC1)C
148	methylcyclopentane	C_6H_{12}	CC1CCCC1
149	ethylcyclopentane	C_7H_{14}	CCC1CCCC1
150	n-propylcyclopentane	C_8H_{16}	CCCC1CCCC1
151	i-propylcyclopentane	C_8H_{16}	CC(C1CCCC1)C
152	n-butylcyclopentane	C_9H_{18}	CCCCC1CCCC1
153	i-butylcyclopentane	C_9H_{18}	CC(C)CC1CCCC1
154	pentylcyclopentane	$C_{10}H_{20}$	CCCCCC1CCCC1
155	hexylcyclopentane	$C_{11}H_{22}$	CCCCCCC1CCCC1
156	heptylcyclopentane	$C_{12}H_{24}$	CCCCCCCC1CCCC1
157	octylcyclopentane	$C_{13}H_{26}$	CCCCCCCCC1CCCC1
158	nonylcyclopentane	$C_{14}H_{28}$	CCCCCCCCCC1CCCC1
159	decylcyclopentane	$C_{15}H_{30}$	CCCCCCCCCCC1CCCC1
160	dodecylcyclopentane	$C_{17}H_{34}$	CCCCCCCCCCCCC1CCCC1
161	tridecylcyclopentane	$C_{18}H_{36}$	CCCCCCCCCCCCCC1CCCC1
162	tetradecylcyclopentane	$C_{19}H_{38}$	CCCCCCCCCCCCCCC1CCCC1
163	pentadecylcyclopentane	$C_{20}H_{40}$	CCCCCCCCCCCCCCCC1CCCC1
164	1-methyl-4-isopropylcyclohexane	$C_{10}H_{20}$	CC(C1CCC(C)CC1)C
165	1,1-dimethylcyclopentane	C_7H_{14}	CC1(C)CCCC1
166	1-trans-3-dimethylcyclopentane	C_7H_{14}	C[C@H]1C[C@H](C)CC1
167	1-cis-3-dimethylcyclopentane	C_7H_{14}	C[C@@H]1C[C@H](C)CC1
168	1-trans-2-dimethylcyclopentane	C_7H_{14}	C[C@H]1[C@H](C)CCC1
169	1,1-dimethylcyclohexane	C_8H_{16}	CC1(C)CCCCC1
170	1-cis-2-dimethylcyclohexane	C_8H_{16}	C[C@@H]1[C@H](C)CCCC1

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
171	1-trans-2-dimethylcyclohexane	C_8H_{16}	C[C@H]1[C@H](C)CCCC1
172	1-cis,3-dimethylcyclohexane	C_8H_{16}	C[C@@H]1C[C@H](C)CCC1
173	1-trans,3-dimethylcyclohexane	C_8H_{16}	C[C@H]1C[C@H](C)CCC1
174	1-cis-4dimethylcyclohexane	C_8H_{16}	C[C@@H]1CC[C@H](C)CC1
175	1-trans,4-dimethylcyclohexane	C_8H_{16}	C[C@H]1CC[C@H](C)CC1
176	cyclopentene	C_5H_8	C1=CCCC1
177	cyclohexene	C_6H_{10}	C1=CCCCC1
178	benzene	C_6H_6	C1=CC=CC=C1
179	toluene	C_7H_8	CC1=CC=CC=C1
180	ethylbenzene	C_8H_{10}	CCC1=CC=CC=C1
181	n-propylbenzene	C_9H_{12}	CCCC1=CC=CC=C1
182	i-propylbenzene (cumene)	C_9H_{12}	CC(C1=CC=CC=C1)C
183	butylbenzene	$C_{10}H_{14}$	CCCCC1=CC=CC=C1
184	o-xylene	C_8H_{10}	CC1=CC=CC=C1C
185	m-xylene	C_8H_{10}	CC1=CC=CC(C)=C1
186	p-xylene	C_8H_{10}	CC1=CC=C(C)C=C1
187	pentylbenzene	$C_{11}H_{16}$	CCCCCC1=CC=CC=C1
188	hexylbenzene	$C_{12}H_{18}$	CCCCCCC1=CC=CC=C1
189	heptylbenzene	$C_{13}H_{20}$	CCCCCCCC1=CC=CC=C1
190	octylbenzene	$C_{14}H_{22}$	CCCCCCCCC1=CC=CC=C1
191	nonylbenzene	$C_{15}H_{24}$	CCCCCCCCCC1=CC=CC=C1
192	decylbenzene	$C_{16}H_{26}$	CCCCCCCCCCC1=CC=CC=C1
193	undecylbenzene	$C_{17}H_{28}$	CCCCCCCCCCCC1=CC=CC=C1
194	dodecylbenzene	$C_{18}H_{30}$	CCCCCCCCCCCCC1=CC=CC=C1
195	tridecylbenzene	$C_{19}H_{32}$	CCCCCCCCCCCCCC1=CC=CC=C1

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
196	α-methylstyrene	C_9H_{10}	C=C(C)C1=CC=CC=C1
197	m-ethyltoluene	C_9H_{12}	CC1=CC=CC(CC)=C1
198	o-ethyltoluene	C_9H_{12}	CC1=CC=CC=C1CC
199	p-ethyltoluene	C_9H_{12}	CC1=CC=C(CC)C=C1
200	2-methyl-4-ethylhexane	C_9H_{20}	CC(C)CC(CC)CC
201	3-methyl-5-ethylheptane	$C_{10}H_{22}$	CCC(C)CC(CC)CC
202	1-methyl-1-ethylcyclohexane	C_9H_{18}	CCC1(C)CCCCC1
203	1-cis-methyl-3-ethylcyclohexane	C_9H_{18}	CC[C@@H]1C[C@H](C)CCC1
204	1-trans-methyl-2-ethylcyclohexane	C_9H_{18}	CC[C@H]1[C@H](C)CCCC1
205	1-ethyl-2,3-dimethylcyclohexane	$C_{10}H_{20}$	CC1C(C)C(CC)CCC1
206	1-methyl-ethyl-benzene	C_9H_{12}	CC(C1=CC=CC=C1)C
207	1-methyl-3-ethyl-benzene	C_9H_{12}	CCC1=CC(C)=CC=C1
208	1-methyl-4-ethyl-benzene	C_9H_{12}	CCC1=CC=C(C)C=C1
209	1-methyl-2-ethyl-benzene	C_9H_{12}	CCC1=CC=CC=C1C
210	1,2,3-trimethylbenzene	C_9H_{12}	CC1=C(C)C(C)=CC=C1
211	1,2,4-trimethylbenzene	C_9H_{12}	CC1=CC=C(C)C(C)=C1
212	1,3,5-trimethylbenzene	C_9H_{12}	CC1=CC(C)=CC(C)=C1
213	1,2-methyl-isopropylbenzene	$C_{10}H_{14}$	CC1=CC=CC=C1C(C)C
214	1,3-methyl-isopropylbenzene	$C_{10}H_{14}$	CC1=CC(C(C)C)=CC=C1
215	1,2,3,4-tetrahydronaphthalene	$C_{10}H_{12}$	C12=C(CCCC2)C=CC=C1
216	tert-butylbenzene	$C_{10}H_{14}$	CC(C1=CC=CC=C1)(C)C
217	i-butylbenzene	$C_{10}H_{14}$	CC(C)CC1=CC=CC=C1
218	diphenylmethane	$C_{13}H_{12}$	C1(CC2=CC=CC=C2)=CC=CC=C1
219	sec-butylbenzene	$C_{10}H_{14}$	CCC(C1=CC=CC=C1)C
220	p-diethylbenzene	$C_{10}H_{14}$	CCC1=CC=C(CC)C=C1

ID de base de données	Nom *IUPAC*	Formules chimiques	Chaînes des écritures *SMILES*
221	diphenyl (biphenyl)	$C_{12}H_{10}$	C1(C2=CC=CC=C2)=CC=CC=C1
222	naphthalene	$C_{10}H_8$	C12=CC=CC=C1C=CC=C2
223	2,2,3,3-tetramethylbutane	C_8H_{18}	CC(C)(C)C(C)(C)C

Annexe C.

Tableau C.1. Caractéristiques des descripteurs pertinentes sélectionnés pour *Tb*.

Descripteurs	Catégorie	Description	*t*-value	*P*-Value
STN	topological descriptors	spanning tree number (log)	21,09052	0,00
TIE	topological descriptors	E-state topological parameter	13,74772	$6{,}041\ 10^{-31}$
IVDM	information indices	mean information content on the vertex degree magnitude	19,95857	0,00
Yindex	information indices	Balaban Y index	-3,27603	$1{,}235\ 10^{-3}$
IC3	information indices	information content index (neighborhood symmetry of 3-order)	10,38857	$1{,}343\ 10^{-20}$
CIC3	information indices	complementary information content (neighborhood symmetry of 3-order)	11,24742	$3{,}421\ 10^{-23}$
EEig07x	edge adjacency indices	Eigenvalue 07from edge adj. matrix weighted by edge degrees	6,90539	$6{,}094\ 10^{-11}$
GGI4	topological charge indices	topological charge index of order 4	-6,35177	$1{,}338\ 10^{-9}$
JGI1	topological charge indices	mean topological charge index of order1	2,84976	$4{,}826\ 10^{-3}$
Mor10v	3D-MoRSE descriptors	3D-MoRSE - signal 10 / weighted by atomic van der Waals volumes	-6,36366	$1{,}255\ 10^{-9}$
Mor16e	3D-MoRSE descriptors	3D-MoRSE - signal 16 / weighted by atomic Sanderson electronegativities	3,17347	$1{,}737\ 10^{-3}$
ISH	GETAWAY descriptors	standardized information content on the leverage equality	4,62401	$6{,}64\ 10^{-6}$
H5e	GETAWAY descriptors	H autocorrelation of lag 5 / weighted by Sanderson electronegativity	10,42748	$1{,}027\ 10^{-20}$
R3e	GETAWAY descriptors	R autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities	-6,10791	$4{,}958\ 10^{-9}$
R4e	GETAWAY descriptors	R autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities	-8,67764	$1{,}227\ 10^{-15}$
nCt	Functional group counts	number of total tertiary C(sp3)	-8,59915	$2{,}031\ 10^{-15}$

Résumé de régression pour la variable dépendante : *Tb* (K) ; $N = 223$; $R = 0{,}9995$; $R^2 = 0{,}9989$; R^2 ajusté = 0,9988 ; F (16,206) = 11942 (Valeur statistique de Fisher) ; $P < 0{,}0000$ (Valeur de probabilité) ; Erreur Std d'estimation : 3,3188 K.

Tableau C.2. Caractéristiques des descripteurs pertinentes sélectionnés pour *d20*.

Descripteurs	Catégorie	Description	*t*-value	*P*-Value
GNar	topological descriptors	Narumi geometric topological index	9,8530	$5{,}368\ 10^{-19}$
J	topological descriptors	Balaban distance connectivity index	2,8475	$4{,}855\ 10^{-3}$
piPC03	Walk and path counts	Molecular multiple path count of order 03	23,6495	0,00
X3Av	connectivity indices	average valence connectivity index chi-3	6,4477	$7{,}992\ 10^{-10}$
IVDE	information indices	mean information content on the vertex degree equality	-8,5350	$3{,}129\ 10^{-15}$
HDcpx	information indices	graph distance complexity index (log)	10,2783	$2{,}976\ 10^{-20}$
Yindex	information indices	Balaban Y index	-10,7656	$1{,}036\ 10^{-21}$
IC3	information indices	information content index (neighborhood symmetry of 3-order)	-17,1642	$1{,}599\ 10^{-41}$
CIC3	information indices	complementary information content (neighborhood symmetry of 3-order)	-20,4706	0,00
GGI5	topological charge indices	topological charge index of order 5	3,4135	$7{,}730\ 10^{-4}$
JGI1	topological charge indices	mean topological charge index of order1	7,5165	$1{,}719\ 10^{-12}$
SPH	Geometrical descriptors	Spherosity	-3,9838	$9{,}422\ 10^{-5}$
RDF030m	RDF descriptors	Radial Distribution Function - 3.0 / weighted by atomic masses	8,2985	$1{,}401\ 10^{-14}$
RDF055m	RDF descriptors	Radial Distribution Function - 5.5 / weighted by atomic masses	-9,2169	$3{,}747\ 10^{-17}$
L3p	WHIM descriptors	3rd component size directional WHIM index / weighted by atomic polarizabilities	-4,3278	$2{,}353\ 10^{-5}$
HATS3u	GETAWAY descriptors	leverage-weighted autocorrelation of lag 3 / unweighted	-3,1875	$1{,}660\ 10^{-3}$

Résumé de régression pour la variable dépendante : *d20* (-) ; $N = 222$; $R = 0{,}9947$; $R^2 = 0{,}9894$; R^2 ajusté = 0,9885 ; F (16,205) = 1193,2 (Valeur statistique de Fisher) ; $P < 0{,}0000$ (Valeur de probabilité) ; Erreur Std d'estimation : 0,00885 (-).

Tableau C.3. Caractéristiques des descripteurs pertinentes sélectionnés pour *Tc*.

Descripteurs	Catégorie	Description	*t*-value	*P*-Value
piPC07	Walk and path counts	Molecular multiple path count of order 07	9,9241	$2{,}619\ 10^{-19}$
H5e	GETAWAY descriptors	H autocorrelation of lag 5 / weighted by Sanderson electronegativity	9,3649	$1{,}156\ 10^{-17}$
MWC10	Walk and path counts	molecular walk count of order 10	38,7568	0
Xindex	Information indices	Balaban X index	-15,5738	$5{,}774\ 10^{-37}$
SPH	Geometrical descriptors	spherosity	6,2239	$2{,}558\ 10^{-9}$
D/D	topological descriptors	distance/detour index	8,8550	$3{,}385\ 10^{-16}$
Mor11v	3D-MoRSE descriptors	signal 11 / weighted by van der Waals volume	8,3354	$9{,}69\ 10^{-15}$
Mor15u	3D-MoRSE descriptors	signal 15 / unweighted	-8,5351	$2{,}7\ 10^{-15}$
nCt	Functional group counts	number of total tertiary C(sp3)	-5,2924	$3{,}01\ 10^{-7}$
HATS4u	GETAWAY descriptors	leverage-weighted autocorrelation of lag 4 / unweighted	-3,8064	$1{,}846\ 10^{-4}$

Résumé de régression pour la variable dépendante : *Tc* (K) ; $N = 223$; $R = 0{,}9935$; $R^2 = 0{,}9869$; R^2 ajusté = 0,9864 ; F (10,212) = 1609,6 (Valeur statistique de Fisher) ; $P < 0{,}0000$ (Valeur de probabilité) ; Erreur Std d'estimation : 11,269 K.

Tableau C.4. Caractéristiques des descripteurs pertinentes sélectionnés pour *Pc*.

Descripteurs	Catégorie	Description	*t*-value	*P*-Value
CIC0	Information indices	complementary information content (neighborhood symmetry of 0-order)	-31,9897	0
P2u	WHIM descriptors	2nd component shape directional WHIM index / unweighted	7,0086	$3{,}484\ 10^{-11}$
JGT	2D autocorrelations	global topological charge index	-11,1597	$7{,}345\ 10^{-23}$
RDF035u	RDF descriptors	Radial Distribution Function - 3.5 / unweighted	-3,7752	$2{,}099\ 10^{-4}$
PJI3	Geometrical descriptors	3D Petijean shape index	-3,5734	$4{,}403\ 10^{-4}$
RDF030v	RDF descriptors	Radial Distribution Function - 3.0 / weighted by atomic van der Waals volumes	9,1984	$4{,}473\ 10^{-17}$
R2m	GETAWAY descriptors	R autocorrelation of lag 2 / weighted by atomic masses	14,7241	$7{,}265\ 10^{-34}$
BAC	topological descriptors	Balaban centric index	7,7187	$5{,}235\ 10^{-13}$
RDF055m	RDF descriptors	Radial Distribution Function - 5.5 / weighted by atomic masses	-4,2201	$3{,}681\ 10^{-5}$
EEig06r	edge adjacency indices	Eigenvalue 06 from edge adj. matrix weighted by resonance integrals	-5,0029	$1{,}220\ 10^{-6}$
Vindex	information indices	Balaban V index	6,1781	$3{,}488\ 10^{-9}$
R1e	GETAWAY descriptors	R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities	-8,1583	$3{,}507\ 10^{-14}$
H0e	GETAWAY descriptors	H autocorrelation of lag 0 / weighted by atomic Sanderson electronegativities	4,9835	$1{,}335\ 10^{-6}$
R5e	GETAWAY descriptors	R autocorrelation of lag 5 / weighted by atomic Sanderson electronegativities	4,9838	$1{,}333\ 10^{-6}$
L3u	WHIM descriptors	3rd component size directional WHIM index / unweighted	3,3996	$8{,}125\ 10^{-4}$
X4Av	connectivity indices	average valence connectivity index chi-4	3,6809	$2{,}978\ 10^{-4}$
GGI5	topological charge indices	topological charge index of order 5	3,3490	$9{,}667\ 10^{-4}$

Résumé de régression pour la variable dépendante : *Pc* (Pa) ; $N = 221$; $R = 0{,}9945$; $R^2 = 0{,}9891$; R^2 ajusté = 0,9882 ; F (17,203) = 1086,3 (Valeur statistique de Fisher) ; $P < 0{,}0000$ (Valeur de probabilité) ; Erreur Std d'estimation : 95414,14 Pa.

Tableau C.5. Caractéristiques des descripteurs pertinentes sélectionnés pour *Vc*.

Descriptors	Category	Description	*t*-value	*P*-Value
ADDD	geometrical descriptors	average distance/distance degree	117,5301	0
BIC1	Information indices	bond information content (neighborhood symmetry of 1-order)	7,5585	$1{,}245\ 10^{-12}$
RDF045u	RDF descriptors	Radial Distribution Function - 4.5 / unweighted	-4,7549	$3{,}685\ 10^{-6}$
Mor07u	3D-MoRSE descriptors	3D-MoRSE - signal 07 / unweighted	9,7583	$8{,}648\ 10^{-19}$
Mor30v	3D-MoRSE descriptors	3D-MoRSE - signal 30 / weighted by atomic van der Waals volumes	-3,9579	$1{,}034\ 10^{-4}$
EEig09x	edge adjacency indices	Eigenvalue 09from edge adj. matrix weighted by edge degrees	4,8210	$2{,}738\ 10^{-6}$
JGI1	topological charge indices	mean topological charge index of order1	7,1600	$1{,}334\ 10^{-11}$
HATS3p	GETAWAY descriptors	leverage-weighted autocorrelation of lag 3 / weighted by atomic polarizabilities	-4,5773	$8{,}057\ 10^{-6}$
ESpm09d	edge adjacency indices	Spectral moment 09 from edge adj. matrix weighted by dipole moments	-4,8656	$2{,}238\ 10^{-6}$
X2A	Connectivity indices	average valence connectivity index chi-2	-4,2442	$3{,}292\ 10^{-5}$
H1e	GETAWAY descriptors	H autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities	4,5479	$9{,}154\ 10^{-6}$
Mor08p	3D-MoRSE descriptors	3D-MoRSE - signal 08 / weighted by atomic polarizabilities	4,2276	$3{,}524\ 10^{-5}$

Résumé de régression pour la variable dépendante : *Vc* (cm^3/mol) ; $N = 223$; $R = 0{,}9990$; $R^2 = 0{,}9980$; R^2 ajusté = 0,9979 ; F (12,210) = 8689,2 (Valeur statistique de Fisher) ; $P < 0{,}0000$ (Valeur de probabilité) ; Erreur Std d'estimation : 15,334 cm^3/mol.

Annexe D.

Tableau D.1. Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Tb*.

	STN	***TIE***	***IVDM***	***Yindex***	***IC3***	***CIC3***	***EEig07x***	***GGI4***	***JGI1***	***Mor10v***	***Mor16e***	***ISH***	***H5e***	***R3e***	***R4e***	***nCt***
STN	1,000	-0,225	0,119	-0,266	0,154	-0,219	0,031	0,375	-0,281	-0,019	-0,338	-0,059	-0,330	0,116	-0,164	0,051
TIE		1,000	0,737	-0,342	-0,192	0,715	0,710	-0,056	-0,333	0,740	0,543	0,093	0,735	0,392	0,125	-0,182
IVDM			1,000	-0,519	0,217	0,511	0,735	0,264	-0,485	0,714	**0,746**	0,332	0,741	0,517	0,164	-0,094
Yindex				1,000	-0,070	-0,238	-0,532	-0,265	0,656	-0,507	-0,418	-0,087	-0,163	-0,369	0,352	0,164
IC3					1,000	-0,695	-0,134	0,479	-0,020	-0,126	0,082	0,724	-0,042	0,000	0,064	0,220
CIC3						1,000	0,589	-0,270	-0,213	0,648	0,515	-0,323	0,653	0,435	0,218	-0,166
EEig07x							1,000	-0,070	-0,530	0,624	0,619	0,033	0,591	0,311	-0,168	-0,392
GGI4								1,000	-0,036	-0,074	0,115	0,227	0,001	-0,061	0,018	0,245
JGI1									1,000	-0,558	-0,374	0,034	-0,132	-0,434	0,619	0,252
Mor10v										1,000	0,539	0,159	0,603	0,721	0,052	0,046
Mor16e											1,000	0,211	0,721	0,371	0,087	-0,112
ISH												1,000	0,234	0,184	0,279	0,205
H5e													1,000	0,374	0,391	0,007
R3e														1,000	0,330	0,186
R4e															1,000	0,368
nCt																1,000

Tableau D.2. Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *d20*.

	GNar	***J***	***piPC03***	***X3Av***	***IVDE***	***HDcpx***	***Yindex***	***IC3***	***CIC3***	***GGI5***	***JGI1***	***SPH***	***RDF030m***	***RDF055m***	***L3p***	***HATS3u***
GNar	1,000	-0,507	0,746	-0,203	-0,348	0,579	-0,514	0,184	0,141	0,351	-0,668	0,429	-0,159	0,199	-0,239	-0,033
J		1,000	-0,168	0,406	0,328	0,210	0,514	-0,009	0,193	-0,129	0,588	-0,291	0,499	-0,118	0,486	-0,450
piPC03			1,000	-0,207	-0,022	**0,747**	-0,247	0,361	0,073	0,400	-0,333	0,433	0,001	0,262	-0,044	-0,541
X3Av				1,000	0,052	0,200	0,403	0,071	0,173	-0,060	-0,008	0,053	0,220	-0,140	0,115	-0,123
IVDE					1,000	-0,078	0,444	0,495	-0,580	0,034	0,655	-0,318	0,495	0,100	0,336	-0,254
HDcpx						1,000	-0,301	0,348	0,369	0,431	-0,253	0,471	0,124	0,186	0,128	-0,720
Yindex							1,000	-0,069	-0,240	-0,440	0,657	-0,572	0,350	-0,165	0,362	0,014
IC3								1,000	-0,693	0,408	-0,021	0,227	0,165	0,237	0,130	-0,307
CIC3									1,000	-0,043	-0,213	0,154	-0,062	-0,121	-0,011	-0,200
GGI5										1,000	-0,301	0,378	0,089	0,259	0,007	-0,276
JGI1											1,000	-0,686	0,436	-0,012	0,608	-0,177
SPH												1,000	-0,416	0,049	-0,728	-0,389
RDF030m													1,000	0,141	0,568	-0,188
RDF055m														1,000	0,183	-0,132
L3p															1,000	-0,142
HATS3u																1,000

Tableau D.3. Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Tc*.

	D/D	***MWC10***	***piPC07***	***Xindex***	***SPH***	***Mor15u***	***Mor11v***	***HATS4u***	***H5e***	***nCt***
D/D	1,000	0,263	0,452	-0,390	0,456	-0,701	0,224	-0,518	**0,745**	-0,202
MWC10		1,000	0,444	-0,400	0,077	-0,092	0,232	-0,070	0,400	0,204
piPC07			1,000	-0,733	0,540	-0,119	0,689	-0,563	0,300	-0,288
Xindex				1,000	-0,551	0,268	-0,492	0,534	-0,223	0,072
SPH					1,000	-0,372	0,549	-0,733	0,301	-0,229
Mor15u						1,000	-0,014	0,467	-0,638	-0,084
Mor11v							1,000	-0,496	0,242	-0,398
HATS4u								1,000	-0,445	0,186
H5e									1,000	0,007
nCt										1,000

Tableau D.4. Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Pc*.

	BAC	***X4Av***	***Vindex***	***CIC0***	***EEig06r***	***GGI5***	***JGT***	***PJI3***	***RDF035u***	***RDF055m***	***RDF030v***	***L3u***	***P2u***	***H0e***	***R2m***	***R1e***	***R5e***
BAC	1,000	0,384	-0,058	0,690	0,420	0,030	0,074	0,456	0,167	-0,100	0,586	0,248	-0,431	0,064	-0,257	0,017	0,443
X4Av		1,000	-0,294	0,517	0,041	0,066	-0,066	0,437	0,178	-0,042	0,525	0,188	-0,375	0,125	-0,119	0,143	0,348
Vindex			1,000	-0,546	-0,435	-0,412	0,326	-0,200	0,170	-0,188	-0,281	0,085	0,466	-0,086	-0,405	-0,173	-0,098
CIC0				1,000	0,743	0,377	-0,300	0,541	-0,025	0,085	0,722	0,181	-0,703	0,205	0,183	0,187	0,363
EEig06r					1,000	0,316	-0,579	0,382	-0,417	0,155	0,297	-0,136	-0,678	0,095	0,190	-0,127	-0,076
GGI5						1,000	-0,190	0,335	-0,143	0,255	0,229	0,016	-0,467	0,133	0,243	0,253	0,036
JGT							1,000	-0,005	0,714	0,046	0,078	0,549	0,459	0,133	0,104	0,549	0,470
PJI3								1,000	-0,012	0,032	0,352	0,135	-0,595	0,185	-0,082	0,090	0,236
RDF035u									1,000	0,085	0,456	**0,752**	0,385	0,135	0,282	0,660	0,712
RDF055m										1,000	0,085	0,185	-0,092	0,113	0,401	0,357	0,053
RDF030v											1,000	0,390	-0,263	0,168	0,123	0,385	0,647
L3u												1,000	0,017	0,287	0,393	0,712	0,513
P2u													1,000	-0,110	0,056	0,157	0,074
H0e														1,000	0,208	0,464	0,213
R2m															1,000	0,689	0,137
R1e																1,000	0,578
R5e																	1,000

Tableau D.05. Matrice des coefficients de corrélation des descripteurs sélectionnés pour le modèle de *Vc*.

	X2A	***BIC1***	***EEig09x***	***ESpm09d***	***JGI1***	***ADDD***	***RDF045u***	***Mor07u***	***Mor30v***	***Mor08p***	***H1e***	***HATS3p***
X2A	1,000	0,245	0,061	-0,235	0,251	-0,149	-0,375	-0,298	-0,119	0,449	-0,385	0,055
BIC1		1,000	-0,309	0,367	0,139	-0,518	-0,020	0,117	0,153	-0,127	0,064	0,293
EEig09x			1,000	-0,424	-0,297	0,719	-0,120	0,372	-0,194	0,415	-0,264	-0,338
ESpm09d				1,000	0,610	-0,424	0,481	-0,249	0,210	-0,642	**0,732**	0,254
JGI1					1,000	-0,419	0,075	-0,668	-0,071	-0,322	0,437	0,268
ADDD						1,000	0,149	0,436	-0,338	0,352	-0,088	-0,677
RDF045u							1,000	0,239	0,271	-0,427	0,616	-0,239
Mor07u								1,000	0,157	-0,063	-0,035	-0,176
Mor30v									1,000	-0,268	0,006	0,396
Mor08p										1,000	-0,710	-0,406
H1e											1,000	0,070
HATS3p												1,000

Annexe E.

Tableau E.1. Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Tb ainsi leurs ordres de partition dans les bases d'apprentissage et de test.

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
1	0,000000	0,1340	1,000000	0,000000	0,000000	3,000000	0,00000	0,000000	0,000000	0,035000	0,081000	0,000000	0,000000	1,442000	0,000000	0,000000	Test
2	0,000000	0,7280	1,500000	0,000000	1,371000	2,088000	0,00000	0,000000	0,250000	0,065000	0,119000	0,579000	0,000000	1,777000	0,812000	0,000000	Train
3	0,000000	1,8020	1,918000	3,132000	1,842000	1,965000	0,00000	0,000000	0,167000	0,057000	0,208000	0,500000	0,000000	1,913000	1,182000	0,000000	Test
4	0,000000	1,9600	1,792000	0,000000	1,140000	2,667000	0,00000	0,000000	0,500000	0,032000	0,174000	0,406000	0,000000	1,828000	1,910000	1,000000	Train
5	0,000000	3,3100	2,250000	2,340000	2,352000	1,736000	0,00000	0,000000	0,125000	0,106000	0,295000	0,655000	0,000000	1,918000	1,321000	0,000000	Train
6	0,000000	3,6080	2,156000	3,916000	2,660000	1,427000	0,00000	0,000000	0,375000	0,014000	0,284000	0,828000	0,163000	1,874000	1,917000	1,000000	Train
7	0,000000	5,2540	2,522000	1,958000	2,446000	1,875000	0,00000	0,080000	0,100000	0,197000	0,414000	0,613000	0,063000	2,017000	1,414000	0,000000	Test
8	0,000000	7,6340	2,752000	1,730000	2,566000	1,957000	0,00000	0,080000	0,083000	0,250000	0,537000	0,554000	0,089000	2,065000	1,491000	0,000000	Train
9	0,000000	10,4460	2,950000	1,574000	2,565000	2,135000	-0,84800	0,080000	0,071000	0,264000	0,676000	0,667000	0,288000	2,135000	1,555000	0,000000	Train
10	0,000000	13,7000	3,125000	1,460000	2,531000	2,327000	-0,53200	0,080000	0,063000	0,300000	0,810000	0,720000	0,440000	2,170000	1,612000	0,000000	Train
11	0,000000	17,3830	3,281000	1,372000	2,483000	2,517000	-0,17600	0,080000	0,056000	0,367000	0,950000	0,699000	0,623000	2,224000	1,664000	0,000000	Train
12	0,000000	21,5090	3,422000	1,301000	2,432000	2,698000	0,16900	0,080000	0,050000	0,430000	1,077000	0,737000	0,772000	2,252000	1,711000	0,000000	Train
13	0,000000	26,0670	3,550000	1,242000	2,379000	2,869000	0,48200	0,080000	0,045000	0,468000	1,209000	0,721000	0,922000	2,293000	1,752000	0,000000	Train
14	0,000000	31,0590	3,668000	1,193000	2,328000	3,030000	0,75900	0,080000	0,042000	0,499000	1,321000	0,751000	1,037000	2,315000	1,790000	0,000000	Train
15	0,000000	36,4880	3,777000	1,151000	2,279000	3,181000	1,00000	0,080000	0,038000	0,545000	1,433000	0,737000	1,160000	2,347000	1,823000	0,000000	Train
16	0,000000	42,3600	3,879000	1,114000	2,232000	3,322000	1,20900	0,080000	0,036000	0,605000	1,520000	0,761000	1,252000	2,364000	1,853000	0,000000	Train
17	0,000000	48,6480	3,974000	1,082000	2,188000	3,455000	1,39000	0,080000	0,033000	0,661000	1,601000	0,750000	1,347000	2,390000	1,880000	0,000000	Test
18	0,000000	55,3900	4,063000	1,053000	2,147000	3,581000	1,54700	0,080000	0,031000	0,699000	1,658000	0,770000	1,425000	2,404000	1,905000	0,000000	Train
19	0,000000	62,5700	4,146000	1,027000	2,109000	3,699000	1,68400	0,080000	0,029000	0,733000	1,703000	0,760000	1,504000	2,425000	1,927000	0,000000	Train
20	0,000000	70,1600	4,225000	1,004000	2,072000	3,811000	1,80300	0,080000	0,028000	0,784000	1,724000	0,777000	1,568000	2,437000	1,948000	0,000000	Train
21	0,000000	78,2140	4,301000	0,983000	2,038000	3,916000	1,90800	0,080000	0,026000	0,845000	1,731000	0,769000	1,631000	2,454000	1,966000	0,000000	Train
22	0,000000	86,6830	4,372000	0,964000	2,006000	4,017000	2,00000	0,080000	0,025000	0,894000	1,717000	0,740000	1,686000	2,464000	1,984000	0,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
23	0,000000	95,6220	4,440000	0,946000	1,975000	4,112000	2,08100	0,080000	0,024000	0,928000	1,687000	0,776000	1,739000	2,479000	2,000000	0,000000	Test
24	0,000000	104,9490	4,505000	0,929000	1,947000	4,203000	2,15300	0,080000	0,023000	0,969000	1,642000	0,789000	1,785000	2,487000	2,015000	0,000000	Train
25	0,000000	114,7390	4,567000	0,914000	1,920000	4,290000	2,21800	0,080000	0,022000	1,024000	1,579000	0,782000	1,831000	2,500000	2,028000	0,000000	Test
26	0,000000	124,9600	4,627000	0,900000	1,894000	4,373000	2,27500	0,080000	0,021000	1,079000	1,505000	0,759000	1,869000	2,507000	2,041000	0,000000	Train
27	0,000000	135,6170	4,684000	0,887000	1,870000	4,452000	2,32600	0,080000	0,020000	1,124000	1,417000	0,787000	1,909000	2,518000	2,053000	0,000000	Train
28	0,000000	158,2330	4,792000	0,863000	1,825000	4,601000	2,41400	0,080000	0,019000	1,207000	1,210000	0,792000	1,977000	2,533000	2,075000	0,000000	Train
29	0,000000	208,7290	4,986000	0,823000	1,748000	4,867000	2,54600	0,080000	0,016000	1,396000	0,730000	0,725000	2,088000	2,559000	2,111000	0,000000	Test
30	0,000000	266,1170	5,158000	0,790000	1,684000	5,097000	2,63800	0,080000	0,014000	1,590000	0,249000	0,764000	2,175000	2,580000	2,139000	0,000000	Train
31	0,000000	4,0380	2,000000	0,000000	0,811000	3,276000	0,00000	0,000000	0,750000	-0,062000	0,246000	0,311000	0,000000	1,759000	3,144000	0,000000	Test
32	0,000000	5,6400	2,446000	2,636000	3,009000	1,313000	0,00000	0,000000	0,300000	0,100000	0,375000	0,693000	0,222000	1,856000	1,853000	1,000000	Train
33	0,000000	5,7130	2,446000	3,135000	2,709000	1,613000	0,00000	0,000000	0,300000	0,056000	0,412000	0,742000	0,421000	1,861000	1,888000	1,000000	Train
34	0,000000	6,3540	2,322000	5,184000	2,320000	2,002000	0,00000	0,000000	0,600000	-0,095000	0,248000	0,871000	0,349000	1,795000	2,889000	0,000000	Train
35	0,000000	6,1280	2,371000	3,971000	1,571000	2,751000	0,00000	0,000000	0,400000	-0,070000	0,282000	0,355000	0,503000	1,850000	2,371000	2,000000	Train
36	0,000000	8,1100	2,689000	2,107000	3,295000	1,229000	0,00000	0,160000	0,250000	0,165000	0,564000	1,000000	0,253000	1,936000	1,825000	1,000000	Train
37	0,000000	8,1920	2,689000	2,460000	3,382000	1,142000	0,00000	0,080000	0,250000	0,178000	0,542000	0,898000	0,498000	1,904000	1,839000	1,000000	Train
38	0,000000	8,9860	2,585000	3,210000	2,696000	1,828000	0,00000	0,000000	0,500000	0,043000	0,386000	0,758000	0,427000	1,761000	2,660000	0,000000	Train
39	0,000000	8,8010	2,626000	3,253000	2,903000	1,621000	0,00000	0,000000	0,333000	0,023000	0,481000	1,000000	0,763000	1,790000	2,215000	2,000000	Train
40	0,000000	8,6200	2,626000	2,672000	2,044000	2,479000	0,00000	0,000000	0,417000	0,169000	0,576000	0,351000	0,401000	1,737000	2,199000	2,000000	Train
41	0,000000	9,1540	2,585000	4,096000	2,566000	1,957000	0,00000	0,000000	0,500000	-0,071000	0,249000	0,695000	0,908000	1,842000	2,729000	0,000000	Train
42	0,000000	8,2480	2,689000	2,933000	2,195000	2,328000	0,00000	0,000000	0,250000	0,130000	0,710000	0,516000	0,412000	1,711000	1,677000	1,000000	Train
43	0,000000	9,7180	2,522000	4,656000	2,315000	2,208000	0,00000	0,000000	0,583000	-0,190000	0,268000	0,593000	0,799000	1,744000	3,082000	1,000000	Train
44	0,000000	11,0140	2,896000	1,812000	3,383000	1,318000	-0,83300	0,160000	0,214000	0,192000	0,656000	0,833000	0,349000	1,986000	1,843000	1,000000	Train
45	0,000000	11,1100	2,896000	2,049000	3,459000	1,241000	-0,81400	0,160000	0,214000	0,203000	0,726000	0,833000	0,521000	1,962000	1,830000	1,000000	Test
46	0,000000	11,1000	2,896000	2,165000	3,075000	1,626000	-0,80200	0,080000	0,214000	0,262000	0,677000	0,750000	0,534000	1,895000	1,784000	1,000000	Train
47	0,000000	14,5150	3,078000	2,115000	2,969000	1,889000	-0,61800	0,160000	0,188000	0,293000	0,972000	0,860000	0,403000	1,816000	1,587000	1,000000	Train
48	0,000000	14,4880	3,078000	2,272000	3,047000	1,811000	-0,70000	0,240000	0,188000	0,316000	0,963000	0,667000	0,358000	1,791000	1,528000	1,000000	Train
49	0,000000	12,0540	2,807000	2,436000	3,007000	1,694000	-0,79800	0,240000	0,429000	0,090000	0,550000	0,917000	0,390000	1,840000	2,525000	0,000000	Train
50	0,000000	11,8300	2,842000	2,610000	3,190000	1,511000	-0,77600	0,160000	0,286000	0,132000	0,598000	1,000000	0,754000	1,833000	2,087000	2,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
51	0,000000	11,7250	2,842000	2,451000	3,190000	1,511000	-0,78600	0,160000	0,357000	0,232000	0,745000	0,917000	0,632000	1,795000	2,114000	2,000000	Train
52	0,000000	11,6070	2,842000	2,125000	2,200000	2,501000	-0,81400	0,320000	0,357000	0,129000	0,746000	0,500000	0,461000	1,913000	2,117000	2,000000	Train
53	0,000000	12,2440	2,807000	3,071000	3,200000	1,501000	-0,74600	0,080000	0,429000	0,108000	0,420000	0,917000	0,959000	1,837000	2,548000	0,000000	Test
54	0,000000	11,9380	2,842000	2,887000	2,815000	1,885000	-0,74900	0,080000	0,286000	0,155000	0,716000	0,667000	1,026000	1,829000	2,126000	2,000000	Train
55	0,000000	15,5620	3,000000	2,021000	3,132000	1,725000	-0,28300	0,240000	0,375000	0,124000	0,558000	0,930000	0,501000	1,887000	2,470000	0,000000	Train
56	0,000000	15,3060	3,031000	2,175000	3,434000	1,424000	-0,46800	0,240000	0,250000	0,168000	0,749000	0,930000	0,786000	1,878000	2,051000	2,000000	Train
57	0,000000	15,1700	3,031000	2,179000	3,434000	1,424000	-0,53200	0,160000	0,313000	0,309000	0,816000	0,930000	0,680000	1,781000	2,023000	2,000000	Train
58	0,000000	15,0340	3,031000	1,821000	2,547000	2,311000	-0,30300	0,160000	0,313000	0,201000	0,615000	0,720000	0,549000	1,881000	2,054000	2,000000	Train
59	0,000000	15,7820	3,000000	2,449000	3,444000	1,414000	-0,47600	0,240000	0,375000	0,117000	0,587000	0,860000	1,023000	1,924000	2,463000	0,000000	Train
60	0,000000	15,4170	3,031000	2,471000	3,444000	1,414000	-0,51800	0,160000	0,250000	0,229000	0,829000	0,930000	1,029000	1,846000	2,031000	2,000000	Train
61	0,000000	15,2980	3,031000	2,282000	3,099000	1,759000	-0,44100	0,240000	0,313000	0,252000	0,857000	0,930000	0,655000	1,808000	2,021000	2,000000	Train
62	0,000000	19,5100	3,170000	1,764000	3,202000	1,798000	0,00000	0,240000	0,333000	0,155000	0,688000	0,880000	0,585000	1,939000	2,422000	0,000000	Test
63	0,000000	19,2190	3,197000	1,884000	3,522000	1,478000	-0,22800	0,240000	0,222000	0,159000	0,851000	0,880000	0,921000	1,934000	2,034000	2,000000	Test
64	0,000000	19,0650	3,197000	1,937000	3,647000	1,353000	0,00000	0,240000	0,278000	0,310000	1,020000	0,940000	0,785000	1,855000	2,007000	2,000000	Train
65	0,000000	19,0430	3,197000	1,887000	3,350000	1,650000	0,00000	0,320000	0,278000	0,219000	0,960000	1,000000	0,787000	1,943000	2,020000	2,000000	Test
66	0,000000	19,7630	3,170000	2,063000	3,531000	1,469000	-0,26300	0,240000	0,333000	0,112000	0,612000	0,797000	1,160000	1,969000	2,421000	0,000000	Train
67	0,000000	19,1870	3,197000	1,932000	3,156000	1,844000	-0,37000	0,320000	0,278000	0,165000	0,914000	0,579000	1,055000	1,970000	2,057000	2,000000	Train
68	0,000000	11,1570	2,896000	2,481000	2,988000	1,713000	-0,75700	0,160000	0,214000	0,239000	0,693000	0,719000	0,381000	1,766000	1,596000	1,000000	Train
69	0,000000	13,0440	2,753000	3,670000	2,978000	1,723000	-0,71100	0,000000	0,500000	-0,062000	0,449000	1,000000	1,114000	1,704000	2,825000	1,000000	Train
70	0,000000	12,7260	2,753000	3,005000	2,670000	2,031000	-0,76200	0,000000	0,571000	0,009000	0,261000	1,000000	0,457000	1,811000	2,861000	1,000000	Train
71	0,000000	13,1340	2,753000	4,013000	3,094000	1,607000	-0,68100	0,000000	0,500000	-0,123000	0,296000	0,917000	1,384000	1,743000	2,851000	1,000000	Train
72	0,000000	12,6120	2,788000	3,265000	1,985000	2,716000	-0,73200	0,000000	0,357000	0,072000	0,389000	1,000000	1,106000	1,696000	2,431000	3,000000	Train
73	0,000000	11,8960	2,842000	3,082000	2,969000	1,732000	-0,69400	0,000000	0,286000	0,213000	0,822000	0,917000	0,607000	1,788000	2,029000	2,000000	Train
74	0,000000	12,4120	2,807000	3,750000	2,458000	2,242000	-0,61800	0,000000	0,429000	0,043000	0,653000	0,833000	1,039000	1,861000	2,470000	0,000000	Train
75	1,609000	8,3910	2,828000	2,914000	3,407000	1,178000	-0,67800	0,000000	0,375000	0,324000	-0,336000	1,000000	0,458000	2,131000	2,880000	1,000000	Test
76	1,609000	8,2730	2,828000	2,469000	3,407000	1,178000	-0,74600	0,000000	0,438000	0,375000	-0,218000	0,917000	0,207000	2,179000	2,894000	1,000000	Train
77	1,609000	8,0910	2,858000	2,341000	2,730000	1,855000	-0,76800	0,000000	0,313000	0,237000	0,066000	0,583000	0,452000	2,349000	2,457000	3,000000	Train
78	1,609000	8,1520	2,858000	2,511000	2,730000	1,855000	-0,73200	0,000000	0,250000	0,494000	-0,141000	0,917000	0,533000	2,204000	2,467000	3,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
79	0,000000	16,7070	2,953000	2,912000	3,244000	1,614000	-0,43100	0,240000	0,438000	0,023000	0,595000	0,930000	1,058000	1,739000	2,619000	1,000000	Train
80	0,000000	16,4670	2,953000	2,658000	3,244000	1,614000	-0,38600	0,240000	0,500000	0,093000	0,562000	1,000000	0,657000	1,860000	2,704000	1,000000	Test
81	0,000000	16,2940	2,953000	2,324000	2,969000	1,889000	0,00000	0,480000	0,500000	0,054000	0,725000	1,000000	0,555000	1,813000	2,661000	1,000000	Train
82	0,000000	16,0530	2,983000	2,527000	2,630000	2,228000	-0,42800	0,320000	0,375000	0,178000	0,914000	0,930000	0,858000	1,726000	2,274000	3,000000	Train
83	1,792000	10,3440	3,017000	2,060000	3,634000	1,121000	-0,28200	0,333000	0,389000	0,184000	0,488000	0,790000	0,358000	2,209000	2,934000	1,000000	Train
84	0,000000	16,0980	3,000000	3,634000	1,995000	2,863000	-0,61800	0,000000	0,375000	0,213000	1,057000	0,439000	1,328000	1,823000	2,323000	0,000000	Train
85	0,000000	18,4260	2,875000	4,278000	2,371000	2,487000	-0,42200	0,000000	0,563000	-0,235000	0,357000	0,763000	1,722000	1,656000	3,283000	0,000000	Train
86	0,000000	17,6950	2,906000	3,535000	2,658000	2,200000	-0,35200	0,000000	0,500000	0,022000	0,160000	1,000000	1,214000	1,748000	2,988000	2,000000	Train
87	0,000000	17,7220	2,875000	3,164000	1,597000	3,261000	0,00000	0,000000	0,688000	0,061000	0,635000	0,860000	0,515000	1,880000	3,339000	0,000000	Train
88	0,000000	17,9070	2,906000	3,955000	2,357000	2,501000	-0,44900	0,000000	0,500000	-0,100000	0,543000	0,720000	1,706000	1,647000	2,941000	2,000000	Train
89	0,000000	14,3570	3,078000	1,622000	3,417000	1,440000	-0,44600	0,160000	0,188000	0,227000	0,772000	0,930000	0,490000	2,047000	1,858000	1,000000	Train
90	0,000000	14,4690	3,078000	1,787000	3,417000	1,440000	-0,46400	0,160000	0,188000	0,197000	0,822000	1,000000	0,690000	2,025000	1,845000	1,000000	Train
91	0,000000	14,4540	3,078000	1,907000	3,470000	1,388000	-0,57300	0,160000	0,188000	0,252000	0,867000	1,000000	0,659000	1,969000	1,793000	1,000000	Train
92	0,000000	18,1350	3,239000	1,489000	3,367000	1,633000	0,00000	0,160000	0,167000	0,272000	0,868000	0,819000	0,632000	2,096000	1,897000	1,000000	Train
93	0,000000	18,2670	3,239000	1,608000	3,429000	1,571000	-0,33000	0,160000	0,167000	0,252000	0,935000	0,880000	0,810000	2,065000	1,860000	1,000000	Train
94	0,000000	18,2450	3,239000	1,751000	3,367000	1,633000	0,00000	0,240000	0,167000	0,264000	1,053000	0,759000	0,726000	2,011000	1,803000	1,000000	Train
95	0,000000	22,4830	3,384000	1,560000	3,593000	1,536000	0,22700	0,160000	0,150000	0,355000	1,066000	0,842000	0,946000	2,061000	1,833000	1,000000	Train
96	0,000000	18,3200	3,239000	1,853000	3,108000	1,892000	-0,37000	0,160000	0,167000	0,347000	0,888000	0,940000	0,565000	1,882000	1,602000	1,000000	Test
97	0,000000	20,8050	3,128000	2,600000	3,632000	1,368000	-0,37300	0,320000	0,444000	0,112000	0,623000	1,000000	1,299000	1,939000	2,613000	1,000000	Train
98	0,000000	22,7580	3,059000	3,456000	2,684000	2,316000	0,00000	0,240000	0,500000	-0,099000	0,573000	0,880000	1,610000	1,655000	2,996000	0,000000	Train
99	0,000000	0,4750	1,500000	0,000000	2,156000	1,014000	0,00000	0,000000	0,250000	-0,071000	-0,040000	1,000000	0,000000	1,051000	0,693000	0,000000	Test
100	0,000000	1,4820	1,918000	3,132000	2,855000	0,730000	0,00000	0,000000	0,167000	-0,003000	0,093000	1,000000	0,002000	1,563000	0,880000	0,000000	Train
101	0,000000	2,9900	2,250000	2,340000	3,190000	0,717000	0,00000	0,000000	0,125000	0,003000	0,118000	1,000000	0,042000	1,767000	1,189000	0,000000	Train
102	0,000000	4,9210	2,522000	1,958000	3,461000	0,709000	0,00000	0,080000	0,100000	0,015000	0,255000	1,000000	0,113000	1,806000	1,328000	0,000000	Train
103	0,000000	7,2840	2,752000	1,730000	3,690000	0,703000	0,00000	0,080000	0,083000	0,091000	0,355000	1,000000	0,197000	1,909000	1,427000	0,000000	Train
104	0,000000	10,0780	2,950000	1,574000	3,720000	0,865000	-0,82700	0,080000	0,071000	0,174000	0,490000	0,917000	0,286000	1,958000	1,494000	0,000000	Train
105	0,000000	13,3090	3,125000	1,460000	3,708000	1,047000	-0,47800	0,080000	0,063000	0,204000	0,612000	1,000000	0,446000	2,030000	1,562000	0,000000	Train
106	0,000000	16,9750	3,281000	1,372000	3,615000	1,292000	-0,09400	0,080000	0,056000	0,216000	0,753000	0,940000	0,577000	2,069000	1,610000	0,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
107	0,000000	21,0780	3,422000	1,301000	3,507000	1,537000	0,26900	0,080000	0,050000	0,269000	0,885000	1,000000	0,726000	2,124000	1,664000	0,000000	Train
108	0,000000	25,6150	3,550000	1,242000	3,398000	1,772000	0,59300	0,080000	0,045000	0,347000	1,027000	1,000000	0,867000	2,156000	1,703000	0,000000	Train
109	0,000000	30,5870	3,668000	1,193000	3,294000	1,991000	0,87400	0,080000	0,042000	0,399000	1,154000	1,000000	0,992000	2,199000	1,746000	0,000000	Test
110	0,000000	35,9950	3,777000	1,151000	3,196000	2,197000	1,11600	0,080000	0,038000	0,422000	1,288000	1,000000	1,102000	2,225000	1,778000	0,000000	Test
111	0,000000	41,8460	3,879000	1,114000	3,104000	2,388000	1,32400	0,080000	0,036000	0,456000	1,397000	0,966000	1,203000	2,259000	1,813000	0,000000	Train
112	0,000000	48,1240	3,974000	1,082000	3,018000	2,567000	1,50100	0,080000	0,033000	0,518000	1,503000	0,969000	1,295000	2,280000	1,840000	0,000000	Train
113	0,000000	54,8420	4,063000	1,053000	2,939000	2,733000	1,65400	0,080000	0,031000	0,582000	1,595000	0,971000	1,378000	2,307000	1,868000	0,000000	Train
114	0,000000	61,9940	4,146000	1,027000	2,866000	2,889000	1,78600	0,080000	0,029000	0,625000	1,663000	0,947000	1,454000	2,325000	1,890000	0,000000	Train
115	0,000000	69,5690	4,225000	1,004000	2,797000	3,035000	1,90000	0,080000	0,028000	0,654000	1,719000	0,926000	1,522000	2,348000	1,914000	0,000000	Train
116	0,000000	77,5990	4,301000	0,983000	2,734000	3,173000	2,00000	0,080000	0,026000	0,697000	1,752000	0,977000	1,586000	2,362000	1,933000	0,000000	Train
117	0,000000	1,0630	1,918000	3,132000	1,922000	1,400000	0,00000	0,000000	0,167000	-0,107000	-0,064000	0,500000	0,000000	0,736000	0,502000	0,000000	Train
118	0,000000	1,2070	1,918000	3,132000	1,792000	1,792000	0,00000	0,000000	0,167000	0,008000	-0,097000	0,500000	0,494000	1,183000	0,777000	0,000000	Train
119	0,000000	1,2070	1,918000	3,132000	1,792000	1,792000	0,00000	0,000000	0,167000	-0,046000	0,049000	0,500000	0,151000	1,045000	0,832000	0,000000	Train
120	0,000000	1,5040	1,792000	0,000000	1,686000	1,899000	0,00000	0,000000	0,500000	-0,097000	0,167000	0,750000	0,000000	0,961000	1,794000	0,000000	Test
121	0,000000	2,5880	2,156000	3,916000	3,027000	0,673000	0,00000	0,000000	0,375000	-0,139000	0,073000	0,828000	0,123000	0,879000	1,312000	0,000000	Test
122	0,000000	5,5340	2,371000	3,971000	2,822000	1,348000	0,00000	0,000000	0,400000	-0,065000	0,410000	0,693000	0,336000	1,365000	1,992000	1,000000	Train
123	0,000000	4,5170	2,371000	3,971000	1,224000	2,946000	0,00000	0,000000	0,400000	-0,117000	0,015000	0,355000	0,842000	0,797000	1,878000	0,000000	Train
124	0,000000	4,3610	2,522000	1,958000	3,419000	0,751000	0,00000	0,080000	0,100000	0,058000	0,038000	0,871000	0,323000	1,706000	1,115000	0,000000	Test
125	0,000000	4,3610	2,522000	1,958000	3,419000	0,751000	0,00000	0,080000	0,100000	-0,020000	0,153000	0,871000	0,174000	1,696000	1,165000	0,000000	Test
126	0,000000	5,3610	2,446000	2,636000	2,975000	1,195000	0,00000	0,000000	0,300000	0,036000	0,147000	1,000000	0,023000	1,753000	1,834000	1,000000	Train
127	0,000000	12,1710	2,753000	3,005000	2,833000	1,752000	-0,73400	0,000000	0,571000	-0,302000	0,361000	1,000000	0,289000	1,681000	2,858000	0,000000	Train
128	0,000000	10,8840	2,753000	3,005000	2,469000	2,116000	-0,46800	0,000000	0,571000	-0,012000	0,012000	0,917000	0,238000	1,435000	2,406000	0,000000	Test
129	0,000000	3,0590	2,156000	3,916000	3,006000	0,901000	0,00000	0,000000	0,375000	-0,090000	0,121000	0,828000	0,189000	1,386000	1,593000	0,000000	Train
130	0,000000	2,5970	2,156000	3,916000	2,423000	1,484000	0,00000	0,000000	0,375000	-0,017000	0,053000	0,590000	0,413000	0,987000	1,521000	0,000000	Train
131	0,000000	3,2860	2,156000	3,916000	2,606000	1,301000	0,00000	0,000000	0,375000	-0,007000	0,178000	0,828000	0,001000	1,556000	1,653000	1,000000	Train
132	0,000000	4,7350	2,371000	3,971000	2,156000	1,844000	0,00000	0,000000	0,400000	-0,161000	0,180000	0,613000	0,551000	0,793000	1,556000	0,000000	Train
133	0,000000	6,1210	2,322000	5,184000	2,210000	1,960000	0,00000	0,000000	0,600000	-0,125000	0,213000	0,693000	0,192000	1,581000	2,600000	0,000000	Train
134	0,000000	4,3400	2,446000	2,636000	2,822000	1,348000	0,00000	0,000000	0,300000	-0,106000	0,122000	0,693000	0,368000	1,270000	1,392000	0,000000	Test

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
135	0,000000	4,5250	2,522000	1,958000	2,500000	1,500000	0,00000	0,080000	0,100000	-0,081000	0,129000	0,613000	0,094000	1,602000	1,227000	0,000000	Train
136	1,099000	0,9800	1,585000	0,000000	0,918000	2,252000	0,00000	0,000000	0,000000	0,280000	0,037000	0,000000	0,000000	1,703000	0,000000	0,000000	Train
137	1,386000	1,7420	2,000000	4,000000	0,918000	2,667000	0,00000	0,000000	0,000000	0,405000	0,140000	0,000000	0,000000	2,196000	0,691000	0,000000	Train
138	1,609000	2,7220	2,322000	3,125000	0,918000	2,989000	0,00000	0,000000	0,000000	0,412000	0,000000	0,655000	0,000000	2,639000	1,246000	0,000000	Train
139	1,792000	3,9200	2,585000	2,056000	0,918000	3,252000	0,00000	0,000000	0,000000	0,214000	0,020000	0,000000	0,000000	2,745000	1,737000	0,000000	Train
140	1,946000	5,3350	2,807000	1,814000	0,918000	3,474000	-0,80200	0,000000	0,000000	0,154000	0,276000	0,307000	0,202000	2,861000	2,157000	0,000000	Test
141	2,079000	6,9690	3,000000	1,488000	0,918000	3,667000	-0,41400	0,000000	0,000000	0,412000	0,530000	0,583000	0,424000	2,862000	2,243000	0,000000	Train
142	1,792000	5,6240	2,753000	2,022000	2,856000	1,536000	-1,00000	0,111000	0,143000	0,165000	0,158000	0,796000	0,015000	2,566000	2,173000	1,000000	Train
143	1,792000	7,2980	2,953000	1,782000	2,804000	1,781000	-0,66200	0,151000	0,125000	0,217000	0,280000	1,000000	0,239000	2,479000	2,059000	1,000000	Train
144	1,792000	9,1600	3,128000	1,506000	3,208000	1,547000	0,00000	0,151000	0,111000	0,375000	0,387000	1,000000	0,279000	2,468000	2,004000	1,000000	Train
145	1,792000	9,6380	3,086000	1,764000	3,087000	1,667000	0,00000	0,191000	0,167000	0,221000	0,331000	0,790000	0,616000	2,331000	2,309000	2,000000	Test
146	1,792000	11,2600	3,284000	1,278000	3,073000	1,834000	0,00000	0,151000	0,100000	0,411000	0,559000	0,940000	0,274000	2,472000	1,982000	1,000000	Train
147	1,792000	11,8480	3,246000	1,608000	3,323000	1,584000	0,00000	0,191000	0,150000	0,394000	0,514000	0,940000	0,817000	2,304000	2,190000	2,000000	Test
148	1,609000	4,1560	2,522000	2,803000	2,795000	1,375000	0,00000	0,000000	0,167000	0,450000	0,011000	0,742000	0,051000	2,503000	1,854000	1,000000	Train
149	1,609000	5,6120	2,753000	2,172000	2,856000	1,536000	-0,84600	0,125000	0,143000	0,459000	0,008000	1,000000	0,326000	2,444000	1,814000	1,000000	Train
150	1,609000	7,2600	2,953000	1,667000	3,157000	1,428000	-0,61800	0,125000	0,125000	0,575000	0,063000	1,000000	0,324000	2,437000	1,771000	1,000000	Train
151	1,609000	7,6700	2,906000	2,073000	3,022000	1,563000	-0,61800	0,250000	0,188000	0,468000	0,127000	0,917000	0,700000	2,300000	2,144000	2,000000	Train
152	1,609000	9,1410	3,128000	1,343000	3,106000	1,649000	-0,53000	0,125000	0,111000	0,600000	0,104000	1,000000	0,356000	2,449000	1,768000	1,000000	Train
153	1,609000	9,5400	3,086000	1,555000	3,292000	1,463000	-0,44600	0,195000	0,222000	0,610000	0,208000	1,000000	0,467000	2,255000	2,037000	2,000000	Test
154	1,609000	11,2430	3,284000	1,135000	3,048000	1,859000	-0,15200	0,125000	0,100000	0,629000	0,164000	1,000000	0,508000	2,474000	1,781000	1,000000	Train
155	1,609000	13,5650	3,425000	0,993000	3,075000	1,970000	0,30600	0,125000	0,091000	0,700000	0,266000	0,895000	0,588000	2,487000	1,801000	1,000000	Test
156	1,609000	16,1080	3,554000	0,892000	3,037000	2,133000	0,73800	0,125000	0,083000	0,762000	0,360000	0,954000	0,749000	2,516000	1,833000	1,000000	Train
157	1,609000	18,8630	3,671000	0,817000	2,982000	2,303000	1,09900	0,125000	0,077000	0,785000	0,463000	0,958000	0,869000	2,525000	1,858000	1,000000	Test
158	1,609000	21,8420	3,780000	0,759000	2,921000	2,471000	1,38800	0,125000	0,071000	0,816000	0,578000	0,962000	1,006000	2,549000	1,889000	1,000000	Train
159	1,609000	25,0360	3,882000	0,712000	2,860000	2,632000	1,61800	0,125000	0,067000	0,881000	0,692000	0,851000	1,101000	2,554000	1,910000	1,000000	Train
160	1,609000	32,0860	4,065000	0,643000	2,740000	2,932000	1,61800	0,125000	0,059000	0,983000	0,921000	0,903000	1,294000	2,577000	1,954000	1,000000	Train
161	1,609000	35,9320	4,149000	0,616000	2,684000	3,071000	1,69800	0,125000	0,056000	1,012000	1,027000	0,847000	1,382000	2,594000	1,976000	1,000000	Train
162	1,609000	39,9940	4,228000	0,593000	2,631000	3,202000	1,83700	0,125000	0,053000	1,063000	1,119000	0,867000	1,452000	2,595000	1,990000	1,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
163	1,609000	44,2890	4,303000	0,573000	2,580000	3,327000	1,95100	0,125000	0,050000	1,123000	1,202000	0,844000	1,529000	2,609000	2,009000	1,000000	Train
164	1,792000	12,2750	3,209000	1,597000	2,731000	2,176000	0,00000	0,382000	0,250000	0,219000	0,492000	0,736000	1,091000	2,295000	2,503000	3,000000	Train
165	1,609000	6,1260	2,664000	3,059000	2,606000	1,786000	-0,81700	0,000000	0,357000	0,316000	-0,197000	0,898000	0,161000	2,291000	2,739000	0,000000	Test
166	1,609000	5,9200	2,700000	2,393000	2,797000	1,596000	-0,77600	0,000000	0,286000	0,510000	-0,105000	0,796000	0,184000	2,395000	2,235000	2,000000	Test
167	1,609000	5,9200	2,700000	2,393000	2,797000	1,596000	-0,77600	0,000000	0,286000	0,473000	-0,009000	0,796000	0,078000	2,383000	2,214000	2,000000	Test
168	1,609000	5,9700	2,700000	2,704000	2,797000	1,596000	-0,80200	0,000000	0,214000	0,489000	0,047000	0,796000	0,139000	2,291000	2,222000	2,000000	Train
169	1,792000	7,9050	2,875000	2,215000	2,709000	1,876000	-0,50400	0,222000	0,313000	0,131000	0,327000	0,750000	0,341000	2,412000	2,868000	0,000000	Train
170	1,792000	7,7280	2,906000	2,091000	2,855000	1,730000	-0,41400	0,222000	0,188000	0,181000	0,266000	0,833000	0,505000	2,479000	2,431000	2,000000	Train
171	1,792000	7,7280	2,906000	2,091000	2,855000	1,730000	-0,41400	0,222000	0,188000	0,104000	0,252000	0,667000	0,227000	2,387000	2,437000	2,000000	Test
172	1,792000	7,6640	2,906000	1,945000	3,105000	1,480000	-0,53200	0,222000	0,250000	0,174000	0,293000	0,750000	0,018000	2,374000	2,401000	2,000000	Train
173	1,792000	7,6640	2,906000	1,945000	3,105000	1,480000	-0,53200	0,222000	0,250000	0,237000	0,335000	0,917000	0,326000	2,458000	2,429000	2,000000	Train
174	1,792000	7,6550	2,906000	1,848000	2,355000	2,230000	-0,56200	0,302000	0,250000	0,166000	0,218000	0,750000	0,317000	2,468000	2,470000	2,000000	Train
175	1,792000	7,6550	2,906000	1,848000	2,355000	2,230000	-0,56200	0,302000	0,250000	0,160000	0,317000	0,500000	0,170000	2,508000	2,545000	2,000000	Test
176	1,609000	2,3490	2,322000	3,125000	2,470000	1,231000	0,00000	0,000000	0,000000	0,381000	-0,011000	0,655000	0,000000	2,198000	0,870000	0,000000	Train
177	1,792000	3,4870	2,585000	2,056000	2,500000	1,500000	0,00000	0,000000	0,000000	0,174000	0,148000	0,355000	0,000000	2,396000	1,352000	0,000000	Train
178	1,792000	2,4140	2,585000	2,056000	1,000000	2,585000	0,00000	0,000000	0,000000	-0,052000	0,220000	0,000000	0,000000	1,111000	0,525000	0,000000	Train
179	1,792000	3,6800	2,753000	2,022000	2,873000	1,034000	-0,50000	0,111000	0,143000	-0,068000	0,302000	0,796000	0,000000	1,121000	1,178000	0,000000	Test
180	1,792000	5,1540	2,953000	1,782000	3,197000	0,973000	-0,50000	0,151000	0,125000	-0,098000	0,205000	0,750000	0,034000	1,642000	1,140000	0,000000	Train
181	1,792000	6,8620	3,128000	1,506000	3,463000	0,929000	0,28300	0,151000	0,111000	-0,095000	0,312000	0,860000	0,098000	1,862000	1,361000	0,000000	Train
182	1,792000	7,2930	3,086000	1,764000	3,046000	1,346000	0,00000	0,191000	0,167000	0,079000	0,331000	0,860000	0,231000	1,493000	1,525000	1,000000	Train
183	1,792000	8,7620	3,284000	1,278000	3,689000	0,896000	0,50000	0,151000	0,100000	-0,063000	0,464000	0,797000	0,173000	1,825000	1,417000	0,000000	Train
184	1,792000	5,2320	2,906000	2,091000	2,642000	1,528000	-0,17200	0,222000	0,188000	-0,070000	0,420000	0,667000	0,590000	0,986000	1,346000	0,000000	Train
185	1,792000	5,1670	2,906000	1,945000	2,864000	1,306000	-0,23200	0,222000	0,250000	-0,039000	0,334000	0,520000	0,000000	0,970000	1,282000	0,000000	Train
186	1,792000	5,1810	2,906000	1,848000	2,197000	1,973000	-0,24800	0,302000	0,250000	-0,082000	0,412000	0,500000	0,093000	1,121000	1,477000	0,000000	Train
187	1,792000	10,8720	3,425000	1,108000	3,884000	0,871000	0,50000	0,151000	0,091000	-0,043000	0,533000	0,822000	0,382000	1,934000	1,461000	0,000000	Train
188	1,792000	13,1960	3,554000	0,982000	3,923000	0,984000	0,71800	0,151000	0,083000	0,027000	0,630000	0,861000	0,386000	1,912000	1,477000	0,000000	Train
189	1,792000	15,7320	3,671000	0,888000	3,923000	1,122000	0,95600	0,151000	0,077000	0,093000	0,703000	0,860000	0,600000	2,008000	1,532000	0,000000	Train
190	1,792000	18,4870	3,780000	0,816000	3,850000	1,320000	1,23200	0,151000	0,071000	0,122000	0,850000	0,887000	0,676000	2,007000	1,567000	0,000000	Test

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
191	1,792000	21,4610	3,882000	0,759000	3,760000	1,525000	1,49200	0,151000	0,067000	0,155000	0,976000	0,829000	0,837000	2,077000	1,609000	0,000000	Train
192	1,792000	24,6450	3,976000	0,713000	3,665000	1,727000	1,71200	0,151000	0,063000	0,214000	1,133000	0,938000	0,893000	2,071000	1,631000	0,000000	Train
193	1,792000	28,0480	4,065000	0,675000	3,571000	1,920000	1,89200	0,151000	0,059000	0,271000	1,275000	0,885000	1,022000	2,132000	1,669000	0,000000	Train
194	1,792000	31,6760	4,149000	0,644000	3,480000	2,105000	2,04100	0,151000	0,056000	0,313000	1,428000	0,820000	1,086000	2,131000	1,692000	0,000000	Test
195	1,792000	35,5120	4,228000	0,617000	3,394000	2,279000	2,16200	0,151000	0,053000	0,353000	1,562000	0,842000	1,195000	2,182000	1,725000	0,000000	Train
196	1,792000	6,7040	3,086000	1,764000	3,326000	0,922000	0,40600	0,191000	0,167000	-0,072000	0,335000	0,860000	0,147000	1,145000	1,320000	0,000000	Train
197	1,792000	6,8660	3,086000	1,735000	3,749000	0,643000	-0,07100	0,342000	0,222000	-0,030000	0,251000	0,790000	0,029000	1,437000	1,239000	0,000000	Test
198	1,792000	6,9140	3,086000	1,894000	3,559000	0,834000	-0,10700	0,262000	0,167000	-0,147000	0,266000	1,000000	0,307000	1,436000	1,271000	0,000000	Train
199	1,792000	6,8880	3,086000	1,630000	3,273000	1,120000	-0,04400	0,342000	0,222000	-0,089000	0,309000	0,860000	0,211000	1,542000	1,416000	0,000000	Test
200	0,000000	15,2390	3,031000	2,452000	3,237000	1,621000	-0,61800	0,320000	0,313000	0,251000	1,036000	0,763000	0,547000	1,699000	1,838000	2,000000	Train
201	0,000000	19,2760	3,197000	2,287000	3,202000	1,798000	-0,40300	0,320000	0,278000	0,323000	0,917000	0,880000	1,085000	1,925000	1,951000	2,000000	Train
202	1,792000	9,9410	3,059000	2,048000	3,310000	1,445000	-0,29500	0,262000	0,278000	0,177000	0,363000	0,930000	0,864000	2,366000	2,670000	0,000000	Train
203	1,792000	9,6100	3,086000	1,735000	3,736000	1,019000	-0,34400	0,342000	0,222000	0,265000	0,459000	1,000000	0,282000	2,358000	2,280000	2,000000	Train
204	1,792000	9,6780	3,086000	1,894000	3,356000	1,399000	-0,41400	0,262000	0,167000	0,210000	0,470000	1,000000	0,423000	2,251000	2,224000	2,000000	Train
205	1,792000	12,4450	3,209000	1,895000	3,565000	1,342000	-0,32800	0,453000	0,200000	0,302000	0,508000	0,940000	1,065000	2,255000	2,412000	3,000000	Train
206	1,792000	7,2930	3,086000	1,764000	3,046000	1,346000	0,00000	0,191000	0,167000	0,079000	0,331000	0,860000	0,231000	1,493000	1,525000	1,000000	Train
207	1,792000	6,8660	3,086000	1,735000	3,749000	0,643000	-0,07100	0,342000	0,222000	-0,029000	0,249000	0,790000	0,029000	1,437000	1,239000	0,000000	Train
208	1,792000	6,8880	3,086000	1,630000	3,273000	1,120000	-0,04400	0,342000	0,222000	-0,089000	0,307000	0,860000	0,212000	1,542000	1,416000	0,000000	Train
209	1,792000	6,9140	3,086000	1,894000	3,559000	0,834000	-0,10700	0,262000	0,167000	-0,117000	0,348000	1,000000	0,330000	1,439000	1,280000	0,000000	Train
210	1,792000	7,0200	3,044000	2,090000	2,653000	1,739000	-0,15000	0,333000	0,222000	-0,033000	0,489000	0,790000	1,032000	0,892000	1,438000	0,000000	Train
211	1,792000	6,9560	3,044000	1,921000	2,843000	1,549000	-0,17100	0,413000	0,278000	-0,039000	0,481000	1,000000	0,485000	0,932000	1,441000	0,000000	Train
212	1,792000	6,8640	3,044000	1,903000	2,128000	2,264000	-0,23200	0,333000	0,333000	0,035000	0,314000	0,290000	0,000000	0,824000	1,243000	0,000000	Test
213	1,792000	9,3520	3,209000	1,869000	3,407000	1,178000	0,00000	0,302000	0,200000	0,007000	0,385000	0,797000	0,359000	1,236000	1,547000	1,000000	Test
214	1,792000	9,2960	3,209000	1,708000	3,574000	1,011000	0,00000	0,462000	0,250000	0,093000	0,357000	0,759000	0,205000	1,325000	1,563000	1,000000	Train
215	3,555000	7,4720	3,300000	1,477000	3,096000	1,364000	0,66900	0,284000	0,091000	0,253000	0,294000	0,699000	0,082000	2,054000	1,564000	0,000000	Test
216	1,792000	10,1690	3,184000	1,847000	2,750000	1,835000	0,00000	0,249000	0,300000	-0,030000	0,348000	0,819000	0,482000	1,472000	2,204000	0,000000	Train
217	1,792000	9,2440	3,246000	1,435000	3,324000	1,261000	0,32600	0,311000	0,200000	-0,092000	0,351000	1,000000	0,164000	1,882000	1,847000	1,000000	Train
218	3,584000	10,4140	3,682000	0,928000	2,823000	1,820000	1,29200	0,302000	0,107000	-0,291000	0,057000	0,584000	0,162000	1,752000	1,429000	0,000000	Train

ID	STN	TIE	IVDM	Yindex	IC3	CIC3	EEig07x	GGI4	JGI1	Mor10v	Mor16e	ISH	H5e	R3e	R4e	nCt	Partition
219	1,792000	9,3740	3,246000	1,608000	3,657000	0,928000	0,50000	0,191000	0,150000	0,102000	0,446000	0,940000	0,405000	1,609000	1,510000	1,000000	Train
220	1,792000	8,8290	3,246000	1,463000	2,689000	1,896000	0,50000	0,382000	0,200000	-0,096000	0,200000	0,413000	0,426000	1,684000	1,385000	0,000000	Train
221	3,584000	8,6380	3,565000	1,122000	2,482000	1,978000	0,90100	0,062000	0,077000	0,207000	0,411000	0,843000	0,128000	1,591000	1,078000	0,000000	Train
222	3,555000	6,2840	3,300000	1,477000	2,281000	1,889000	0,88200	0,284000	0,091000	0,000000	0,019000	0,458000	0,077000	1,089000	0,803000	0,000000	Train
223	0,000000	14,2910	2,664000	4,976000	1,140000	3,560000	-0,64600	0,000000	0,643000	-0,347000	0,306000	0,270000	1,215000	1,683000	3,592000	0,000000	Train

Tableau E.2. Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour d20 ainsi leurs ordres de partition dans les bases d'apprentissage et de test.

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
1	1,000000	1,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	3,000000	0,000000	0,000000	0,372000	0,016000	0,000000	0,282000	2,055000	Test
2	1,260000	1,633000	0,000000	0,000000	0,918000	0,662000	0,000000	1,371000	2,088000	0,000000	0,250000	0,530000	0,031000	0,000000	0,299000	1,735000	Train
3	1,414000	1,975000	0,693000	0,500000	1,000000	0,906000	3,132000	1,842000	1,965000	0,000000	0,167000	0,655000	0,051000	0,006000	0,309000	1,395000	Test
4	1,316000	2,324000	0,000000	0,000000	0,811000	0,929000	0,000000	1,140000	2,667000	0,000000	0,500000	0,550000	0,043000	0,000000	0,363000	1,252000	Train
5	1,516000	2,191000	1,099000	0,354000	0,971000	1,049000	2,340000	2,352000	1,736000	0,000000	0,125000	0,736000	0,070000	0,000000	0,315000	1,136000	Train
6	1,431000	2,540000	1,099000	0,408000	1,371000	1,067000	3,916000	2,660000	1,427000	0,000000	0,375000	0,625000	1,044000	0,006000	0,377000	1,018000	Train
7	1,587000	2,339000	1,386000	0,319000	0,918000	1,148000	1,958000	2,446000	1,875000	0,000000	0,100000	0,794000	0,090000	0,001000	0,319000	1,013000	Test
8	1,641000	2,447000	1,609000	0,302000	0,863000	1,222000	1,730000	2,566000	1,957000	0,056000	0,083000	0,836000	0,110000	0,001000	0,322000	0,890000	Train
9	1,682000	2,530000	1,792000	0,291000	0,811000	1,281000	1,574000	2,565000	2,135000	0,056000	0,071000	0,866000	0,130000	0,001000	0,325000	0,821000	Train
10	1,714000	2,595000	1,946000	0,285000	0,764000	1,330000	1,460000	2,531000	2,327000	0,056000	0,063000	0,889000	0,150000	0,002000	0,327000	0,745000	Train
11	1,741000	2,648000	2,079000	0,280000	0,722000	1,371000	1,372000	2,483000	2,517000	0,056000	0,056000	0,907000	0,170000	0,002000	0,328000	0,697000	Train
12	1,763000	2,691000	2,197000	0,276000	0,684000	1,407000	1,301000	2,432000	2,698000	0,056000	0,050000	0,921000	0,190000	0,002000	0,329000	0,644000	Train
13	1,782000	2,727000	2,303000	0,273000	0,650000	1,438000	1,242000	2,379000	2,869000	0,056000	0,045000	0,932000	0,210000	0,002000	0,330000	0,608000	Train
14	1,798000	2,758000	2,398000	0,271000	0,619000	1,466000	1,193000	2,328000	3,030000	0,056000	0,042000	0,941000	0,230000	0,003000	0,331000	0,568000	Train
15	1,811000	2,785000	2,485000	0,269000	0,592000	1,491000	1,151000	2,279000	3,181000	0,056000	0,038000	0,948000	0,249000	0,003000	0,332000	0,540000	Train
16	1,823000	2,808000	2,565000	0,267000	0,567000	1,514000	1,114000	2,232000	3,322000	0,056000	0,036000	0,954000	0,269000	0,003000	0,333000	0,509000	Train
17	1,834000	2,828000	2,639000	0,266000	0,544000	1,534000	1,082000	2,188000	3,455000	0,056000	0,033000	0,959000	0,289000	0,004000	0,333000	0,486000	Test
18	1,843000	2,846000	2,708000	0,265000	0,523000	1,553000	1,053000	2,147000	3,581000	0,056000	0,031000	0,964000	0,309000	0,004000	0,334000	0,461000	Train
19	1,852000	2,862000	2,773000	0,264000	0,503000	1,571000	1,027000	2,109000	3,699000	0,056000	0,029000	0,967000	0,329000	0,004000	0,334000	0,442000	Train
20	1,859000	2,877000	2,833000	0,263000	0,485000	1,587000	1,004000	2,072000	3,811000	0,056000	0,028000	0,970000	0,349000	0,005000	0,335000	0,421000	Train
21	1,866000	2,890000	2,890000	0,262000	0,469000	1,602000	0,983000	2,038000	3,916000	0,056000	0,026000	0,973000	0,369000	0,005000	0,335000	0,405000	Train
22	1,872000	2,901000	2,944000	0,262000	0,454000	1,617000	0,964000	2,006000	4,017000	0,056000	0,025000	0,975000	0,389000	0,005000	0,335000	0,388000	Train
23	1,878000	2,912000	2,996000	0,261000	0,439000	1,630000	0,946000	1,975000	4,112000	0,056000	0,024000	0,977000	0,409000	0,006000	0,336000	0,374000	Test
24	1,883000	2,922000	3,045000	0,260000	0,426000	1,643000	0,929000	1,947000	4,203000	0,056000	0,023000	0,979000	0,428000	0,006000	0,336000	0,359000	Train
25	1,888000	2,931000	3,091000	0,260000	0,414000	1,654000	0,914000	1,920000	4,290000	0,056000	0,022000	0,981000	0,448000	0,006000	0,336000	0,347000	Test
26	1,892000	2,939000	3,135000	0,259000	0,402000	1,666000	0,900000	1,894000	4,373000	0,056000	0,021000	0,982000	0,468000	0,007000	0,337000	0,335000	Train

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
27	1,896000	2,947000	3,178000	0,259000	0,391000	1,676000	0,887000	1,870000	4,452000	0,056000	0,020000	0,984000	0,488000	0,007000	0,337000	0,324000	Train
28	1,903000	2,961000	3,258000	0,258000	0,371000	1,696000	0,863000	1,825000	4,601000	0,056000	0,019000	0,986000	0,528000	0,007000	0,337000	0,304000	Train
29	1,915000	2,983000	3,401000	0,257000	0,337000	1,731000	0,823000	1,748000	4,867000	0,056000	0,016000	0,989000	0,607000	0,009000	0,338000	0,271000	Test
30	1,924000	3,000000	3,526000	0,256000	0,310000	1,761000	0,790000	1,684000	5,097000	0,056000	0,014000	0,991000	0,687000	0,010000	0,338000	0,244000	Train
31	1,320000	3,024000	0,000000	0,000000	0,722000	1,084000	0,000000	0,811000	3,276000	0,000000	0,750000	0,000000	0,055000	0,000000	1,065000	0,832000	Test
32	1,513000	2,627000	1,386000	0,289000	1,459000	1,159000	2,636000	3,009000	1,313000	0,000000	0,300000	0,702000	1,064000	0,035000	0,393000	0,849000	Train
33	1,513000	2,754000	1,609000	0,348000	1,459000	1,166000	3,135000	2,709000	1,613000	0,000000	0,300000	0,663000	2,046000	0,000000	0,427000	0,846000	Train
34	1,414000	3,168000	1,386000	0,354000	1,252000	1,177000	5,184000	2,320000	2,002000	0,000000	0,600000	0,239000	1,283000	0,006000	0,935000	0,734000	Train
35	1,442000	2,993000	1,609000	0,333000	0,918000	1,173000	3,971000	1,571000	2,751000	0,000000	0,400000	0,668000	1,290000	0,012000	0,368000	0,828000	Train
36	1,575000	2,678000	1,609000	0,284000	1,449000	1,230000	2,107000	3,295000	1,229000	0,000000	0,250000	0,766000	1,084000	0,199000	0,395000	0,784000	Train
37	1,575000	2,832000	1,792000	0,296000	1,449000	1,236000	2,460000	3,382000	1,142000	0,000000	0,250000	0,742000	2,066000	0,035000	0,417000	0,749000	Train
38	1,486000	3,154000	1,609000	0,250000	1,379000	1,243000	3,210000	2,696000	1,828000	0,000000	0,500000	0,436000	1,303000	0,040000	0,870000	0,611000	Train
39	1,511000	3,144000	1,946000	0,297000	1,379000	1,246000	3,253000	2,903000	1,621000	0,000000	0,333000	0,699000	2,296000	0,018000	0,406000	0,682000	Train
40	1,511000	2,953000	1,609000	0,236000	1,379000	1,239000	2,672000	2,044000	2,479000	0,000000	0,417000	0,465000	2,051000	0,040000	0,844000	0,586000	Train
41	1,486000	3,360000	1,946000	0,319000	1,379000	1,250000	4,096000	2,566000	1,957000	0,000000	0,500000	0,413000	2,511000	0,000000	0,842000	0,688000	Train
42	1,575000	2,992000	1,946000	0,289000	1,449000	1,244000	2,933000	2,195000	2,328000	0,000000	0,250000	0,691000	3,057000	0,085000	0,432000	0,656000	Train
43	1,426000	3,541000	1,946000	0,289000	1,149000	1,254000	4,656000	2,315000	2,208000	0,000000	0,583000	0,349000	2,509000	0,012000	0,860000	0,598000	Train
44	1,622000	2,716000	1,792000	0,277000	1,406000	1,286000	1,812000	3,383000	1,318000	0,111000	0,214000	0,812000	1,104000	0,198000	0,403000	0,714000	Train
45	1,622000	2,862000	1,946000	0,291000	1,406000	1,292000	2,049000	3,459000	1,241000	0,056000	0,214000	0,787000	2,086000	0,199000	0,434000	0,684000	Test
46	1,622000	2,920000	1,946000	0,260000	1,406000	1,294000	2,165000	3,075000	1,626000	0,056000	0,214000	0,794000	2,086000	0,070000	0,411000	0,647000	Train
47	1,661000	3,092000	2,197000	0,265000	1,352000	1,345000	2,115000	2,969000	1,889000	0,111000	0,188000	0,790000	3,103000	0,336000	0,439000	0,552000	Train
48	1,661000	3,175000	2,197000	0,246000	1,352000	1,349000	2,272000	3,047000	1,811000	0,056000	0,188000	0,790000	3,104000	0,319000	0,417000	0,551000	Train
49	1,542000	3,112000	1,792000	0,256000	1,406000	1,295000	2,436000	3,007000	1,694000	0,000000	0,429000	0,580000	1,323000	0,289000	0,808000	0,593000	Train
50	1,565000	3,171000	2,079000	0,269000	1,500000	1,300000	2,610000	3,190000	1,511000	0,000000	0,286000	0,760000	2,316000	0,151000	0,402000	0,624000	Train
51	1,565000	3,099000	1,946000	0,262000	1,500000	1,298000	2,451000	3,190000	1,511000	0,000000	0,357000	0,566000	3,052000	0,185000	0,825000	0,556000	Train
52	1,565000	2,928000	1,792000	0,264000	1,500000	1,292000	2,125000	2,200000	2,501000	0,000000	0,357000	0,781000	2,078000	0,583000	0,390000	0,681000	Train
53	1,542000	3,373000	2,079000	0,269000	1,406000	1,304000	3,071000	3,200000	1,501000	0,000000	0,429000	0,545000	2,531000	0,040000	0,808000	0,585000	Test
54	1,565000	3,292000	2,197000	0,282000	1,500000	1,304000	2,887000	2,815000	1,885000	0,000000	0,286000	0,746000	3,302000	0,031000	0,411000	0,612000	Train

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
55	1,587000	3,073000	1,946000	0,255000	1,392000	1,339000	2,021000	3,132000	1,725000	0,167000	0,375000	0,686000	1,342000	0,281000	0,758000	0,562000	Train
56	1,608000	3,155000	2,197000	0,269000	1,530000	1,344000	2,175000	3,434000	1,424000	0,111000	0,250000	0,800000	2,336000	0,309000	0,421000	0,573000	Train
57	1,608000	3,151000	2,079000	0,236000	1,530000	1,345000	2,179000	3,434000	1,424000	0,111000	0,313000	0,654000	3,107000	0,217000	0,806000	0,485000	Train
58	1,608000	2,915000	1,946000	0,260000	1,530000	1,337000	1,821000	2,547000	2,311000	0,222000	0,313000	0,712000	2,098000	0,236000	0,718000	0,558000	Train
59	1,587000	3,330000	2,197000	0,271000	1,392000	1,347000	2,449000	3,444000	1,414000	0,056000	0,375000	0,654000	2,551000	0,290000	0,758000	0,566000	Train
60	1,608000	3,325000	2,303000	0,262000	1,530000	1,349000	2,471000	3,444000	1,414000	0,056000	0,250000	0,789000	3,322000	0,163000	0,418000	0,549000	Train
61	1,608000	3,223000	2,197000	0,275000	1,530000	1,347000	2,282000	3,099000	1,759000	0,056000	0,313000	0,636000	4,084000	0,359000	0,778000	0,491000	Train
62	1,625000	3,044000	2,079000	0,254000	1,361000	1,377000	1,764000	3,202000	1,798000	0,167000	0,333000	0,758000	1,362000	0,281000	0,719000	0,538000	Test
63	1,644000	3,130000	2,303000	0,267000	1,522000	1,382000	1,884000	3,522000	1,478000	0,167000	0,222000	0,839000	2,356000	0,308000	0,418000	0,545000	Test
64	1,644000	3,160000	2,197000	0,241000	1,522000	1,383000	1,937000	3,647000	1,353000	0,111000	0,278000	0,722000	3,092000	0,378000	0,792000	0,475000	Train
65	1,644000	3,124000	2,197000	0,252000	1,522000	1,382000	1,887000	3,350000	1,650000	0,111000	0,278000	0,839000	3,099000	0,597000	0,414000	0,548000	Test
66	1,625000	3,277000	2,303000	0,268000	1,361000	1,384000	2,063000	3,531000	1,469000	0,167000	0,333000	0,736000	2,571000	0,282000	0,719000	0,537000	Train
67	1,644000	3,168000	2,303000	0,283000	1,522000	1,383000	1,932000	3,156000	1,844000	0,167000	0,278000	0,829000	4,081000	0,541000	0,433000	0,551000	Train
68	1,622000	3,074000	2,079000	0,265000	1,406000	1,300000	2,481000	2,988000	1,713000	0,000000	0,214000	0,750000	3,082000	0,182000	0,424000	0,601000	Train
69	1,488000	3,623000	2,197000	0,261000	1,549000	1,310000	3,670000	2,978000	1,723000	0,000000	0,500000	0,472000	3,513000	0,046000	0,824000	0,509000	Train
70	1,488000	3,389000	1,792000	0,204000	1,549000	1,303000	3,005000	2,670000	2,031000	0,000000	0,571000	0,491000	1,605000	0,183000	0,813000	0,544000	Train
71	1,488000	3,708000	2,303000	0,275000	1,549000	1,312000	4,013000	3,094000	1,607000	0,000000	0,500000	0,475000	3,738000	0,026000	0,794000	0,542000	Train
72	1,510000	3,464000	2,197000	0,263000	0,954000	1,307000	3,265000	1,985000	2,716000	0,000000	0,357000	0,496000	3,488000	0,058000	0,796000	0,497000	Train
73	1,565000	3,355000	2,197000	0,249000	1,500000	1,307000	3,082000	2,969000	1,732000	0,000000	0,286000	0,447000	4,046000	0,093000	0,861000	0,553000	Train
74	1,542000	3,583000	2,303000	0,285000	1,406000	1,311000	3,750000	2,458000	2,242000	0,000000	0,429000	0,493000	4,040000	0,006000	0,769000	0,606000	Train
75	1,769000	2,542000	2,639000	0,215000	1,811000	1,314000	2,914000	3,407000	1,178000	0,000000	0,375000	0,381000	0,493000	0,005000	0,842000	0,732000	Test
76	1,769000	2,435000	2,485000	0,201000	1,811000	1,308000	2,469000	3,407000	1,178000	0,000000	0,438000	0,404000	0,577000	0,042000	0,896000	0,748000	Train
77	1,795000	2,388000	2,565000	0,210000	1,561000	1,308000	2,341000	2,730000	1,855000	0,000000	0,313000	0,595000	0,569000	0,125000	0,578000	0,874000	Train
78	1,795000	2,436000	2,639000	0,221000	1,561000	1,311000	2,511000	2,730000	1,855000	0,000000	0,250000	0,573000	0,582000	0,016000	0,599000	0,785000	Train
79	1,537000	3,589000	2,303000	0,244000	1,658000	1,353000	2,912000	3,244000	1,614000	0,000000	0,438000	0,589000	3,533000	0,315000	0,784000	0,477000	Train
80	1,537000	3,467000	2,079000	0,237000	1,658000	1,350000	2,658000	3,244000	1,614000	0,000000	0,500000	0,575000	2,518000	0,255000	0,808000	0,520000	Test
81	1,537000	3,281000	1,946000	0,245000	1,658000	1,345000	2,324000	2,969000	1,889000	0,000000	0,500000	0,617000	2,316000	0,643000	0,781000	0,519000	Train
82	1,558000	3,377000	2,197000	0,248000	1,352000	1,349000	2,527000	2,630000	2,228000	0,000000	0,375000	0,595000	3,334000	0,450000	0,795000	0,464000	Train

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
83	1,793000	2,411000	2,565000	0,204000	1,753000	1,355000	2,060000	3,634000	1,121000	0,000000	0,389000	0,482000	3,219000	0,072000	0,813000	0,658000	Train
84	1,587000	3,825000	2,565000	0,250000	1,392000	1,362000	3,634000	1,995000	2,863000	0,000000	0,375000	0,560000	5,570000	0,001000	0,716000	0,553000	Train
85	1,470000	4,145000	2,565000	0,243000	1,224000	1,364000	4,278000	2,371000	2,487000	0,000000	0,563000	0,274000	4,962000	0,053000	1,190000	0,436000	Train
86	1,489000	3,878000	2,398000	0,237000	1,224000	1,359000	3,535000	2,658000	2,200000	0,000000	0,500000	0,521000	2,085000	0,185000	0,769000	0,464000	Train
87	1,470000	3,746000	1,946000	0,177000	1,224000	1,355000	3,164000	1,597000	3,261000	0,000000	0,688000	0,358000	3,279000	0,160000	1,054000	0,509000	Train
88	1,489000	4,014000	2,565000	0,248000	1,224000	1,362000	3,955000	2,357000	2,501000	0,000000	0,500000	0,339000	5,115000	0,052000	1,087000	0,426000	Train
89	1,661000	2,747000	1,946000	0,273000	1,352000	1,333000	1,622000	3,417000	1,440000	0,111000	0,188000	0,849000	1,124000	0,198000	0,401000	0,670000	Train
90	1,661000	2,877000	2,079000	0,285000	1,352000	1,337000	1,787000	3,417000	1,440000	0,111000	0,188000	0,832000	2,106000	0,199000	0,428000	0,645000	Train
91	1,661000	2,955000	2,079000	0,262000	1,352000	1,340000	1,907000	3,470000	1,388000	0,056000	0,188000	0,829000	2,106000	0,234000	0,422000	0,613000	Train
92	1,692000	2,773000	2,079000	0,269000	1,295000	1,373000	1,489000	3,367000	1,633000	0,111000	0,167000	0,874000	1,144000	0,199000	0,406000	0,625000	Train
93	1,692000	2,886000	2,197000	0,281000	1,295000	1,376000	1,608000	3,429000	1,571000	0,111000	0,167000	0,860000	2,126000	0,199000	0,433000	0,599000	Train
94	1,692000	2,998000	2,197000	0,263000	1,295000	1,381000	1,751000	3,367000	1,633000	0,056000	0,167000	0,852000	2,126000	0,399000	0,444000	0,568000	Train
95	1,718000	2,973000	2,303000	0,259000	1,241000	1,413000	1,560000	3,593000	1,536000	0,111000	0,150000	0,883000	2,146000	0,234000	0,423000	0,542000	Train
96	1,692000	3,087000	2,303000	0,263000	1,295000	1,383000	1,853000	3,108000	1,892000	0,111000	0,167000	0,832000	3,122000	0,291000	0,433000	0,532000	Test
97	1,578000	3,642000	2,398000	0,256000	1,685000	1,393000	2,600000	3,632000	1,368000	0,056000	0,444000	0,655000	3,739000	1,198000	0,754000	0,502000	Train
98	1,516000	4,102000	2,639000	0,223000	1,371000	1,400000	3,456000	2,684000	2,316000	0,000000	0,500000	0,435000	4,982000	0,341000	1,109000	0,392000	Train
99	1,260000	1,633000	0,000000	0,000000	0,918000	0,662000	0,000000	2,156000	1,014000	0,000000	0,250000	0,804000	0,018000	0,000000	0,114000	1,237000	Test
100	1,414000	1,975000	1,099000	0,289000	1,000000	0,906000	3,132000	2,855000	0,730000	0,000000	0,167000	0,648000	0,048000	0,001000	0,315000	1,394000	Train
101	1,516000	2,191000	1,386000	0,246000	0,971000	1,049000	2,340000	3,190000	0,717000	0,000000	0,125000	0,728000	0,068000	0,026000	0,320000	1,213000	Train
102	1,587000	2,339000	1,609000	0,254000	0,918000	1,148000	1,958000	3,461000	0,709000	0,000000	0,100000	0,751000	0,088000	0,002000	0,375000	0,991000	Train
103	1,641000	2,447000	1,792000	0,253000	0,863000	1,222000	1,730000	3,690000	0,703000	0,056000	0,083000	0,805000	0,108000	0,002000	0,371000	0,899000	Train
104	1,682000	2,530000	1,946000	0,252000	0,811000	1,281000	1,574000	3,720000	0,865000	0,056000	0,071000	0,835000	0,127000	0,002000	0,385000	0,796000	Train
105	1,714000	2,595000	2,079000	0,252000	0,764000	1,330000	1,460000	3,708000	1,047000	0,056000	0,063000	0,866000	0,147000	0,002000	0,381000	0,738000	Train
106	1,741000	2,648000	2,197000	0,252000	0,722000	1,371000	1,372000	3,615000	1,292000	0,056000	0,056000	0,887000	0,167000	0,003000	0,384000	0,676000	Train
107	1,763000	2,691000	2,303000	0,251000	0,684000	1,407000	1,301000	3,507000	1,537000	0,056000	0,050000	0,905000	0,187000	0,003000	0,381000	0,634000	Train
108	1,782000	2,727000	2,398000	0,251000	0,650000	1,438000	1,242000	3,398000	1,772000	0,056000	0,045000	0,918000	0,207000	0,003000	0,382000	0,590000	Train
109	1,798000	2,758000	2,485000	0,251000	0,619000	1,466000	1,193000	3,294000	1,991000	0,056000	0,042000	0,930000	0,227000	0,004000	0,379000	0,559000	Test
110	1,811000	2,785000	2,565000	0,251000	0,592000	1,491000	1,151000	3,196000	2,197000	0,056000	0,038000	0,939000	0,247000	0,004000	0,378000	0,526000	Test

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
111	1,823000	2,808000	2,639000	0,251000	0,567000	1,514000	1,114000	3,104000	2,388000	0,056000	0,036000	0,946000	0,267000	0,004000	0,377000	0,500000	Train
112	1,834000	2,828000	2,708000	0,251000	0,544000	1,534000	1,082000	3,018000	2,567000	0,056000	0,033000	0,952000	0,287000	0,005000	0,376000	0,474000	Train
113	1,843000	2,846000	2,773000	0,251000	0,523000	1,553000	1,053000	2,939000	2,733000	0,056000	0,031000	0,958000	0,306000	0,005000	0,374000	0,453000	Train
114	1,852000	2,862000	2,833000	0,251000	0,503000	1,571000	1,027000	2,866000	2,889000	0,056000	0,029000	0,962000	0,326000	0,005000	0,373000	0,432000	Train
115	1,859000	2,877000	2,890000	0,251000	0,485000	1,587000	1,004000	2,797000	3,035000	0,056000	0,028000	0,966000	0,346000	0,006000	0,372000	0,415000	Train
116	1,866000	2,890000	2,944000	0,251000	0,469000	1,602000	0,983000	2,734000	3,173000	0,056000	0,026000	0,969000	0,366000	0,006000	0,371000	0,397000	Train
117	1,414000	1,975000	1,609000	0,167000	1,000000	0,906000	3,132000	1,922000	1,400000	0,000000	0,167000	1,000000	0,012000	0,006000	0,000000	0,518000	Train
118	1,414000	1,975000	1,099000	0,333000	1,000000	0,906000	3,132000	1,792000	1,792000	0,000000	0,167000	0,768000	0,064000	0,000000	0,171000	0,965000	Train
119	1,414000	1,975000	1,099000	0,333000	1,000000	0,906000	3,132000	1,792000	1,792000	0,000000	0,167000	0,810000	0,029000	0,004000	0,171000	0,835000	Train
120	1,316000	2,324000	0,000000	0,000000	0,811000	0,929000	0,000000	1,686000	1,899000	0,000000	0,500000	0,773000	0,213000	0,000000	0,171000	0,592000	Test
121	1,431000	2,540000	1,946000	0,174000	1,371000	1,067000	3,916000	3,027000	0,673000	0,000000	0,375000	0,913000	0,450000	0,006000	0,075000	0,499000	Test
122	1,442000	2,993000	1,946000	0,246000	0,918000	1,173000	3,971000	2,822000	1,348000	0,000000	0,400000	0,281000	0,998000	0,001000	0,907000	0,592000	Train
123	1,442000	2,993000	2,197000	0,250000	0,918000	1,173000	3,971000	1,224000	2,946000	0,000000	0,400000	0,786000	0,311000	0,004000	0,233000	0,290000	Train
124	1,587000	2,339000	1,792000	0,230000	0,918000	1,148000	1,958000	3,419000	0,751000	0,000000	0,100000	0,755000	0,109000	0,032000	0,334000	0,941000	Test
125	1,587000	2,339000	1,792000	0,230000	0,918000	1,148000	1,958000	3,419000	0,751000	0,000000	0,100000	0,794000	0,079000	0,082000	0,340000	0,949000	Test
126	1,513000	2,627000	1,609000	0,213000	1,459000	1,159000	2,636000	2,975000	1,195000	0,000000	0,300000	0,618000	1,066000	0,017000	0,450000	0,887000	Train
127	1,488000	3,389000	1,946000	0,166000	1,549000	1,303000	3,005000	2,833000	1,752000	0,000000	0,571000	0,494000	2,251000	0,123000	0,773000	0,560000	Train
128	1,488000	3,389000	2,398000	0,144000	1,549000	1,303000	3,005000	2,469000	2,116000	0,000000	0,571000	0,520000	1,637000	0,153000	0,766000	0,466000	Test
129	1,431000	2,540000	1,386000	0,302000	1,371000	1,067000	3,916000	3,006000	0,901000	0,000000	0,375000	0,611000	0,204000	0,000000	0,410000	0,831000	Train
130	1,431000	2,540000	1,609000	0,289000	1,371000	1,067000	3,916000	2,423000	1,484000	0,000000	0,375000	0,747000	0,613000	0,002000	0,253000	0,533000	Train
131	1,431000	2,540000	1,609000	0,236000	1,371000	1,067000	3,916000	2,606000	1,301000	0,000000	0,375000	0,554000	0,050000	0,001000	0,432000	0,914000	Train
132	1,442000	2,993000	2,303000	0,182000	0,918000	1,173000	3,971000	2,156000	1,844000	0,000000	0,400000	0,880000	0,826000	0,013000	0,123000	0,324000	Train
133	1,414000	3,168000	1,946000	0,204000	1,252000	1,177000	5,184000	2,210000	1,960000	0,000000	0,600000	0,242000	0,055000	0,001000	0,907000	0,703000	Train
134	1,513000	2,627000	1,946000	0,204000	1,459000	1,159000	2,636000	2,822000	1,348000	0,000000	0,300000	0,740000	0,089000	0,070000	0,351000	0,616000	Test
135	1,587000	2,339000	1,792000	0,192000	0,918000	1,148000	1,958000	2,500000	1,500000	0,000000	0,100000	0,710000	0,085000	0,856000	0,395000	0,981000	Train
136	2,000000	2,250000	0,000000	0,000000	0,000000	0,693000	0,000000	0,918000	2,252000	0,000000	0,000000	0,246000	0,007000	0,000000	0,368000	2,416000	Train
137	2,000000	2,000000	1,609000	0,250000	0,000000	0,916000	4,000000	0,918000	2,667000	0,000000	0,000000	0,363000	0,280000	0,000000	0,396000	1,983000	Train
138	2,000000	2,083000	1,792000	0,250000	0,000000	1,071000	3,125000	0,918000	2,989000	0,000000	0,000000	0,507000	0,401000	0,000000	0,386000	1,769000	Train

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
139	2,000000	2,000000	1,946000	0,250000	0,000000	1,162000	2,056000	0,918000	3,252000	0,000000	0,000000	0,534000	1,871000	0,000000	0,430000	1,484000	Train
140	2,000000	2,042000	2,079000	0,250000	0,000000	1,241000	1,814000	0,918000	3,474000	0,000000	0,000000	0,535000	2,306000	0,000000	0,499000	1,286000	Test
141	2,000000	2,000000	2,197000	0,250000	0,000000	1,296000	1,488000	0,918000	3,667000	0,000000	0,000000	0,481000	2,117000	0,008000	0,634000	1,124000	Train
142	1,919000	2,123000	2,197000	0,237000	1,149000	1,241000	2,022000	2,856000	1,536000	0,000000	0,143000	0,631000	1,894000	0,032000	0,420000	1,141000	Train
143	1,929000	2,125000	2,398000	0,230000	1,061000	1,299000	1,782000	2,804000	1,781000	0,063000	0,125000	0,687000	2,895000	0,222000	0,435000	0,964000	Train
144	1,937000	2,078000	2,485000	0,220000	0,986000	1,344000	1,506000	3,208000	1,547000	0,090000	0,111000	0,757000	2,915000	0,188000	0,430000	0,862000	Train
145	1,876000	2,228000	2,565000	0,216000	1,436000	1,350000	1,764000	3,087000	1,667000	0,125000	0,167000	0,722000	3,141000	0,301000	0,424000	0,800000	Test
146	1,943000	2,017000	2,565000	0,224000	0,922000	1,382000	1,278000	3,073000	1,834000	0,090000	0,100000	0,803000	2,935000	0,353000	0,434000	0,782000	Train
147	1,888000	2,240000	2,708000	0,221000	1,371000	1,391000	1,608000	3,323000	1,584000	0,153000	0,150000	0,760000	4,147000	0,307000	0,435000	0,702000	Test
148	1,906000	2,184000	2,079000	0,235000	1,252000	1,171000	2,803000	2,795000	1,375000	0,000000	0,167000	0,598000	0,541000	0,003000	0,402000	1,342000	Train
149	1,919000	2,140000	2,303000	0,228000	1,149000	1,240000	2,172000	2,856000	1,536000	0,000000	0,143000	0,677000	1,230000	0,158000	0,402000	1,111000	Train
150	1,929000	2,058000	2,398000	0,217000	1,061000	1,294000	1,667000	3,157000	1,428000	0,080000	0,125000	0,742000	1,250000	0,160000	0,418000	0,975000	Train
151	1,861000	2,242000	2,485000	0,213000	1,500000	1,302000	2,073000	3,022000	1,563000	0,000000	0,188000	0,712000	2,496000	0,177000	0,394000	0,886000	Train
152	1,937000	1,979000	2,485000	0,222000	0,986000	1,338000	1,343000	3,106000	1,649000	0,080000	0,111000	0,796000	1,269000	0,426000	0,413000	0,878000	Train
153	1,876000	2,113000	2,485000	0,207000	1,436000	1,344000	1,555000	3,292000	1,463000	0,160000	0,222000	0,586000	2,229000	0,792000	0,799000	0,734000	Test
154	1,943000	1,911000	2,565000	0,224000	0,922000	1,377000	1,135000	3,048000	1,859000	0,080000	0,100000	0,834000	1,289000	0,426000	0,425000	0,812000	Train
155	1,948000	1,857000	2,639000	0,226000	0,866000	1,410000	0,993000	3,075000	1,970000	0,080000	0,091000	0,865000	1,309000	0,426000	0,419000	0,750000	Test
156	1,953000	1,813000	2,708000	0,228000	0,817000	1,440000	0,892000	3,037000	2,133000	0,080000	0,083000	0,886000	1,329000	0,427000	0,428000	0,705000	Train
157	1,956000	1,778000	2,773000	0,229000	0,773000	1,467000	0,817000	2,982000	2,303000	0,080000	0,077000	0,905000	1,349000	0,427000	0,422000	0,658000	Test
158	1,959000	1,750000	2,833000	0,231000	0,735000	1,491000	0,759000	2,921000	2,471000	0,080000	0,071000	0,918000	1,369000	0,427000	0,429000	0,625000	Train
159	1,962000	1,727000	2,890000	0,232000	0,700000	1,513000	0,712000	2,860000	2,632000	0,080000	0,067000	0,930000	1,389000	0,428000	0,424000	0,588000	Train
160	1,966000	1,692000	2,996000	0,234000	0,640000	1,553000	0,643000	2,740000	2,932000	0,080000	0,059000	0,946000	1,429000	0,429000	0,424000	0,531000	Train
161	1,968000	1,679000	3,045000	0,235000	0,614000	1,570000	0,616000	2,684000	3,071000	0,080000	0,056000	0,952000	1,448000	0,429000	0,428000	0,508000	Train
162	1,970000	1,668000	3,091000	0,235000	0,591000	1,586000	0,593000	2,631000	3,202000	0,080000	0,053000	0,958000	1,468000	0,429000	0,424000	0,484000	Train
163	1,971000	1,658000	3,135000	0,236000	0,569000	1,601000	0,573000	2,580000	3,327000	0,080000	0,050000	0,962000	1,488000	0,429000	0,427000	0,465000	Train
164	1,835000	2,260000	2,708000	0,215000	1,571000	1,389000	1,597000	2,731000	2,176000	0,236000	0,250000	0,626000	4,444000	1,149000	0,686000	0,664000	Train
165	1,811000	2,400000	2,303000	0,217000	1,379000	1,251000	3,059000	2,606000	1,786000	0,000000	0,357000	0,287000	0,441000	0,002000	0,904000	0,946000	Test
166	1,842000	2,257000	2,303000	0,217000	1,557000	1,246000	2,393000	2,797000	1,596000	0,000000	0,286000	0,652000	0,589000	0,014000	0,435000	1,060000	Test

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
167	1,842000	2,257000	2,303000	0,217000	1,557000	1,246000	2,393000	2,797000	1,596000	0,000000	0,286000	0,614000	0,631000	0,017000	0,493000	1,040000	Test
168	1,842000	2,325000	2,398000	0,229000	1,557000	1,250000	2,704000	2,797000	1,596000	0,000000	0,214000	0,642000	0,707000	0,004000	0,415000	0,982000	Train
169	1,834000	2,328000	2,398000	0,221000	1,299000	1,305000	2,215000	2,709000	1,876000	0,000000	0,313000	0,353000	3,196000	0,032000	0,870000	0,867000	Train
170	1,861000	2,279000	2,485000	0,231000	1,500000	1,305000	2,091000	2,855000	1,730000	0,000000	0,188000	0,474000	3,801000	0,032000	0,698000	0,911000	Train
171	1,861000	2,279000	2,485000	0,231000	1,500000	1,305000	2,091000	2,855000	1,730000	0,000000	0,188000	0,671000	2,522000	0,064000	0,418000	0,890000	Test
172	1,861000	2,231000	2,398000	0,220000	1,500000	1,303000	1,945000	3,105000	1,480000	0,000000	0,250000	0,693000	1,918000	0,052000	0,415000	0,883000	Train
173	1,861000	2,231000	2,398000	0,220000	1,500000	1,303000	1,945000	3,105000	1,480000	0,000000	0,250000	0,499000	3,197000	0,052000	0,690000	0,897000	Train
174	1,861000	2,192000	2,398000	0,230000	1,500000	1,300000	1,848000	2,355000	2,230000	0,000000	0,250000	0,444000	3,197000	0,034000	0,778000	0,896000	Train
175	1,861000	2,192000	2,398000	0,230000	1,500000	1,300000	1,848000	2,355000	2,230000	0,000000	0,250000	0,705000	1,918000	0,024000	0,410000	0,951000	Test
176	2,000000	2,083000	2,197000	0,182000	0,000000	1,071000	3,125000	2,470000	1,231000	0,000000	0,000000	0,650000	0,387000	0,000000	0,250000	1,800000	Train
177	2,000000	2,000000	2,303000	0,193000	0,000000	1,162000	2,056000	2,500000	1,500000	0,000000	0,000000	0,645000	1,282000	0,000000	0,315000	1,511000	Train
178	2,000000	2,000000	3,056000	0,111000	0,000000	1,162000	2,056000	1,000000	2,585000	0,000000	0,000000	1,000000	0,012000	0,000000	0,000000	0,648000	Train
179	1,919000	2,123000	3,248000	0,118000	1,149000	1,241000	2,022000	2,873000	1,034000	0,000000	0,143000	0,939000	0,183000	0,018000	0,060000	0,541000	Test
180	1,929000	2,125000	3,359000	0,125000	1,061000	1,299000	1,782000	3,197000	0,973000	0,063000	0,125000	0,788000	0,139000	0,195000	0,311000	0,892000	Train
181	1,937000	2,078000	3,393000	0,126000	0,986000	1,344000	1,506000	3,463000	0,929000	0,090000	0,111000	0,828000	0,160000	0,050000	0,319000	0,861000	Train
182	1,876000	2,228000	3,458000	0,122000	1,436000	1,350000	1,764000	3,046000	1,346000	0,125000	0,167000	0,602000	1,982000	0,134000	0,662000	0,631000	Train
183	1,943000	2,017000	3,426000	0,138000	0,922000	1,382000	1,278000	3,689000	0,896000	0,090000	0,100000	0,782000	0,179000	0,381000	0,512000	0,697000	Train
184	1,861000	2,279000	3,458000	0,130000	1,500000	1,305000	2,091000	2,642000	1,528000	0,000000	0,188000	0,909000	0,660000	0,037000	0,102000	0,361000	Train
185	1,861000	2,231000	3,409000	0,117000	1,500000	1,303000	1,945000	2,864000	1,306000	0,000000	0,250000	0,918000	0,355000	0,133000	0,102000	0,343000	Train
186	1,861000	2,192000	3,409000	0,122000	1,500000	1,300000	1,848000	2,197000	1,973000	0,000000	0,250000	0,922000	0,354000	0,018000	0,102000	0,431000	Train
187	1,948000	1,958000	3,458000	0,147000	0,866000	1,414000	1,108000	3,884000	0,871000	0,090000	0,091000	0,829000	0,200000	0,383000	0,490000	0,687000	Train
188	1,953000	1,906000	3,489000	0,154000	0,817000	1,444000	0,982000	3,923000	0,984000	0,090000	0,083000	0,824000	0,219000	0,382000	0,616000	0,611000	Train
189	1,956000	1,862000	3,519000	0,161000	0,773000	1,470000	0,888000	3,923000	1,122000	0,090000	0,077000	0,860000	0,239000	0,382000	0,586000	0,601000	Train
190	1,959000	1,825000	3,548000	0,166000	0,735000	1,493000	0,816000	3,850000	1,320000	0,090000	0,071000	0,865000	0,259000	0,382000	0,671000	0,552000	Test
191	1,962000	1,794000	3,577000	0,171000	0,700000	1,515000	0,759000	3,760000	1,525000	0,090000	0,067000	0,889000	0,278000	0,383000	0,641000	0,541000	Train
192	1,964000	1,767000	3,604000	0,176000	0,669000	1,535000	0,713000	3,665000	1,727000	0,090000	0,063000	0,895000	0,298000	0,383000	0,701000	0,504000	Train
193	1,966000	1,745000	3,631000	0,180000	0,640000	1,554000	0,675000	3,571000	1,920000	0,090000	0,059000	0,912000	0,319000	0,383000	0,673000	0,492000	Train
194	1,968000	1,727000	3,657000	0,183000	0,614000	1,571000	0,644000	3,480000	2,105000	0,090000	0,056000	0,920000	0,338000	0,384000	0,676000	0,463000	Test

ID	GNar	J	piPC03	X3Av	IVDE	HDcpx	Yindex	IC3	CIC3	GGI5	JGI1	SPH	RDF030m	RDF055m	L3p	HATS3u	Partition
195	1,970000	1,711000	3,683000	0,186000	0,591000	1,587000	0,617000	3,394000	2,279000	0,090000	0,053000	0,931000	0,358000	0,385000	0,657000	0,452000	Train
196	1,876000	2,228000	3,548000	0,106000	1,436000	1,350000	1,764000	3,326000	0,922000	0,125000	0,167000	0,729000	0,622000	0,407000	0,461000	0,544000	Train
197	1,876000	2,232000	3,504000	0,124000	1,436000	1,350000	1,735000	3,749000	0,643000	0,063000	0,222000	0,819000	0,310000	0,391000	0,318000	0,624000	Test
198	1,876000	2,297000	3,548000	0,126000	1,436000	1,354000	1,894000	3,559000	0,834000	0,063000	0,167000	0,788000	0,602000	0,207000	0,323000	0,620000	Train
199	1,876000	2,180000	3,504000	0,127000	1,436000	1,346000	1,630000	3,273000	1,120000	0,118000	0,222000	0,823000	0,309000	0,195000	0,337000	0,671000	Test
200	1,608000	3,307000	2,197000	0,245000	1,530000	1,350000	2,452000	3,237000	1,621000	0,000000	0,313000	0,588000	4,065000	0,377000	0,805000	0,462000	Train
201	1,644000	3,412000	2,398000	0,259000	1,522000	1,391000	2,287000	3,202000	1,798000	0,111000	0,278000	0,681000	5,332000	0,878000	0,728000	0,506000	Train
202	1,852000	2,366000	2,639000	0,224000	1,224000	1,354000	2,048000	3,310000	1,445000	0,063000	0,278000	0,480000	4,424000	0,166000	0,807000	0,768000	Train
203	1,876000	2,232000	2,565000	0,219000	1,436000	1,350000	1,735000	3,736000	1,019000	0,063000	0,222000	0,740000	2,919000	0,250000	0,423000	0,789000	Train
204	1,876000	2,297000	2,639000	0,216000	1,436000	1,354000	1,894000	3,356000	1,399000	0,063000	0,167000	0,710000	3,528000	0,266000	0,433000	0,745000	Train
205	1,835000	2,424000	2,833000	0,212000	1,571000	1,397000	1,895000	3,565000	1,342000	0,063000	0,200000	0,620000	5,428000	0,277000	0,679000	0,630000	Train
206	1,876000	2,228000	3,458000	0,122000	1,436000	1,350000	1,764000	3,046000	1,346000	0,125000	0,167000	0,602000	1,982000	0,134000	0,662000	0,631000	Train
207	1,876000	2,232000	3,504000	0,124000	1,436000	1,350000	1,735000	3,749000	0,643000	0,063000	0,222000	0,819000	0,311000	0,405000	0,318000	0,624000	Train
208	1,876000	2,180000	3,504000	0,127000	1,436000	1,346000	1,630000	3,273000	1,120000	0,118000	0,222000	0,823000	0,310000	0,195000	0,337000	0,671000	Train
209	1,876000	2,297000	3,548000	0,126000	1,436000	1,354000	1,894000	3,559000	0,834000	0,063000	0,167000	0,788000	0,625000	0,208000	0,323000	0,619000	Train
210	1,817000	2,413000	3,631000	0,134000	1,585000	1,357000	2,090000	2,653000	1,739000	0,000000	0,222000	0,898000	1,125000	0,147000	0,133000	0,273000	Train
211	1,817000	2,346000	3,590000	0,128000	1,585000	1,354000	1,921000	2,843000	1,549000	0,000000	0,278000	0,908000	0,832000	0,132000	0,133000	0,260000	Train
212	1,817000	2,341000	3,548000	0,114000	1,585000	1,355000	1,903000	2,128000	2,264000	0,000000	0,333000	0,911000	0,526000	0,339000	0,133000	0,194000	Test
213	1,835000	2,396000	3,631000	0,121000	1,571000	1,396000	1,869000	3,407000	1,178000	0,125000	0,200000	0,647000	2,121000	0,178000	0,620000	0,428000	Test
214	2,169000	1,925000	3,707000	0,126000	0,722000	1,393000	1,477000	3,096000	1,364000	0,000000	0,091000	0,843000	1,386000	0,027000	0,217000	0,972000	Train
215	1,813000	2,389000	3,548000	0,117000	1,685000	1,394000	1,847000	2,750000	1,835000	0,188000	0,300000	0,637000	1,995000	0,126000	0,625000	0,539000	Test
216	1,888000	2,131000	3,426000	0,124000	1,371000	1,386000	1,435000	3,324000	1,261000	0,118000	0,200000	0,759000	1,234000	0,212000	0,465000	0,707000	Train
217	2,129000	1,687000	4,034000	0,105000	0,619000	1,478000	0,928000	2,823000	1,820000	0,014000	0,107000	0,688000	0,026000	4,436000	0,765000	0,652000	Train
218	1,888000	2,240000	3,519000	0,141000	1,371000	1,391000	1,608000	3,657000	0,928000	0,153000	0,150000	0,634000	2,988000	0,149000	0,735000	0,586000	Train
219	1,888000	2,174000	3,590000	0,131000	1,371000	1,386000	1,463000	2,689000	1,896000	0,181000	0,200000	0,740000	0,264000	0,372000	0,624000	0,616000	Train
220	2,140000	1,800000	4,086000	0,094000	0,650000	1,455000	1,122000	2,482000	1,978000	0,139000	0,077000	1,000000	0,529000	0,269000	0,000000	0,702000	Train
221	2,169000	1,925000	4,123000	0,092000	0,722000	1,393000	1,477000	2,281000	1,889000	0,000000	0,091000	1,000000	0,054000	0,014000	0,000000	0,388000	Train
222	1,414000	4,020000	2,303000	0,250000	0,811000	1,317000	4,976000	1,140000	3,560000	0,000000	0,643000	0,106000	3,734000	0,018000	1,291000	0,493000	Train

Tableau E.3. Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Tc ainsi leurs ordres de partition dans les bases d'apprentissage et de test.

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
1	1,0000	1,09900	0,000000	0,000000	0,372000	-0,13900	-0,142000	0,000000	0,000000	0,000000	Test
2	3,0000	4,57500	0,000000	1,704000	0,530000	-0,21900	-0,133000	0,968000	0,000000	0,000000	Train
3	6,0000	6,14600	0,000000	1,328000	0,655000	-0,31700	-0,118000	1,054000	0,000000	0,000000	Test
4	6,0000	6,88000	0,000000	1,496000	0,550000	-0,17600	-0,110000	1,813000	0,000000	1,000000	Train
5	10,0000	7,03400	0,000000	1,175000	0,736000	-0,35500	-0,093000	0,883000	0,000000	0,000000	Train
6	10,0000	7,67100	0,000000	1,327000	0,625000	-0,33000	-0,108000	1,414000	0,163000	1,000000	Train
7	15,0000	7,59100	0,000000	1,084000	0,794000	-0,36800	-0,044000	0,784000	0,063000	0,000000	Test
8	21,0000	7,98500	0,000000	1,019000	0,836000	-0,44900	0,004000	0,707000	0,089000	0,000000	Train
9	28,0000	8,27700	0,693000	0,971000	0,866000	-0,54000	0,068000	0,649000	0,288000	0,000000	Train
10	36,0000	8,50700	1,099000	0,932000	0,889000	-0,56900	0,123000	0,606000	0,440000	0,000000	Train
11	45,0000	8,69500	1,386000	0,899000	0,907000	-0,58500	0,188000	0,566000	0,623000	0,000000	Train
12	55,0000	8,85300	1,609000	0,872000	0,921000	-0,64800	0,242000	0,536000	0,772000	0,000000	Train
13	66,0000	8,99000	1,792000	0,848000	0,932000	-0,73300	0,298000	0,506000	0,922000	0,000000	Train
14	78,0000	9,11000	1,946000	0,827000	0,941000	-0,78800	0,345000	0,482000	1,037000	0,000000	Train
15	91,0000	9,21700	2,079000	0,808000	0,948000	-0,81700	0,388000	0,459000	1,160000	0,000000	Train
16	105,0000	9,31400	2,197000	0,792000	0,954000	-0,85800	0,421000	0,439000	1,252000	0,000000	Train
17	120,0000	9,40200	2,303000	0,777000	0,959000	-0,92400	0,449000	0,420000	1,347000	0,000000	Test
18	136,0000	9,48300	2,398000	0,763000	0,964000	-0,99600	0,466000	0,404000	1,425000	0,000000	Train
19	153,0000	9,55900	2,485000	0,750000	0,967000	-1,04600	0,476000	0,387000	1,504000	0,000000	Train
20	171,0000	9,62800	2,565000	0,739000	0,970000	-1,07600	0,475000	0,373000	1,568000	0,000000	Train
21	190,0000	9,69400	2,639000	0,728000	0,973000	-1,12200	0,467000	0,359000	1,631000	0,000000	Train
22	210,0000	9,75500	2,708000	0,718000	0,975000	-1,19600	0,449000	0,347000	1,686000	0,000000	Train
23	231,0000	9,81300	2,773000	0,709000	0,977000	-1,26400	0,422000	0,335000	1,739000	0,000000	Test
24	253,0000	9,86700	2,833000	0,700000	0,979000	-1,30100	0,387000	0,325000	1,785000	0,000000	Train
25	276,0000	9,91900	2,890000	0,692000	0,981000	-1,33300	0,343000	0,314000	1,831000	0,000000	Test
26	300,0000	9,96800	2,944000	0,684000	0,982000	-1,39600	0,290000	0,305000	1,869000	0,000000	Train
27	325,0000	10,01500	2,996000	0,677000	0,984000	-1,47200	0,229000	0,295000	1,909000	0,000000	Train
28	378,0000	10,10300	3,091000	0,664000	0,986000	-1,55300	0,085000	0,279000	1,977000	0,000000	Train
29	496,0000	10,25800	3,258000	0,641000	0,989000	-1,77700	-0,286000	0,251000	2,088000	0,000000	Test
30	630,0000	10,39200	3,401000	0,622000	0,991000	-1,99500	-0,743000	0,228000	2,175000	0,000000	Train
31	10,0000	8,54100	0,000000	1,531000	0,000000	-0,00800	-0,073000	2,473000	0,000000	0,000000	Test
32	15,0000	8,13200	0,000000	1,201000	0,702000	-0,35800	-0,077000	1,089000	0,222000	1,000000	Train
33	15,0000	8,25500	0,000000	1,244000	0,663000	-0,44900	-0,099000	1,103000	0,421000	1,000000	Train
34	15,0000	9,00000	0,000000	1,410000	0,239000	-0,09300	-0,070000	1,836000	0,349000	0,000000	Train
35	15,0000	8,60600	0,000000	1,340000	0,668000	-0,23400	-0,077000	1,436000	0,503000	2,000000	Train
36	21,0000	8,41500	0,000000	1,108000	0,766000	-0,34000	-0,032000	0,929000	0,253000	1,000000	Train
37	21,0000	8,58500	0,000000	1,159000	0,742000	-0,47500	-0,047000	0,901000	0,498000	1,000000	Train
38	21,0000	9,25200	0,000000	1,284000	0,436000	-0,06300	-0,068000	1,441000	0,427000	0,000000	Train
39	21,0000	8,99400	0,000000	1,271000	0,699000	-0,36600	-0,108000	1,091000	0,763000	2,000000	Train
40	21,0000	8,80300	0,000000	1,207000	0,465000	-0,37400	-0,101000	1,149000	0,401000	2,000000	Train
41	21,0000	9,38800	0,000000	1,348000	0,413000	-0,11700	-0,057000	1,450000	0,908000	0,000000	Train

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
42	21,0000	8,72300	0,000000	1,210000	0,691000	-0,76400	-0,179000	0,739000	0,412000	1,000000	Train
43	21,0000	9,58000	0,000000	1,415000	0,349000	-0,05400	-0,052000	1,626000	0,799000	1,000000	Train
44	28,0000	8,63600	0,000000	1,038000	0,812000	-0,48800	-0,006000	0,831000	0,349000	1,000000	Train
45	28,0000	8,79600	0,000000	1,086000	0,787000	-0,51300	-0,007000	0,791000	0,521000	1,000000	Test
46	28,0000	8,85400	0,000000	1,103000	0,794000	-0,54400	0,001000	0,733000	0,534000	1,000000	Train
47	36,0000	9,13900	0,000000	1,089000	0,790000	-0,82700	-0,047000	0,537000	0,403000	1,000000	Train
48	36,0000	9,19300	0,000000	1,112000	0,790000	-0,67600	-0,031000	0,503000	0,358000	1,000000	Train
49	28,0000	9,38600	0,000000	1,179000	0,580000	0,00100	-0,003000	1,227000	0,390000	0,000000	Train
50	28,0000	9,20200	0,000000	1,190000	0,760000	-0,35800	-0,043000	0,888000	0,754000	2,000000	Train
51	28,0000	9,10100	0,000000	1,166000	0,566000	-0,43000	-0,064000	0,950000	0,632000	2,000000	Train
52	28,0000	8,91200	0,000000	1,112000	0,781000	-0,21300	-0,001000	0,943000	0,461000	2,000000	Train
53	28,0000	9,58900	0,000000	1,259000	0,545000	-0,06300	-0,030000	1,175000	0,959000	0,000000	Test
54	28,0000	9,30400	0,000000	1,228000	0,746000	-0,47900	-0,113000	0,894000	1,026000	2,000000	Train
55	36,0000	9,49200	0,000000	1,096000	0,686000	-0,22300	0,029000	1,113000	0,501000	0,000000	Train
56	36,0000	9,33800	0,000000	1,116000	0,800000	-0,43600	0,012000	0,786000	0,786000	2,000000	Train
57	36,0000	9,28700	0,000000	1,112000	0,654000	-0,58600	-0,011000	0,797000	0,680000	2,000000	Train
58	36,0000	9,05200	0,000000	1,040000	0,712000	-0,64600	0,028000	0,881000	0,549000	2,000000	Train
59	36,0000	9,69700	0,000000	1,174000	0,654000	-0,06600	0,033000	1,027000	1,023000	0,000000	Train
60	36,0000	9,47400	0,000000	1,166000	0,789000	-0,53700	-0,052000	0,743000	1,029000	2,000000	Train
61	36,0000	9,34500	0,000000	1,135000	0,636000	-0,74300	-0,046000	0,765000	0,655000	2,000000	Train
62	45,0000	9,57100	1,386000	1,030000	0,758000	-0,22500	0,108000	1,012000	0,585000	0,000000	Test
63	45,0000	9,43900	1,099000	1,052000	0,839000	-0,56600	0,057000	0,715000	0,921000	2,000000	Test
64	45,0000	9,40500	1,099000	1,059000	0,722000	-0,59800	0,053000	0,720000	0,785000	2,000000	Train
65	45,0000	9,34500	1,099000	1,048000	0,839000	-0,53200	0,086000	0,692000	0,787000	2,000000	Test
66	45,0000	9,77800	0,693000	1,100000	0,736000	-0,30100	0,082000	0,941000	1,160000	0,000000	Train
67	45,0000	9,39600	0,693000	1,062000	0,829000	-0,74100	0,049000	0,712000	1,055000	2,000000	Train
68	28,0000	8,97700	0,000000	1,150000	0,750000	-0,63400	-0,106000	0,603000	0,381000	1,000000	Train
69	28,0000	9,82700	0,000000	1,343000	0,472000	-0,15200	-0,072000	1,269000	1,114000	1,000000	Train
70	28,0000	9,63400	0,000000	1,269000	0,491000	0,14900	-0,053000	1,339000	0,457000	1,000000	Train
71	28,0000	9,88400	0,000000	1,368000	0,475000	-0,04500	-0,050000	1,261000	1,384000	1,000000	Train
72	28,0000	9,52100	0,000000	1,287000	0,496000	-0,27700	-0,133000	1,059000	1,106000	3,000000	Train
73	28,0000	9,33600	0,000000	1,244000	0,447000	-0,84600	-0,160000	0,839000	0,607000	2,000000	Train
74	28,0000	9,72400	0,000000	1,322000	0,493000	-0,65400	-0,235000	1,058000	1,039000	0,000000	Train
75	16,4830	10,46300	0,000000	0,935000	0,381000	0,28400	-0,231000	1,488000	0,458000	1,000000	Test
76	17,0830	10,37000	0,000000	0,904000	0,404000	-0,12800	-0,308000	1,537000	0,207000	1,000000	Train
77	16,5830	10,21200	0,000000	0,886000	0,595000	0,07800	-0,194000	1,281000	0,452000	3,000000	Train
78	16,2830	10,26000	0,000000	0,901000	0,573000	-0,18900	-0,368000	1,259000	0,533000	3,000000	Train
79	36,0000	9,94600	0,000000	1,256000	0,589000	-0,11200	-0,001000	1,046000	1,058000	1,000000	Train
80	36,0000	9,80100	0,000000	1,218000	0,575000	-0,27800	-0,058000	1,123000	0,657000	1,000000	Test
81	36,0000	9,63800	0,000000	1,162000	0,617000	-0,05000	0,024000	1,116000	0,555000	1,000000	Train
82	36,0000	9,54700	0,000000	1,186000	0,595000	-0,41100	-0,065000	0,889000	0,858000	3,000000	Train
83	21,1330	10,38000	0,000000	0,840000	0,482000	-0,34500	-0,042000	1,297000	0,358000	1,000000	Train
84	36,0000	10,02100	0,000000	1,317000	0,560000	-1,13100	-0,402000	0,839000	1,328000	0,000000	Train
85	36,0000	10,47800	0,000000	1,428000	0,274000	0,15100	-0,029000	1,285000	1,722000	0,000000	Train

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
86	36,0000	10,17300	0,000000	1,345000	0,521000	0,28200	-0,071000	1,198000	1,214000	2,000000	Train
87	36,0000	10,17500	0,000000	1,309000	0,358000	-0,60400	-0,287000	1,386000	0,515000	0,000000	Train
88	36,0000	10,28000	0,000000	1,384000	0,339000	-0,19900	-0,134000	1,131000	1,706000	2,000000	Train
89	36,0000	8,80600	1,099000	0,984000	0,849000	-0,54400	0,062000	0,754000	0,490000	1,000000	Train
90	36,0000	8,95700	0,693000	1,025000	0,832000	-0,64100	0,030000	0,720000	0,690000	1,000000	Train
91	36,0000	9,02300	0,693000	1,048000	0,829000	-0,58100	0,054000	0,663000	0,659000	1,000000	Train
92	45,0000	8,95100	1,386000	0,941000	0,874000	-0,52800	0,122000	0,703000	0,632000	1,000000	Train
93	45,0000	9,08600	1,386000	0,976000	0,860000	-0,65100	0,091000	0,666000	0,810000	1,000000	Train
94	45,0000	9,17100	1,099000	1,007000	0,852000	-0,57100	0,098000	0,607000	0,726000	1,000000	Train
95	55,0000	9,26400	1,609000	0,957000	0,883000	-0,67100	0,146000	0,571000	0,946000	1,000000	Train
96	45,0000	9,26100	1,099000	1,034000	0,832000	-0,84700	-0,037000	0,500000	0,565000	1,000000	Test
97	45,0000	10,03900	0,000000	1,208000	0,655000	-0,49000	-0,095000	0,949000	1,299000	1,000000	Train
98	45,0000	10,57700	0,000000	1,348000	0,435000	0,28300	0,013000	1,043000	1,610000	0,000000	Train
99	3,0000	4,57500	0,000000	1,704000	0,804000	0,11100	-0,042000	1,135000	0,000000	0,000000	Test
100	6,0000	6,14600	0,000000	1,328000	0,648000	-0,22700	-0,076000	0,971000	0,002000	0,000000	Train
101	10,0000	7,03400	0,000000	1,175000	0,728000	-0,34800	-0,078000	0,998000	0,042000	0,000000	Train
102	15,0000	7,59100	0,000000	1,084000	0,751000	-0,35700	-0,070000	0,866000	0,113000	0,000000	Train
103	21,0000	7,98500	0,000000	1,019000	0,805000	-0,41400	-0,038000	0,776000	0,197000	0,000000	Train
104	28,0000	8,27700	1,099000	0,971000	0,835000	-0,49800	0,007000	0,696000	0,286000	0,000000	Train
105	36,0000	8,50700	1,386000	0,932000	0,866000	-0,57400	0,059000	0,644000	0,446000	0,000000	Train
106	45,0000	8,69500	1,609000	0,899000	0,887000	-0,61200	0,119000	0,596000	0,577000	0,000000	Train
107	55,0000	8,85300	1,792000	0,872000	0,905000	-0,63600	0,179000	0,560000	0,726000	0,000000	Train
108	66,0000	8,99000	1,946000	0,848000	0,918000	-0,69200	0,240000	0,526000	0,867000	0,000000	Train
109	78,0000	9,11000	2,079000	0,827000	0,930000	-0,76800	0,297000	0,500000	0,992000	0,000000	Test
110	91,0000	9,21700	2,197000	0,808000	0,939000	-0,82900	0,352000	0,474000	1,102000	0,000000	Test
111	105,0000	9,31400	2,303000	0,792000	0,946000	-0,86000	0,398000	0,453000	1,203000	0,000000	Train
112	120,0000	9,40200	2,398000	0,777000	0,952000	-0,90200	0,440000	0,432000	1,295000	0,000000	Train
113	136,0000	9,48300	2,485000	0,763000	0,958000	-0,97000	0,473000	0,415000	1,378000	0,000000	Train
114	153,0000	9,55900	2,565000	0,750000	0,962000	-1,03100	0,498000	0,397000	1,454000	0,000000	Train
115	171,0000	9,62800	2,639000	0,739000	0,966000	-1,08000	0,514000	0,382000	1,522000	0,000000	Train
116	190,0000	9,69400	2,708000	0,728000	0,969000	-1,12000	0,521000	0,367000	1,586000	0,000000	Train
117	6,0000	6,14600	0,000000	1,328000	1,000000	0,16000	0,035000	0,450000	0,000000	0,000000	Train
118	6,0000	6,14600	0,000000	1,328000	0,768000	0,31800	-0,052000	0,830000	0,494000	0,000000	Train
119	6,0000	6,14600	0,000000	1,328000	0,810000	0,06500	-0,057000	0,829000	0,151000	0,000000	Train
120	6,0000	6,88000	0,000000	1,496000	0,773000	-0,10300	-0,044000	2,299000	0,000000	0,000000	Test
121	10,0000	7,67100	0,000000	1,327000	0,913000	0,01500	0,023000	1,445000	0,123000	0,000000	Test
122	15,0000	8,60600	0,000000	1,340000	0,281000	-0,50200	-0,066000	1,396000	0,336000	1,000000	Train
123	15,0000	8,60600	0,000000	1,340000	0,786000	0,26800	0,055000	1,327000	0,842000	0,000000	Train
124	15,0000	7,59100	0,000000	1,084000	0,755000	-0,27200	-0,090000	0,774000	0,323000	0,000000	Test
125	15,0000	7,59100	0,000000	1,084000	0,794000	-0,29700	-0,104000	0,761000	0,174000	0,000000	Test
126	15,0000	8,13200	0,000000	1,201000	0,618000	-0,38200	-0,122000	1,223000	0,023000	1,000000	Train
127	28,0000	9,63400	0,000000	1,269000	0,494000	-0,23100	0,067000	1,502000	0,289000	0,000000	Train
128	28,0000	9,63400	0,000000	1,269000	0,520000	0,08900	-0,120000	1,260000	0,238000	0,000000	Test
129	10,0000	7,67100	0,000000	1,327000	0,611000	-0,21800	-0,036000	1,418000	0,189000	0,000000	Train

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
130	10,0000	7,67100	0,000000	1,327000	0,747000	-0,04100	-0,081000	1,380000	0,413000	0,000000	Train
131	10,0000	7,67100	0,000000	1,327000	0,554000	-0,28700	-0,079000	1,432000	0,001000	1,000000	Train
132	15,0000	8,60600	0,000000	1,340000	0,880000	-0,02100	0,014000	1,190000	0,551000	0,000000	Train
133	15,0000	9,00000	0,000000	1,410000	0,242000	-0,12900	-0,007000	1,864000	0,192000	0,000000	Train
134	15,0000	8,13200	0,000000	1,201000	0,740000	-0,16200	0,052000	0,959000	0,368000	0,000000	Test
135	15,0000	7,59100	0,000000	1,084000	0,710000	-0,33500	-0,058000	0,946000	0,094000	0,000000	Train
136	1,5000	8,03000	0,000000	2,250000	0,246000	0,20800	-0,167000	0,000000	0,000000	0,000000	Train
137	3,3330	8,31800	0,000000	1,333000	0,363000	0,02000	-0,125000	0,873000	0,000000	0,000000	Train
138	4,5830	8,54100	0,000000	1,086000	0,507000	0,07200	-0,265000	1,362000	0,000000	0,000000	Train
139	7,2000	8,72300	0,000000	0,910000	0,534000	-0,53800	-0,017000	1,280000	0,000000	0,000000	Train
140	9,2170	8,87800	0,000000	0,830000	0,535000	-0,72900	0,018000	1,251000	0,202000	0,000000	Test
141	12,6100	9,01100	2,197000	0,753000	0,481000	-0,70500	-0,200000	1,107000	0,424000	0,000000	Train
142	11,0670	9,33600	0,000000	0,864000	0,631000	-0,49100	0,008000	1,295000	0,015000	1,000000	Train
143	16,2570	9,58900	1,099000	0,796000	0,687000	-0,57000	0,027000	1,011000	0,239000	1,000000	Train
144	22,6860	9,73000	1,609000	0,731000	0,757000	-0,61200	0,110000	0,859000	0,279000	1,000000	Train
145	22,4480	9,95700	1,609000	0,779000	0,722000	-0,39700	0,051000	0,983000	0,616000	2,000000	Test
146	30,2970	9,81700	1,946000	0,675000	0,803000	-0,63800	0,152000	0,761000	0,274000	1,000000	Train
147	29,8760	10,12200	1,946000	0,741000	0,760000	-0,53700	0,076000	0,806000	0,817000	2,000000	Test
148	7,8830	9,24400	0,000000	0,980000	0,598000	-0,03800	-0,280000	1,459000	0,051000	1,000000	Train
149	12,4830	9,52200	0,000000	0,868000	0,677000	-0,32400	-0,273000	1,136000	0,326000	1,000000	Train
150	18,2930	9,67200	1,099000	0,776000	0,742000	-0,29300	-0,238000	0,932000	0,324000	1,000000	Train
151	18,0830	9,91500	0,000000	0,835000	0,712000	-0,47700	-0,282000	1,107000	0,700000	2,000000	Train
152	25,2570	9,76500	1,609000	0,702000	0,796000	-0,35900	-0,220000	0,817000	0,356000	1,000000	Train
153	25,1020	9,92400	1,609000	0,744000	0,586000	-0,49100	-0,218000	0,940000	0,467000	2,000000	Test
154	33,3400	9,83500	1,946000	0,645000	0,834000	-0,42700	-0,193000	0,740000	0,508000	1,000000	Train
155	42,5180	9,89300	2,197000	0,599000	0,865000	-0,52300	-0,159000	0,682000	0,588000	1,000000	Test
156	52,7730	9,94500	2,303000	0,563000	0,886000	-0,62200	-0,125000	0,638000	0,749000	1,000000	Train
157	64,0910	9,99300	2,398000	0,533000	0,905000	-0,64500	-0,082000	0,598000	0,869000	1,000000	Test
158	76,4630	10,03900	2,485000	0,508000	0,918000	-0,66000	-0,046000	0,567000	1,006000	1,000000	Train
159	89,8800	10,08300	2,565000	0,487000	0,930000	-0,72600	-0,002000	0,536000	1,101000	1,000000	Train
160	119,8290	10,16500	2,708000	0,454000	0,946000	-0,85800	0,069000	0,486000	1,294000	1,000000	Train
161	136,3510	10,20300	2,773000	0,440000	0,952000	-0,89000	0,096000	0,466000	1,382000	1,000000	Train
162	153,9000	10,24100	2,833000	0,429000	0,958000	-0,93900	0,121000	0,445000	1,452000	1,000000	Train
163	172,4730	10,27700	2,890000	0,418000	0,962000	-1,00500	0,138000	0,428000	1,529000	1,000000	Train
164	29,3140	10,24400	1,609000	0,751000	0,626000	-0,47000	-0,031000	0,984000	1,091000	3,000000	Train
165	12,1830	10,01900	0,000000	0,961000	0,287000	0,09500	-0,287000	1,727000	0,161000	0,000000	Test
166	11,9830	9,75600	0,000000	0,913000	0,652000	-0,15000	-0,347000	1,398000	0,184000	2,000000	Test
167	11,9830	9,75600	0,000000	0,913000	0,614000	-0,15200	-0,295000	1,389000	0,078000	2,000000	Test
168	11,6830	9,81800	0,000000	0,933000	0,642000	-0,18900	-0,249000	1,368000	0,139000	2,000000	Train
169	15,9330	10,05800	0,000000	0,867000	0,353000	-0,48800	-0,065000	1,495000	0,341000	0,000000	Train
170	15,3620	9,85800	0,693000	0,848000	0,474000	-0,60300	-0,086000	1,232000	0,505000	2,000000	Train
171	15,3620	9,85800	0,693000	0,848000	0,671000	-0,43600	0,026000	1,204000	0,227000	2,000000	Test
172	15,6000	9,79800	0,000000	0,834000	0,693000	-0,38100	0,042000	1,179000	0,018000	2,000000	Train
173	15,6000	9,79800	0,000000	0,834000	0,499000	-0,57700	-0,090000	1,246000	0,326000	2,000000	Train

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
174	15,9330	9,76600	0,000000	0,823000	0,444000	-0,62400	-0,037000	1,297000	0,317000	2,000000	Train
175	15,9330	9,76600	0,000000	0,823000	0,705000	-0,41800	0,068000	1,313000	0,170000	2,000000	Test
176	4,5830	8,54100	0,000000	1,086000	0,650000	0,05400	-0,237000	1,235000	0,000000	0,000000	Train
177	7,2000	8,72300	0,000000	0,910000	0,645000	-0,47000	-0,041000	1,220000	0,000000	0,000000	Train
178	7,2000	8,72300	0,000000	0,910000	1,000000	0,12500	0,259000	0,507000	0,000000	0,000000	Train
179	11,0670	9,33600	0,000000	0,864000	0,939000	0,16000	0,246000	1,092000	0,000000	0,000000	Test
180	16,2570	9,58900	2,784000	0,796000	0,788000	0,17300	0,292000	0,760000	0,034000	0,000000	Train
181	22,6860	9,73000	3,270000	0,731000	0,828000	0,09300	0,416000	0,835000	0,098000	0,000000	Train
182	22,4480	9,95700	3,446000	0,779000	0,602000	-0,02600	0,143000	0,918000	0,231000	1,000000	Train
183	30,2970	9,81700	3,498000	0,675000	0,782000	0,10400	0,397000	0,741000	0,173000	0,000000	Train
184	15,3620	9,85800	2,151000	0,848000	0,909000	0,07400	0,214000	0,872000	0,590000	0,000000	Train
185	15,6000	9,79800	0,000000	0,834000	0,918000	0,17400	0,216000	0,824000	0,000000	0,000000	Train
186	15,9330	9,76600	0,000000	0,823000	0,922000	0,18700	0,219000	1,058000	0,093000	0,000000	Train
187	39,0520	9,88400	3,626000	0,628000	0,829000	0,05200	0,493000	0,685000	0,382000	0,000000	Train
188	48,9250	9,93900	3,703000	0,589000	0,824000	0,03800	0,483000	0,633000	0,386000	0,000000	Train
189	59,8950	9,98900	3,727000	0,556000	0,860000	-0,01900	0,551000	0,593000	0,600000	0,000000	Train
190	71,9460	10,03500	3,751000	0,528000	0,865000	-0,10600	0,555000	0,560000	0,676000	0,000000	Test
191	85,0670	10,07900	3,774000	0,505000	0,889000	-0,13300	0,597000	0,531000	0,837000	0,000000	Train
192	99,2480	10,12100	3,797000	0,485000	0,895000	-0,15300	0,612000	0,505000	0,893000	0,000000	Train
193	114,4810	10,16100	3,819000	0,468000	0,912000	-0,22200	0,634000	0,482000	1,022000	0,000000	Train
194	130,7610	10,20000	3,841000	0,453000	0,920000	-0,30800	0,651000	0,460000	1,086000	0,000000	Test
195	148,0810	10,23700	3,862000	0,440000	0,931000	-0,36100	0,658000	0,441000	1,195000	0,000000	Train
196	22,4480	9,95700	3,841000	0,779000	0,729000	0,24900	0,240000	0,935000	0,147000	0,000000	Train
197	21,5050	9,98600	3,056000	0,781000	0,819000	0,15500	0,282000	0,635000	0,029000	0,000000	Test
198	21,0520	10,06600	3,169000	0,800000	0,788000	0,10500	0,273000	0,657000	0,307000	0,000000	Train
199	22,1240	9,95700	2,784000	0,768000	0,823000	0,19000	0,293000	0,801000	0,211000	0,000000	Test
200	36,0000	9,38300	0,000000	1,158000	0,588000	-0,95500	-0,135000	0,664000	0,547000	2,000000	Train
201	45,0000	9,58400	0,000000	1,132000	0,681000	-0,74300	-0,144000	0,637000	1,085000	2,000000	Train
202	22,1240	10,28700	1,099000	0,823000	0,480000	-0,49600	-0,057000	1,171000	0,864000	0,000000	Train
203	21,5050	9,98600	1,386000	0,781000	0,740000	-0,43700	0,084000	0,959000	0,282000	2,000000	Train
204	21,0520	10,06600	1,386000	0,800000	0,710000	-0,55200	0,005000	0,907000	0,423000	2,000000	Train
205	26,7290	10,42700	1,792000	0,797000	0,620000	-0,51100	-0,004000	0,873000	1,065000	3,000000	Train
206	22,4480	9,95700	3,446000	0,779000	0,602000	-0,02600	0,143000	0,918000	0,231000	1,000000	Train
207	21,5050	9,98600	3,056000	0,781000	0,819000	0,16100	0,281000	0,635000	0,029000	0,000000	Train
208	22,1240	9,95700	2,784000	0,768000	0,823000	0,19100	0,292000	0,801000	0,212000	0,000000	Train
209	21,0520	10,06600	3,169000	0,800000	0,788000	0,06300	0,271000	0,657000	0,330000	0,000000	Train
210	20,3240	10,26700	2,784000	0,838000	0,898000	-0,02900	0,166000	0,739000	1,032000	0,000000	Train
211	20,8950	10,19700	2,151000	0,819000	0,908000	0,08200	0,172000	0,735000	0,485000	0,000000	Train
212	20,8000	10,17500	0,000000	0,817000	0,911000	0,16700	0,172000	0,540000	0,000000	0,000000	Test
213	27,7430	10,36300	3,663000	0,788000	0,647000	0,00200	0,157000	0,771000	0,359000	1,000000	Test
214	28,4100	10,26900	3,726000	0,767000	0,693000	-0,03900	0,113000	0,797000	0,205000	1,000000	Train
215	14,8590	10,57000	4,375000	0,634000	0,843000	-0,23100	0,159000	0,920000	0,082000	0,000000	Test
216	29,6380	10,45800	3,841000	0,788000	0,637000	0,09800	0,189000	1,169000	0,482000	0,000000	Train
217	30,1140	9,96800	3,596000	0,709000	0,759000	0,19800	0,416000	1,010000	0,164000	1,000000	Train

ID	D/D	MWC10	piPC07	Xindex	SPH	Mor15u	Mor11v	HATS4u	H5e	nCt	Partition
218	43,3950	10,53300	5,030000	0,497000	0,688000	0,58600	0,802000	0,758000	0,162000	0,000000	Train
219	29,8760	10,12200	3,726000	0,741000	0,634000	-0,04300	0,138000	0,735000	0,405000	1,000000	Train
220	29,3140	10,12200	3,641000	0,726000	0,740000	0,13600	0,364000	0,653000	0,426000	0,000000	Train
221	33,8690	10,61300	5,433000	0,546000	1,000000	0,38400	0,389000	0,637000	0,128000	0,000000	Train
222	14,8590	10,57000	5,614000	0,634000	1,000000	0,51600	0,276000	0,411000	0,077000	0,000000	Train
223	28,0000	10,25500	0,000000	1,474000	0,106000	0,10700	-0,020000	1,642000	1,215000	0,000000	Train

Tableau E.4. Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Pc ainsi leurs ordres de partition dans les bases d'apprentissage et de test.

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
1	4,00000	0,000000	0,000000	2,189000	0,00000	0,000000	0,000000	0,417000	0,00000	0,000000	0,205000	0,396000	0,209000	2,676000	0,053000	0,967000	0,000000	Test
2	5,00000	0,000000	2,042000	2,614000	0,00000	0,000000	0,250000	0,986000	1,52200	0,000000	0,386000	0,432000	0,234000	2,688000	0,108000	1,259000	0,000000	Train
3	8,00000	0,000000	1,089000	2,944000	0,00000	0,000000	0,278000	0,608000	1,52200	0,006000	0,635000	0,453000	0,164000	2,691000	0,124000	1,298000	0,455000	Test
4	10,00000	0,000000	1,592000	2,944000	0,00000	0,000000	0,500000	0,986000	4,56500	0,000000	0,544000	0,507000	0,425000	2,699000	0,185000	1,462000	0,000000	Train
5	9,00000	0,354000	0,853000	3,213000	0,00000	0,000000	0,262000	0,967000	1,52200	0,000000	0,885000	0,466000	0,136000	2,695000	0,120000	1,332000	0,706000	Train
6	13,00000	0,000000	1,106000	3,213000	0,00000	0,000000	0,486000	0,607000	6,34800	0,006000	1,845000	0,531000	0,296000	2,703000	0,184000	1,515000	0,864000	Train
7	12,00000	0,250000	0,737000	3,441000	0,00000	0,000000	0,237000	0,730000	1,52200	0,001000	1,134000	0,475000	0,103000	2,696000	0,135000	1,358000	0,807000	Test
8	13,00000	0,226000	0,666000	3,637000	-0,80200	0,056000	0,213000	0,997000	1,52200	0,001000	1,384000	0,482000	0,085000	2,698000	0,147000	1,385000	0,864000	Train
9	16,00000	0,213000	0,617000	3,810000	-0,41400	0,056000	0,192000	0,779000	1,52200	0,001000	1,634000	0,487000	0,068000	2,699000	0,157000	1,396000	0,912000	Train
10	17,00000	0,206000	0,580000	3,964000	0,00000	0,056000	0,174000	0,989000	1,52200	0,002000	1,883000	0,492000	0,058000	2,700000	0,166000	1,408000	0,944000	Train
11	20,00000	0,201000	0,552000	4,104000	0,38200	0,056000	0,159000	0,824000	1,52200	0,002000	2,133000	0,495000	0,048000	2,701000	0,174000	1,415000	0,982000	Train
12	21,00000	0,198000	0,528000	4,231000	0,71500	0,056000	0,145000	0,999000	1,52200	0,002000	2,382000	0,498000	0,041000	2,701000	0,181000	1,423000	1,012000	Train
13	24,00000	0,195000	0,509000	4,348000	1,00000	0,056000	0,134000	0,846000	1,52200	0,002000	2,632000	0,500000	0,035000	2,702000	0,188000	1,428000	1,045000	Train
14	25,00000	0,193000	0,492000	4,456000	1,24100	0,056000	0,117000	0,994000	1,52200	0,003000	2,881000	0,502000	0,031000	2,702000	0,193000	1,434000	1,070000	Train
15	28,00000	0,191000	0,478000	4,557000	1,44500	0,056000	0,104000	0,870000	1,52200	0,003000	3,131000	0,504000	0,027000	2,702000	0,198000	1,437000	1,098000	Train
16	29,00000	0,190000	0,465000	4,651000	1,61800	0,056000	0,095000	0,999000	1,52200	0,003000	3,380000	0,506000	0,024000	2,703000	0,203000	1,441000	1,120000	Train
17	32,00000	0,189000	0,454000	4,739000	1,76500	0,056000	0,087000	0,882000	1,52200	0,004000	3,630000	0,507000	0,021000	2,703000	0,207000	1,444000	1,144000	Test
18	33,00000	0,188000	0,444000	4,823000	1,89100	0,056000	0,080000	0,997000	1,52200	0,004000	3,879000	0,508000	0,019000	2,703000	0,210000	1,447000	1,162000	Train
19	36,00000	0,187000	0,435000	4,901000	2,00000	0,056000	0,075000	0,897000	1,52200	0,004000	4,129000	0,509000	0,017000	2,704000	0,214000	1,449000	1,182000	Train
20	37,00000	0,187000	0,427000	4,976000	2,09400	0,056000	0,070000	1,000000	1,52200	0,005000	4,378000	0,510000	0,016000	2,704000	0,217000	1,451000	1,198000	Train
21	40,00000	0,186000	0,419000	5,047000	2,17600	0,056000	0,066000	0,905000	1,52200	0,005000	4,628000	0,511000	0,014000	2,704000	0,220000	1,453000	1,216000	Train
22	41,00000	0,185000	0,412000	5,115000	2,24700	0,056000	0,062000	0,998000	1,52200	0,005000	4,877000	0,512000	0,013000	2,704000	0,222000	1,455000	1,229000	Train
23	44,00000	0,185000	0,406000	5,179000	2,31000	0,056000	0,059000	0,915000	1,52200	0,006000	5,127000	0,513000	0,012000	2,704000	0,224000	1,456000	1,244000	Test
24	45,00000	0,184000	0,400000	5,241000	2,36500	0,056000	0,056000	1,000000	1,52200	0,006000	5,377000	0,513000	0,011000	2,704000	0,226000	1,458000	1,256000	Train
25	48,00000	0,184000	0,394000	5,300000	2,41400	0,056000	0,053000	0,920000	1,52200	0,006000	5,626000	0,514000	0,010000	2,704000	0,229000	1,459000	1,269000	Test

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
26	49,00000	0,184000	0,389000	5,357000	2,45800	0,056000	0,050000	0,998000	1,52200	0,007000	5,876000	0,514000	0,009000	2,705000	0,230000	1,460000	1,279000	Train
27	52,00000	0,183000	0,384000	5,412000	2,49700	0,056000	0,048000	0,927000	1,52200	0,007000	6,125000	0,515000	0,009000	2,705000	0,232000	1,461000	1,291000	Train
28	56,00000	0,183000	0,375000	5,516000	2,56400	0,056000	0,044000	0,931000	1,52200	0,007000	6,624000	0,516000	0,007000	2,705000	0,235000	1,463000	1,310000	Train
29	64,00000	0,182000	0,360000	5,703000	2,66300	0,056000	0,038000	0,939000	1,52200	0,009000	7,622000	0,517000	0,006000	2,705000	0,240000	1,466000	1,343000	Test
30	72,00000	0,181000	0,348000	5,869000	2,73200	0,056000	0,034000	0,946000	1,52200	0,010000	8,620000	0,518000	0,005000	2,705000	0,244000	1,469000	1,369000	Train
31	17,00000	0,000000	1,553000	3,213000	0,00000	0,000000	0,750000	0,986000	9,13100	0,000000	0,679000	1,276000	0,333000	2,710000	0,269000	1,576000	0,000000	Test
32	14,00000	0,289000	0,882000	3,441000	0,00000	0,000000	0,472000	0,966000	5,70000	0,035000	2,094000	0,554000	0,213000	2,706000	0,178000	1,546000	0,945000	Train
33	14,00000	0,289000	0,955000	3,441000	0,00000	0,000000	0,465000	0,965000	8,10800	0,000000	3,159000	0,594000	0,214000	2,706000	0,167000	1,538000	1,452000	Train
34	20,00000	0,000000	1,203000	3,441000	0,00000	0,000000	0,695000	0,608000	13,01700	0,006000	1,942000	1,135000	0,260000	2,713000	0,263000	1,675000	1,154000	Train
35	20,00000	0,000000	1,088000	3,441000	0,00000	0,000000	0,548000	0,608000	9,97300	0,012000	2,033000	0,526000	0,390000	2,705000	0,204000	1,554000	1,485000	Train
36	17,00000	0,204000	0,758000	3,637000	-0,77500	0,000000	0,440000	0,730000	5,70000	0,199000	2,344000	0,560000	0,158000	2,707000	0,189000	1,560000	0,928000	Train
37	17,00000	0,232000	0,826000	3,637000	-0,74000	0,000000	0,438000	0,729000	7,45900	0,035000	3,409000	0,586000	0,162000	2,707000	0,173000	1,557000	1,409000	Train
38	21,00000	0,250000	0,970000	3,637000	-0,70400	0,000000	0,677000	0,967000	11,10900	0,040000	2,191000	1,081000	0,195000	2,715000	0,258000	1,712000	1,092000	Train
39	21,00000	0,236000	0,966000	3,637000	-0,67800	0,000000	0,534000	0,966000	10,88100	0,018000	3,400000	0,574000	0,302000	2,709000	0,184000	1,592000	1,857000	Train
40	21,00000	0,236000	0,875000	3,637000	-0,73200	0,000000	0,605000	0,966000	10,09200	0,040000	3,211000	1,052000	0,191000	2,715000	0,230000	1,718000	1,067000	Train
41	21,00000	0,250000	1,074000	3,637000	-0,61800	0,000000	0,660000	0,967000	16,90300	0,000000	3,204000	1,035000	0,204000	2,714000	0,233000	1,662000	1,981000	Train
42	19,00000	0,289000	0,899000	3,637000	-0,61800	0,000000	0,424000	0,738000	10,30000	0,085000	4,347000	0,603000	0,447000	2,713000	0,146000	1,631000	1,750000	Train
43	29,00000	0,000000	1,159000	3,637000	-0,60200	0,000000	0,731000	0,608000	18,42400	0,012000	3,181000	1,063000	0,343000	2,716000	0,273000	1,730000	2,044000	Train
44	18,00000	0,201000	0,680000	3,810000	-0,27200	0,111000	0,405000	0,996000	5,70000	0,198000	2,593000	0,573000	0,119000	2,708000	0,197000	1,567000	0,935000	Train
45	18,00000	0,189000	0,734000	3,810000	-0,41400	0,056000	0,410000	0,995000	7,45900	0,199000	3,659000	0,609000	0,123000	2,709000	0,187000	1,577000	1,339000	Test
46	18,00000	0,226000	0,756000	3,810000	-0,50800	0,056000	0,416000	0,995000	6,81000	0,070000	3,659000	0,582000	0,137000	2,708000	0,173000	1,574000	1,369000	Train
47	23,00000	0,198000	0,738000	3,964000	-0,15000	0,111000	0,372000	0,800000	9,99100	0,336000	4,855000	0,616000	0,253000	2,716000	0,179000	1,695000	1,477000	Train
48	23,00000	0,196000	0,767000	3,964000	0,00000	0,056000	0,366000	0,789000	8,99500	0,319000	4,856000	0,592000	0,341000	2,717000	0,173000	1,695000	1,429000	Train
49	24,00000	0,177000	0,825000	3,810000	0,00000	0,000000	0,638000	0,730000	11,10900	0,289000	2,441000	1,017000	0,151000	2,717000	0,266000	1,722000	0,987000	Train
50	24,00000	0,197000	0,849000	3,810000	-0,41400	0,000000	0,518000	0,729000	10,23200	0,151000	3,650000	0,572000	0,237000	2,711000	0,192000	1,618000	1,712000	Train
51	24,00000	0,194000	0,820000	3,810000	-0,38200	0,000000	0,556000	0,729000	11,85100	0,185000	4,526000	1,029000	0,151000	2,716000	0,223000	1,717000	1,418000	Train
52	24,00000	0,167000	0,755000	3,810000	0,00000	0,000000	0,570000	0,730000	9,87800	0,583000	3,553000	0,557000	0,202000	2,710000	0,213000	1,609000	1,020000	Train
53	24,00000	0,213000	0,931000	3,810000	-0,45300	0,000000	0,626000	0,730000	14,99500	0,040000	3,454000	1,017000	0,155000	2,715000	0,236000	1,691000	1,829000	Test

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
54	24,00000	0,201000	0,897000	3,810000	-0,41400	0,000000	0,519000	0,730000	11,78800	0,031000	4,768000	0,582000	0,232000	2,710000	0,182000	1,608000	2,130000	Train
55	25,00000	0,181000	0,730000	3,964000	0,00000	0,167000	0,594000	0,997000	11,10900	0,281000	2,690000	0,966000	0,114000	2,717000	0,269000	1,712000	0,953000	Train
56	25,00000	0,172000	0,758000	3,964000	0,00000	0,111000	0,496000	0,996000	10,23200	0,309000	3,899000	0,596000	0,175000	2,712000	0,203000	1,638000	1,606000	Train
57	25,00000	0,202000	0,757000	3,964000	0,00000	0,111000	0,536000	0,995000	11,75000	0,217000	4,763000	1,017000	0,123000	2,717000	0,217000	1,723000	1,347000	Train
58	25,00000	0,187000	0,678000	3,964000	0,00000	0,222000	0,537000	0,997000	9,87800	0,236000	3,803000	0,921000	0,113000	2,717000	0,240000	1,708000	0,931000	Train
59	25,00000	0,166000	0,818000	3,964000	0,00000	0,056000	0,595000	0,997000	14,99500	0,290000	3,703000	0,966000	0,122000	2,716000	0,251000	1,704000	1,672000	Train
60	25,00000	0,190000	0,816000	3,964000	-0,29600	0,056000	0,499000	0,996000	11,14000	0,163000	5,017000	0,593000	0,186000	2,711000	0,186000	1,629000	1,973000	Train
61	25,00000	0,170000	0,780000	3,964000	-0,33800	0,056000	0,528000	0,784000	14,19000	0,359000	5,740000	0,982000	0,176000	2,720000	0,231000	1,777000	1,676000	Train
62	28,00000	0,180000	0,663000	4,104000	0,00000	0,167000	0,551000	0,779000	11,10900	0,281000	2,940000	0,925000	0,092000	2,717000	0,274000	1,705000	0,938000	Test
63	28,00000	0,175000	0,689000	4,104000	0,00000	0,167000	0,469000	0,779000	10,23200	0,308000	4,149000	0,595000	0,136000	2,713000	0,211000	1,643000	1,550000	Test
64	28,00000	0,184000	0,699000	4,104000	0,00000	0,111000	0,502000	0,778000	11,20600	0,378000	5,025000	1,004000	0,095000	2,718000	0,225000	1,721000	1,289000	Train
65	28,00000	0,179000	0,688000	4,104000	0,00000	0,111000	0,505000	0,778000	10,98800	0,597000	5,117000	0,590000	0,136000	2,712000	0,211000	1,637000	1,302000	Test
66	28,00000	0,163000	0,700000	4,104000	0,00000	0,167000	0,508000	0,779000	13,39600	0,541000	6,182000	0,611000	0,127000	2,712000	0,212000	1,636000	1,615000	Train
67	22,00000	0,221000	0,814000	3,810000	-0,61800	0,000000	0,395000	0,565000	8,77800	0,182000	4,497000	0,597000	0,331000	2,715000	0,162000	1,667000	1,588000	Train
68	30,00000	0,204000	1,031000	3,810000	-0,32600	0,000000	0,702000	0,966000	17,95300	0,046000	4,516000	1,033000	0,271000	2,718000	0,254000	1,755000	2,236000	Train
69	30,00000	0,204000	0,932000	3,810000	0,00000	0,000000	0,788000	0,958000	16,88500	0,183000	2,860000	1,020000	0,248000	2,718000	0,295000	1,757000	1,159000	Train
70	30,00000	0,204000	1,068000	3,810000	-0,41400	0,000000	0,706000	0,967000	21,35600	0,026000	4,443000	0,994000	0,278000	2,718000	0,241000	1,734000	2,663000	Train
71	30,00000	0,192000	0,964000	3,810000	-0,34300	0,000000	0,568000	0,964000	15,91400	0,058000	4,467000	1,003000	0,270000	2,718000	0,231000	1,760000	2,219000	Train
72	26,00000	0,246000	0,923000	3,810000	-0,61800	0,000000	0,503000	0,935000	18,57300	0,093000	6,161000	1,077000	0,283000	2,721000	0,219000	1,790000	2,122000	Train
73	26,00000	0,250000	1,019000	3,810000	-0,61800	0,000000	0,618000	0,967000	26,09900	0,006000	5,608000	0,957000	0,324000	2,719000	0,233000	1,753000	2,560000	Train
74	10,00000	0,136000	0,746000	3,667000	-0,22400	0,000000	0,606000	0,560000	16,35800	0,005000	2,091000	1,084000	0,379000	2,721000	0,337000	1,760000	1,844000	Train
75	10,00000	0,137000	0,698000	3,667000	0,00000	0,000000	0,683000	0,876000	19,20300	0,042000	2,514000	1,148000	0,220000	2,720000	0,358000	1,744000	1,044000	Test
76	10,00000	0,141000	0,678000	3,667000	-0,32500	0,000000	0,544000	0,548000	18,58700	0,125000	2,290000	0,773000	0,285000	2,718000	0,345000	1,711000	1,440000	Train
77	10,00000	0,134000	0,699000	3,667000	-0,28600	0,000000	0,490000	0,877000	18,66700	0,016000	2,598000	0,805000	0,308000	2,718000	0,309000	1,712000	1,799000	Train
78	33,00000	0,173000	0,908000	3,964000	0,00000	0,000000	0,682000	0,729000	17,30500	0,315000	4,766000	0,992000	0,218000	2,720000	0,256000	1,776000	2,010000	Train
79	33,00000	0,170000	0,863000	3,964000	0,00000	0,000000	0,722000	0,723000	20,17000	0,255000	4,202000	1,017000	0,197000	2,719000	0,289000	1,767000	1,402000	Train
80	33,00000	0,144000	0,798000	3,964000	0,00000	0,000000	0,739000	0,730000	15,28700	0,643000	3,650000	0,988000	0,186000	2,719000	0,284000	1,772000	1,049000	Test
81	33,00000	0,169000	0,832000	3,964000	0,00000	0,000000	0,602000	0,730000	14,66100	0,450000	4,757000	1,004000	0,210000	2,720000	0,239000	1,782000	1,665000	Train

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
82	10,00000	0,148000	0,621000	3,837000	0,00000	0,000000	0,631000	0,731000	18,41900	0,072000	4,353000	1,056000	0,288000	2,722000	0,354000	1,786000	1,506000	Train
83	33,00000	0,250000	0,998000	3,964000	-0,61800	0,000000	0,571000	0,967000	35,29500	0,001000	8,012000	0,903000	0,427000	2,721000	0,197000	1,754000	3,000000	Train
84	41,00000	0,177000	1,114000	3,964000	0,00000	0,000000	0,782000	0,967000	27,33200	0,053000	5,659000	1,441000	0,251000	2,721000	0,273000	1,792000	3,128000	Train
85	41,00000	0,167000	1,011000	3,964000	0,00000	0,000000	0,723000	0,958000	22,58200	0,185000	4,069000	0,976000	0,327000	2,719000	0,280000	1,776000	2,418000	Train
86	41,00000	0,177000	0,960000	3,964000	0,00000	0,000000	0,926000	0,957000	30,87400	0,160000	6,432000	1,283000	0,240000	2,720000	0,351000	1,796000	1,279000	Train
87	41,00000	0,167000	1,064000	3,964000	0,00000	0,000000	0,727000	0,967000	28,46600	0,052000	6,253000	1,326000	0,271000	2,721000	0,252000	1,797000	3,115000	Train
88	21,00000	0,196000	0,626000	3,964000	0,00000	0,111000	0,372000	0,779000	5,70000	0,198000	2,842000	0,573000	0,094000	2,709000	0,206000	1,569000	0,949000	Train
89	21,00000	0,189000	0,667000	3,964000	-0,22500	0,111000	0,380000	0,779000	7,45900	0,199000	3,907000	0,604000	0,095000	2,709000	0,195000	1,575000	1,312000	Train
90	21,00000	0,198000	0,693000	3,964000	0,00000	0,056000	0,385000	0,778000	6,81000	0,234000	3,908000	0,597000	0,104000	2,709000	0,183000	1,580000	1,324000	Train
91	22,00000	0,193000	0,586000	4,104000	0,00000	0,111000	0,342000	0,988000	5,70000	0,199000	3,093000	0,581000	0,074000	2,709000	0,213000	1,570000	0,971000	Train
92	22,00000	0,187000	0,618000	4,104000	0,24100	0,111000	0,352000	0,987000	7,45900	0,199000	4,158000	0,613000	0,076000	2,709000	0,203000	1,578000	1,294000	Train
93	22,00000	0,179000	0,652000	4,104000	0,00000	0,056000	0,360000	0,987000	6,81000	0,399000	4,158000	0,625000	0,082000	2,710000	0,190000	1,585000	1,277000	Train
94	25,00000	0,195000	0,602000	4,231000	0,44700	0,111000	0,333000	0,825000	6,81000	0,234000	4,407000	0,602000	0,068000	2,710000	0,198000	1,585000	1,294000	Train
95	26,00000	0,197000	0,677000	4,104000	0,38200	0,111000	0,348000	0,790000	8,77800	0,291000	4,996000	0,611000	0,191000	2,717000	0,191000	1,700000	1,426000	Train
96	34,00000	0,154000	0,847000	4,104000	0,00000	0,056000	0,682000	0,986000	23,77600	1,198000	5,586000	0,956000	0,160000	2,719000	0,277000	1,756000	1,956000	Test
97	44,00000	0,167000	0,999000	4,104000	0,00000	0,000000	0,763000	0,730000	25,42400	0,341000	5,909000	1,355000	0,216000	2,723000	0,271000	1,822000	2,795000	Train
98	5,00000	0,000000	2,042000	2,252000	0,00000	0,000000	0,250000	0,970000	0,74500	0,000000	0,219000	0,176000	0,254000	2,688000	0,091000	1,095000	0,000000	Train
99	8,00000	0,000000	1,089000	2,667000	0,00000	0,000000	0,278000	0,564000	3,55700	0,001000	0,372000	0,432000	0,184000	2,701000	0,144000	1,366000	0,377000	Test
100	9,00000	0,204000	0,853000	2,989000	0,00000	0,000000	0,262000	0,891000	3,73900	0,026000	0,621000	0,448000	0,141000	2,704000	0,158000	1,387000	0,532000	Train
101	12,00000	0,174000	0,737000	3,252000	0,00000	0,000000	0,237000	0,822000	3,74400	0,002000	0,871000	0,529000	0,111000	2,707000	0,157000	1,413000	0,723000	Train
102	13,00000	0,180000	0,666000	3,474000	-0,80200	0,056000	0,213000	0,935000	3,73900	0,002000	1,120000	0,527000	0,088000	2,707000	0,168000	1,427000	0,821000	Train
103	16,00000	0,179000	0,617000	3,667000	-0,41400	0,056000	0,192000	0,782000	3,74400	0,002000	1,370000	0,553000	0,073000	2,708000	0,177000	1,448000	0,872000	Train
104	17,00000	0,178000	0,580000	3,837000	0,00000	0,056000	0,174000	0,937000	3,74400	0,002000	1,620000	0,550000	0,060000	2,709000	0,186000	1,453000	0,921000	Train
105	20,00000	0,178000	0,552000	3,989000	0,38200	0,056000	0,159000	0,881000	3,74400	0,003000	1,869000	0,558000	0,051000	2,709000	0,193000	1,462000	0,954000	Train
106	21,00000	0,178000	0,528000	4,126000	0,71500	0,056000	0,145000	0,954000	3,74400	0,003000	2,119000	0,556000	0,043000	2,709000	0,200000	1,465000	0,991000	Train
107	24,00000	0,178000	0,509000	4,252000	1,00000	0,056000	0,134000	0,860000	3,74400	0,003000	2,368000	0,559000	0,037000	2,709000	0,205000	1,471000	1,020000	Train
108	25,00000	0,178000	0,492000	4,367000	1,24100	0,056000	0,117000	0,956000	3,74400	0,004000	2,618000	0,557000	0,032000	2,709000	0,210000	1,473000	1,051000	Train
109	28,00000	0,178000	0,478000	4,474000	1,44500	0,056000	0,104000	0,910000	3,73900	0,004000	2,867000	0,557000	0,028000	2,709000	0,214000	1,476000	1,076000	Test

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
110	29,00000	0,178000	0,465000	4,574000	1,61800	0,056000	0,095000	0,965000	3,74400	0,004000	3,117000	0,556000	0,025000	2,709000	0,219000	1,477000	1,102000	Test
111	32,00000	0,177000	0,454000	4,667000	1,76500	0,056000	0,087000	0,898000	3,74400	0,005000	3,366000	0,556000	0,022000	2,709000	0,222000	1,480000	1,124000	Train
112	33,00000	0,177000	0,444000	4,754000	1,89100	0,056000	0,080000	0,966000	3,73900	0,005000	3,616000	0,554000	0,020000	2,709000	0,225000	1,480000	1,146000	Train
113	36,00000	0,177000	0,435000	4,837000	2,00000	0,056000	0,075000	0,929000	3,74400	0,005000	3,865000	0,554000	0,018000	2,709000	0,228000	1,482000	1,164000	Train
114	37,00000	0,177000	0,427000	4,915000	2,09400	0,056000	0,070000	0,972000	3,74400	0,006000	4,115000	0,553000	0,016000	2,709000	0,230000	1,482000	1,183000	Train
115	40,00000	0,177000	0,419000	4,989000	2,17600	0,056000	0,066000	0,920000	3,74400	0,006000	4,364000	0,553000	0,015000	2,709000	0,233000	1,484000	1,199000	Train
116	8,00000	0,000000	1,089000	2,351000	0,00000	0,000000	0,278000	0,601000	1,18800	0,006000	0,145000	0,000000	0,219000	1,807000	0,092000	0,915000	0,211000	Train
117	8,00000	0,000000	1,089000	2,667000	0,00000	0,000000	0,278000	0,492000	1,40500	0,000000	0,387000	0,264000	0,307000	2,700000	0,141000	1,362000	0,861000	Train
118	8,00000	0,000000	1,089000	2,667000	0,00000	0,000000	0,278000	0,634000	0,17200	0,004000	0,360000	0,264000	0,156000	2,689000	0,097000	1,211000	0,608000	Train
119	10,00000	0,000000	1,592000	2,667000	0,00000	0,000000	0,500000	0,994000	3,96000	0,000000	0,953000	0,264000	0,405000	2,698000	0,147000	1,318000	0,000000	Train
120	13,00000	0,000000	1,106000	2,739000	0,00000	0,000000	0,486000	0,601000	5,33100	0,006000	0,881000	0,122000	0,385000	2,706000	0,154000	1,327000	0,678000	Test
121	20,00000	0,000000	1,088000	3,252000	0,00000	0,000000	0,548000	0,558000	13,36800	0,001000	1,754000	1,117000	0,263000	2,720000	0,228000	1,706000	1,459000	Test
122	20,00000	0,000000	1,088000	3,252000	0,00000	0,000000	0,548000	0,645000	7,56000	0,004000	0,990000	0,357000	0,442000	2,703000	0,148000	1,518000	1,807000	Train
123	12,00000	0,167000	0,737000	3,252000	0,00000	0,000000	0,237000	0,831000	6,15800	0,032000	0,730000	0,471000	0,187000	2,713000	0,199000	1,562000	0,784000	Train
124	12,00000	0,167000	0,737000	3,252000	0,00000	0,000000	0,237000	0,681000	2,85800	0,082000	0,762000	0,478000	0,098000	2,702000	0,163000	1,435000	0,686000	Test
125	14,00000	0,167000	0,882000	3,252000	0,00000	0,000000	0,472000	0,896000	7,97400	0,017000	1,889000	0,647000	0,221000	2,714000	0,212000	1,573000	0,734000	Test
126	30,00000	0,151000	0,932000	3,667000	0,00000	0,000000	0,788000	0,823000	16,59600	0,123000	3,513000	0,970000	0,275000	2,723000	0,332000	1,758000	0,914000	Train
127	30,00000	0,144000	0,932000	3,667000	0,00000	0,000000	0,788000	0,936000	13,54500	0,153000	2,743000	0,970000	0,255000	2,720000	0,291000	1,767000	0,992000	Train
128	13,00000	0,000000	1,106000	2,989000	0,00000	0,000000	0,486000	0,527000	6,90600	0,000000	0,945000	0,552000	0,306000	2,712000	0,200000	1,587000	0,836000	Test
129	13,00000	0,000000	1,106000	2,989000	0,00000	0,000000	0,486000	0,647000	4,80400	0,002000	1,552000	0,374000	0,319000	2,704000	0,154000	1,480000	1,039000	Train
130	13,00000	0,000000	1,106000	2,989000	0,00000	0,000000	0,486000	0,557000	7,58100	0,001000	0,446000	0,629000	0,292000	2,710000	0,184000	1,515000	0,661000	Train
131	20,00000	0,000000	1,088000	3,046000	0,00000	0,000000	0,548000	0,612000	7,97200	0,013000	1,353000	0,198000	0,385000	2,707000	0,165000	1,432000	1,420000	Train
132	20,00000	0,000000	1,203000	3,252000	0,00000	0,000000	0,695000	0,557000	12,65400	0,001000	0,509000	1,117000	0,281000	2,720000	0,266000	1,648000	0,879000	Train
133	14,00000	0,204000	0,882000	3,252000	0,00000	0,000000	0,472000	0,888000	6,48600	0,070000	0,403000	0,486000	0,243000	2,710000	0,187000	1,602000	0,953000	Train
134	12,00000	0,118000	0,737000	3,046000	0,00000	0,000000	0,237000	0,757000	5,96100	0,856000	0,607000	0,573000	0,122000	2,715000	0,188000	1,472000	0,647000	Test
135	0,00000	0,000000	4,500000	2,252000	0,00000	0,000000	0,000000	0,393000	0,00000	0,000000	0,084000	0,567000	0,374000	2,696000	0,101000	1,314000	0,000000	Train
136	0,00000	0,000000	1,231000	2,667000	0,00000	0,000000	0,000000	0,255000	0,10000	0,000000	1,154000	0,594000	0,394000	2,701000	0,194000	1,418000	0,000000	Train
137	0,00000	0,177000	0,920000	2,989000	0,00000	0,000000	0,000000	0,260000	2,56600	0,000000	1,809000	0,579000	0,418000	2,705000	0,275000	1,486000	0,000000	Train

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
138	0,00000	0,177000	0,684000	3,252000	-1,00000	0,000000	0,000000	0,263000	9,15300	0,000000	2,249000	0,629000	0,422000	2,711000	0,298000	1,584000	0,537000	Train
139	0,00000	0,177000	0,604000	3,474000	-0,80200	0,000000	0,000000	0,329000	14,30100	0,000000	3,591000	0,703000	0,398000	2,716000	0,323000	1,677000	1,105000	Train
140	0,00000	0,177000	0,522000	3,667000	-0,41400	0,000000	0,000000	0,524000	23,60800	0,008000	3,466000	0,849000	0,398000	2,723000	0,351000	1,768000	1,542000	Test
141	2,00000	0,163000	0,643000	3,474000	-0,41400	0,000000	0,347000	0,516000	9,16700	0,032000	2,544000	0,616000	0,308000	2,712000	0,299000	1,604000	0,929000	Train
142	3,00000	0,160000	0,573000	3,667000	0,00000	0,063000	0,314000	0,849000	10,29000	0,222000	3,844000	0,633000	0,247000	2,714000	0,275000	1,630000	1,322000	Train
143	4,00000	0,157000	0,507000	3,837000	0,00000	0,090000	0,279000	0,796000	9,64100	0,188000	4,094000	0,627000	0,180000	2,715000	0,273000	1,637000	1,250000	Train
144	6,00000	0,150000	0,558000	3,837000	0,00000	0,125000	0,397000	0,843000	12,66700	0,301000	4,032000	0,620000	0,273000	2,715000	0,267000	1,642000	1,636000	Train
145	5,00000	0,151000	0,451000	3,989000	0,38200	0,090000	0,252000	0,795000	9,64100	0,353000	4,344000	0,631000	0,145000	2,716000	0,281000	1,657000	1,192000	Test
146	7,00000	0,145000	0,520000	3,989000	0,24900	0,153000	0,388000	0,786000	13,63300	0,307000	5,400000	0,633000	0,226000	2,717000	0,261000	1,666000	1,843000	Train
147	2,00000	0,161000	0,793000	3,252000	-0,75700	0,000000	0,341000	0,602000	5,70200	0,003000	2,453000	0,594000	0,296000	2,708000	0,291000	1,542000	0,476000	Test
148	3,00000	0,158000	0,655000	3,474000	-0,61800	0,000000	0,316000	0,923000	5,73600	0,158000	3,487000	0,595000	0,235000	2,710000	0,270000	1,569000	0,997000	Train
149	4,00000	0,155000	0,549000	3,667000	-0,41400	0,080000	0,290000	0,722000	5,72800	0,160000	3,737000	0,612000	0,168000	2,712000	0,274000	1,596000	0,981000	Train
150	6,00000	0,148000	0,621000	3,667000	-0,26000	0,000000	0,385000	0,921000	8,00800	0,177000	4,979000	0,586000	0,277000	2,712000	0,268000	1,602000	1,419000	Train
151	5,00000	0,149000	0,474000	3,837000	0,09100	0,080000	0,265000	0,962000	5,72800	0,426000	3,986000	0,606000	0,134000	2,713000	0,280000	1,611000	0,955000	Train
152	7,00000	0,147000	0,518000	3,837000	0,00000	0,160000	0,414000	0,719000	10,27500	0,792000	4,896000	1,035000	0,152000	2,721000	0,310000	1,745000	1,038000	Train
153	6,00000	0,152000	0,419000	3,989000	0,65800	0,080000	0,243000	0,814000	5,72800	0,426000	4,235000	0,620000	0,101000	2,714000	0,287000	1,619000	0,953000	Test
154	7,00000	0,154000	0,379000	4,126000	1,12500	0,080000	0,224000	0,960000	5,73600	0,426000	4,485000	0,613000	0,084000	2,715000	0,293000	1,624000	0,956000	Train
155	8,00000	0,156000	0,349000	4,252000	1,47800	0,080000	0,207000	0,830000	5,73200	0,427000	4,734000	0,623000	0,066000	2,715000	0,299000	1,626000	0,979000	Test
156	9,00000	0,157000	0,325000	4,367000	1,61800	0,080000	0,192000	0,974000	5,72800	0,427000	4,984000	0,616000	0,056000	2,715000	0,302000	1,628000	0,999000	Train
157	10,00000	0,158000	0,306000	4,474000	1,61800	0,080000	0,171000	0,871000	5,72800	0,427000	5,235000	0,623000	0,046000	2,716000	0,306000	1,627000	1,027000	Test
158	11,00000	0,159000	0,291000	4,574000	1,61800	0,080000	0,156000	0,973000	5,73600	0,428000	5,483000	0,617000	0,040000	2,716000	0,307000	1,626000	1,050000	Train
159	13,00000	0,161000	0,267000	4,754000	1,94100	0,080000	0,134000	0,981000	5,72800	0,429000	5,983000	0,617000	0,030000	2,716000	0,311000	1,623000	1,097000	Train
160	14,00000	0,162000	0,258000	4,837000	2,05200	0,080000	0,126000	0,901000	5,73600	0,429000	6,232000	0,622000	0,026000	2,716000	0,312000	1,621000	1,121000	Train
161	15,00000	0,162000	0,249000	4,915000	2,13700	0,080000	0,118000	0,980000	5,73200	0,429000	6,482000	0,617000	0,023000	2,716000	0,312000	1,619000	1,138000	Train
162	16,00000	0,163000	0,242000	4,989000	2,20500	0,080000	0,111000	0,905000	5,72800	0,429000	6,732000	0,620000	0,020000	2,716000	0,313000	1,617000	1,160000	Train
163	11,00000	0,139000	0,525000	3,989000	0,00000	0,236000	0,540000	0,697000	20,78000	1,149000	5,570000	0,905000	0,234000	2,723000	0,325000	1,778000	1,776000	Train
164	5,00000	0,148000	0,781000	3,474000	-0,61800	0,000000	0,557000	0,590000	14,78900	0,002000	1,905000	1,139000	0,267000	2,719000	0,351000	1,708000	0,882000	Train
165	5,00000	0,149000	0,706000	3,474000	-0,73200	0,000000	0,502000	0,877000	10,70800	0,014000	2,698000	0,626000	0,256000	2,711000	0,305000	1,597000	0,794000	Test

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
166	5,00000	0,149000	0,706000	3,474000	-0,73200	0,000000	0,502000	0,867000	9,39700	0,017000	2,820000	0,690000	0,255000	2,713000	0,315000	1,632000	0,790000	Test
167	5,00000	0,148000	0,741000	3,474000	-0,66500	0,000000	0,427000	0,554000	9,20300	0,004000	3,123000	0,607000	0,424000	2,712000	0,282000	1,605000	1,202000	Test
168	5,00000	0,151000	0,654000	3,667000	0,00000	0,000000	0,558000	0,516000	19,35900	0,032000	4,058000	1,120000	0,272000	2,721000	0,365000	1,752000	1,360000	Train
169	5,00000	0,150000	0,633000	3,667000	-0,41400	0,000000	0,457000	0,516000	19,78100	0,032000	4,587000	0,913000	0,329000	2,721000	0,344000	1,753000	1,717000	Train
170	5,00000	0,150000	0,633000	3,667000	-0,41400	0,000000	0,457000	0,516000	11,12400	0,064000	3,345000	0,613000	0,375000	2,714000	0,290000	1,636000	1,574000	Train
171	5,00000	0,158000	0,612000	3,667000	-0,26300	0,000000	0,462000	0,731000	9,18100	0,052000	2,838000	0,609000	0,322000	2,714000	0,292000	1,637000	1,198000	Test
172	5,00000	0,158000	0,612000	3,667000	-0,26300	0,000000	0,462000	0,516000	16,88400	0,052000	4,081000	0,905000	0,305000	2,721000	0,347000	1,750000	1,301000	Train
173	5,00000	0,143000	0,597000	3,667000	0,00000	0,000000	0,526000	0,552000	17,28000	0,034000	4,081000	1,008000	0,260000	2,721000	0,355000	1,758000	1,288000	Train
174	5,00000	0,143000	0,597000	3,667000	0,00000	0,000000	0,526000	0,513000	9,18100	0,024000	2,838000	0,603000	0,231000	2,711000	0,314000	1,606000	1,155000	Train
175	0,00000	0,118000	0,920000	2,739000	0,00000	0,000000	0,000000	0,307000	1,99000	0,000000	1,530000	0,398000	0,439000	2,708000	0,280000	1,398000	0,000000	Test
176	0,00000	0,127000	0,684000	3,046000	-1,00000	0,000000	0,000000	0,280000	8,69500	0,000000	2,196000	0,478000	0,419000	2,714000	0,303000	1,525000	0,409000	Train
177	0,00000	0,064000	0,684000	2,585000	-1,00000	0,000000	0,000000	0,281000	2,39200	0,000000	0,012000	0,000000	0,500000	1,828000	0,267000	1,091000	0,137000	Train
178	2,00000	0,067000	0,643000	2,910000	-0,41400	0,000000	0,347000	0,510000	2,27300	0,018000	0,658000	0,106000	0,326000	2,716000	0,261000	1,385000	0,576000	Train
179	3,00000	0,071000	0,573000	3,179000	0,00000	0,063000	0,314000	0,847000	3,23800	0,195000	0,646000	0,434000	0,244000	2,735000	0,333000	1,711000	0,935000	Test
180	4,00000	0,078000	0,507000	3,407000	0,00000	0,090000	0,279000	0,760000	3,42100	0,050000	0,898000	0,447000	0,175000	2,735000	0,357000	1,692000	0,876000	Train
181	6,00000	0,070000	0,558000	3,407000	0,00000	0,125000	0,397000	0,847000	9,45500	0,134000	2,298000	0,882000	0,212000	2,739000	0,308000	1,771000	1,197000	Train
182	5,00000	0,079000	0,451000	3,605000	0,38200	0,090000	0,252000	0,764000	3,42100	0,381000	1,145000	0,688000	0,133000	2,739000	0,362000	1,735000	0,912000	Train
183	5,00000	0,066000	0,633000	3,179000	-0,41400	0,000000	0,457000	0,510000	2,41800	0,037000	1,696000	0,176000	0,413000	2,718000	0,249000	1,494000	1,439000	Train
184	5,00000	0,073000	0,612000	3,179000	-0,26300	0,000000	0,462000	0,696000	2,15300	0,133000	1,303000	0,176000	0,322000	2,716000	0,248000	1,507000	0,707000	Train
185	5,00000	0,064000	0,597000	3,179000	0,00000	0,000000	0,526000	0,502000	2,15400	0,018000	1,301000	0,176000	0,222000	2,712000	0,274000	1,449000	0,665000	Train
186	6,00000	0,087000	0,407000	3,780000	0,78500	0,090000	0,230000	0,787000	3,42500	0,383000	1,397000	0,660000	0,102000	2,738000	0,366000	1,708000	0,938000	Train
187	7,00000	0,093000	0,373000	3,936000	1,20100	0,090000	0,212000	0,963000	3,42100	0,382000	1,644000	0,813000	0,080000	2,740000	0,368000	1,716000	0,908000	Train
188	8,00000	0,099000	0,345000	4,077000	1,54000	0,090000	0,196000	0,819000	3,42500	0,382000	1,893000	0,775000	0,064000	2,739000	0,375000	1,709000	0,954000	Train
189	9,00000	0,103000	0,323000	4,206000	1,80100	0,090000	0,172000	0,897000	3,42100	0,382000	2,143000	0,876000	0,052000	2,739000	0,380000	1,726000	0,962000	Train
190	10,00000	0,107000	0,305000	4,324000	2,00000	0,090000	0,156000	0,858000	3,42500	0,383000	2,393000	0,839000	0,043000	2,738000	0,381000	1,712000	1,002000	Test
191	11,00000	0,111000	0,290000	4,434000	2,00000	0,090000	0,144000	0,991000	3,42100	0,383000	2,642000	0,909000	0,036000	2,738000	0,381000	1,712000	1,008000	Train
192	12,00000	0,114000	0,277000	4,535000	2,00000	0,090000	0,134000	0,860000	3,42100	0,383000	2,894000	0,875000	0,031000	2,737000	0,381000	1,703000	1,046000	Train
193	13,00000	0,117000	0,267000	4,631000	2,00000	0,090000	0,125000	0,948000	3,42100	0,384000	3,141000	0,894000	0,027000	2,737000	0,381000	1,706000	1,055000	Train

ID	BAC	X4Av	Vindex	CIC0	EEig06r	GGI5	JGT	PJI3	RDF035u	RDF055m	RDF030v	L3u	P2u	H0e	R2m	R1e	R5e	Partition
194	14,00000	0,120000	0,257000	4,720000	2,12100	0,090000	0,118000	0,894000	3,42500	0,385000	3,391000	0,873000	0,024000	2,736000	0,381000	1,697000	1,088000	Test
195	6,00000	0,060000	0,558000	3,250000	0,00000	0,125000	0,397000	0,861000	7,01700	0,407000	1,474000	0,627000	0,230000	2,744000	0,296000	1,725000	1,074000	Train
196	6,00000	0,072000	0,558000	3,407000	0,20500	0,063000	0,438000	0,781000	3,12300	0,391000	1,291000	0,450000	0,266000	2,732000	0,308000	1,758000	0,959000	Train
197	6,00000	0,077000	0,582000	3,407000	0,14900	0,063000	0,433000	0,846000	6,59000	0,207000	1,503000	0,459000	0,357000	2,734000	0,312000	1,754000	1,436000	Test
198	6,00000	0,069000	0,540000	3,407000	0,00000	0,118000	0,475000	0,789000	3,11800	0,195000	1,286000	0,469000	0,172000	2,728000	0,329000	1,703000	0,984000	Train
199	29,00000	0,184000	0,810000	3,964000	0,00000	0,000000	0,510000	0,569000	14,27300	0,377000	5,748000	1,015000	0,296000	2,724000	0,215000	1,824000	1,540000	Test
200	30,00000	0,170000	0,775000	4,104000	0,18600	0,111000	0,488000	0,997000	19,17500	0,878000	7,166000	0,921000	0,187000	2,722000	0,233000	1,792000	1,935000	Train
201	6,00000	0,147000	0,608000	3,837000	0,00000	0,063000	0,536000	0,843000	22,29100	0,166000	5,320000	1,046000	0,234000	2,722000	0,329000	1,760000	2,017000	Train
202	6,00000	0,149000	0,558000	3,837000	0,20500	0,063000	0,438000	0,730000	10,30400	0,250000	4,138000	0,618000	0,244000	2,715000	0,278000	1,652000	1,469000	Train
203	6,00000	0,154000	0,582000	3,837000	0,14900	0,063000	0,433000	0,849000	11,38300	0,266000	4,712000	0,631000	0,359000	2,716000	0,256000	1,667000	1,795000	Train
204	11,00000	0,142000	0,578000	3,989000	0,43900	0,063000	0,478000	0,730000	21,48100	0,277000	6,657000	0,896000	0,254000	2,723000	0,306000	1,786000	2,342000	Train
205	6,00000	0,070000	0,558000	3,407000	0,00000	0,125000	0,397000	0,847000	9,45500	0,134000	2,298000	0,882000	0,212000	2,739000	0,308000	1,771000	1,197000	Train
206	6,00000	0,072000	0,558000	3,407000	0,20500	0,063000	0,438000	0,781000	3,12600	0,405000	1,292000	0,451000	0,266000	2,732000	0,308000	1,758000	0,959000	Train
207	6,00000	0,069000	0,540000	3,407000	0,00000	0,118000	0,475000	0,789000	3,12000	0,195000	1,290000	0,469000	0,172000	2,728000	0,329000	1,703000	0,984000	Train
208	6,00000	0,077000	0,582000	3,407000	0,14900	0,063000	0,433000	0,847000	4,96800	0,208000	1,647000	0,459000	0,357000	2,734000	0,311000	1,761000	1,467000	Train
209	10,00000	0,069000	0,622000	3,407000	-0,19900	0,000000	0,498000	0,697000	2,64700	0,147000	2,722000	0,226000	0,412000	2,718000	0,234000	1,546000	1,886000	Train
210	10,00000	0,069000	0,597000	3,407000	0,00000	0,000000	0,548000	0,502000	2,33800	0,132000	2,342000	0,226000	0,302000	2,714000	0,249000	1,553000	1,294000	Train
211	10,00000	0,080000	0,595000	3,407000	0,26800	0,000000	0,549000	0,453000	2,03700	0,339000	1,947000	0,226000	0,485000	2,714000	0,233000	1,602000	0,705000	Train
212	11,00000	0,077000	0,570000	3,605000	0,15600	0,125000	0,476000	0,847000	9,35000	0,178000	2,972000	0,832000	0,329000	2,739000	0,277000	1,833000	1,481000	Test
213	11,00000	0,069000	0,544000	3,605000	0,22400	0,125000	0,490000	0,748000	9,33400	0,123000	2,949000	0,838000	0,207000	2,735000	0,283000	1,808000	1,137000	Test
214	0,00000	0,076000	0,457000	3,465000	0,38200	0,000000	0,223000	0,844000	7,18000	0,027000	2,300000	0,353000	0,319000	2,730000	0,362000	1,616000	1,051000	Train
215	11,00000	0,067000	0,568000	3,605000	0,00000	0,188000	0,556000	0,847000	12,79800	0,126000	2,393000	0,838000	0,265000	2,739000	0,333000	1,797000	1,356000	Test
216	7,00000	0,078000	0,486000	3,605000	0,00000	0,118000	0,394000	0,760000	7,61000	0,212000	2,007000	0,626000	0,205000	2,739000	0,403000	1,767000	0,867000	Train
217	0,00000	0,059000	0,328000	3,645000	2,00000	0,014000	0,214000	0,631000	6,02500	4,436000	0,026000	0,995000	0,207000	2,767000	0,548000	1,901000	0,886000	Train
218	7,00000	0,073000	0,520000	3,605000	0,24900	0,153000	0,388000	0,760000	10,84800	0,149000	3,611000	0,955000	0,168000	2,740000	0,317000	1,788000	1,513000	Train
219	9,00000	0,072000	0,498000	3,605000	0,41100	0,181000	0,441000	0,630000	4,08800	0,372000	1,274000	0,797000	0,137000	2,734000	0,341000	1,796000	1,237000	Train
220	0,00000	0,053000	0,373000	3,465000	2,00000	0,139000	0,202000	0,722000	4,60200	0,269000	0,529000	0,000000	0,235000	2,766000	0,504000	1,822000	0,952000	Train
221	40,00000	0,000000	1,201000	3,810000	0,00000	0,000000	0,810000	0,608000	25,35400	0,018000	4,397000	1,546000	0,298000	2,719000	0,304000	1,774000	2,698000	Train

Tableau E.5. Valeurs des descripteurs pertinentes utilisées pour l'élaboration du modèle de RQSP-RNA pour Vc ainsi leurs ordres de partition dans les bases d'apprentissage et de test.

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
1	0,000000	0,000000	0,00000	0,00000	0,000000	6,9710	0,00000	0,10100	0,023000	0,246000	0,554000	0,298000	Test
2	0,707000	0,330000	0,00000	0,00000	0,250000	9,8580	0,00700	0,28500	0,021000	0,239000	0,893000	0,294000	Train
3	0,500000	0,310000	0,00000	0,00000	0,167000	12,8150	0,43000	0,34000	0,022000	0,294000	0,959000	0,250000	Test
4	0,577000	0,296000	0,00000	6,23600	0,500000	12,6720	0,02200	0,17400	0,012000	0,165000	1,191000	0,296000	Train
5	0,451000	0,290000	0,00000	0,00000	0,125000	15,7880	0,61500	0,09500	0,021000	0,218000	1,042000	0,202000	Train
6	0,451000	0,319000	0,00000	6,94000	0,375000	15,3980	0,66700	0,13800	0,007000	0,209000	1,288000	0,241000	Train
7	0,427000	0,272000	0,00000	0,00000	0,100000	18,8220	0,80000	0,05200	0,016000	0,210000	1,065000	0,180000	Test
8	0,412000	0,258000	0,00000	0,00000	0,083000	21,8920	0,98600	0,36800	0,005000	0,241000	1,103000	0,161000	Train
9	0,402000	0,246000	0,00000	0,00000	0,071000	25,0080	1,17100	0,67900	-0,003000	0,217000	1,116000	0,149000	Train
10	0,395000	0,235000	0,00000	0,00000	0,063000	28,1580	1,35600	0,81200	-0,003000	0,273000	1,135000	0,138000	Train
11	0,390000	0,226000	-0,90200	0,00000	0,056000	31,3450	1,54100	0,79200	-0,004000	0,275000	1,147000	0,130000	Train
12	0,386000	0,219000	-0,68300	0,00000	0,050000	34,5590	1,72700	0,74600	-0,010000	0,271000	1,159000	0,122000	Train
13	0,383000	0,212000	-0,41400	0,00000	0,045000	37,8030	1,91200	0,87400	-0,020000	0,298000	1,164000	0,116000	Train
14	0,380000	0,206000	-0,13600	0,00000	0,042000	41,0710	2,09700	1,14800	-0,023000	0,268000	1,175000	0,110000	Train
15	0,378000	0,201000	0,13200	0,00000	0,038000	44,3620	2,28200	1,33300	-0,023000	0,284000	1,178000	0,105000	Train
16	0,376000	0,196000	0,38200	0,00000	0,036000	47,6730	2,46800	1,35900	-0,030000	0,288000	1,185000	0,100000	Train
17	0,374000	0,192000	0,61000	0,00000	0,033000	51,0040	2,65300	1,35500	-0,038000	0,281000	1,189000	0,096000	Test
18	0,373000	0,188000	0,81500	0,00000	0,031000	54,3520	2,83800	1,45100	-0,042000	0,313000	1,193000	0,092000	Train
19	0,372000	0,185000	1,00000	0,00000	0,029000	57,7160	3,02300	1,64900	-0,043000	0,305000	1,196000	0,089000	Train
20	0,371000	0,181000	1,16500	0,00000	0,028000	61,0940	3,20900	1,84700	-0,048000	0,320000	1,201000	0,085000	Train
21	0,370000	0,178000	1,31300	0,00000	0,026000	64,4860	3,39400	1,92900	-0,056000	0,332000	1,202000	0,082000	Train
22	0,369000	0,176000	1,44500	0,00000	0,025000	67,8900	3,57900	1,94100	-0,061000	0,320000	1,206000	0,079000	Train
23	0,368000	0,173000	1,56300	0,00000	0,024000	71,3070	3,76400	2,01400	-0,063000	0,340000	1,208000	0,077000	Test
24	0,368000	0,171000	1,67000	0,00000	0,023000	74,7350	3,94900	2,19000	-0,067000	0,333000	1,210000	0,074000	Train
25	0,367000	0,169000	1,76500	0,00000	0,022000	78,1730	4,13500	2,37400	-0,075000	0,339000	1,212000	0,072000	Test
26	0,366000	0,166000	1,85200	0,00000	0,021000	81,6200	4,32000	2,47900	-0,080000	0,355000	1,214000	0,070000	Train
27	0,366000	0,164000	1,92900	0,00000	0,020000	85,0760	4,50500	2,51700	-0,082000	0,349000	1,215000	0,068000	Train
28	0,365000	0,161000	2,06400	0,00000	0,019000	92,0150	4,87600	2,72600	-0,093000	0,371000	1,218000	0,064000	Train
29	0,363000	0,155000	2,26900	0,00000	0,016000	105,9800	5,61700	3,14700	-0,112000	0,394000	1,223000	0,057000	Test
30	0,362000	0,150000	2,41400	0,00000	0,014000	120,0430	6,35800	3,64700	-0,129000	0,412000	1,226000	0,052000	Train
31	0,500000	0,272000	0,00000	9,88700	0,750000	15,4520	0,04300	-0,23000	-0,004000	0,024000	1,510000	0,317000	Test
32	0,437000	0,311000	0,00000	7,07100	0,300000	18,3320	3,33900	-0,03300	0,005000	0,107000	1,344000	0,193000	Train
33	0,384000	0,311000	0,00000	7,45900	0,300000	18,2060	1,22800	-0,19700	0,004000	0,131000	1,364000	0,196000	Train
34	0,416000	0,296000	0,00000	10,14100	0,600000	17,9920	0,46600	-0,42300	-0,003000	0,070000	1,589000	0,260000	Train
35	0,415000	0,272000	0,00000	8,45400	0,400000	18,0910	0,87400	-0,25800	0,007000	0,199000	1,356000	0,220000	Train
36	0,423000	0,298000	0,00000	7,08600	0,250000	21,3010	3,42300	0,01900	0,005000	0,128000	1,362000	0,174000	Train
37	0,384000	0,298000	0,00000	7,54800	0,250000	21,2190	3,90200	-0,16900	-0,001000	0,095000	1,370000	0,166000	Train
38	0,414000	0,293000	0,00000	10,16800	0,500000	20,8850	4,34200	-0,55100	-0,025000	-0,084000	1,628000	0,207000	Train
39	0,376000	0,293000	0,00000	8,67900	0,333000	20,8600	1,90300	-0,57600	-0,006000	0,055000	1,443000	0,176000	Train
40	0,432000	0,293000	0,00000	7,93200	0,417000	20,7930	5,87200	-0,18500	-0,011000	-0,052000	1,633000	0,177000	Train
41	0,359000	0,293000	0,00000	10,36400	0,500000	20,6530	0,65100	-0,91500	-0,002000	-0,015000	1,592000	0,221000	Train

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
42	0,348000	0,298000	0,00000	7,87700	0,250000	20,7630	5,66300	0,22000	-0,015000	0,139000	1,530000	0,145000	Train
43	0,391000	0,277000	0,00000	10,63900	0,583000	20,5780	0,89600	-0,91100	-0,009000	0,049000	1,665000	0,230000	Train
44	0,413000	0,285000	0,00000	7,08600	0,214000	24,3750	3,60800	0,30200	-0,006000	0,144000	1,379000	0,159000	Train
45	0,379000	0,285000	0,00000	7,55700	0,214000	24,2430	3,99100	0,21900	-0,003000	0,157000	1,401000	0,151000	Test
46	0,383000	0,285000	0,00000	7,63000	0,214000	24,2670	6,57900	0,20500	-0,009000	0,101000	1,402000	0,138000	Train
47	0,353000	0,273000	0,00000	7,94900	0,188000	26,6360	9,06600	0,28500	0,009000	0,187000	1,609000	0,118000	Train
48	0,356000	0,273000	0,00000	8,00500	0,188000	26,5760	11,18400	-0,03500	0,022000	0,057000	1,614000	0,113000	Train
49	0,407000	0,285000	0,00000	10,17000	0,429000	23,7860	5,98800	-0,38800	-0,010000	-0,086000	1,643000	0,188000	Train
50	0,376000	0,291000	0,00000	8,71300	0,286000	23,7930	5,55100	-0,47600	-0,014000	0,010000	1,473000	0,152000	Train
51	0,393000	0,291000	0,00000	8,17200	0,357000	23,6390	7,24700	-0,26800	-0,008000	-0,063000	1,638000	0,158000	Train
52	0,421000	0,291000	0,00000	7,79400	0,357000	23,8120	6,67000	0,04200	0,018000	0,000000	1,446000	0,164000	Train
53	0,363000	0,285000	0,00000	10,38700	0,429000	23,6380	4,52900	-0,84200	-0,028000	-0,101000	1,609000	0,181000	Test
54	0,346000	0,291000	0,00000	8,86900	0,286000	23,7220	2,98100	-0,70600	-0,023000	-0,023000	1,457000	0,151000	Train
55	0,402000	0,276000	0,00000	10,17000	0,375000	26,8530	6,17300	-0,07900	-0,021000	-0,046000	1,635000	0,175000	Train
56	0,374000	0,285000	0,00000	8,71600	0,250000	26,8060	5,64000	-0,05100	-0,012000	0,095000	1,507000	0,140000	Train
57	0,391000	0,285000	0,00000	8,21700	0,313000	26,6960	9,97700	0,11300	0,014000	-0,021000	1,659000	0,134000	Train
58	0,413000	0,285000	0,00000	7,77900	0,313000	26,8280	6,37600	0,25800	-0,016000	0,064000	1,634000	0,154000	Train
59	0,362000	0,276000	0,00000	10,38900	0,375000	26,6100	6,16700	-0,31900	-0,020000	-0,064000	1,610000	0,167000	Train
60	0,350000	0,285000	0,00000	8,89600	0,250000	26,7110	6,63600	-0,24700	-0,034000	-0,027000	1,492000	0,131000	Train
61	0,363000	0,285000	0,00000	8,36600	0,313000	26,3240	8,80400	-0,15000	0,020000	0,075000	1,731000	0,139000	Train
62	0,397000	0,267000	-0,88500	10,17000	0,333000	29,9470	6,35900	0,26100	-0,024000	-0,168000	1,634000	0,165000	Test
63	0,372000	0,277000	-0,87600	8,71600	0,222000	29,9100	5,82500	0,27800	-0,017000	0,023000	1,514000	0,132000	Test
64	0,388000	0,277000	-0,87100	8,22200	0,278000	29,7600	10,01500	0,56900	-0,010000	-0,066000	1,653000	0,127000	Train
65	0,387000	0,277000	-0,87400	8,10400	0,278000	29,8440	9,89300	0,61700	0,005000	-0,049000	1,502000	0,128000	Test
66	0,361000	0,267000	-0,86800	10,38900	0,333000	29,7380	6,35700	-0,01600	-0,029000	-0,079000	1,605000	0,156000	Train
67	0,360000	0,277000	-0,87200	8,26000	0,278000	29,7700	7,78500	0,39800	0,000000	0,030000	1,498000	0,132000	Train
68	0,353000	0,285000	0,00000	7,94300	0,214000	23,6750	7,33900	0,05500	0,018000	0,083000	1,567000	0,127000	Train
69	0,368000	0,294000	0,00000	10,69400	0,500000	23,3530	4,39200	-1,13900	-0,027000	-0,106000	1,703000	0,185000	Train
70	0,416000	0,294000	0,00000	10,26700	0,571000	23,1990	10,40100	-0,28800	-0,007000	0,057000	1,690000	0,194000	Train
71	0,350000	0,294000	0,00000	10,80700	0,500000	23,2390	2,92400	-1,34600	-0,020000	-0,075000	1,685000	0,191000	Train
72	0,372000	0,255000	0,00000	9,30900	0,357000	23,2410	5,17600	-0,74000	-0,025000	-0,049000	1,707000	0,165000	Train
73	0,353000	0,291000	0,00000	8,89400	0,286000	22,8920	10,84500	0,44000	-0,017000	0,235000	1,760000	0,155000	Train
74	0,319000	0,285000	0,00000	10,56500	0,429000	23,0080	1,09600	-1,07200	-0,007000	-0,028000	1,716000	0,183000	Train
75	0,315000	0,332000	0,00000	10,93500	0,375000	22,4750	13,08300	1,59800	0,125000	-0,485000	1,719000	0,250000	Test
76	0,334000	0,332000	0,00000	10,60300	0,438000	22,7540	9,43700	1,42600	0,129000	-0,538000	1,673000	0,250000	Train
77	0,321000	0,314000	0,00000	9,48500	0,313000	22,5160	8,86300	1,53200	0,127000	-0,579000	1,587000	0,243000	Train
78	0,308000	0,314000	0,00000	9,71500	0,250000	22,6010	10,03800	1,72500	0,150000	-0,394000	1,608000	0,228000	Train
79	0,369000	0,293000	0,00000	10,70000	0,438000	26,2460	8,42500	-0,91000	-0,025000	-0,136000	1,734000	0,162000	Train
80	0,389000	0,293000	0,00000	10,29400	0,500000	25,9960	10,61100	-0,16200	0,008000	-0,074000	1,714000	0,174000	Test
81	0,408000	0,293000	0,00000	10,21800	0,500000	26,2330	7,48100	-0,33800	0,004000	-0,191000	1,725000	0,173000	Train
82	0,385000	0,273000	0,00000	8,95900	0,375000	26,1440	9,23000	-0,49600	-0,009000	-0,112000	1,745000	0,146000	Train
83	0,336000	0,316000	-1,00000	10,54200	0,389000	25,0500	6,69300	0,10800	-0,055000	-0,391000	1,745000	0,219000	Train
84	0,291000	0,276000	0,00000	10,74800	0,375000	25,4970	1,30300	-1,52700	-0,013000	-0,171000	1,740000	0,156000	Train
85	0,345000	0,265000	0,00000	11,72900	0,563000	25,8530	5,21000	-2,06900	-0,043000	-0,201000	1,782000	0,184000	Train

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
86	0,367000	0,265000	0,00000	10,84500	0,500000	25,6670	12,31400	-0,20500	-0,083000	0,025000	1,730000	0,176000	Train
87	0,408000	0,265000	0,00000	10,94800	0,688000	25,3190	12,16600	-0,00500	-0,072000	-0,402000	1,753000	0,199000	Train
88	0,344000	0,265000	0,00000	11,14400	0,500000	25,7400	5,20900	-1,75700	-0,040000	-0,164000	1,790000	0,169000	Train
89	0,405000	0,273000	0,00000	7,08600	0,188000	27,4840	3,79800	0,56400	-0,014000	0,060000	1,383000	0,148000	Train
90	0,376000	0,273000	0,00000	7,55700	0,188000	27,3700	4,17200	0,49500	-0,008000	0,114000	1,396000	0,142000	Train
91	0,380000	0,273000	0,00000	7,63900	0,188000	27,3400	6,66000	0,59900	-0,007000	0,105000	1,409000	0,130000	Train
92	0,400000	0,263000	-0,89700	7,08600	0,167000	30,6560	3,97900	0,63300	-0,018000	0,161000	1,386000	0,140000	Train
93	0,374000	0,263000	-0,89000	7,55700	0,167000	30,5220	4,35700	0,60000	-0,016000	0,107000	1,407000	0,133000	Train
94	0,377000	0,263000	-0,87900	7,64700	0,167000	30,4500	6,74400	0,84100	-0,002000	0,184000	1,427000	0,121000	Train
95	0,374000	0,253000	-0,65800	7,63900	0,150000	33,6830	7,02200	0,66500	-0,017000	0,084000	1,418000	0,116000	Train
96	0,353000	0,263000	-0,87200	7,94900	0,167000	29,7160	8,17800	0,47900	0,011000	0,110000	1,615000	0,114000	Test
97	0,353000	0,287000	-0,83700	10,49200	0,444000	28,8390	11,75100	-0,17800	0,001000	0,026000	1,694000	0,156000	Train
98	0,349000	0,267000	-0,79800	11,73600	0,500000	28,7300	10,54000	-1,79000	-0,050000	-0,281000	1,826000	0,156000	Train
99	0,707000	0,667000	0,00000	3,05000	0,250000	8,0710	0,00000	0,07300	0,006000	0,289000	0,726000	0,236000	Test
100	0,500000	0,620000	0,00000	5,25400	0,167000	10,9750	2,53900	0,73800	0,039000	0,223000	1,078000	0,267000	Train
101	0,451000	0,584000	0,00000	5,79400	0,125000	13,9260	2,52000	0,68100	0,039000	0,319000	1,122000	0,237000	Train
102	0,427000	0,530000	0,00000	5,92400	0,100000	16,9070	2,61100	0,46300	0,032000	0,216000	1,178000	0,195000	Train
103	0,412000	0,484000	0,00000	5,94000	0,083000	19,9430	2,80300	0,30300	0,029000	0,221000	1,191000	0,175000	Train
104	0,402000	0,447000	0,00000	5,94000	0,071000	23,0180	2,98100	0,53000	0,020000	0,257000	1,215000	0,157000	Train
105	0,395000	0,416000	0,00000	5,94000	0,063000	26,1370	3,16700	0,94800	0,007000	0,212000	1,221000	0,146000	Train
106	0,390000	0,390000	-0,89200	5,94000	0,056000	29,2920	3,35200	1,17400	0,007000	0,288000	1,233000	0,136000	Train
107	0,386000	0,369000	-0,65200	5,94000	0,050000	32,4820	3,53700	1,15000	0,008000	0,282000	1,238000	0,128000	Train
108	0,383000	0,350000	-0,36500	5,94000	0,045000	35,7000	3,72200	1,05700	0,001000	0,276000	1,245000	0,121000	Train
109	0,380000	0,334000	-0,07200	5,94000	0,042000	38,9470	3,90800	1,11300	-0,008000	0,315000	1,246000	0,115000	Test
110	0,378000	0,320000	0,20800	5,94000	0,038000	42,2190	4,10000	1,37700	-0,012000	0,270000	1,251000	0,109000	Test
111	0,376000	0,307000	0,46500	5,94000	0,036000	45,5120	4,27800	1,64700	-0,013000	0,295000	1,251000	0,104000	Train
112	0,374000	0,296000	0,69700	5,94000	0,033000	48,8260	4,46300	1,72100	-0,017000	0,297000	1,254000	0,099000	Train
113	0,373000	0,286000	0,90500	5,94000	0,031000	52,1600	4,65500	1,67300	-0,026000	0,283000	1,255000	0,095000	Train
114	0,372000	0,278000	1,09000	5,94000	0,029000	55,5100	4,83400	1,71400	-0,030000	0,327000	1,257000	0,091000	Train
115	0,371000	0,269000	1,25400	5,94000	0,028000	58,8760	5,01900	1,90200	-0,032000	0,309000	1,257000	0,088000	Train
116	0,370000	0,262000	1,40000	5,94000	0,026000	62,2570	5,20400	2,13700	-0,036000	0,329000	1,259000	0,085000	Train
117	0,500000	0,396000	0,00000	0,00000	0,167000	9,2580	3,16400	0,82300	0,023000	0,261000	0,776000	0,110000	Train
118	0,500000	0,500000	0,00000	5,25400	0,167000	10,6130	0,00000	-0,26200	-0,008000	0,392000	0,998000	0,198000	Train
119	0,500000	0,500000	0,00000	5,25400	0,167000	11,0900	1,83000	0,04300	0,026000	0,303000	0,839000	0,157000	Train
120	0,577000	0,546000	0,00000	8,19200	0,500000	10,9830	0,00300	0,00500	0,017000	0,205000	1,047000	0,225000	Test
121	0,451000	0,601000	0,00000	7,83600	0,375000	11,9290	3,20700	0,66700	0,030000	0,174000	1,077000	0,185000	Test
122	0,415000	0,504000	0,00000	9,76700	0,400000	16,3210	3,63500	0,83700	0,070000	-0,014000	1,631000	0,221000	Train
123	0,415000	0,294000	0,00000	9,92000	0,400000	16,3820	6,58100	0,22200	0,025000	0,244000	1,269000	0,130000	Train
124	0,427000	0,504000	0,00000	6,46600	0,100000	16,2330	4,67500	0,52900	0,037000	0,375000	1,358000	0,202000	Test
125	0,427000	0,504000	0,00000	6,46600	0,100000	16,9630	8,08000	0,65000	0,051000	0,351000	1,153000	0,184000	Test
126	0,437000	0,530000	0,00000	7,59100	0,300000	16,4310	2,76000	0,68600	0,024000	0,262000	1,379000	0,221000	Train
127	0,416000	0,454000	0,00000	10,54300	0,571000	21,3820	5,26400	0,25800	0,035000	-0,063000	1,703000	0,220000	Train
128	0,416000	0,406000	0,00000	10,95000	0,571000	21,3510	6,57600	-0,10400	0,123000	-0,041000	1,717000	0,183000	Test
129	0,451000	0,556000	0,00000	8,73700	0,375000	13,6820	2,29800	0,69600	0,052000	0,042000	1,407000	0,240000	Train

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
130	0,451000	0,432000	0,00000	8,69000	0,375000	13,6660	2,96600	-0,03700	0,056000	0,315000	1,231000	0,170000	Train
131	0,451000	0,549000	0,00000	8,09400	0,375000	13,7870	3,28100	0,91800	0,050000	0,087000	1,339000	0,243000	Train
132	0,415000	0,527000	0,00000	9,26900	0,400000	14,7690	5,70700	0,55800	0,044000	0,179000	1,186000	0,146000	Train
133	0,416000	0,493000	0,00000	10,73600	0,600000	16,3820	3,41400	0,49900	0,072000	-0,036000	1,599000	0,273000	Train
134	0,437000	0,466000	0,00000	8,81100	0,300000	16,4290	7,79700	0,75400	0,059000	0,224000	1,413000	0,165000	Test
135	0,427000	0,527000	0,00000	6,69500	0,100000	15,0110	2,52800	1,30400	0,063000	0,369000	1,264000	0,214000	Train
136	0,354000	0,290000	0,00000	6,23600	0,000000	8,6110	0,00000	0,93300	0,065000	-0,022000	0,961000	0,350000	Train
137	0,354000	0,256000	0,00000	0,00000	0,000000	11,2300	0,00000	0,73200	0,146000	0,223000	1,052000	0,343000	Train
138	0,354000	0,235000	0,00000	5,88900	0,000000	14,4490	0,00000	0,89200	0,178000	-0,052000	1,197000	0,353000	Train
139	0,354000	0,220000	0,00000	0,00000	0,000000	16,8840	0,04300	1,52800	0,017000	-0,078000	1,362000	0,317000	Train
140	0,354000	0,209000	0,00000	4,84400	0,000000	19,4870	3,06800	1,64400	-0,046000	-0,144000	1,513000	0,298000	Test
141	0,354000	0,200000	0,00000	0,00000	0,000000	21,5780	4,38700	3,19500	0,140000	-0,550000	1,668000	0,274000	Train
142	0,343000	0,296000	0,00000	7,73000	0,143000	19,7600	2,55100	1,23900	-0,002000	-0,064000	1,412000	0,269000	Train
143	0,324000	0,277000	0,00000	8,08000	0,125000	22,4470	5,48300	1,15500	0,006000	-0,166000	1,486000	0,220000	Train
144	0,329000	0,262000	-1,00000	8,13400	0,111000	25,4230	8,15800	1,15700	-0,004000	-0,190000	1,498000	0,193000	Train
145	0,331000	0,294000	-1,00000	8,98300	0,167000	25,1060	6,65500	0,61900	-0,006000	-0,376000	1,519000	0,194000	Test
146	0,331000	0,250000	-0,84600	8,13900	0,100000	28,3860	8,24400	1,65900	-0,012000	-0,166000	1,529000	0,176000	Train
147	0,315000	0,280000	-0,79600	9,13000	0,150000	27,9460	7,99700	0,45100	-0,017000	-0,389000	1,556000	0,166000	Test
148	0,341000	0,320000	0,00000	7,94300	0,167000	17,2880	4,62600	0,77600	0,167000	-0,007000	1,306000	0,303000	Train
149	0,320000	0,296000	0,00000	8,23900	0,143000	20,0110	7,30000	0,99200	0,174000	0,037000	1,375000	0,245000	Train
150	0,327000	0,277000	0,00000	8,28600	0,125000	22,9580	9,71600	0,97800	0,189000	-0,039000	1,415000	0,211000	Train
151	0,329000	0,309000	0,00000	9,05700	0,188000	22,6710	10,24700	0,94300	0,148000	0,054000	1,435000	0,211000	Train
152	0,329000	0,262000	-0,92800	8,29000	0,111000	25,9220	9,90600	0,90200	0,187000	-0,037000	1,437000	0,190000	Train
153	0,344000	0,294000	-0,92300	8,63900	0,222000	25,3610	15,27100	0,99800	0,198000	-0,113000	1,684000	0,191000	Test
154	0,331000	0,250000	-0,61800	8,29000	0,100000	29,0040	10,10000	1,09500	0,173000	-0,036000	1,452000	0,177000	Train
155	0,333000	0,239000	-0,61800	8,29000	0,091000	32,1130	10,28800	1,43400	0,165000	-0,059000	1,459000	0,165000	Test
156	0,335000	0,230000	-0,41400	8,29000	0,083000	35,2920	10,46600	1,56900	0,169000	-0,020000	1,465000	0,156000	Train
157	0,336000	0,223000	-0,11200	8,29000	0,077000	38,4950	10,65400	1,60700	0,166000	-0,007000	1,470000	0,147000	Test
158	0,337000	0,216000	0,21900	8,29000	0,071000	41,7460	10,83300	1,56900	0,156000	0,001000	1,471000	0,140000	Train
159	0,338000	0,210000	0,53700	8,29000	0,067000	45,0170	11,02600	1,59200	0,147000	0,018000	1,472000	0,133000	Train
160	0,340000	0,200000	1,06900	8,29000	0,059000	51,6420	11,39600	2,09600	0,142000	0,011000	1,471000	0,121000	Train
161	0,341000	0,195000	1,28200	8,29000	0,056000	54,9880	11,56700	2,17800	0,142000	0,013000	1,468000	0,117000	Train
162	0,341000	0,191000	1,46300	8,29000	0,053000	58,3490	11,76000	2,18100	0,131000	0,012000	1,468000	0,111000	Train
163	0,342000	0,188000	1,61800	8,29000	0,050000	61,7290	11,94400	2,21100	0,120000	0,025000	1,464000	0,107000	Train
164	0,328000	0,294000	-0,77000	9,24800	0,250000	27,3350	7,24400	0,40700	-0,023000	-0,397000	1,729000	0,190000	Train
165	0,337000	0,314000	0,00000	10,47000	0,357000	19,8920	6,79600	1,67200	0,151000	-0,543000	1,632000	0,295000	Test
166	0,336000	0,327000	0,00000	8,71300	0,286000	20,1220	7,23400	1,16900	0,153000	-0,174000	1,398000	0,263000	Test
167	0,336000	0,327000	0,00000	8,71300	0,286000	20,1030	8,60600	1,16600	0,164000	-0,075000	1,454000	0,264000	Test
168	0,321000	0,327000	0,00000	9,09800	0,214000	20,0490	11,76500	0,84200	0,154000	0,019000	1,442000	0,251000	Train
169	0,339000	0,295000	0,00000	10,42900	0,313000	22,1470	2,58400	0,63600	-0,025000	-0,234000	1,688000	0,267000	Train
170	0,324000	0,309000	0,00000	8,98300	0,188000	22,0120	2,57300	0,80800	-0,017000	-0,173000	1,665000	0,251000	Train
171	0,324000	0,309000	0,00000	8,98300	0,188000	22,4620	5,05700	0,82700	-0,020000	-0,056000	1,488000	0,227000	Test
172	0,338000	0,309000	0,00000	8,53500	0,250000	22,6680	4,81800	0,65100	-0,021000	-0,182000	1,473000	0,222000	Train
173	0,338000	0,309000	0,00000	8,53500	0,250000	22,1240	4,41900	0,98900	-0,028000	-0,207000	1,665000	0,248000	Train

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
174	0,337000	0,309000	0,00000	8,45400	0,250000	22,0330	4,12200	0,93500	-0,001000	-0,157000	1,677000	0,256000	Train
175	0,337000	0,309000	0,00000	8,45400	0,250000	22,6650	5,05600	0,95200	-0,021000	-0,034000	1,408000	0,243000	Test
176	0,354000	0,537000	0,00000	7,51300	0,000000	12,6830	0,07600	0,97000	0,031000	-0,009000	1,102000	0,365000	Train
177	0,354000	0,489000	0,00000	7,27300	0,000000	15,2800	0,12300	1,18800	-0,016000	0,068000	1,308000	0,329000	Train
178	0,354000	0,256000	0,00000	0,00000	0,000000	11,8470	0,02400	2,03300	0,013000	-0,082000	1,044000	0,167000	Train
179	0,343000	0,490000	0,00000	8,40300	0,143000	14,7480	2,37400	1,87300	0,010000	-0,064000	1,162000	0,179000	Test
180	0,324000	0,509000	0,00000	8,85200	0,125000	17,3680	4,41100	2,21700	0,036000	-0,204000	1,654000	0,230000	Train
181	0,329000	0,513000	-0,67900	8,92200	0,111000	20,2910	7,17300	2,37300	0,028000	-0,086000	1,639000	0,225000	Train
182	0,331000	0,492000	-0,57500	9,78100	0,167000	20,0000	11,11600	2,09600	0,012000	-0,163000	1,807000	0,199000	Train
183	0,331000	0,493000	-0,50000	8,93200	0,100000	23,1780	7,27600	2,48500	0,033000	-0,039000	1,713000	0,197000	Train
184	0,324000	0,500000	0,00000	9,65300	0,188000	17,3920	4,85600	1,75600	0,012000	-0,062000	1,293000	0,138000	Train
185	0,338000	0,500000	0,00000	9,18600	0,250000	17,6940	4,37100	1,70800	0,009000	-0,074000	1,280000	0,132000	Train
186	0,337000	0,500000	0,00000	9,11700	0,250000	17,6800	4,72300	1,76400	0,001000	-0,049000	1,191000	0,166000	Train
187	0,333000	0,471000	-0,29500	8,93300	0,091000	26,2190	7,45800	2,22500	0,030000	-0,105000	1,681000	0,188000	Train
188	0,335000	0,449000	0,18400	8,93300	0,083000	29,2670	7,64700	2,16000	0,027000	-0,004000	1,711000	0,174000	Train
189	0,336000	0,428000	0,50000	8,93300	0,077000	32,4060	7,83500	2,40000	0,015000	-0,066000	1,686000	0,167000	Train
190	0,337000	0,409000	0,50000	8,93300	0,071000	35,5610	8,01400	2,58000	0,007000	-0,103000	1,701000	0,157000	Test
191	0,338000	0,393000	0,58800	8,93300	0,067000	38,7760	8,20200	2,79300	0,000000	-0,036000	1,675000	0,152000	Train
192	0,339000	0,377000	0,77700	8,93300	0,063000	42,0070	8,39100	2,91700	-0,003000	-0,094000	1,683000	0,144000	Train
193	0,340000	0,363000	0,96700	8,93300	0,059000	45,2790	8,57300	2,89600	-0,008000	-0,036000	1,661000	0,139000	Train
194	0,341000	0,351000	1,16900	8,93300	0,056000	48,5670	8,75800	2,97800	-0,013000	-0,011000	1,662000	0,132000	Test
195	0,341000	0,339000	1,36100	8,93300	0,053000	51,8870	8,94000	3,14900	-0,018000	-0,045000	1,644000	0,127000	Train
196	0,331000	0,520000	-0,53400	9,30700	0,167000	18,4940	3,96600	2,31000	0,060000	-0,246000	1,768000	0,199000	Train
197	0,322000	0,522000	-0,52000	9,42100	0,222000	20,3110	6,40600	1,93000	0,032000	-0,209000	1,707000	0,173000	Test
198	0,312000	0,522000	-0,52100	9,85700	0,167000	19,8290	6,74800	2,07400	0,012000	-0,211000	1,719000	0,177000	Train
199	0,321000	0,522000	-0,52000	9,36200	0,222000	20,2970	6,76000	2,17000	0,025000	-0,163000	1,616000	0,197000	Test
200	0,368000	0,285000	0,00000	8,41100	0,313000	25,9830	10,70900	-0,00700	-0,045000	-0,012000	1,826000	0,126000	Train
201	0,343000	0,277000	-0,84400	8,56600	0,278000	28,7010	11,41400	0,15300	-0,022000	0,035000	1,754000	0,130000	Train
202	0,310000	0,279000	-1,00000	10,62100	0,278000	24,7370	4,62500	0,34900	-0,023000	-0,362000	1,720000	0,223000	Train
203	0,322000	0,294000	-1,00000	8,71300	0,222000	25,4380	7,75300	0,76700	-0,015000	-0,216000	1,514000	0,191000	Train
204	0,312000	0,294000	-1,00000	9,15300	0,167000	25,1270	8,58300	0,75900	-0,018000	-0,216000	1,559000	0,181000	Train
205	0,303000	0,294000	-0,70500	9,70100	0,200000	27,6400	10,24800	0,30100	-0,036000	-0,340000	1,736000	0,175000	Train
206	0,331000	0,492000	-0,57500	9,78100	0,167000	20,0000	11,11600	2,09600	0,012000	-0,163000	1,807000	0,199000	Train
207	0,322000	0,522000	-0,52000	9,42100	0,222000	20,3100	6,41300	1,92800	0,032000	-0,207000	1,707000	0,173000	Train
208	0,321000	0,522000	-0,52000	9,36200	0,222000	20,2960	6,76000	2,16900	0,025000	-0,162000	1,616000	0,197000	Train
209	0,312000	0,522000	-0,52100	9,85700	0,167000	19,8330	7,68000	2,01900	0,012000	-0,226000	1,719000	0,177000	Train
210	0,312000	0,464000	-0,50000	10,25400	0,222000	20,1480	6,95600	1,61200	0,014000	-0,080000	1,350000	0,116000	Train
211	0,323000	0,464000	-0,50000	9,95700	0,278000	20,3960	6,84200	1,63000	0,002000	-0,072000	1,329000	0,114000	Train
212	0,335000	0,464000	-0,50000	9,67600	0,333000	20,6650	6,02100	1,53400	0,008000	-0,113000	1,383000	0,085000	Test
213	0,321000	0,498000	-0,50000	10,36200	0,200000	22,5280	12,13100	1,70600	-0,017000	-0,083000	1,902000	0,154000	Test
214	0,329000	0,498000	-0,50000	10,04900	0,250000	23,0350	12,92500	1,79000	0,016000	-0,236000	1,854000	0,160000	Train
215	0,292000	0,504000	-0,42000	10,12100	0,091000	21,7660	3,76100	1,93300	0,074000	-0,337000	1,524000	0,238000	Test
216	0,336000	0,469000	-0,50000	11,36600	0,300000	22,6370	11,93500	1,17200	0,012000	-0,288000	1,847000	0,203000	Train
217	0,344000	0,493000	-0,50000	9,20600	0,200000	22,5160	9,38200	2,53100	0,039000	-0,272000	1,744000	0,220000	Train

ID	X2A	BIC1	EEig09x	ESpm09d	JGI1	ADDD	RDF045u	Mor07u	Mor30v	Mor08p	H1e	HATS3p	Partition
218	0,308000	0,365000	0,50000	9,85100	0,107000	24,1540	14,34600	5,63600	0,043000	-0,441000	2,008000	0,230000	Train
219	0,315000	0,493000	-0,50000	9,90000	0,150000	22,7850	11,35900	1,97500	-0,019000	-0,119000	1,819000	0,175000	Train
220	0,309000	0,480000	-0,50000	9,55900	0,200000	22,8590	8,79600	2,59600	0,064000	-0,279000	1,770000	0,180000	Train
221	0,300000	0,278000	0,50000	9,34000	0,077000	22,0640	1,08800	2,34600	0,010000	-0,528000	1,863000	0,225000	Train
222	0,292000	0,302000	-0,11800	9,38400	0,091000	18,3550	4,51600	2,81500	0,153000	-0,337000	1,301000	0,132000	Train
223	0,375000	0,246000	0,00000	11,64600	0,643000	23,1940	1,33300	-1,69500	-0,022000	-0,038000	1,747000	0,225000	Train

Article scientifique, résumés et attestations de participation à d'autres travaux de recherche

Attestation de publication et support d'un article scientifique classe "A"

ELSEVIER

Journal Of Molecular Graphics And Modelling

Certificate of publication for the article titled:

"QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods"

Authored by:

ROUBEHIE FISSA Mohamed, LAHIOUEL Yasmina, KHAOUANE Latifa, & HANINI Salah

Published in:

Volume 87C, 2019, Pages 109-120

Journal of Molecular Graphics and Modelling 87 (2019) 109–120

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods

Mohamed Roubehie Fissa [a, *], Yasmina Lahiouel [a], Latifa Khaouane [b], Salah Hanini [b]

[a] Laboratory of Silicates, Polymers and Nanocomposites (LSPN), Université 8 Mai 1945 Guelma, BP 401, Guelma, 24000, Algeria
[b] Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Médéa, Médéa, 26000, Algeria

ARTICLE INFO

Article history:
Received 11 September 2018
Received in revised form
8 November 2018
Accepted 27 November 2018
Available online 29 November 2018

Keywords:
Pure hydrocarbons
QSPR-ANN
MLR
Normal boiling point
Relative liquid density
Descriptors

ABSTRACT

This work aimed to predict the normal boiling point temperature *(Tb)* and relative liquid density *(d20)* of petroleum fractions and pure hydrocarbons, through a multi-layer perceptron artificial neural network (MLP-ANN) based on the molecular descriptors. A set of 223 and 222 diverse data points for *Tb* and *d20* were respectively used to build two quantitative structure property relationships-artificial neural network (QSPR-ANN) models. For each model, the total database was divided respectively into two subsets: 80% for the training set and 20% for the test set. A total of 1666 descriptors were calculated, and the statistical reduction methodology, based on the Multiple Linear Regression (MLR) method, has been adopted. The Quasi-Newton back propagation (BFGS) algorithm was applied in order to train the ANN. A comparison was made between the outcomes of obtained QSPR-ANN models and other well-known correlations for each property. The two best QSPR-ANN models result showed a good accuracy confirmed by the high determination coefficient (R^2) values and the low mean absolute percentage error (*MAPE)* values ranging from 0.9999 to 0.9931 and from 0.5797 to 0.2600%, respectively for both best models (*Tb* and *d20* models). Furthermore, the comparison between our models and the other quantitative structure property relationships (QSPR) models shows that the QSPR-ANN models provided better results. This computational approach can be applied in the petroleum engineering for an accurate determination of *Tb* and *d20* of pure hydrocarbons.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

There have been a fair amount of researches on chemical, biochemical and environmental processes, which are an important component of scientific and industrial developments. Nonetheless, in the petroleum industry field, thermodynamic and physicochemical properties can be considered as the key element in the development and control of this industry [1], since it can explain a lot of phenomena that could help both of industrialists and scientists, in order to have an idea about the nature of their materials, with a view to finding and applying focused ways to transform and transport them [2,3]; with lower cost, in higher quality and safety [4,5]. Hydrocarbons are the largest constituents of petroleum fractions [6], it is known that a fundamental composition for any petroleum fraction is pure molecules. In particular, we find the pure hydrocarbons which are made up of hydrogen and carbon [7,8]. *Tb* and *d20* are the most important physicochemical and thermodynamic properties with a significant impact on hydrocarbons [6,9–12]. Despite the development studies in this area, the dataset on these properties are not currently available as needed due to several reasons, two of which are major: the expensive and time-demanding laboratory testing [13,14]. In recent years, the quantitative structure activity/property relationship (QSAR/QSPR) approaches have become the most used areas in the development of mathematical models designed to predict the physicochemical and thermodynamic properties of different chemical species [15–19].

QSAR/QSPR are chemoinformatic techniques that establish a relationship between the molecular structure and the defined activity/property [20,21]. To apply these techniques, it is necessary to extract maximum structural information through the best available molecular descriptors. A molecular descriptor is defined as a result of a mathematical and logical procedure. A standardized experiment that converts encrypted chemical information represented by a symbolic expression of a molecule into a useful number in order to be able to quantify the molecule [14,21]. Several important QSPR studies have been designed to predict the different properties of

* Corresponding author.
E-mail address: roubehiefissa.mohamed@univ-guelma.dz (M. Roubehie Fissa).

https://doi.org/10.1016/j.jmgm.2018.11.013
1093-3263/© 2018 Elsevier Inc. All rights reserved.

pure chemicals in different fields [22].

All those QSPR studies focused on the distinction between the molecule on the same type or the same family; for example, on the alkane molecules among them [23–27]. But if we get deep about these molecules, we find no talking about the isomer of position, for example, *cis-trans, meta, ortho* and *para*. It is known that is difficult or even impossible to find a molecular descriptor that would differentiate between all the isomers. However, in this study, we try to find a group of descriptors that can together distinguish among those isomers.

Usually the QSPR models should be constructed following these fundamental three steps: **(1)** choice of a set of molecular properties, *Tb* and *d20* in this study; **(2)** extraction of molecular structure information by calculation and choice of the best relevant molecular descriptors; **(3)** link the first and second steps by using a mathematical or numerical learning method [28]. Accordingly, this study aimed to develop a QSPR models in order to predict the *Tb* and *d20* of petroleum fractions and pure hydrocarbons. 223 and 222 of pure hydrocarbons molecules were used as dataset for each parameter. So, for these molecular structures, hundreds of descriptors were first calculated. Then, a selection operation was performed using statistical methods to obtain just a few relevant molecular descriptors. Finally, the relationship between the structures represented by the molecular descriptors and the properties (*Tb* and *d20*) was applied using an MLP-ANN. The steps cited above were respected to develop QSPR models, and the final best models were tested by different statistical coefficients and error types, for estimation of *Tb* and *d20* of pure hydrocarbons and petroleum fractions with high degree of precision.

2. Methodology

In the development of the QSPR models, several steps were used as described below and summarized in Fig. 1.

2.1. Database collection

It is known that the quality of the experimental database is the foundation stone for any successful mathematical model [29,30]. However, collecting these data with the appropriate quantity and quality is a very difficult and heavy task. The database used in this study was collected from several sources containing 223 and 222 different pure hydrocarbons molecules and petroleum fractions for *Tb* and *d20*, respectively [23,25,31,32]. These compounds consist of various categories saturated and unsaturated hydrocarbons representing main industrially important groups such as: n-paraffin, iso-paraffin, cyclo-paraffin, simple olefin, aromatic olefin and alkyne. The majority of those categories have more than one isomer of position such as, *cis*, *trans*, *meta*, *ortho* and *para*. The entire database used in this study was shown in Table S1.

2.2. Molecular descriptors calculation

One of the most important steps in the construction of the QSPR models is the quantification of structure information of the studied molecules [29], which are called molecular descriptors. Actually, there exist more than 11145 molecular descriptors that can be used to solve several problems in different field such as chemistry, biology and other related sciences [30,33]. In the course of this study 1666 molecular descriptors were calculated online by E-Dragon 1.0 software [34]. 20 different classes are displayed in Table 1.

The calculation of these descriptors was preceded by a very interesting step which is the generating of Simplified Molecular Input-Line Entry System (SMILES) notation with use of the "ChemDraw" software.

2.3. Pretreatment for descriptors reduction and selection

One of the challenges in this study is the selection of the relevant descriptors, hence we used a statistical methodology to reduce the number of descriptors and obtain selected descriptors used in building the QSPR models. The methodology was focused on the following points: **(1)** Eliminate all descriptors with error value that has at least one value "-999" indicated by E-Dragon software; **(2)** Omit descriptors with a same value, otherwise we look for *min* and *max* values for each descriptor; **(3)** Remove descriptors with a same value for more than 75%, by comparing the three first quartile; **(4)** Omit descriptors with a relative standard deviation (*RSD*) less than 0.05. In the two last steps, we used "STATISTICA" software [35]; **(5)** Remove descriptors with an intercorrelation Pearson coefficient *(R)* more than 0.75. We considered the intercorrelation between descriptors and output, in order to retain only relevant descriptors. In this stage, we used "Matlab" software [36] in order to calculate *R*, compare and eliminate each unwanted descriptor manually. At least, an important number of descriptors was retained as shown in Table 1. We reduced more than 90% of descriptors number in the preceding step (*RSD* step) for each parameter (*Tb* and *d20*); **(6)** Finally, we use an MLR forward stepwise regression method to remove descriptors with *P*-value (Probability value) higher than 0.005 by using "STATISTICA" software [35]. Table 2 and Table 3 show the *P*-value, *t*-value (Student's t-test) and the definition of the selected descriptors with summary of the MLR regression for each parameter.

2.4. Artificial neural network model's construction

After the pretreatment we reduced the number of descriptors as possible, we obtained the actual relevant descriptors for each parameter, to the construction phase of the ANN models. This is another challenge regarding the large parameter of choice in the ANN; that, we used different choices in order to obtain the best models for each parameter. The dataset was randomly split into two subsets (80%) for the training phase and (20%) for the testing phase of the model [37,38]; for the type of ANN, we used an MLP-ANN. A Quasi Newton back propagation BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm was used as training algorithm [39]. The activation function is performed by choosing several types of functions in both layers: hidden and output. Consequently, at the end of execution, the software gives the best possible combination between the two activation functions of concerned layer, we used the sum of squares (*SOS*) as error function. Relating the hidden layer numbers, as there exists one hidden layer in STATISTICA Software, version 8, we used only one hidden layer. But, regarding the number of neurons per hidden layer we generally varied the number from 3 to 20 neurons, to ensure that the variable of the neuron models do not exceed the size of database; so quite simply we follow this rule: ***(Number of input neurons * Number of hidden neurons) + (Number of hidden neurons * Number of output neurons) ≤ (Size of database)***

Several executions were made for each parameter to obtain the final best models. The execution has a significant number of the iterations; sometimes overtaking 100000 iterations per execution.

3. Results and discussions

3.1. Architectures of the best QSPR-ANN models obtained for each parameter

After these steps of collection, reduction, construction and

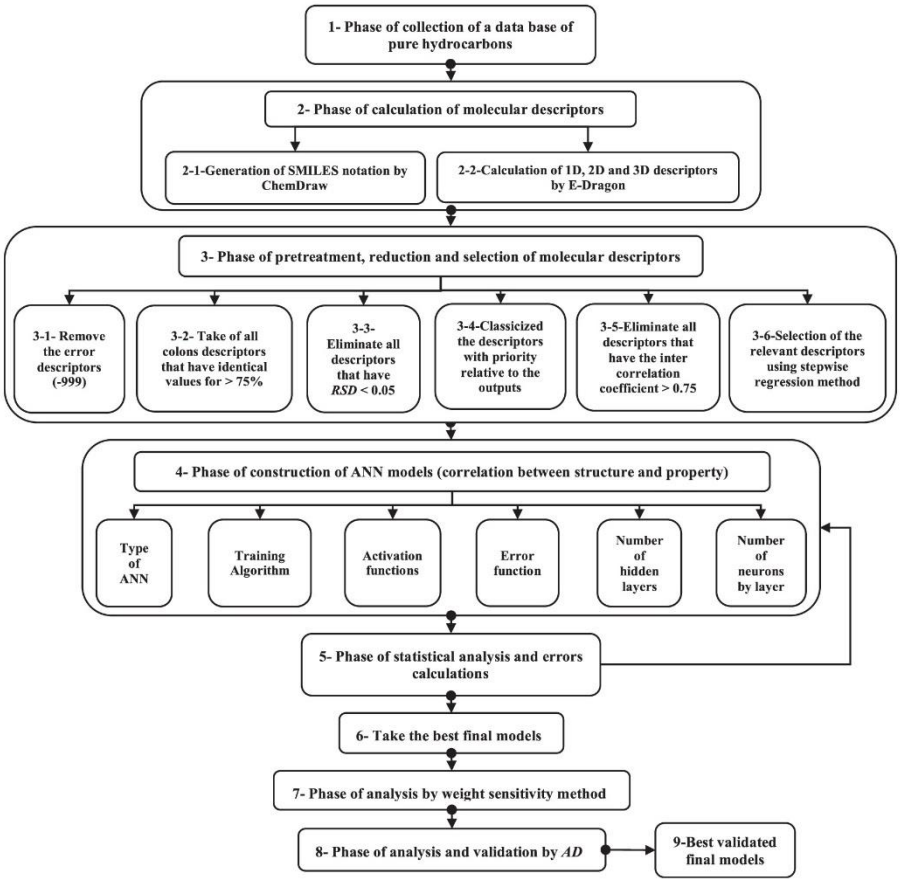


Fig. 1. Flow diagram of the methodology followed in this study.

calculation; only the best models were selected for testing their performance, by calculating various errors. The architecture of the best QSPR-ANN models obtained for each parameter with some primary statistic parameters are represented in Table 4. The *SOS* error has value of 10^{-6} for *Tb* model and 10^{-5} for *d20* model, for both test and training set. The *R* values show a high performance of the QSPR-ANN models.

3.2. Application of mathematical equation for the QSPR-ANN models

The QSPR-ANN architectures models are (16-12-1) for *Tb* model and (16-10-1) for *d20* model as shown in Table 4. So, the network has 16 inputs ($\boldsymbol{E_i}$, $i = 1\ to 16$) and one output ($\boldsymbol{S}$) for each model. Concerning the hidden layer, twelve neurons appeared in *Tb* model and ten neurons in *d20* model. For the activation functions frequently, we found *"Exponential"* in the first range for both hidden layers in *Tb* and *d20* models, *"Tanh"* and *"Sine"* in the second range; only for output layers, respectively for *Tb* and *d20* models, their mathematical definitions are given in Table 5, (see Eqs. (1), (2) and (3)) [35,40].

Each hidden neuron computes activation function, sending results ($\boldsymbol{Z_j}$: $j = 1\ to 12$; for *Tb* model and $j = 1\ to 10$; for *d20* model) to the output layer's neuron which finally gives the response of the network ($\boldsymbol{S}$). The output signal of each hidden neuron ($\boldsymbol{Z_j}$) is calculated by the Eq. (4) and Eq. (6) for *Tb* and *d20* models, respectively. The output of the network is given by Eq. (5) for *Tb* model and Eq. (7) for *d20* model.

$$\boldsymbol{Z_j} = f^H\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right] = e^{-\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right]} \tag{4}$$

$$\boldsymbol{S} = f^O\left[\sum_{j=1}^{12} w_{1j}^H \boldsymbol{Z_j} + b_1^O\right] = \frac{e^{+\left[\sum_{j=1}^{12} w_{1j}^H \boldsymbol{Z_J} + b_1^O\right]} - e^{-\left[\sum_{j=1}^{12} w_{1j}^H \boldsymbol{Z_J} + b_1^O\right]}}{e^{+\left[\sum_{j=1}^{12} w_{1j}^H \boldsymbol{Z_J} + b_1^O\right]} + e^{-\left[\sum_{j=1}^{12} w_{1j}^H \boldsymbol{Z_J} + b_1^O\right]}} \tag{5}$$

Table 1
Evolution of the number of descriptors during the pretreatment procedure by classes for *Tb* and *d20*.

Name of descriptors classes	(0)	(1)	(2)	(3)	(4)	(5)		(6)	
						For *Tb*	For *d20*	For *Tb*	For *d20*
(1) constitutional descriptors;*	48	48	33	22	20	0	2	0	0
(2) topological descriptors;*	119	119	80	74	72	6	6	2	2
(3) walk and path counts;*	47	47	47	41	41	1	1	0	1
(4) connectivity indices;*	33	33	33	33	33	7	8	0	1
(5) information indices;*	47	37	37	37	33	7	6	4	5
(6) two dimensional (2D) autocorrelations;*	96	96	60	48	48	1	2	0	0
(7) edge adjacency indices;*	107	107	106	88	88	6	6	1	0
(8) Burden eigenvalue descriptors;*	64	64	64	64	64	1	1	0	0
(9) topological charge indices;*	21	21	21	15	7	5	5	2	2
(10) Eigenvalue based indices;*	44	44	39	39	39	1	0	0	0
(11) Randic molecular profiles;*	41	41	41	41	41	0	0	0	0
(12) geometrical descriptors;*	74	74	38	34	31	4	5	0	1
(13) RDF descriptors;*	150	150	150	95	95	9	8	0	2
(14) 3D-MORSE descriptors;*	160	160	160	160	159	19	22	2	0
(15) WHIM descriptors;*	99	99	99	99	90	8	9	0	1
(16) GET-AWAY descriptors;*	197	197	197	195	121	13	10	4	1
(17) functional group counts;*	154	154	14	3	3	1	1	1	0
(18) atom-centered fragments;*	120	115	11	5	5	0	1	0	0
(19) charge descriptors;*	14	0	0	0	0	0	0	0	0
(20) molecular properties;*	31	31	16	12	11	0	0	0	0
Total of molecular descriptors	**1666**	**1637**	**1246**	**1105**	**1001**	**89**	**93**	**16**	**16**

(0): After calculated by E-Dragon [35]; *(1)*: After omitted the errors value (−999); *(2)* After eliminated the (max = min) value; *(3)* After eliminated the repeated value more than (>75%); *(4)* After removed the absolute value of *RSD* less than (<0.05); *(5)* After *R* intercorrelation; *(6)* After forward stepwise MLR method.

Table 2
Characteristics of selected descriptors for *Tb*.

Descriptors	Category	Description	*t*-value	*P*-Value
STN	topological descriptors	spanning tree number (log)	21.09052	0.00
TIE	topological descriptors	E-state topological parameter	13.74772	$6.041\ 10^{-31}$
IVDM	information indices	mean information content on the vertex degree magnitude	19.95857	0.00
Yindex	information indices	Balaban Y index	−3.27603	$1.235\ 10^{-3}$
IC3	information indices	information content index (neighborhood symmetry of 3-order)	10.38857	$1.343\ 10^{-20}$
CIC3	information indices	complementary information content (neighborhood symmetry of 3-order)	11.24742	$3.421\ 10^{-23}$
EEig07x	edge adjacency indices	Eigenvalue 07from edge adj. matrix weighted by edge degrees	6.90539	$6.094\ 10^{-11}$
GGI4	topological charge indices	topological charge index of order 4	−6.35177	$1.338\ 10^{-9}$
JGI1	topological charge indices	mean topological charge index of order1	2.84976	$4.826\ 10^{-3}$
Mor10v	3D-MoRSE descriptors	3D-MoRSE - signal 10/weighted by atomic van der Waals volumes	−6.36366	$1.255\ 10^{-9}$
Mor16e	3D-MoRSE descriptors	3D-MoRSE - signal 16/weighted by atomic Sanderson electronegativities	3.17347	$1.737\ 10^{-3}$
ISH	GETAWAY descriptors	standardized information content on the leverage equality	4.62401	$6.64\ 10^{-6}$
H5e	GETAWAY descriptors	H autocorrelation of lag 5/weighted by Sanderson electronegativity	10.42748	$1.027\ 10^{-20}$
R3e	GETAWAY descriptors	R autocorrelation of lag 3/weighted by atomic Sanderson electronegativities	−6.10791	$4.958\ 10^{-9}$
R4e	GETAWAY descriptors	R autocorrelation of lag 4/weighted by atomic Sanderson electronegativities	−8.67764	$1.227\ 10^{-15}$
nCt	Functional group counts	number of total tertiary C (sp3)	−8.59915	$2.031\ 10^{-15}$

Regression **S**ummary for Dependent Variable: *Tb*(K); *N*=223; $R = 0.9995$; $R^2 = 0.9989$; Adjusted $R^2 = 0.9988$; $F(16.206) = 11942$ (Fisher statistic value); $P<0.0000$ (Probability value); Std. Error of estimate: 3.3188.

$$\boldsymbol{Z_j} = f^H\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right] = e^{-\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right]} \quad (6)$$

$$\boldsymbol{S} = f^O\left[\sum_{j=1}^{10} w_{1j}^H \boldsymbol{Z_j} + b_1^O\right] = \sin\left[\sum_{j=1}^{10} w_{1j}^H \boldsymbol{Z_j} + b_1^O\right] \quad (7)$$

$\boldsymbol{f^H}$ is the activation function of hidden layer (relating the input with hidden layer), $\boldsymbol{f^O}$ the activation function of output layer (relating the hidden with output layer), $\boldsymbol{S}$ the output of the network (outputs parameters: *Tb* and *d20*), $\boldsymbol{Z}$ the output signal of the hidden neurons, $\boldsymbol{w_{ij}^I}$ are the weights of input layer (connections between the input and hidden neurons). $\boldsymbol{w_{1j}^H}$ the weights of hidden layer (connections between the hidden and output neurons), $\boldsymbol{E_i}$ the input neurons (in this case the relevant molecular descriptors), $\boldsymbol{b_j^H}$ are the bias of hidden layer (bias of hidden neurons), $\boldsymbol{b_1^O}$ is the bias of output layer (bias of output neuron), $\boldsymbol{i}$ the number of input ($i = 1\ to 16$ for each model), and $\boldsymbol{j}$ is the number of hidden neurons ($j = 1\ to 12$; for *Tb* model and $j = 1\ to 10$; for *d20* model).

The final mathematical models for predicting *Tb* and *d20* of pure hydrocarbons are obtained by using the MLP-ANN method and are given by Eq. (5) for *Tb* and Eq. (7) for *d20*. The values of the weights and bias of *Tb* and *d20* models for each layer are given in Table S2 and Table S3 respectively.

3.3. Estimation of physicochemical proprieties of pure hydrocarbons by QSPR-ANN models

In order to estimate *Tb* and *d20*, we use the previous equations (Eq. (5) and Eq. (7)) by introducing the values of the inputs which are the values of the relevant descriptors, ranked from the first descriptor to the last one; for *Tb* model (*STN, TIE, IVDM, Yindex, IC3, CIC3, EEig07x, GGI4, JGI1, Mor10v, Mor16e, ISH, H5e, R3e, R4e* and *nCt*) (see Table 2), also for *d20* model (*GNar, J, piPC03, X3Av, IVDE, HDcpx,*

Table 3
Characteristics of selected descriptors for *d20*.

Descriptors	Category	Description	*t*-value	*P*-Value
GNar	topological descriptors	Narumi geometric topological index	9.8530	$5.368\ 10^{-19}$
J	topological descriptors	Balaban distance connectivity index	2.8475	$4.855\ 10^{-3}$
piPC03	Walk and path counts	Molecular multiple path count of order 03	23.6495	0.00
X3Av	connectivity indices	average valence connectivity index chi-3	6.4477	$7.992\ 10^{-10}$
IVDE	information indices	mean information content on the vertex degree equality	−8.5350	$3.129\ 10^{-15}$
HDcpx	information indices	graph distance complexity index (log)	10.2783	$2.976\ 10^{-20}$
Yindex	information indices	Balaban Y index	−10.7656	$1.036\ 10^{-21}$
IC3	information indices	information content index (neighborhood symmetry of 3-order)	−17.1642	$1.599\ 10^{-41}$
CIC3	information indices	complementary information content (neighborhood symmetry of 3-order)	−20.4706	0.00
GGI5	topological charge indices	topological charge index of order 5	3.4135	$7.730\ 10^{-4}$
JGI1	topological charge indices	mean topological charge index of order1	7.5165	$1.719\ 10^{-12}$
SPH	Geometrical descriptors	Spherosity	−3.9838	$9.422\ 10^{-5}$
RDF030m	RDF descriptors	Radial Distribution Function - 3.0/weighted by atomic masses	8.2985	$1.401\ 10^{-14}$
RDF055m	RDF descriptors	Radial Distribution Function - 5.5/weighted by atomic masses	−9.2169	$3.747\ 10^{-17}$
L3p	WHIM descriptors	3rd component size directional WHIM index/weighted by atomic polarizabilities	−4.3278	$2.353\ 10^{-5}$
HATS3u	GETAWAY descriptors	leverage-weighted autocorrelation of lag 3/unweighted	−3.1875	$1.660\ 10^{-3}$

Regression Summary for Dependent Variable: *d20*(−); $N = 222$; $R = 0.9947$; $R^2 = 0.9894$; Adjusted $R^2 = 0.9885$; $F\ (16.205) = 1193.2$ (Fisher statistic value); $P < 0.0000$ (Probability value); Std. Error of estimate: 0.00885.

Table 4
Architectures of the best QSPR-ANN models obtained for each parameter.

Parameter	Layers	Number of Neurons	Activation Function	Correlation coefficient *R* and *SOS* Value	
Tb	Input Layer	16	***(none)***	*R* training	0.9997
	Hidden Layer	12	*Exponential*		
	Output Layer	1	*Tanh*	*R* test	0.9999
	Training Algorithm	BFGS 65		*SOS* training	$8.0\ 10^{-6}$
	Error Function	Sum of Squares (*SOS*)		*SOS* test	$3.0\ 10^{-6}$
d20	Input Layer	16	***(none)***	*R* training	0.9965
	Hidden Layer	10	*Exponential*		
	Output Layer	1	*Sine*	*R* test	0.9993
	Training Algorithm	BFGS 47		*SOS* training	$7.4\ 10^{-5}$
	Error Function	Sum of Squares (*SOS*)		*SOS* test	$2.9\ 10^{-5}$

Table 5
Activation functions appeared in this study [35,40].

	Activation functions		
	Exponential	Hyperbolic Tangent	Sinus
Notation in STATISTICA	*Exponential*	*Tanh*	*Sine*
Notation in Matlab	*exp*	*tanh*	*sin*
Mathematical Formula	e^{-n} (1)	$\frac{e^{+n}-e^{-n}}{e^{+n}+e^{-n}}$ (2)	$\sin(n)$ (3)

(*e*): The exponential function; (*n*): The variables of the ANN models (see Eqs. (4) and (6)).

Yindex, IC3, CIC3, GGI5, JGI1, SPH, RDF030m, RDF055m, L3p and *HATS3u*) (see Table 3). In addition to the use of weights and bias connections of both hidden and output layers given in Table S2 and Table S3 for *Tb* and *d20*, respectively.

The visual comparison between the experimental and calculated values obtained by using the developed QSPR-ANN models, shown in Fig. 2a, Fig. 2b and Fig. 2c for *Tb* model, as well as in Fig. 3a, Fig. 3b and Fig. 2c for *d20* model, and this for the whole dataset, training and test, respectively. A tight concentration of data points around the unit-slope line (shown by the dashed line) reveals that the QSPR-ANN models can accurately predict the *Tb* and *d20* values. It was confirmed that the proposed models give satisfactory results, through giving an accuracy agreement with the vector values approaching the ideal (i.e. $\alpha = 1$ (slope), $\beta = 0$ (y intercept), $R = 1$) in graphs fitting of *Tb* and *d20* profiles, for the whole dataset, training and test (see the equations of regression at the top of each figure). On the basis of the previous figures, the QSPR-ANN models give good results with high R^2 values (see also Table 6). The experimental and calculated values of both models (*Tb* and *d20*) and the splitting way are given in Table S1.

3.4. Analysis of model's statistical parameters

The performances of the QSPR-ANN models are summarized in Table 6, in terms of Root Mean Squared Error (*RMSE*), Standard Error of Prediction (*SEP*), Mean Percentage Error (*MPE*) and *MAPE* corresponding to each model for training, test and whole dataset. In addition to the *R* and R^2, the values of slope (α), intercept (β) and the number of data point (N) in each phase. The mathematical definition of the previous error types, *R* and R^2 are given by Eq. (10) to Eq. (15), respectively.

$$\widehat{y}^{cal} = \sum_{i=1}^{N} y_i^{cal} \Big/ N \tag{8}$$

$$\widehat{y}^{exp} = \sum_{i=1}^{N} y_i^{exp} \Big/ N \tag{9}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \left(y_i^{exp} - y_i^{cal}\right)^2 \Big/ N} \tag{10}$$

$$SEP = (RMSE/\widehat{y}^{exp})*100 \tag{11}$$

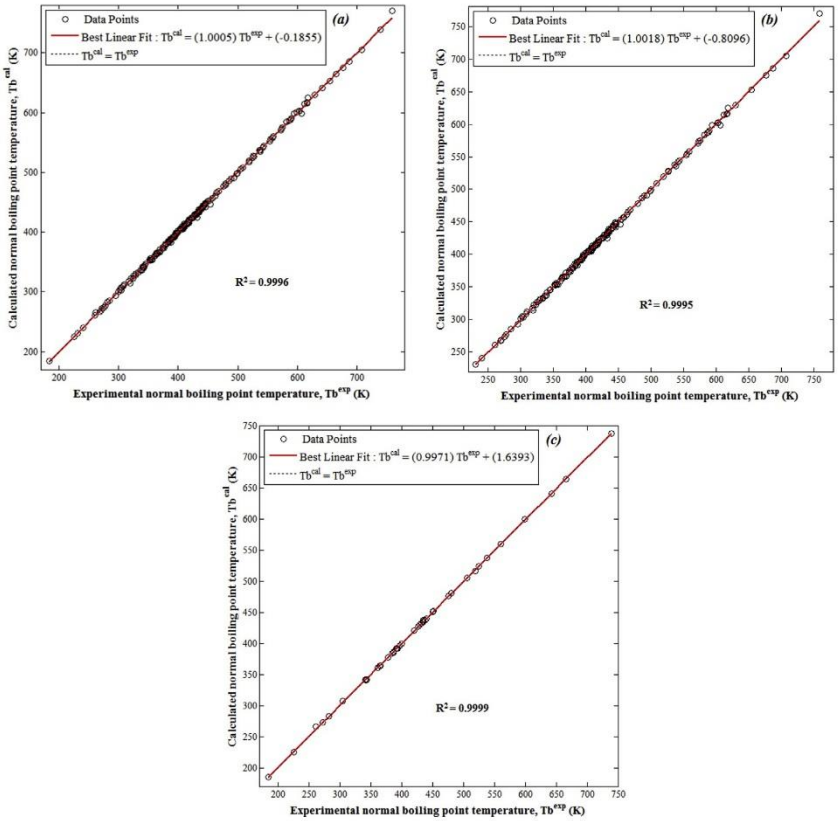


Fig. 2. Comparison between experimental and calculated values of boiling point temperature *Tb* for: ***(a)*** the whole dataset, ***(b)*** training set, and ***(c)*** test set.

$$MPE = (100/N)*\sum_{i=1}^{N}\left(\left|\left(y_i^{exp}-y_i^{cal}\right)\right|/y_i^{cal}\right) \quad (12)$$

$$MAPE = (100/N)*\sum_{i=1}^{N}\left(\left|\left(\left(y_i^{exp}-y_i^{cal}\right)/y_i^{exp}\right)\right|\right) \quad (13)$$

$$R = \frac{\sum_{i=1}^{N}\left(y_i^{exp}-\hat{y}^{exp}\right)\left(y_i^{cal}-\hat{y}^{cal}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i^{exp}-\hat{y}^{exp}\right)^2\sum_{i=1}^{N}\left(y_i^{cal}-\hat{y}^{cal}\right)^2}} \quad (14)$$

$$R^2 = 1-\sum_{i=1}^{N}\left(y_i^{exp}-y_i^{cal}\right)^2\Big/\sum_{i=1}^{N}\left(y_i^{exp}-\hat{y}^{exp}\right)^2 \quad (15)$$

Where $\boldsymbol{N}$ is the number of compounds in the dataset for each phase (training, test or whole dataset), $\boldsymbol{y^{cal}}$ is the calculated (predicted) value of y; for training set or test set, $\boldsymbol{y^{exp}}$ is the experimental (observed) value of y, $\hat{\boldsymbol{y}}^{\boldsymbol{cal}}$ the average of calculated values of y; (see Eq. (8)), and $\hat{\boldsymbol{y}}^{\boldsymbol{exp}}$ the average of experimental values of y; (see Eq. (9)).

According to the results of different statistical coefficients and error types shown in Table 6, the developed models were very interesting, as shows the R^2 values, which is higher than 0.9930 both in *Tb* and *d20* models, for the three-subset phase, training, test and all over the dataset.

The errors types confirmed that the *RMSE* value of *Tb* model ranged from a maximum value of 2.1458 K for the training phase to minimum value of 1.3718 K for the test phase. Nonetheless, the *RMSE* value of all dataset which means the value of the developed *Tb* model nearly the same range which is exactly 2.0168 K, confirming the stability of the model. The same statement of *d20* model, with *RMSE* value comprising between the extreme value of 0.0064 and tiniest value of 0.0040, respectively for training and test phase. The *RMSE* value of all the dataset belonging to the developed *d20* model is almost identical, i.e., equal to 0.0060, approving also the balance between under-learning (under-fitting) and over-learning (over-fitting) on the actual QSPR-ANN models, which have confirmed their stability [41–43].

It is known that the *RMSE* value unit depends on the concerning property calculated (i.e. *RMSE* of *Tb* is in the same unit of *Tb*, K) Thus, it seems that *RMSE* is not a comprehensive indicator for judging a model, since there are big differences in values if using different units. Besides in the same unit it is difficult to identify what the *RMSE* values are really like; small or large values [44]. Consequently, we look for other error types depending on a more

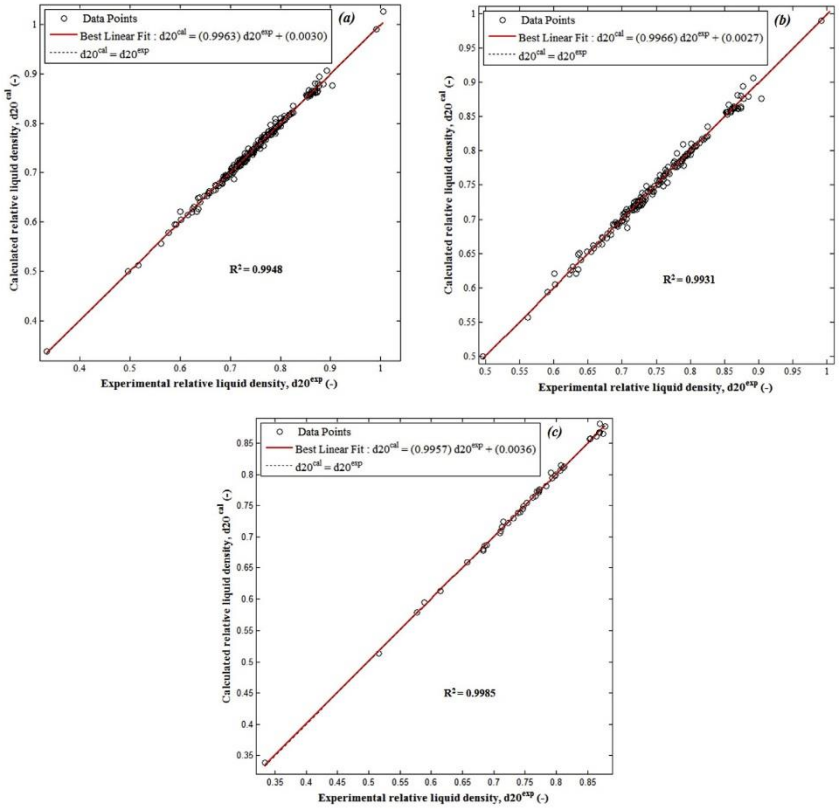


Fig. 3. Comparison between experimental and calculated values of relative liquid density *d20* for: ***(a)*** the whole dataset, ***(b)*** training set, and ***(c)*** test set.

Table 6
Statistical parameters of the best QSPR-ANN models for both properties.

Statistical parameters	*R*	R^2	*(RMSE)*	*(SEP)*	*(MPE)*	*(MAPE)*	*α*	*β*	*N*
Training phase									
Tb^{Train}	0.9997	0.9995	2.1458	0.5009	0.3536	0.3539	1.0018	−0.8096	179
$d20^{Train}$	0.9965	0.9931	0.0064	0.8429	0.5804	0.5797	0.9966	0.0027	178
Test phase									
Tb^{Test}	0.9999	0.9999	1.3718	0.3255	0.2613	0.2600	0.9971	1.6393	44
$d20^{Test}$	0.9993	0.9985	0.0040	0.5381	0.3975	0.3960	0.9957	0.0036	44
Training + Test phase									
Tb^{All}	0.9998	0.9996	2.0168	0.4723	0.3354	0.3353	1.0005	−0.1855	223
$d20^{All}$	0.9974	0.9948	0.0060	0.7930	0.5441	0.5433	0.9963	0.0030	222

representative unit, such as relative errors expressed as a percentage (%).

Considering the mathematical definition of *RMSE* and *SEP* calculated by Eq. (10) and Eq. (11), the *SEP* value is actually the same type with *RMSE* error but with % unit. The *SEP* value approving the preceding result of *RMSE* is, in fact, lower than 1% in *Tb* and *d20* models, with a quite better result in *Tb* model. The *MPE* and *MAPE* values confirm that our models predicting *Tb* and *d20* are more accurate: the *MAPE* values of 0.35 and 0.55% for *Tb* and *d20* models, respectively.

3.5. Sensitivity analysis

In order to appreciate the contribution of the input descriptors on each model we use the weight sensitivity analysis method [45]. This can determinate the impact of each individual input descriptor on the output for each ANN model (*Tb* and *d20* models). This method was initially proposed by Garson [46]. Then taken up by Goh [47], the method essentially includes partitioning connection weights into: **(1)** input-hidden weights and **(2)** hidden-output weights of each hidden and output neurons connected, then gives a significant quantification (relative importance (RI)) of each

input to the output of the ANN [45]. The main steps of this method "weight method" are shown as flowchart in study of Khaouane [48] and Ammi et al. [49].

The contributions of the relevant descriptors on the QSPR-ANN models were established and depicted in Fig. 4a and Fig. 4b for *Tb* and *d20* models, respectively. According to the results of sensitivity analysis and the average value of contribution which is 6.25% for both models, we can classify the descriptors contribution for *Tb* model into three groups: **(1)** high contribution, ranging from 12 to 7% {*H5e*(11.90%) > *R4e*(10.58%) > *nCt*(8.23%) > *R3e*(7.50%)}, **(2)** medium contribution, ranging from 7 to 6% {*IVDM*(6.62%) > *Mor*10*v*(6.31%) > *GGI4*(6.12%)} and **(3)** low contribution, ranging from 6 to 4% {*ISH*(5.33%) > *STN*(5.18%) > *Yindex*(4.93%) > *JGI*1(4.91%) > *TIE*(4.80%) > *EEig*07*x*(4.58%) > *ClC*3 (4.55%) > *Mor*16*e*(4, 54%) > *IC*3(3.92%)}. And in four groups for *d20* model: **(1)** high contribution, ranging from 11 to 8% {*X*3*Av*(10.44%) > *GGI5*(9.74%) > *piPC*03(8.79%) > *IC*3(8.09%)}, **(2)** medium contribution, ranging from 7 to 6% {*SPH*(7.01%) > *Yindex* (6.60%) > *L*3*p*(6.59%) > *IVDE*(6.37%) > *ClC*3(6.26%) > *RDF*030*m* (6.16%)}, **(3)** low contribution, ranging from 5 to 4%

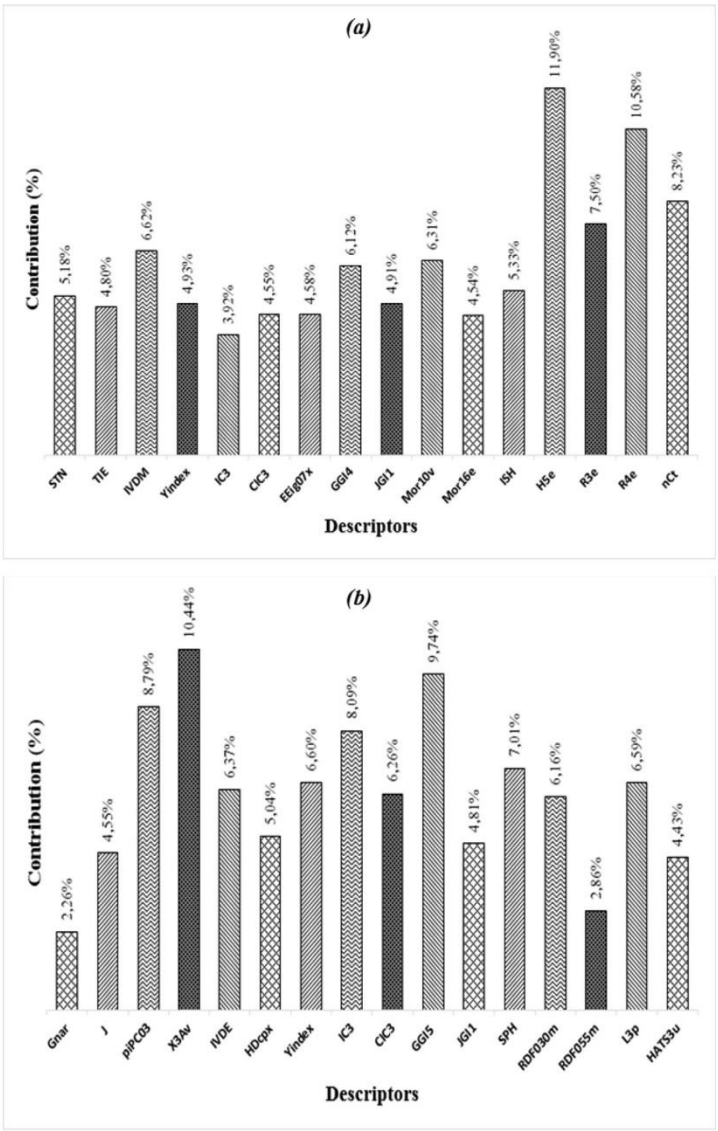


Fig. 4. Histogram depicting the relative contributions of the relevant descriptors for each QSPR-ANN model: ***(a)*** boiling point temperature *Tb*, ***(b)*** relative liquid density *d20*.

{*HDcpx*(5.04%) > *JGI1*(4.81%) > *J*(4.55%) > *HATS3u*(4.43%)}, and **(4)** very low contribution, less than 3% {*RDF055m*(2.86%) > *GNar*(2.26%)}.

Except the two descriptors of *d20* model (*RDF055m* and *GNar*) which do not exceed the 4% of difference with the average contribution value, it is worth mentioning that there is no significant difference between descriptors, especially in the same group classification above, which means that all selected descriptors were important in varying degrees in the development of QSPR models according to the group classification to which they belong.

According to the sensitivity analysis, it is obvious that the most significant descriptor classes that influenced *Tb* model are: *GET-AWAY descriptors* with *RI* = 35.31% by 4 descriptors, *information indices* with *RI* = 20.02% by 4 descriptors too, then *topological charge indices, 3D-MORSE descriptors* and *topological descriptors* with *RI* = 11.03%, 10.85% and 9.98%, respectively, with 2 descriptors for each class, and in the end range with just 1 descriptor for each class we find *functional group counts* and *edge adjacency indices* with *RI* = 8.23% and 4.58%, respectively. Regarding the *d20* model, the most influenced significant descriptors classes are: *information indices* with *RI* = 32.36% by 5 descriptors; *topological charge indices* with *RI* = 14.55% by 2 descriptors; *connectivity indices* with *RI* = 10.44% by just 1 descriptor; *RDF descriptors* with *RI* = 9.02% by 2 descriptors; *walk and path counts* with *RI* = 8.79% by 1 descriptor; *geometrical descriptors* with *RI* = 7.01% by 1 descriptor; *topological descriptors* with *RI* = 6.81% by 2 descriptors; *WHIM descriptors* with *RI* = 6.59% by 1 descriptor; and *GET-AWAY descriptors* with *RI* = 4.43% by 1 descriptor. From the contradictory coincidences between the *Tb* and *d20* models it appears that the *GET-AWAY descriptors* class represents the most important class in *Tb* model and the least important class in *d20* model. However, we found that the same two classes *information indices* and *topological charge indices* are among the most influential descriptor classes in both models.

3.6. Applicability domain

The applicability domain (AD) is defined as the confirmation that a model within its own domain of application has a suitable range of precision within the proposed model application [30]. Generally, the reliability of QSPR predictions is limited to structurally similar chemicals for those used in the construction of the proposed models [29,50,51].

There exist several methods for determining the AD, like distance based, geometrical based, range based, etc. [52–54]. A distance-based method is one of the furthermost well-known methods of AD approach. The principle of the method lies in leverage (***h***) value determination of each compound, which is then compared with the standard leverage (***h****) value [54,55]. The advantage of this method is that it is possible to quantify the model applicability, presenting it on a visual graph called the Williams plot [54].

The ***h*** value within an original variable area for any compound is defined as: $\boldsymbol{h_i} = \boldsymbol{X_i}(\boldsymbol{X^T X})^{-1}\boldsymbol{X_i^T}$, ($\boldsymbol{i} = 1, 2, 3 \ldots \boldsymbol{n}$). where $\boldsymbol{X_i}$ is the row-vector descriptor of the concerned compound, ***X*** the training descriptors matrix values, and ***n*** the whole compounds number. The ***h**** value is generally taken as: $\boldsymbol{h}^* = 3*(\boldsymbol{k}+1)/\boldsymbol{m}$, where ***k*** is the model variables number and ***m*** the training compounds number. If the ***h*** value of any considered compound is greater than the ***h**** value, the predicted value of the compound could be considered as out of the range application of the model. Consequently, the predicted result possibly will not be reliable [54,56–58].

The toolbox developed by Milano Chemometrics and QSAR Research Group was used [52,58]; the AD of the developed QSPR-ANN models is shown in the Williams plot in Fig. 5. The AD is established in a square area within ***h**** value and ±2 of standard residual values. They are considered fairly strict, compared with the rest of calculations in other studies, so that most standard residual values are confined to ±3 [59].

The ***h**** for the *Tb* model (Fig. 5a) is 0.2374 and means that there are only 17 of 223 compounds out of the domain; in other words, 92% of compounds within the AD. 12 and 5 among the 179 and 44 compounds, respectively for training and test sets, are shown, according to their splitting way in Table S1; the 1st, 2nd, 23rd, 64th, 109th, 111th, 112th, 130th, 131st, 132nd, 164th, and 175th for the training compounds and the 1st, 7th, 8th, 17th, and 20th for the test compounds. Therefore, more than 93% and 88% are within the AD area for training and test compounds.

Regarding *d20* model (Fig. 5b) the ***h**** value is 0.2388 which means just 22 of 222 are out of the AD. So about 90% are in the domain; 17 and 5 among the 178 and 44 compounds, respectively, for training and test sets, are according to the splitting way in Table S1 as mentioned above; the 1st, 2nd, 26th, 64th, 98th, 99th, 111th, 112th, 120th, 136th, 142nd, 157th, 162nd, 169th, 171st, 173rd and 177th for the training compounds and the 1st, 2nd, 8th, 17th, and 20th for the test compounds. Consequently, almost 93% and more than 88% are within the scope of the AD, respectively for training and test compounds. This is regardless of the standard residual value taken ±2, which again confirms the effectiveness of the models developed in this study.

3.7. Comparison with different ANN and QSPR models

For further evaluation of our proposed models, and knowing the importance of the results obtained from this study, we try to find some analogous studies with high similarity for making a comparison with them. The comparison is shown with details in Table 7 and Table 8, respectively for *Tb* and *d20* models. We have taken almost different studies which have common methods with our study (QSPR, ANN, etc.). Evaluation of their strengths and weaknesses are quite difficult, given the specificity of each study (datasets, descriptors, modeling approach, different validation strategy, etc.).

Although our models include all the hydrocarbons families, they outperformed all the models mentioned in this comparison, and even surpassed the studies that treat each family individually, as in the Ha et al. [11] Dai et al. [60] studies, They also exceeded those which have almost taken the same family (all families of hydrocarbons) as in the study of Wakeham et al. [25] and the study of Saaidpour & Ghaderi [61]. As expected, the accuracy of our models is best, compared to the studies which used different organic families; including hydrocarbons, as Sola et al. [62] and Varamesh et al. [63]. When we work on a large area, the value of the error increases; by contrast, for chemical families, working with closer compounds leads to smaller errors. This shows that all the models presented in this study outweigh the available studies and gives the most accurate prediction. Notwithstanding the diversity of modeling methods and the differences of chemical compounds families, in this comparison, our models showed good accuracy, in all situations.

Table 7 reveals that the proposed *Tb* model has more than ten times less *RMSE* compared to the QSPR-FF-ANN method in Gharagheizi et al. study [64] and compared to the MLP-ANN, LSSVM and GMDH-NN methods in the Varamesh et al. study, eight times less for RBF-ANN in the same study [63], the same proportions with *MAPE* error in both studies [63,64]. The proposed *Tb* model also has more than four times less *RMSE* compared with Sola et al. study [62], more than two times less in training set and more than seven times less in test set in terms of *RMSE* compared to the study of Saaidpour & Ghaderi [61]. Other studies [11,60] have not calculated

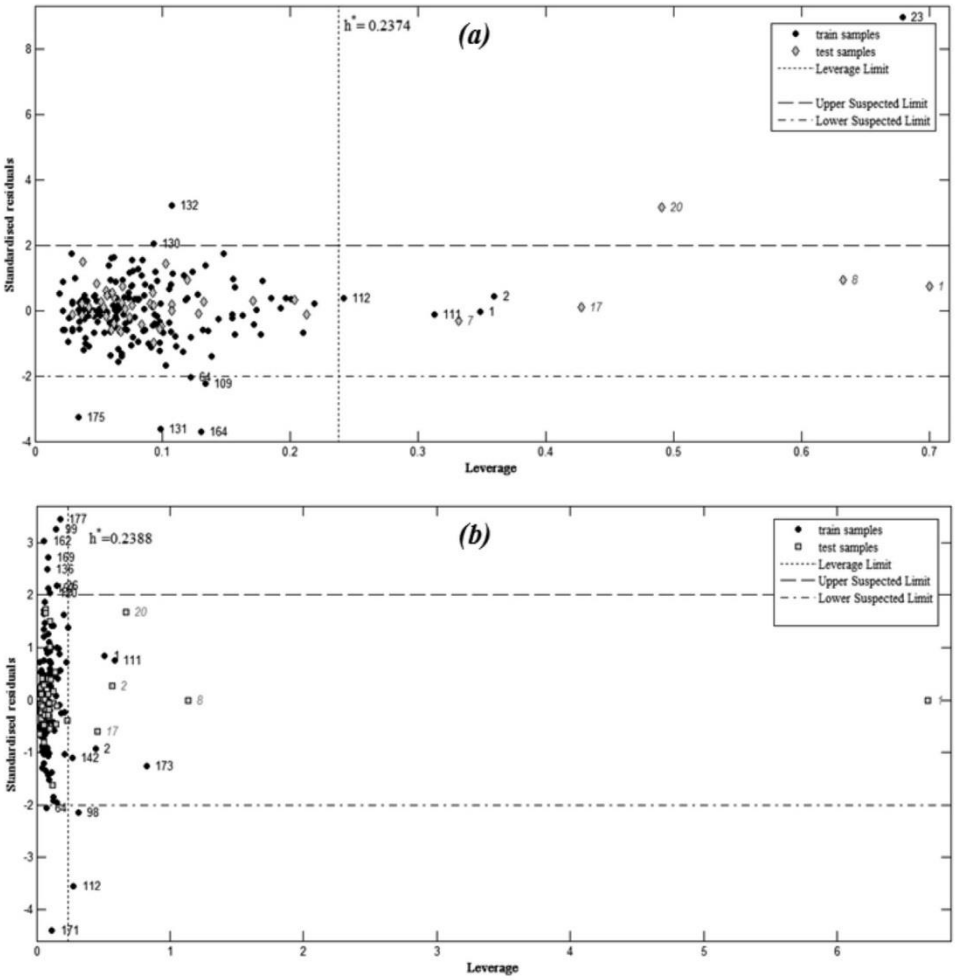


Fig. 5. Williams plot describing the AD for each QSPR-ANN model: ***(a)*** boiling point temperature *Tb*, ***(b)*** relative liquid density *d20*.

the same error types; so, we have evaluated R^2 and noted that its value does not exceed our values of 0.9995 for training, 0.9999 for test and 0.9996 for the whole dataset (the real *Tb* model value).

Concerning the comparison of *d20* model, the Table 8 shows that the proposed *d20* model is the best in comparison with R^2 since the actual studies also have not used the same error types. With the exception of the Wakeham et al. study [25] that slightly exceeded our R^2 value which is of 0.9970, all R^2 values in the Ha et al. study [11] have not exceeded the R^2 values obtained by our *d20* model. All these indicators confirm the efficiency and accuracy of the developed *Tb* and *d20* models.

4. Conclusion

In this study, two accurate QSPR-ANN models, have been developed to predict ***Tb*** and ***d20*** of pure hydrocarbons and petroleum fractions, based on their molecular structure (i.e., QSPR approach). 1666 descriptors were calculated. Several steps were followed, based on statistical and MLR methods for reducing this huge number of descriptors in order to keep only the real relevant descriptors. The experimental dataset has been selected from several sources. The final best QSPR-ANN models were attained by BFGS algorithm with "**16-12-1**″ and "**16-10-1**″ network architectures for ***Tb*** and ***d20***, respectively.

The results of the proposed QSPR-ANN models showed good accuracy according to R^2 value; they may take 0.9999 for Tb^{Test} and 0.9985 for $d20^{Test}$. Moreover, the strength and predictive ability of the models was emphasized by various error types ***(RMSE, SEP, MPE*** and ***MAPE)***; the result was very encouraging 1.3718 K and 0.0040 for the *RMSE* of *Tb* and *d20*, respectively, between 0.2600% and 0.5009%, also 0.3353%–0.8429% for the rest of the error types of *Tb* and *d20*, respectively.

The comparison of the QSPR-ANN models developed in this study with those of similar models in the literature confirmed that

Table 7
Comparison between the presented *Tb* model in this study and previous models.

Parameter	Models	Method	Nature of Hydrocarbons	Training Set				Test Set				Whole Set			
				N	R^2 (–)	*RMSE* (K)	*MAPE* (%)	*N*	R^2 (–)	*RMSE* (K)	*MAPE* (%)	*N*	R^2 (–)	*RMSE* (K)	*MAPE* (%)
Tb	**Present work**	**QSPR MLP-ANN**	**223 different pure hydrocarbons**	**179**	**0.9995**	**2.1458**	**0.3539**	**44**	**0.9999**	**1.3718**	**0.2600**	**223**	**0.9996**	**2.0168**	**0.3353**
	Saaidpour & Ghaderi [61]	**QSPR** MLR	80 crude oil hydrocarbons	64	0.9938	5.8971	(–)	16	(–)	10.8742	(–)	80	(–)	(–)	(–)
	Dai et al. [60]	**QSPR** MLR	80 alkanes	60	0.9993	(–)	(–)	20	(–)	(–)	(–)	(–)	(–)	(–)	(–)
		QSPR MLR	65 unsaturated hydrocarbons	50	0.9910	(–)	(–)	15	(–)	(–)	(–)	(–)	(–)	(–)	(–)
	Ha et al. [11]	**QSPR** MLR	186 pure saturates hydrocarbons	152	(–)	(–)	(–)	34	(–)	(–)	(–)	186	0.9979	(–)	(–)
		QSPR MLR	200 pure aromatic hydrocarbons	139	(–)	(–)	(–)	61	(–)	(–)	(–)	200	0.9960	(–)	(–)
		QSPR MLR	386 pure saturate and aromatic hydrocarbons	291	(–)	(–)	(–)	95	(–)	(–)	(–)	386	0.9947	(–)	(–)
	Sola et al. [62]	**QSPR** MLR	155 pure Organic compounds including hydrocarbons	135	(–)	9.10	(–)	20	(–)	7.33	(–)	155	0.9864	(–)	(–)
	Varamesh et al. [63]	**MLP-ANN**	563 pure Organic compounds including hydrocarbons	450	(–)	21.22	3.24	113	(–)	24.70	4.07	563	(–)	21.95	3.41
		RBF-**ANN**		450	(–)	17.92	2.60	113	(–)	21.87	2.82	563	(–)	18.78	2.65
		LSSVM		450	(–)	21.52	3.09	113	(–)	18.54	3.16	563	(–)	20.95	3.11
		GMDH-**NN**		450	(–)	25.59	3.79	113	(–)	23.98	3.73	563	(–)	25.27	3.78
	Gharagheizi et al. [64]	**QSPR** FF-**ANN**	A large dataset of 17768 pure chemical compounds	**14216**	0.9430	22	3.19	**1776**	0.9470	21	3.05	**17768**	0.9430	22	3.16

Table 8
Comparison between the presented *d20* model in this study and previous models.

Parameter	Models	Method	Nature of Hydrocarbons	Training Set		Test Set		Whole Set	
				N	R^2 (–)	*N*	R^2 (–)	*N*	R^2 (–)
d20	**Present work**	**QSPR MLP-ANN**	**222 different pure hydrocarbons**	**178**	**0.9931**	**44**	**0.9985**	**222**	**0.9948**
	Ha et al. [11]	**QSPR** MLR	186 pure saturates hydrocarbons	152	(–)	34	(–)	186	0.9910
		QSPR MLR	200 pure aromatic hydrocarbons	164	(–)	36	(–)	200	0.9881
		QSPR MLR	**386** pure saturate and aromatic hydrocarbons	**316**	(–)	**70**	(–)	**386**	0.9805
	Wakeham et al. [25]	**QSPR** MLR	219 different pure hydrocarbons	(–)	(–)	(–)	(–)	219	**0.9970**

our models are superior when predicting *Tb* and *d20* as they are providing best performance, therefore they bring best results.

The models are not only directed to particular or specific families such as the alkanes or alkenes but also encompass all hydrocarbon families in the same model. Furthermore, our models are able to distinguish even between isomers of position (i.e., *cis*, *trans*, *meta*, *ortho* and *para*) with a high degree of precision. The two QSPR-ANN models developed in this study are used in the field of industry and petroleum engineering, and also in other fields related to the pure hydrocarbons to estimate their ***Tb*** and ***d20***, only from its molecular structure using some simple calculable molecular descriptors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare that there is no conflict of interests.

Acknowledgments

The authors appreciate the efforts of the (LSPN) and (LBMPT) laboratory team for their encouragement throughout this project. The authors are also thankful to the anonymous reviewers of this manuscript for constructive observations and suggestions.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2018.11.013.

References

[1] M.R. Riazi, Characterization and Properties of Petroleum Fractions, vol. 50, ASTM international, 2005.
[2] C. Tsonopoulos, J.L. Heidman, S. Hwang, Thermodynamic and Transport Properties of Coal Liquids, 1986.
[3] P. Leprince, Petroleum Refining. Vol. 3 Conversion Processes, vol. 3, Editions Technip, 2001.
[4] B.K. Bose, Energy, environment, and advances in power electronics, IEEE Trans. Power Electron. 15 (4) (2000) 688–701.
[5] C. Belghit, Y. Lahiouel, T.A. Albahri, New empirical correlation for estimation of vaporization enthalpy of algerian saharan blend petroleum fractions, Petrol. Sci. Technol. (2018) 1–6.
[6] B.E. Poling, J.M. Prausnitz, O.C. John Paul, R.C. Reid, The Properties of Gases and Liquids, vol. 5, Mcgraw-hill, New York, 2001.
[7] W.D. McCain, The Properties of Petroleum Fluids, PennWell Books, 1990.
[8] M.A. Fahim, T.A. Al-Sahhaf, A. Elkilani, Fundamentals of Petroleum Refining, Elsevier, 2009.
[9] M.R. Riazi, Y.A. Roomi, Use of the refractive index in the estimation of thermophysical properties of hydrocarbons and petroleum mixtures, Ind. Eng. Chem. Res. 40 (8) (2001) 1975–1984.
[10] S.D. Nelson, P.G. Seybold, Molecular structure–property relationships for alkenes, J. Mol. Graph. Model. 20 (1) (2001) 36–53.
[11] Z. Ha, Z. Ring, S. Liu, Quantitative Structure– property relationship (QSPR) models for boiling points, specific gravities, and refraction indices of hydrocarbons, Energy Fuels 19 (1) (2005) 152–163.
[12] P.Y. Chan, C.M. Tong, M.C. Durrant, Estimation of boiling points using density functional theory with polarized continuum model solvent corrections, J. Mol. Graph. Model. 30 (2011) 120–128.
[13] P.R. Duchowicz, E.A. Castro, The importance of the QSAR-QSPR methodology to the theoretical study of pesticides, International Journal of Chemical

Modeling 5 (1) (2013) 35.
[14] A. Mauri, V. Consonni, R. Todeschini, Molecular descriptors, Handbook of Computational Chemistry (2016) 1–29.
[15] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (3) (1996) 1027–1044.
[16] A.R. Katritzky, D.C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M. Karelson, The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors, Curr. Top. Med. Chem. 2 (12) (2002) 1333–1356.
[17] A.R. Katritzky, I.B. Stoyanova-Slavova, D.A. Dobchev, M. Karelson, QSPR modeling of flash points: an update, J. Mol. Graph. Model. 26 (2) (2007) 529–536.
[18] J.A. Morrill, E.F. Byrd, Development of quantitative structure property relationships for predicting the melting point of energetic materials, J. Mol. Graph. Model. 62 (2015) 190–201.
[19] S. Yousefinejad, B. Hemmateenejad, Chemometrics tools in QSAR/QSPR studies: a historical perspective, Chemometr. Intell. Lab. Syst. 149 (2015) 177–204.
[20] D.P. Visco Jr., R.S. Pophale, M.D. Rintoul, J.L. Faulon, Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor, J. Mol. Graph. Model. 20 (6) (2002) 429–438.
[21] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, vol. 11, John Wiley & Sons, New York, 2008.
[22] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, Chem. Rev. 110 (10) (2010) 5714–5789.
[23] M.D. Wessel, P.C. Jurs, Prediction of normal boiling points of hydrocarbons from molecular structure, J. Chem. Inf. Comput. Sci. 35 (1) (1995) 68–76.
[24] R. Zhang, S. Liu, M. Liu, Z. Hu, Neural network-molecular descriptors approach to the prediction of properties of alkenes, Comput. Chem. 21 (5) (1997) 335–341.
[25] W.A. Wakeham, G.S. Cholakov, R.P. Stateva, Liquid density and critical properties of hydrocarbons estimated from molecular structure, J. Chem. Eng. Data 47 (3) (2002) 559–570.
[26] A.A. Toropov, A.P. Toropova, QSPR modeling of alkanes properties based on graph of atomic orbitals, J. Mol. Struct.: Theochem 637 (1–3) (2003) 1–10.
[27] Y. Pan, J. Jiang, Z. Wang, Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, J. Hazard Mater. 147 (1–2) (2007) 424–430.
[28] M. Hamadache, A. Amrane, S. Hanini, O. Benkortbi, Multilayer perceptron model for predicting acute toxicity of fungicides on rats, International Journal of Quantitative Structure-Property Relationships (IJQSPR) 3 (1) (2018) 100–118.
[29] M. Hamadache, O. Benkortbi, S. Hanini, A. Amrane, L. Khaouane, C.S. Moussa, A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction, J. Hazard Mater. 303 (2016a) 28–40.
[30] M. Hamadache, S. Hanini, O. Benkortbi, A. Amrane, L. Khaouane, C.S. Moussa, Artificial neural network-based equation to predict the toxicity of herbicides on rats, Chemometr. Intell. Lab. Syst. 154 (2016b) 7–15.
[31] Y. Lahiouel, Industrial Experimental Dataset on Petroleum Fractions and Pure Hydrocarbons, Refinery of Skikda, Skikda, Algeria, 1995.
[32] KDB, Korean Thermophysical Properties Data Bank–cheric, 2017 (KDB) [Online]. Available: http://www.cheric.org/research/kdb/. (Accessed 2 December 2017).
[33] V.H. Masand, V. Rastija, PyDescriptor: a new PyMOL plugin for calculating thousands of easily understandable molecular descriptors, Chemometr. Intell. Lab. Syst. 169 (2017) 12–18.
[34] VCCLAB, Virtual Computational Chemistry Laboratory, 2005 [Online]. Available: http://www.vcclab.org. (Accessed 15 February 2018).
[35] StatSoft, STATISTICA (Data Analysis Software System) Vs 8.0, StatSoft Inc, Tulsa, OK, 2007.
[36] The MathWorks, Inc USA, Matlab Software, Version 2013b, 2013.
[37] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of size of training sets for the development of predictive QSAR models, Chemometr. Intell. Lab. Syst. 90 (1) (2008) 31–42.
[38] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (10) (2012) 2570–2578.
[39] M.D. Wessel, P.C. Jurs, Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks, Anal. Chem. 66 (15) (1994) 2480–2487.
[40] M.H. Beale, M.T. Hagan, H.B. Demuth, Neural network toolbox™ user's guide, in: R2012a, the MathWorks, Inc., 3 apple hill drive natick, MA 01760-2098, 2012. http://www.mathworks.com.
[41] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford university press, New York, USA, 1995.
[42] G.M. Pigram, T.R. MacDonald, Use of neural network models to predict industrial bioreactor effluent quality, Environ. Sci. Technol. 35 (1) (2001) 157–162.
[43] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (1) (2004) 1–12.
[44] T. Chai, R.R. Draxler, Root mean square error (*RMSE*) or mean absolute error (*MAE*)? –Arguments against avoiding *RMSE* in the literature, Geosci. Model Dev. (GMD) 7 (3) (2014) 1247–1250.
[45] M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecol. Model. 160 (3) (2003) 249–264.
[46] G.D. Garson, Interpreting neural-network connection weights, AI Expet. 6 (4) (1991) 46–51.
[47] A.T.C. Goh, Back-propagation neural networks for modeling complex systems, Artif. Intell. Eng. 9 (3) (1995) 143–151.
[48] L. Khaouane, Etude et modélisation de la biosynthèse des antibiotiques à partir de différentes souches productrices - cas de Pleuromutiline, PhD thesis, Université de Médéa, Algeria, 2013.
[49] Y. Ammi, L. Khaouane, S. Hanini, Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks, Kor. J. Chem. Eng. 32 (11) (2015) 2300–2310.
[50] S. Weaver, M.P. Gleeson, The importance of the domain of applicability in QSAR modeling, J. Mol. Graph. Model. 26 (8) (2008) 1315–1326.
[51] A. Hemmati-Sarapardeh, A. Varamesh, M.M. Husein, K. Karan, On the evaluation of the viscosity of nanofluid systems: modeling and data assessment, Renew. Sustain. Energy Rev. 81 (2018) 313–329.
[52] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (5) (2012) 4791–4810.
[53] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemometr. Intell. Lab. Syst. 145 (2015) 22–29.
[54] L. Zhou, B. Wang, J. Jiang, Y. Pan, Q. Wang, Predicting the gas-liquid critical temperature of binary mixtures based on the quantitative structure property relationship, Chemometr. Intell. Lab. Syst. 167 (2017) 190–195.
[55] K. Kraim, D. Khatmi, Y. Saihi, F. Ferkous, M. Brahimi, Quantitative structure activity relationship for the computational prediction of α-glucosidase inhibitory, Chemometr. Intell. Lab. Syst. 97 (2) (2009) 118–126.
[56] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs, Environ. Health Perspect. 111 (10) (2003) 1361.
[57] P. Gramatica, Principles of QSAR models validation: internal and external, Molecular Informatics 26 (5) (2007) 694–701.
[58] F. Sahigara, D. Ballabio, R. Todeschini, V. Consonni, Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective, Curr. Comput. Aided Drug Des. 10 (2) (2014) 137–147.
[59] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set descriptor space: a review, ATLA-NOTTINGHAM- 33 (5) (2005) 445.
[60] Y.M. Dai, Z.P. Zhu, Z. Cao, Y.F. Zhang, J.L. Zeng, X. Li, Prediction of boiling points of organic compounds by QSPR tools, J. Mol. Graph. Model. 44 (2013) 113–119.
[61] S. Saaidpour, F. Ghaderi, Quantitative modeling of physical properties of crude oil hydrocarbons using Volsurf+ molecular descriptors, Structure 22 (2016) 26.
[62] D. Sola, A. Ferri, M. Banchero, L. Manna, S. Sicardi, QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method, Fluid Phase Equil. 263 (1) (2008) 33–42.
[63] A. Varamesh, A. Hemmati-Sarapardeh, B. Dabir, A.H. Mohammadi, Development of robust generalized models for estimating the normal boiling points of pure chemical compounds, J. Mol. Liq. 242 (2017) 59–69.
[64] F. Gharagheizi, S.A. Mirkhani, P. Ilani-Kashkouli, A.H. Mohammadi, D. Ramjugernath, D. Richon, Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy: application to a very large dataset, Fluid Phase Equil. 354 (2013) 250–258.

Attestations, résumés et poster des conférences internationales avec comité de lecture

الندوة الدولية الثامنة للمحروقات و الكيمياء

The 8th International Symposium on Hydrocarbons and Chimistry

CERTIFICATE OF PARTICIPATION

We hereby certify that

Mohamed Roubehie Fissa

has attended the 8th International Symposium on Hydrocarbons and Chemistry (ISHC 8), which is organized from 8th to 10th in Boumerdes - ALGERIA; with communication entitled:

« QSPR-ANN Estimation models for critical pressure and critical volume of pure hydrocarbons using MLR and MLP-ANN methods»

الأستاذ مراد شارش
رئيس لجنة التنظيم

جامعة بومرداس
الندوة الدولية الثامنة للمحروقات و الكيمياء
كلية المحروقات و الكيمياء

8th International Symposium on Hydrocarbons and Chemistry ISHC8 2019, Boumerdès, 08-10 April, 2019

QSPR-ANN Estimation models for critical pressure and critical volume of pure hydrocarbons using MLR and MLP-ANN methods

Mohamed Roubehie Fissa [1,*], Yasmina Lahiouel [1], Latifa Khaouane [2] and Salah Hanini [2]

[1] Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, 24000 Guelma, Algérie.

[2] Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

[*] Corresponding Author: Email: roubehiefissa.mohamed@univ-guelma.dz

Abstract:

Hydrocarbons and petroleum fractions are considered as one of the energetic materials which it deals frequently in the petroleum industry field. Thermophysical properties such as the critical properties, can be considered as the key element in the development and control of this industry. Despite the progresses in this sector, the dataset on these properties are not currently available as needed. In this study, we have considered dealing with these issues. Two models have been developed in order to predict the critical pressure (Pc) and the critical volume (Vc) of pure hydrocarbons and petroleum fractions, based on their molecular structures, using an Artificial Neuron Network (ANN) method. Several steps were used during the development of the QSPR-ANN models. The effectiveness and accuracy were tested by different coefficient and error types. In order to appreciate the contribution of the input descriptors on each model we use the sensitivity analysis method. In addition, the models were also validated by using the Applicability Domain (AD) method.

Keywords: Process engineering, Energetic materials, Industrial operations, Pure hydrocarbons, Critical properties, QSPR-ANN models.

Modèles RQSP-RNA d'estimation de la pression critique et du volume critique des hydrocarbures purs en utilisant des méthodes RLM et RNA-PMC

Résumé :

Les hydrocarbures et les fractions pétrolières étant considérés comme les matériaux les plus énergétiques fréquemment traités dans le domaine d'industrie pétrolière. Les propriétés thermophysiques telles que les propriétés critiques, peuvent être considérées comme élément clé du développement et du contrôle de cette industrie. Malgré les progrès dans ce secteur, les données sur ces propriétés ne sont pas assez disponibles pour couvrir les besoins. Dans cette étude, nous voulons contribuer à faire face à ces problèmes. Deux modèles ont été développés afin de prédire la pression critique (Pc) et le volume critique (Vc) des hydrocarbures purs et des fractions pétrolières, basé sur leurs structures moléculaires, utilisant la méthode de Réseau de Neurones Artificiels (RNA). Plusieurs étapes ont été utilisées lors de l'élaboration des modèles RQSP-RNA. L'efficacité et la précision ont été testées par différents types de coefficient et d'erreurs. Afin d'apprécier la contribution des descripteurs d'entrée sur chaque modèle, nous utilisons la méthode d'analyse de sensibilité. Les modèles ont également été validés à l'aide de la méthode de Domaine d'Applicabilité (DA).

Mots clés : Génie des procédés, Matériaux énergétiques, Opérations industriel, Hydrocarbures purs, Propriétés critique, Modèles RQSP-RNA.

Summary of the work with the essential results, comments and conclusion

Two Quantitative Structure Property Relationship-Artificial Neuron Network (QSPR-ANN) models were developed, in order to predict the Pc and Vc of pure hydrocarbons and petroleum fractions, the model's steps development are: (1) Collection and pre-treatment of the database, over than 220 hydrocarbons compounds were used for each model, (2) Calculating thousands of molecular descriptors using E-Dragon online software, (3) Selection of the relevant descriptors by following a statistical methodology that contain several steps, the most important is that which based on the Multi Linear Regression (MLR) method, (4) Selection the best parameter for training the ANN in order to obtain the best QSPR-ANN models, (6) Analysis of the performance of each model by the Root Mean Squared Error (RMSE) and the Mean Percentage Error (MPE) in addition to the determination coefficient (R^2), (7) Sensitivity analysis for each model, and finally (8) Validation of the models by using the AD method. The QSPR-ANN model's results showed a good accuracy confirmed by the high R^2 value; that outweighed the **0.995** for both models, in the three-subset data (training, test, and overall). Moreover, the strength and predictive ability of the QSPR-ANN models was emphasized by two error types (RMSE and MPE); the result was very encouraging **0.048** ***MPa*** and **10.635** ***cm³/mol*** respectively, for the RMSE of Pc and Vc. Concerning the MPE values, are not exceed 2.5%, for both Pc and Vc. According the results of sensitivity analysis and the average values of contribution for each model, the descriptors contribution can be classified into three groups: **(1)** high contribution, ranging from 10 to 6% for Pc model and from 16 to 8.50% for Vc model, **(2)** medium contribution, ranging from 8.50 to 4% for Pc model and from 6 to 4.80% for Vc model, and finally **(3)** low contribution which is less than 4% for Pc model and less than 4.80% for Vc model. from this classification groups above, it should be noted that all selected descriptors were important in varying degrees in the development of the QSPR-ANN models. However, the AD results were so encouraging with 87.33% and 88.34% of the total data points within the scope of application, for Pc model and Vc model, respectively. Furthermore, the comparison between our models and the other QSPR models show that the QSPR-ANN models provided the better results. This computational approach can be applied in the petroleum engineering and other related engineering for an accurate determination of Pc and Vc of pure hydrocarbons and petroleum fractions, only from their molecular structure.

Bibliographies :

[1] Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120.

[2] Varamesh, A., Hemmati-Sarapardeh, A., Moraveji, M. K., & Mohammadi, A. H. (2017). Generalized models for predicting the critical properties of pure chemical compounds. *Journal of Molecular Liquids*, *240*, 777-793.

[3] Wakeham, W. A., Cholakov, G. S., & Stateva, R. P. (2002). Liquid density and critical properties of hydrocarbons estimated from molecular structure. *Journal of Chemical & Engineering Data*, *47*(3), 559-570.

[4] Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

[5] Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791-4810.

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Faculté des Sciences et de la Technologie
Département de Génie des Procédés
Université 08 mai 1945 – Guelma

Laboratoire des Silicates, polymères et des Nanocomposites

Université M'Hamed BOUGARA de Boumerdes –UMBB, Faculté des Hydrocarbures et de la Chimie

The 8th International Symposium on Hydrocarbons and Chemistry (ISHC8'2019) -
UMBB – Boumèrdes, on the 08th, 09th and 10th April 2019.

QSPR-ANN Estimation models for critical pressure and critical volume of pure hydrocarbons using MLR and MLP-ANN methods

Presented by : ➢ **Mr. ROUBEHIE FISSA. Mohamed[1,*]**

Co-Auteurs : ➢ **Pr. LAHIOUEL. Yasmina[1], Dr. KHAOUANE. Latifa[2] and Pr. HANINI. Salah[2]**

➢ *Authors Affiliation :*

[1]Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, Guelma 24000, Algérie.

[2]Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

➢ *Corresponding e-mail adresses: roubehiefissa.mohamed@univ-guelma.dz; mohfismasbioprophar@mail.com*



الندوة الدولية الثامنة للمحروقات و الكيمياء

The 8th International Symposium on Hydrocarbons and Chimistry

CERTIFICATE OF PARTICIPATION

We hereby certify that

Mohamed Roubehie Fissa

has attended the 8th International Symposium on Hydrocarbons and Chemistry (ISHC 8), which is organized from 8th to 10th in Boumerdes - ALGERIA; with communication entitled:

« QSPR-ANN and QSPR-MLR estimation models for normal boiling point temperature and critical temperature of pure hydrocarbons and petroleum fractions»

الأستاذ مراد قارش

رئيس لجنة التنظيم

جامعة بومرداس

الندوة الدولية الثامنة للمحروقات و الكيمياء

كلية المحروقات و الكيمياء

8th International Symposium on Hydrocarbons and Chemistry ISHC8 2019, Boumerdès, 08-10 April, 2019

QSPR-ANN and QSPR-MLR estimation models for normal boiling point temperature and critical temperature of pure hydrocarbons and petroleum fractions

Mohamed Roubehie Fissa [1,*], Yasmina Lahiouel [1], Latifa Khaouane [2] and Salah Hanini [2]

[1] Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, 24000 Guelma, Algérie.

[2] Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

[*] Corresponding Author: Email: roubehiefissa.mohamed@univ-guelma.dz

Abstract:

Like many thermophysical properties, the normal boiling point temperature (Tb) and the critical temperature (Tc) play a fundamental role in many engineering and industry fields, such as chemicals, petrochemicals, process engineering, and in petroleum industry as well. Where, they are important in production, separation and treatment of energetic materials. In order to predict the two properties (Tb and Tc) of pure hydrocarbons and petroleum fractions, four robust models were developed based on the molecular structures, through using a Multi Linear Regression (MLR) and an Artificial Neuron Network (ANN) methods, based on the Quantitative Structure Property Relationship (QSPR) approach. Each property has been predicted by two different expressions: QSPR-ANN and QSPR-MLR models. The efficiency and accuracy of the developed models was tested in terms of different coefficient and error types (R^2, RMSE and SEP). The results of the predictions were performed with satisfactory precision.

Keywords: Process engineering, Energetic materials, Pure hydrocarbons, Thermophysical properties, QSPR-ANN/QSPR-MLR models.

Modèles d'estimation RQSP-RNA et RQSP-RLM pour la température de point d'ébullition normale et la température critique des hydrocarbures purs et des fractions pétrolières

Résumé :

Comme beaucoup des propriétés thermophysiques, la température du point d'ébullition normale (Tb) et la température critique (Tc) jouent un rôle fondamental dans plusieurs domaines d'ingénierie et d'industrie, tels que la chimie, la pétrochimie, le génie des procédés et dans l'industrie pétrolière aussi. Où, ils sont importants dans la production, la séparation et le traitement des matériaux énergétiques. Afin de prédire les deux propriétés (Tb et Tc) des hydrocarbures purs et des fractions pétrolières, quatre modèles robustes ont été développés sur la base de structures moléculaires, en utilisant des méthodes de régression linéaire multiple (RLM) et de réseaux de neurones artificiels (RNA), basées sur l'approche Relation Quantitative Structure Propriété (RQSP). Chaque propriété a été prédite par deux expressions différentes : modèles RQSP-RNA et RQSP-RLM. L'efficacité et la précision des modèles développés ont été testées en termes de différent coefficient et types d'erreur (R^2, REQM, ESP). Les résultats des prédictions ont été réalisées avec une précision satisfaisante.

Mots clés : Génie des procédés, Matériaux énergétiques, Hydrocarbures purs, Propriétés thermophysiques, Modèles RQSP-RNA / RQSP-RLM.

Summary of the work with the essential results, comments and conclusion

In order to predict the Tb and Tc of pure hydrocarbons and petroleum fractions, four robust models were developed, by two different expressions in each parameter: Quantitative Structure Property Relationship-Artificial Neuron Network (QSPR-ANN) models and Quantitative Structure Property Relationship-Multi Linear Regression (QSPR-MLR) models. 223 data points were used, 80% of the total database was taken randomly as a training set and the rest, as a test set. The generating of Simplified Molecular Input Line Entry System (SMILES) notation were performed with "ChemDraw" software, and total of 1666 molecular descriptors were calculated using E-Dragon online software, so as to reduce them and retain just the relevant descriptors, a statistical reduction methodology was adopted. In which, about 99% of the primary molecular descriptors number was excluded, to get finally, just 16 and 10 relevant descriptors for Tb and Tc respectively. Thus, the BFGS (Quasi-Newton Back Propagation) algorithm has been applied to train the ANN in order to develop the QSPR-ANN models. On the other hand, the MLR forward stepwise regression method has been applied also to develop the QSPR-MLR models. It was confirmed that the proposed models give satisfactory results with high determination coefficient **(R^2)** values: **0.999** and ***0.998***, respectively for the QSPR-ANN and the *QSPR-MLR* models of ***Tb***, **0.995** and ***0.987***, respectively for the QSPR-ANN and the *QSPR-MLR* models of ***Tc***. The results of the Root Mean Square Error **(RMSE)** showed also good accuracy, with less than **3.2 K** for the both expressions models of **Tb**, and less than ***11.00 K*** for the both expression models of **Tc**. Finally, the Standard Error of Prediction **(SEP)** results confirmed the reliability of the proposed models with: **0.47**% and ***0.75%***, respectively for the QSPR-ANN and the *QSPR-MLR* models of ***Tb***; **1.15%** and ***1.81%***, respectively for the QSPR-ANN and the *QSPR-MLR* models of ***Tc***. As provided, the previous results confirmed the satisfactory ability of the developed models in prediction of *Tb* and *Tc* of pure hydrocarbons and petroleum fractions, with slightly higher performance for the **QSPR-ANN** models compared with those of the **QSPR-MLR** models in both parameters. This computational approach can be applied in the field of industry and petroleum engineering, and also in other fields related to the pure hydrocarbons for accurate determination of *Tb* and *Tc* of pure hydrocarbons and petroleum fractions, only from their molecular structures using some simple calculable molecular descriptors.

Bibliographies :

[1] Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120.

[2] Saaidpour, S., & Ghaderi, F. (2016). Quantitative Modeling of Physical Properties of Crude Oil Hydrocarbons Using Volsurf+ Molecular Descriptors. *structure*, *22*, 26.

[3] Yao, X., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2002). Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemometrics and Intelligent Laboratory Systems*, *62*(2), 217-225.

[4] Varamesh, A., Hemmati-Sarapardeh, A., Dabir, B., & Mohammadi, A. H. (2017). Development of robust generalized models for estimating the normal boiling points of pure chemical compounds. *Journal of Molecular Liquids*, *242*, 59-69.

[5] Gharagheizi, F., Mirkhani, S. A., Ilani-Kashkouli, P., Mohammadi, A. H., Ramjugernath, D., & Richon, D. (2013). Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy: application to a very large dataset. *Fluid Phase Equilibria*, *354*, 250-258.

The 8th International Symposium on Hydrocarbons and Chemistry ISHC8

QSPR-ANN and QSPR-MLR estimation models for normal boiling point temperature and critical temperature of pure hydrocarbons and petroleum fractions

Mohamed Roubehie Fissa [1,*], Yasmina Lahiouel [1], Latifa Khaouane [2] and Salah Hanini [2]

[1] *Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, 24000 Guelma, Algérie.*

[2] *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.*

1. Introduction

The thermophysical properties like the normal boiling point temperature (Tb) and the critical temperature (Tc), play a fundamental role in many engineering and industry fields, such as chemicals, petrochemicals, process engineering, and in the petroleum industry as well [1-5]. Where, they are important in production, separation and treatment of energetic materials [1,2]. In this study, four robust models were developed based on the QSPR approach, in order to predict Tb and Tc of pure hydrocarbons and petroleum fractions. Each property has been predicted by two different expressions: QSPR-ANN and QSPR-MLR models, in which two computational methods were used for each model: a Multi Linear Regression (MLR) and an Artificial Neuron Network (ANN).

2. Methodology

Several steps were used in the development of the QSPR models, as described below:

- Collection of 223 hydrocarbons molecules from several sources to use It as a database.
- Calculation of 1666 molecular descriptors, using E-Dragon software.
- Reduction of the molecular descriptors, by a statistical methodology in order to select the relevant molecular descriptors. The mean is the MLR and intercorrelations steps [1].
- Construction of the QSPR models, For the QSPR-ANN models, 80% of the dataset was used for training and the rest for a test, several activation functions were used, the BFGS was used as training, SOS was used also as errors function. For the QSPR-MLR models. The dataset was splitting in the same way and the MLR forward stepwise regression method was used. All models were performed using STATISTICA Software.
- The performances the developed models were tested in terms of a different coefficient and error types (R^2, RMSE and SEP).

4. Conclusion

In this study, four robust models were developed with two different expressions: QSPR-ANN and QSPR-MLR models, in order to predict Tb and Tc of pure hydrocarbons and petroleum fractions, the efficiency and accuracy of the developed models were tested in terms of a different coefficient and error types (R2, RMSE and SEP). The results of the predictions were performed with satisfactory precision. This computational approach can be applied in the field of industry and petroleum engineering, and also in other fields related to the pure hydrocarbons for accurate determination of Tb and Tc of pure hydrocarbons and petroleum fractions, only from their molecular structures using some simple calculable molecular descriptors.

Acknowledgment

The authors are grateful to the Organizing Committee of ISHC8_2019, UMBB-BOUMERDES, for their efforts in this event.

3. Results and discussions

3.1. Evolution of descriptor's numbers

In order to reduce the number of descriptors, we follow a statistical methodology contain several steps, the evolution of the number of descriptors according to the pre-treatment procedure for each parameter are represented in Table 1.

Table 1. Evolution of the number of descriptors according to the pre-treatment procedure for each parameter

Steps	(0)	(1)	(2)	(3)		(4)	
				Tb	Tc	Tb	Tc
Nbr Des	1666	1105	1001	89	90	16	10

At the end of the selection, about 99% of the number of primary molecular descriptors was excluded, with only 16 and 10 relevant descriptors taken for Tb and Tc, respectively.

3.2. Mathematical definitions for each model

The architectures of the QSPR-ANN models are: (16-12-1) for Tb with "Exponential" and "Tanh" as activation functions respectively, for Hidden (f^H) and Output (f^O) layers, (10-16-1) for Tc with "Tanh" as activation function in both layers (See Eq. (1) and Eq. (2)).

$$e^{-n} \quad (1); \qquad \frac{e^{+n}-e^{-n}}{e^{+n}+e^{-n}} \quad (2);$$

Each hidden neuron computes activation function, sending results to the output layer's neuron which finally gives the response of the network. The output signal of each hidden neuron (Z_j) is calculated by the **Eq. (3)** for *Tb* and *Tc* models. The output of the network (S) is given by **Eq. (4)**. Where n and k are respectively the total number of input and hidden neurons, E_i are the inputs neurons, w_{ij}^I and w_{1j}^H are the weights of input and hidden layers, b_j^H and b_1^O the bias of hidden and output layers.

$$Z_j = f^H\left[\sum_{i=1}^{n} w_{ij}^I E_i + b_j^H\right] \quad (3);$$

$$S = f^O\left[\sum_{j=1}^{k} w_{1j}^H Z_j + b_1^O\right] \quad (4);$$

$$Y = A_0 + \sum_{i=1}^{n} A_i E_i \quad (5);$$

The QSPR-MLR models were developed by applying the MLR forward regression method. MLR consists of a quantitative relationship between a group of predictor variables (E_i) in this case are the relevant descriptors and a response (Y) in this study are the *Tb* and *Tc*, as shown in **Eq. (5)**.

where A_i and A_0 are respectively the coefficients of the selected descriptors and the intercepts of each QSPR-MLR model.

3.3. Estimation of Tb and Tc using the QSPR models

In order to estimate Tb and Tc, we use the previous equations (Eq. (4), Eq. (6) and Eq. (7)) by introducing the values of the inputs and all other coefficients. The visual comparison between the experimental and calculated values obtained by using the developed QSPR-ANN models shown in **Fig. 1a** and in **Fig. 1b** respectively, for Tb and Tc. As well the QSPR-MLR models shown in **Fig. 1c** and in **Fig. 1d** respectively, for Tb and Tc. And this for the whole dataset in both models.

A tight concentration of data points around the unit-slope line (shown by the dashed line), reveals that the QSPR-ANN and the QSPR-MLR models can accurately predict the Tb and Tc values. It was confirmed that the proposed models give satisfactory results, through giving an accuracy agreement with the vector values approaching the ideal (i.e. α = 1 (slope), β = 0 (y intercept), R^2 = 1)

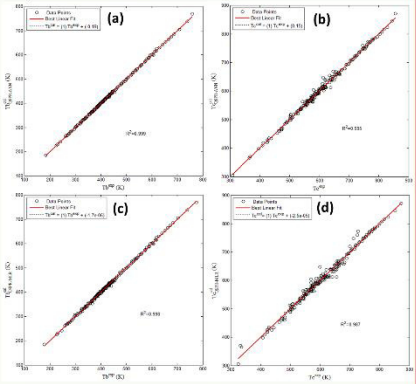


Fig. 1 The experimental and calculated values of Tb and Tc obtained by using the developed QSPR models

3.4. Analysis of the model's performances

The performances of the proposed models are tested by the R2, RMSE and SEP. The **fig 2a** and **fig 2b** show the RMSE and SEP of the QSPR-ANN and the QSPR-MLR models for both properties (Tb and Tc). The results were very encouraging, with RMSE values less than 3.2 K for Tb and less than 11.00 K for Tc, in both expression models. Finally, the SEP results confirmed the reliability of the proposed models with 0.47% and 0.75%, respectively for the QSPR ANN and the QSPR-MLR models of Tb; 1.15% and 1.81%, respectively for the QSPR-ANN and the QSPR-MLR models of Tc.

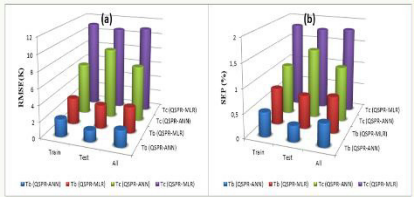


Fig. 2 The values of RMSE and SEP for the two expression models in both parameter.

Note: Refer to the full paper in the symposium proceedings for references [1-5].

Attestations, résumés, posters et support d'article des conférences et séminaires nationales avec comité de lecture

جامعة 8 ماي 1945 قالمة
UNIVERSITE 8 MAI 1945 GUELMA

2nd Conference on Informatics and Applied Mathematics

LabSTIC Université 8 Mai 1945

IAM'2019

Guelma, Algeria, June 12-13, 2019

CERTIFICATE OF PARTICIPATION

The general chair of the 2nd Conference on Informatics and Applied Mathematics (IAM'2019) is pleased to certify that:

Mohamed Roubehie FISSA

Presented a poster communication entitled:

"Comparative study between QSPR-MLP, QSPR-SVR and QSPR-MLR models for estimation of critical temperature of pure hydrocarbons and petroleum fractions"

Author (s): Mohamed Roubehie Fissa, Yasmina Lahiouel, Latifa Khaouane and Salah Hanini

2nd Conference IAM 2019 Informatics and Applied Mathematics

General Chair

Dr. Mohamed Nadjib KOUAHLA

Dr. Mohamed Nadjib KOUAHLA
Enseignant Chercheur, LabSTIC
Université 8 Mai 1945 de Guelma

2ND CONFERENCE ON INFORMATICS AND APPLIED MATHEMATICS (IAM CONFERENCE 2019)

Comparative study between QSPR-MLP, QSPR-SVR and QSPR-MLR models for estimation of critical temperature of pure hydrocarbons and petroleum fractions

Mohamed Roubehie Fissa[1,*], Yasmina Lahiouel[1], Latifa Khaouane[2], and Salah Hanini[2].
[1]Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, Guelma 24000, Algérie.
[2]Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

[*]Corresponding author, e-mail addresses: roubehiefissa.mohamed@univ-guelma.dz; mohfismasbioprophar@gmail.com.

Abstract

In this study, three models based on the Quantitative Structure Property Relationship (QSPR) approach have been developed, in order to estimate the critical temperature of pure hydrocarbons and petroleum fractions. Three computational regression methods were chosen in each model: multilayer perceptron network (MLP), support vector regression (SVR) and multiple linear regression (MLR). Thousands of descriptors were calculated, then just 10 relevant descriptors were selected to develop each QSPR model. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) was adopted as a training algorithm in the QSPR-MLP model, and different activation functions were tested in both layers. The "BuildQSAR" software was used in the development of the QSPR-MLR model. Concerning the QSPR-SVR model, the ε-SVR type was chosen with Gaussian radial basis function (RBF) as kernel function, the values of the width (γ) and the capacity (C) parameters were respectively tested from 0.1 to 2.5 and from 1 to 25 with incremental steps of 0.1 in both parameters, the tube size (ε) was ranging from 0.01 to 0.5 with incremental steps of 0.01. The performances of the proposed models are tested by the correlation coefficient (R), the root mean square error (RMSE) and the mean absolute percentage error (MAPE). The results show that all the proposed models are reliable and acceptable in prediction of Tc of hydrocarbons, so that all the R values in the three models are outweighed the 0.980. In addition, the errors were very acceptable indeed with 6.967 K, 15.179 K and 10.987 K respectively, for the RMSE values of QSPR-MLP, QSPR-SVR and QSPR-MLR. The MAPE values are confirmed the preceding error results, in which they did not exceed the 1.5 % in all models. It should be noted that the QSPR-MLP model is superior to the rest of the models, but this does not mean the inefficiency of the other two models, given the nature of each model, particularly the large number of choices like the types of Kernel in the QSPR-SVR, and without forgetting the *nu*-SVR type.

Keywords: Artificial Intelligence, Computational Methods, Hydrocarbons, QSPR Models, Critical Temperature, Descriptors.

***IAM CONFERENCE 2019*, 24-25 APRIL 2019 / GUELMA UNIVERSITY**

2nd Conference on Informatics and Applied Mathematics
IAM'2019_24-25 April 2019

Comparative study between QSPR-MLP, QSPR-SVR and QSPR-MLR models for estimation of critical temperature of pure hydrocarbons and petroleum fractions

Mohamed Roubehie Fissa [1,*], Yasmina Lahiouel [1], Latifa Khaouane [2] and Salah Hanini [2]

[1] *Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, 24000 Guelma, Algérie.*
[2] *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.*

1. Introduction

The critical properties are widely used in the chemical industry to understand the thermodynamic behaviour of pure compounds or their mixtures [1,2]. In this study, three models based on the Quantitative Structure Property Relationship (QSPR) approach [1,3] have been developed, in order to estimate the critical temperature (Tc) of pure hydrocarbons and petroleum fractions. Three computational regression methods were chosen in each model: multilayer perceptron network (MLP), support vector regression (SVR) and multiple linear regression (MLR). Thousands of descriptors were calculated, then just 10 relevant descriptors were selected to develop each QSPR model [3].

2. Methodology

Several steps were used in the development of the QSPR models, as described in the flowchart below:

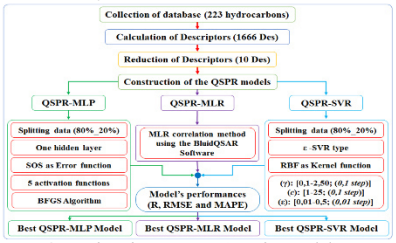


Fig. 1 Flowchart summarizes the modeling methodology

3. Results and discussions

3.1. The reduction methodology of molecular descriptors

In order to reduce the number of descriptors, we follow a statistical methodology contain several steps [3]: (1): Omitted the errors value (-999) and the repeated value more than (> 75%); (2): Removed the absolute value of RSD less than (<0.05); (3): Deleting all the high intercorrelation; (4): use a forward stepwise MLR method. At the end of the selection, about 99% of the primary number (1666) of descriptors was excluded, and only 10 relevant descriptors was taken as input in the QSPR models.

3.2. Mathematical definitions for each model

The architecture of the best QSPR-MLP model is: (10-16-1), with "Tahh" as activation function in both layers Hidden (f^H) and Output (f^O) (see **Eq. (1)**), The mathematical definition of the QSPR-MLP model is given by **Eq. (2)**. Where n and k are respectively the total number of input and hidden neurons, E_i are the inputs neurons, w_{ij}^I and w_{1j}^H are the weights of input and hidden layers, b_j^H and b_1^O the bias of hidden and output layers.

$$f^H(x) = f^O(x) = \frac{e^{+x}-e^{-x}}{e^{+x}+e^{-x}} \quad (1); \qquad Tc_{QSPR_MLP} = f^O\left[\sum_{j=1}^{k} w_{1j}^H \left(f^H\left[\sum_{i=1}^{n} w_{ij}^I E_i + b_j^H\right]\right) + b_1^O\right] \quad (2);$$

The QSPR-MLR models were developed by applying the MLR forward regression method. MLR consists of a quantitative relationship between a group of predictor variables (E_i) and a response (Tc), as shown in **Eq. (3)**. where A_i and A_0 are respectively the coefficients of the selected descriptors and the intercepts of the QSPR-MLR model.

$$Tc_{QSPR_MLR} = A_0 + \sum_{i=1}^{n} A_i E_i \quad (3); \qquad Tc_{QSPR_SVR} = w^T(x)\,\varphi(x) + b \quad (4);$$

The mathematical definition of the QSPR-SVR model is shown in **Eq. (4)**. Where w^T is the transposed vector of output layer, $\varphi(x)$ is kernel function (See **Eq. (5)**) and b is bias function. In order to calculate w and b parameters, we use the cost function, represented in **Eq. (6)**;

$$\varphi(x_i, x_j) = \sum_{j=1}^{k}\sum_{i=1}^{n} exp\left(\frac{-\|x_i - x_j\|^2}{(2*\gamma_j)^2}\right) \quad (5);$$

$$Fcost = \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \quad (6);$$

Which minimized subject to:

$$w^T\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$
$$y_i - w^T\varphi(x_i) - b \le \varepsilon + \xi_i$$
$$\xi_i, \xi_i^* \ge 0, i = 1, \dots, n$$

where x_i is input of i^{th} data and y_i refers to its output. ε is the precision of the function approximations (the tube size), and slack variables ($\xi_i and\ \xi_i^*$) are the acceptable margin of error. C is the capacity constant. (γ_j) represents the kernel width (an adjustable parameter of RBF).

The best QSPR-SVR model was obtained with γ equal to 0.1 (with C = 2.2 and ε = 0.02).

3.3. Estimation of Tc using the QSPR-MLP, QSPR-SVR and QSPR-MLR models

In order to estimate Tc, we use the previous equations. The visual comparison between the experimental and calculated values of *Tc* obtained by using the developed models: QSPR-MLP model shown in **Fig. 2a**, QSPR-SVR model shown in **Fig. 2b**, and QSPR-MLR model shown in **Fig. 2c** And this for the whole dataset.

The **Fig. 2** confirmed that the proposed models give satisfactory results, through giving an accuracy agreement with the vector values approaching the ideal (i.e. α = 1 (slope), β = 0 (γ intercept), R = 1).

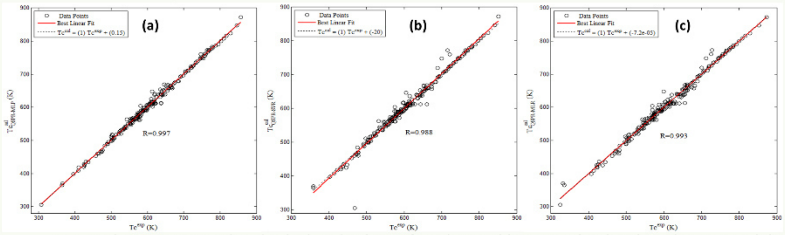


Fig. 2 The experimental and calculated values of Tc obtained by using the developed QSPR models

3.4. Analysis of the model's performances

The performances analysis values of the proposed models are shown in **table 2**. The results of the predictions were performed with satisfactory precision. The R values in the three models are outweighed the 0.980. In addition, the errors were very acceptable indeed, between 6.967-15.179 K for the RMSE values of three models. The MAPE values are confirmed the preceding error results, in which they did not exceed the 1.5 % in all models.

Table 2. Statistical parameters of the best QSPR models for Tc.

Parameters	R	RMSE (K)	MAPE (%)	N
$Tc^{QSPR\text{-}MLP\ (all)}$	0.9974	6.9672	0.7927	223
$Tc^{QSPR\text{-}SVR\ (all)}$	0.9880	15.1790	1.3249	223
$Tc^{QSPR\text{-}MLR\ (all)}$	0.9935	10.9875	1.2654	223

From these results, It should be noted that the QSPR-MLP model is superior to the rest of the models, but this does not mean the inefficiency of the other two models, given the nature of each model, particularly the large number of choices like the types of Kernel in the QSPR-SVR, and without forgetting the *nu*-SVR type.

4. Conclusion

In study, three models were developed with three different expressions: QSPR-MLP, QSPR-SVR and QSPR-MLR, in order to predict Tc of pure hydrocarbons and petroleum fractions, the efficiency and accuracy of the developed models were tested in terms of a different coefficient and error types (R, RMSE and MAPE). The results of the predictions were performed with satisfactory precision. This computational approach can be applied in the field of industry and petroleum engineering, and in other related fields for accurate determination of Tc of pure hydrocarbons and petroleum fractions. only from their molecular structures.

References

[1] Yao, X., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2002). Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemometrics and Intelligent Laboratory Systems, 62*(2), 217-225.
[2] Halali, M. A., Azari, V., Arabloo, M., Mohammadi, A. H., & Bahadori, A. (2016). Application of a radial basis function neural network to estimate pressure gradient in water–oil pipelines. Journal of the Taiwan Institute of Chemical Engineers, 58, 189-202.
[3] Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling, 87*, 109-120.

Acknowledgment: The authors are grateful to the Organizing Committee of IAM_2019, University 08 Mai 1945 GUELMA, for their efforts in this event.

IAM'2019

جامعة 8 ماي 1945 قالمة
UNIVERSITE 8 MAI 1945 GUELMA

2nd Conference on Informatics and Applied Mathematics

LabSTIC
Université 8 Mai 1945

Guelma, Algeria, June 12-13, 2019

CERTIFICATE OF PARTICIPATION

The general chair of the 2nd Conference on Informatics and Applied Mathematics (IAM'2019) is pleased to certify that:

Mohamed Roubehie FISSA

Presented a poster communication entitled:

"QSPR based models for estimation of normal boiling point temperature of pure hydrocarbons and petroleum fractions using RBF-ANN, MLP-ANN and MLR Methods"

Author (s): Mohamed Roubehie Fissa, Yasmina Lahiouel, Latifa Khaouane and Salah Hanini

2nd Conference
IAM 2019
Informatics and Applied Mathematics

وزارة التعليم العالي والبحث العلمي
مخبر العلوم وتكنولوجية الإعلام والاتصال
جامعة 08 ماي 1945 قالمة

General Chair

Dr. Mohamed Nadjib KOUAHLA

Dr. Mohamed Nadjib KOUAHLA
Enseignant Chercheur, LabSTIC
Université 08 Mai 1945 de Guelma

2ND CONFERENCE ON INFORMATICS AND APPLIED MATHEMATICS (IAM CONFERENCE 2019)

QSPR based models for estimation of normal boiling point temperature of pure hydrocarbons and petroleum fractions using RBF-ANN, MLP-ANN and MLR Methods

Mohamed Roubehie Fissa[1,*], Yasmina Lahiouel[1], Latifa Khaouane[2], and Salah Hanini[2].
[1]Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, Guelma 24000, Algérie.
[2]Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

[*]Corresponding author, e-mail addresses: roubehiefissa.mohamed@univ-guelma.dz; mohfismasbioprophar@gmail.com.

Abstract

Quantitative Structure Property Relationship (QSPR) methods, are among the well-known computational methods that resounding success in several areas, such as in chemistry and biology fields, especially when it combines with artificial intelligence methods. In this communication, a QSPR: radial basis function- artificial neural network (RBF-ANN) and multilayer perceptron-artificial neural network (MLP-ANN) models, have been developed in order to predict the boiling point temperature (Tb) of hydrocarbons and petroleum fractions. 223 hydrocarbons molecules were used as dataset. 80% of this data were taken as training set and the rest as test set. 1666 molecular descriptors were calculated online by E-Dragon software. A statistical reduction methodology was adopted, in order to select just the relevant descriptors, the main step was that of the multi linear regression (MLR) method. The "Broyden Fletcher Goldfarb Shanno" (BFGS) and the "Reduced Breadth First Search" (RBFT) algorithms were applied in order to train the ANNs, the sum of square "SOS" was chosen in both models as error function. The topology of the best models is: (16-16-1) for the QSPR-RBF model and (16-12-1) for the QSPR-MLP model. The robustness and accurateness of the developed models was tested in terms of the determination coefficient (R^2), the root mean square error (RMSE) and the standard error of prediction (SEP), The results of the predictions were performed with satisfactory precision. In which: ($R^2 > 0.960$, RMSE < 18 K and SEP < 5%) for the QSPR-RBF model and ($R^2 > 0.990$, RMSE < 2.5 K and SEP < 1%) for the QSPR-MLP model. From these preceding results, it should be noted that the QSPR-MLP model outperforms the QSPR-RBF model in the prediction of Tb of hydrocarbons. This computational approach can be applied in the field of industry and petroleum engineering, and in other related fields for accurate determination of Tb of pure hydrocarbons and petroleum fractions.

Keywords: Artificial Intelligence, Computational Intelligence, Pure Hydrocarbons, QSPR-ANN Models, Normal Boiling Points Temperature, Descriptors.

2nd Conference on Informatics and Applied Mathematics
IAM'2019_24-25 April 2019

QSPR based models for estimation of normal boiling point temperature of pure hydrocarbons and petroleum fractions using RBF-ANN, MLP-ANN and MLR Methods

Mohamed Roubehie Fissa [1,*], Yasmina Lahiouel [1], Latifa Khaouane [2] and Salah Hanini [2]

[1] *Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, 24000 Guelma, Algérie.*
[2] *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.*

1. Introduction

Quantitative Structure Property Relationship (QSPR) methods, are among the well-known computational methods that resounding success in several areas, such as in chemistry and biology fields, especially when it combines with artificial intelligence methods. The thermophysical properties like the normal boiling point temperature (Tb), play a fundamental role in many engineering and industry fields [1-3]. where, they are important in production, separation and treatment of materials [1,2]. In this communication, a QSPR: radial basis function- artificial neural network (RBF-ANN) and multilayer perceptron- artificial neural network (MLP-ANN) models, have been developed in order to predict the Tb of pure hydrocarbons and petroleum fractions.

2. Methodology

Several steps were used in the development of the QSPR models, as described in the flowchart below:

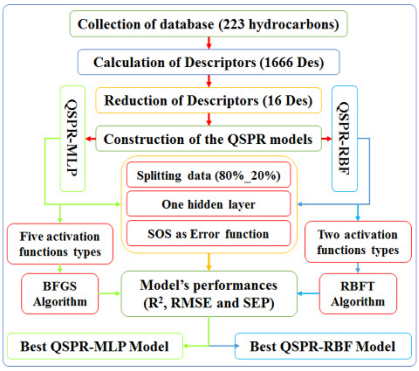


Fig. 1 Flowchart summarizes the modeling methodology

4. Conclusion

In this study, two robust models were developed with two different expressions: QSPR-MLP and QSPR-RBF models, in order to predict Tb of pure hydrocarbons and petroleum fractions, the efficiency and accuracy of the developed models were tested in terms of a different coefficient and error types (R^2, RMSE and SEP). The results of the predictions were performed with satisfactory precision. This computational approach can be applied in the field of industry and petroleum engineering, and in other related fields for accurate determination of Tb of pure hydrocarbons and petroleum fractions. only from their molecular structures using some simple molecular descriptors.

References

[1] Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120.

[2] Yao, X., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2002). Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemometrics and Intelligent Laboratory Systems*, 62(2), 217-225.

[3] Varamesh, A., Hemmati-Sarapardeh, A., Dabir, B., & Mohammadi, A. H. (2017). Development of robust generalized models for estimating the normal boiling points of pure chemical compounds. *Journal of Molecular Liquids*, *242*, 59-69.

Acknowledgment

The authors are grateful to the Organizing Committee of IAM_2019, University 08 Mai 1945 GUELMA, for their efforts in this event.

3. Results and discussions

3.1. Evolution of descriptor's numbers

In order to reduce the number of descriptors, we follow a statistical methodology contain several steps, the evolution of the number of descriptors according to the pre-treatment procedure are represented in Table 1.

Table 1. Evolution of the number of descriptors according to the pre-treatment procedure

Steps	(0)	(1)	(2)	(3)	(4)
Nbr Des	1666	1105	1001	89	16

(0): After calculated by E-Dragon; (1): Omitted the errors value (-999) and the repeated value more than (> 75%); (2): Removed the absolute value of RSD less than (<0.05); (3): Deleting all the high intercorrelation; (4): use a forward stepwise MLR method.

At the end of the selection, about 99% of the number of primary molecular descriptors was excluded, with only 16 relevant descriptors taken for Tb.

3.2. Mathematical definitions for each model

The architectures of the best QSPR models are: (16-12-1) for the QSPR-MLP model and (16-16-1) for the QSPR-RBF model. With "Exponential" and "Gaussian" as Hidden (f^H) activation functions (See Eq. (1-2)), "Tanh" and "Identity" as Output (f^O) activation functions (See Eq. (3-4)), respectively for each model.

$$f^H(x) = e^{-x} \quad (1); \qquad f^H(x) = e^{-x^2} \quad (2);$$

$$f^O(x) = \frac{e^{+x}-e^{-x}}{e^{+x}+e^{-x}} \quad (3); \qquad f^O(x) = x \quad (4);$$

The mathematical definition of the QSPR-MLP model is given by **Eq. (5)**. Where $\boldsymbol{n}$ and $\boldsymbol{k}$ are respectively the total number of input and hidden neurons, $\boldsymbol{E_i}$ are the inputs neurons, $\boldsymbol{w_{ij}^I}$ and $\boldsymbol{w_{ij}^H}$ are the weights of input and hidden layers, $\boldsymbol{b_j^H}$ and $\boldsymbol{b_1^O}$ the bias of hidden and output layers.

$$Tb_{QSPR_MLP} = f^O\left[\sum_{j=1}^{k} w_{1j}^H \left(f^H\left[\sum_{i=1}^{n} w_{ij}^I E_i + b_j^H\right]\right) + b_1^O\right] \quad (5);$$

The mathematical definition of the QSPR-RBF model is shown in **Eq. (6)**. Where $||E_i - C_j||$ the Euclidian distance between input data (E_i) and the radial function centre (C_j), (σ_j) represent the width parameter (always $\sigma_j > 0$). The parameters (C_j,σ_j), must be definite separately at each RBF unit.

$$Tb_{QSPR_RBF} = f^O\left[\sum_{j=1}^{k} w_{1j}^H \left(f^H\left[\sum_{i=1}^{n} \frac{||E_i - C_j||}{2*\sigma_j}\right]\right) + b_1^O\right] \quad (6);$$

3.3. Estimation of Tb using the QSPR-MLP and QSPR-RBF models

In order to estimate Tb, we use the previous equations. The visual comparison between the experimental and calculated values of *Tb* obtained by using the developed QSPR-MLP model shown in **Fig. 2a**, As well as by the QSPR-RBF model shown in **Fig. 2b**. And this for the whole dataset.

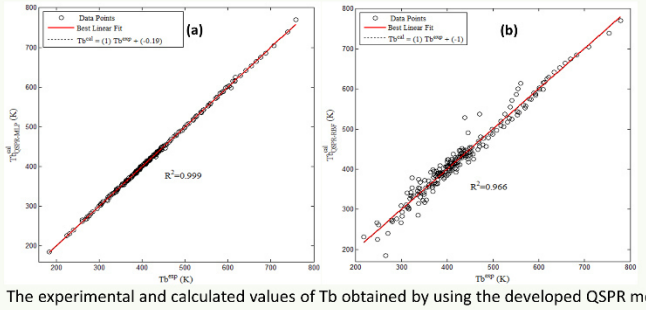


Fig. 2 The experimental and calculated values of Tb obtained by using the developed QSPR models

The **Fig. 2** confirmed that the proposed models give satisfactory results, through giving an accuracy agreement with the vector values approaching the ideal (i.e. α = 1 (slope), β = 0 (γ intercept), R^2 = 1).

3.4. Analysis of the model's performances

The performances of the proposed models are tested by the RMSE shown in **fig 3a** and by the SEP shown in **fig 3b**.

The results of the predictions were performed with satisfactory precision. In which: (RMSE < 2.5 K and SEP < 1%) for the QSPR-MLP model and (RMSE < 18 K and SEP < 5%) for the QSPR-RBF model.

From these results, it should be noted that the QSPR-MLP model outperforms the QSPR-RBF model in the prediction of Tb of hydrocarbons.

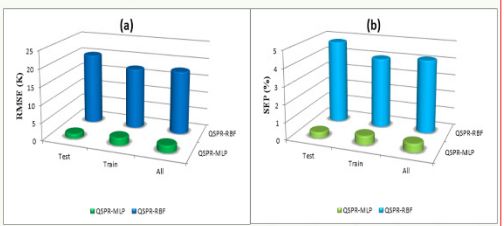


Fig. 3 The values of RMSE and SEP for the both expression models.

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
High School of Technological Teaching Enset-Skikda

CERTIFICATE OF ATTENDANCE

Ref. :94/DPGRS/ENSET/2018

IS HEREBY GRANTED TO

Mohamed Roubehie Fissa

AUTHOR(S):MOHAMED ROUBEHIE FISSA, YASMINA LAHIOUEL, LATIFA KHAOUANE AND SALAH HANINI

PAPER # 73

Robust QSPR-ANN Estimation Model for Normal Boiling Points Temperature of Pure Hydrocarbons and Petroleum Fractions.

FOR PARTICIPATION AS «ORALE » PRESENTATION IN

1er Séminaire National de physique et chimie des matériaux

November 28-29, 2018 – ENSET – Azzaba-Skikda

Coordinatrice du Séminaire
Dr BEDBOUDI Hayette

Robust QSPR-ANN Estimation Model for Normal Boiling Point Temperature of Pure Hydrocarbons and Petroleum Fractions

Mohamed Roubehie Fissa[1,*], Yasmina Lahiouel[1].
[1]Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN).
Université 8 mai 1945 Guelma.
Guelma 24000, Algérie.
[*]Corresponding author, e-mail addresses:
roubehiefissa.mohamed@univ-guelma.dz;
mohfismasbioprophar@gmail.com

Latifa Khaouane[2], Salah Hanini[2].
[2]Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT).
Université Yahia Farès de Médéa.
Médéa 26000, Algérie.

***Abstract*—This work aimed to predict the normal boiling point temperature (*Tb*) of petroleum fractions and pure hydrocarbons, by using a multi-layer perceptron artificial neural network (MLP-ANN) based on the molecular descriptors. A set of 223 different data points from saturated and unsaturated hydrocarbons representing main industrially important groups was used to build a quantitative structure-property relationship artificial neuron network (QSPR-ANN) model. The total database was divided randomly into two subsets: 80% for the training set and 20% for the test set. 1666 descriptors were calculated, in order to reduce and select this huge number of molecular descriptors, and retain acceptable number of molecular relevant descriptors. A statistical reduction methodology based on the multi linear regression (MLR) method has been adopted, and 99% of the initial molecular descriptors were excluded through the use of the adopted reduction methodology. Thus, the Quasi-Newton back propagation (Broyden-Fletcher-Goldfarb-Shanno (BFGS)) algorithm was applied to train the ANN. The architecture of the developed QSPR-ANN model is: (16) input neurons, (12) hidden neurons, and (1) output neuron. The QSPR-ANN model's results showed a good accuracy confirmed, by the high values of determination coefficient (R^2); that outweighed the 0.999, and the low values: of root mean square error (*RMSE*) and standard error of prediction (*SEP*), which are respectively, lower than 2.200 K and 1%. The weight sensitivity analysis method has been used, in order to know the contribution of the input descriptors on our QSPR-ANN model. Almost all the selected descriptors used to build the QSPR-ANN model has good contribution percentage; which confirmed the importance of all descriptors used in this study. Besides that, the applicability domain (AD) method was used with strict standard residual values ($h^* = \pm 2$). However, the results were so encouraging with more than 90 % of the total data points within the scope of application. This computational approach can be applied in the petroleum engineering for a precise determination of *Tb* of pure hydrocarbons and petroleum fractions.**

***Keywords*—Computational Approach, Pure Hydrocarbons, Petroleum Fractions, QSPR-ANN model, Normal Boiling Point, Descriptors.**

I. INTRODUCTION

The prediction of physicochemical properties such as boiling point temperature (*Tb*) is a fundamental goal of computational chemistry. Nonetheless, it can be considered one of the most important and necessary keys for the development and control of the petroleum industry [1]. Since it can explain a lot of phenomena that can help the industrialists and researchers both, to know and understand the nature of their materials and then, to find and apply the perfect ways to transport and transform them [2,3]. The *Tb* can be defined as the temperature at which the vapor pressure of a pure liquid reaches 760 mmHg [4,5]. Despite the progress of studies in this area the *Tb* of hydrocarbons and petroleum fractions, as other physicochemical properties, is still not available as needed, due to several reasons. The expensive and time-demanding laboratory testing are the most reasons [6,7]. It is obvious to note that generally the *Tb* of a compound is related to its molecular structure, but the nature of the relationship is subtle and often difficult to predict [4,5].

In recent years, the QSAR/QSPR approaches have become the most used areas in the development of mathematical models designed to predict the physicochemical properties of different chemical species [8,9]. QSAR/QSPR are chemoinformatic methods that create a relationship between the molecular structure and the concerned activity/property [10,11]. The application of these methods required an extraction of a maximum structural information by using the best available molecular descriptors. A molecular descriptor is defined as an outcome of a mathematical and logical process. A standardized experiment that translates encoded chemical data information denoted by a symbolic expression of a molecule into a suitable number aimed to quantify the molecule [7,11]. There exist many important previous QSPR studies intended to predict diverse properties of pure chemicals in different fields [12]. It should be noted that most of the preceding QSPR studies dealt with distinction between the molecule on the same type or the same family (i.e. alkane

molecules among them) [13,14]. Distinction between isomer of position (i.e. *cis-trans...*) has been ignored in most of these studies. It is hard to find one molecular descriptor which can distinguish between all the isomers. However, in this study, we found a group of descriptors which are able to do this.

Usually the QSPR models must be created following this important three stages: **(1)** select and find a set of molecular properties (*Tb* in this study); **(2)** extract molecular structure information by calculation and selection of best relevant molecular descriptors; **(3)** link both preceding steps through the use of a mathematical learning methods [15]. Accordingly, this study aimed to generate a robust QSPR-ANN model able to predict the *Tb* of petroleum fractions and pure hydrocarbons. The previous steps were respected to develop QSPR-ANN model. The model was tested by two statistical coefficients (R and R^2) and two error types (*RMSE* and *SEP*). The QSPR-ANN model shows a precise determination of *Tb* of pure hydrocarbons and petroleum fractions.

II. MATERIALS AND METHODS

In order to develop a QSPR-ANN model, several steps were used as described below and summarized in **Fig. 1.**

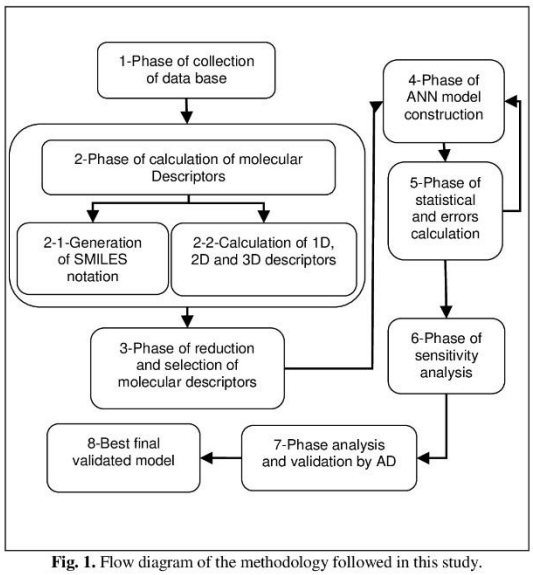


Fig. 1. Flow diagram of the methodology followed in this study.

A. Data set collection

The dataset used in this study was collected from several sources containing 223 different pure hydrocarbons molecules and petroleum fractions for *Tb* [13,16,17]. These compounds consist of various categories saturated and unsaturated hydrocarbons representing main industrially important groups. The majority of those categories have more than one isomer of position. A part of database used in this study was indicated in "*Appendix A*".

B. Molecular descriptors calculation

The construction of QSPR models need the quantification of structural information of the studied molecules [18], which are called molecular descriptors. In this study 1666 molecular descriptors distributed over 20 different classes, were calculated online by E-Dragon 1.0 software. The calculation of these descriptors was preceded by a very important step, which is the generation of Simplified Molecular Input-Line Entry System (SMILES) notation.

C. Selection of relevant descriptors

One of the hard tasks in this study is the selection of relevant descriptors. Since it has been done manually, so it takes a lot of time and effort, regarding the large number of molecules and the huge number of descriptors. A statistical methodology was used to reduce the number of descriptors, and get just the selected descriptors, in order to use it in the construction of QSPR-ANN model. The methodology was focused on the following points: **(1)** Eliminate descriptors containing the uncalculated error values; **(2)** Remove descriptors with a same value for more than 75%; **(3)** Omit descriptors with a relative standard deviation (*RSD*) less than 0.05; **(4)** Remove descriptors with an intercorrelation Pearson coefficient more than 0.75, in this stage the comparison and elimination of each unwanted descriptor was done manually; **(5)** Finally, use a MLR (Multiple Linear Regression) method to remove descriptors with *P*-value (Probability value) more than 0.005. In the end, an important number of descriptors were retained as shown in **Table I**.

TABLE I. EVOLUTION OF THE NUMBER OF DESCRIPTORS ACCORDING TO THE SELECTION PROCEDURE

Reduction phases	*(0)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
Number of molecular descriptors	1666	1637	1105	1001	89	**16**

D. Constraction of the QSPR-ANN model

After the selection procedure shown in **Table I**, more than 99% of the initial molecular descriptors were reduced (from 1666 to 16 descriptors). Thus, we obtained just 16 real relevant descriptors, to be used in construction phase of the QSPR-ANN model. This is another task regarding the large selection's parameter in the artificial neuron network (ANN); so that, we used different choices in order to obtain the best QSPR-ANN model. The dataset was randomly split into two subsets (80%) for the training phase and (20%) for the testing phase. We used MLP-ANN as a type of ANN. A BFGS algorithm was used as a training algorithm [18]. The activation function is done by choosing several types of functions in both layers, at the end of execution we obtain the best possible combination between the two activation functions of concerned layer and we used the sum of squares "*SOS*" as error function, relating the hidden layer numbers, actually we use one hidden layer. However, we varied the number of neurons per hidden layer from 3 to 20 neurons. Several executions were made to obtain the final best model, generally we take from 1000 to 100000 iterations per execution.

III. RESULTS AND DISCUSSION

A. Architectures of the best obtained QSPR-ANN model

After these steps of collection, reduction, construction and calculation; we choose only the best model for testing their performance, by calculating two error types. The architecture of the best QSPR-ANN model obtained for *Tb* with some primary statistic parameters are represented in **Table II**. The "*SOS*" error has value of 10^{-6} for both test and training set. The Pearson correlation coefficient values *R* shows a high performance of the QSPR-ANN model.

TABLE II. ARCHITECTURES OF THE BEST OBTAINED QSPR-ANN (*TB*) MODEL.

Nbr of neurons by layer	*Inputs:*16	*Hidden:*12	*Output:*1
Activation functions	***(none)***	*Exponential*	*Tanh*
R values	R^{train}= 0.9997	R^{test}= 0.9999	Training algorithm: *BFGS 65*
SOS values	SOS^{train} = 8.0 10^{-6}	SOS^{test}= 3.0 10^{-6}	

B. Application of QSPR-ANN model equation for Tb estimation

As shown in **Table II**, the architecture of the QSPR-ANN model is (16-12-1). So, the network has 16 inputs ($\mathbf{E}_i$, *i= 1 to 16)*, 12 hidden neurons *(j= 1 to 12)* and one output *(**S**)*. The activation functions used in this study are: the *"Exponential"* as hidden layer function and the *"Hyperbolic Tangent"* as an output layer function. Their mathematical definitions are given respectively in **Eq. (1)** and **Eq. (2)**.

$$f^H = e^{-x} \quad (1)$$

$$f^O = (e^{+x} - e^{-x}) / (e^{+x} + e^{-x}) \quad (2)$$

Each hidden neuron calculates activation function, transfer the results ($\mathbf{Z}_j$, *j= 1 to 12)* to the output layer's neuron which finally gives the response of the network *(**S**)*. The output signal of each hidden neuron ($\mathbf{Z}_j$) is calculated by the **Eq. (3)** and the output of the network is given by **Eq. (4)**, (***x*** are the variables of **Eq. (3)** and **Eq. (4)**).

$$\mathbf{Z}_j = f^H [\sum_{i=1}^{16} w^I_{ij} E_i + b^H_j] \quad (3)$$

$$S = f^O [\sum_{j=1}^{12} w^H_{1j} Z_j + b^O_1] \quad (4)$$

Where f^H is the activation function of hidden layer, f^O the activation function of output layer, ***S*** is the output of the network (output parameter: *Tb*), **Z** is the output signal of the hidden neurons, $\mathbf{W}^I_{ij}$ are the weights of input layer, $\mathbf{W}^H_{1j}$ the weights of hidden layer, $\mathbf{E}_i$ is the inputs neurons, in this case the relevant molecular descriptors, b^H_j are the bias of hidden layer, $\mathbf{b}^O_1$ is the bias of output layer, ***i*** is the number of input *(i= 1 to 16)*, and ***j*** is the number of hidden neurons *(j= 1 to 12)*. The final QSPR-ANN model for predicting the *Tb* of pure hydrocarbons is given in **Eq. (4)**. A part of weight's and bias's values of *Tb* model for each layer are given in *"Appendix **B**"*.

C. Estimation of Tb of pure hydrocarbons by QSPR-ANN model

The introduction of input values, which are the relevant descriptors values, according to the appropriate order (*STN, TIE, IVDM, Yindex, IC3, CIC3, EEig07x, GGI4, JGI1, Mor10v, Mor16e, ISH, H5e, R3e, R4e* and *nCt*), in addition to the weights ($\mathbf{W}^I_{ij}$ and $\mathbf{W}^H_{1j}$) and bias ($\mathbf{b}^H_j$ and $\mathbf{b}^O_1$) connections values of both layers, allowed us to predict the *Tb* of pure hydrocarbons using **Eq. (4)**, (A part this data is given in *"Appendix **B**"*).

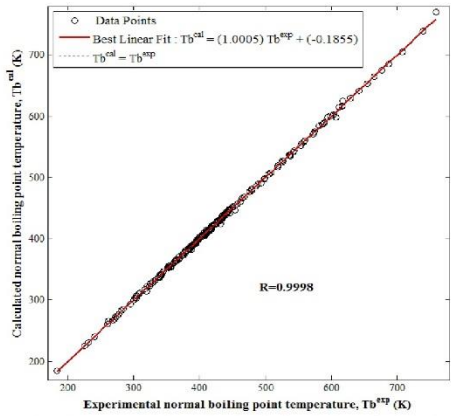


Fig. 2. Comparison between experimental and calculated values of *Tb* for the whole dataset.

The comparison between experimental and calculated values of *Tb*, obtained using the developed QSPR-ANN model, as shown in **Fig.2,** confirmed that the proposed model gives satisfactory results. The accuracy of the regression is high, with good agreement vector values (see the unit slope line equation above **Fig.2**. And the *R* value) in fitting *Tb* profile, for entire dataset, (Apart of experimental and calculated values of *Tb* model and their splitting way are given in *"Appendix **A**"*).

D. Statistical Analysis of the QSPR-ANN Model

In order to evaluate the QSPR-ANN model, we tested the performance of *Tb* model; by two error's types, which are: The Root Mean Squared Error (RMSE) and Standard Error of Prediction (SEP), for training, test and whole dataset. **Table III** summarize the result of this error types in addition to *R* and R^2 values, and the number of data point (*N*) in each phase.

TABLE III. STATISTICAL PARAMETERS OF THE BEST QSPR-ANN MODEL.

Statistical parameters	***Training set***	***Test set***	***Whole set***
N	179	44	223
R	0.9997	0.9999	0.9998
R^2	0.9995	0.9999	0.9996
RMSE (K)	2.15	1.37	2.02
SEP (%)	0.50	0.33	0.47

The mathematical definition of the previous error types is given by **Eq. (7)** and **Eq. (8)**, respectively.

$$\hat{y}^{exp} = \sum_{i=1}^{N} y_i^{exp} / N \quad (6)$$

$$RMSE = (\sum_{i=1}^{N} (y_i^{exp} - y_i^{cal})^2 / N)^{1/2} \quad (7)$$

$$SEP = (RMSE / \hat{y}^{exp}) * 100 \quad (8)$$

Where N is the number of compounds in the dataset for each phase (training, test or whole dataset), y^{cal} is the calculated value of y; for training set or test set, y^{exp} is the experimental value of y; and $\hat{y}^{exp}$ is the average of experimental values of y (see **Eq. (6)**).

According to the results of different statistical coefficients and error types shown in **Table III**, the developed QSPR-ANN model was very interesting, with R and R^2 values, higher than 0.999 in the three-subset phases.

The error types showed that the *RMSE* values of *Tb* model are ranged from 1.3718 K to 2.1458 K. Nonetheless, the *RMSE* values of the three phases are in same range; within 2 K (see **Table III**), which confirms the stability and balance of the developed QSPR-ANN model. The same results with *SEP* values, confirming previous *RMSE* results, less than 1% in the three-subset phases (see **Table III**). The *RMSE* and *SEP* results confirmed the accuracy of our model for predicting *Tb*.

E. Sensitivity analysis of the QSPR-ANN Model

The weight sensitivity analysis method was used [19]. In order to appreciate the contribution (relative importance (RI)) of the input descriptors on the developed model and to determine the impact of each individual input on the output of the model. This method was initially proposed by Garson [20], then taken up by Goh [21].

TABLE IV. THE RI OF THE INPUT DESCRIPTORS ON THE DEVELOPED QSPR-ANN MODEL.

Descriptors order	Descriptors symbol	RI (%) of Descriptors
1^{st}	*STN*	5.18
2^{nd}	*TIE*	4.80
3^{rd}	*IVDM*	6.62
4^{th}	*Yindex*	4.93
5^{th}	*IC3*	3.92
6^{th}	*CIC3*	4.55
7^{th}	*EEig07x*	4.58
8^{th}	*GGI4*	6.12
9^{th}	*JGI1*	4.91
10^{th}	*Mor10v*	6.31
11^{th}	*Mor16e*	4.54
12^{th}	*ISH*	5.33
13^{th}	*H5e*	11.90
14^{th}	*R3e*	7.50
15^{th}	*R4e*	10.58
16^{th}	*nCt*	8.23

ARI = 6.25 %; ARI = sum of RI (100%) /Nbr of Descriptors (16)

The contribution of each descriptor on the QSPR-ANN model has been summarized in **Table IV**. According to these results, and the average relative importance (ARI) value (see the foot note of **Table IV**), we can classify the descriptor's contribution for *Tb* model into three groups: **(1)** high contribution, ranging from 12 to 7% *{H5e > R4e > nCt > R3e}*, **(2)** medium contribution, ranging from 7 to 6% *{IVDM > Mor10v > GGI4}* and **(3)** low contribution, ranging from 6 to 4% *{ISH > STN > Yindex > JGI1 > TIE > EEig07x > CIC3 > Mor16e > IC3}*.

The lowest contribution descriptor shown in **Table IV**, do not exceed the 2.5% of difference with the ARI value. This means that all selected descriptors were important, in varying degrees, in the development of QSPR-ANN model according to their classification's group.

F. Applicability Domain

The applicability domain (AD) of a QSPR model, is defined as the physicochemical or structural spaces, a confirmation on which state whether the model's assumptions are met, and for which chemicals the model can be reliably applied [22]. A distance-based method is one of the interested methods of AD approach, their principle based on the leverage value (h) determination of each compound, which is then compared with the standard leverage value (h^*) [23,24].

The h value is defined as: $(h_i = X_i (X^T X)^{-1} X^T_i)$ (i = 1, 2, 3…n). Such as X_i is the row-vector descriptor of the concerned compound, X the matrix values of training descriptors, and n the whole compounds number. Usually the h^* value is taken as: $h^* = 3*(k+1)/m$. Such as k is the variable numbers of the model and m the training compounds number. If the h value of any considered compound is greater than the h^* value, the predicted result possibly will not be reliable since the predicted value could be considered out of model's AD. [22,23,25,26].

The toolbox developed by Milano Chemometrics and QSAR Research Group has been used [26,27]. As shown in **Fig. 3**, the Williams plot summarizes the AD results of the obtained QSPR-ANN model. The AD is established in a square zone within h^* value and ±2 of the standard residual values. Therefore, it is considered quite strict, compared to other studies, so most standard residual values are limited to ±3 [28].

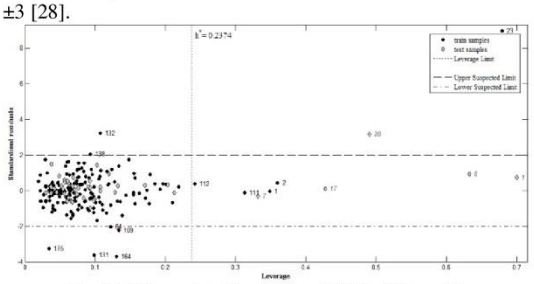


Fig. 3. Williams plot of the developed QSPR-ANN model.

The value of h^* obtained for our model is 0.2374, which means that 92% of the compounds are within the AD. Thus,

there are only 17 of 223 compounds out of range; 12 for the training set and 5 for the test set (See *"Appendix A"* with a line below). This is regardless of the standard residual value taken (±2), which again confirms the effectiveness of the QSPR-ANN model developed in this study.

IV. CONCLUSION

An accurate QSPR-ANN model has been developed, in order to predict *Tb* of pure hydrocarbons and petroleum fractions. A structural information extraction process was performed to quantify the studied molecules. As a result, 1666 descriptors were calculated. Several statistical steps have been followed, in order to reduce the important number of descriptors and retain only the relevant descriptors. The experimental dataset was selected from several sources.

The final best QSPR-ANN model was achieved by BFGS algorithm with "16-12-1" network architecture. The QSPR-ANN model's results showed good accuracy according to R^2 value; they take 0.9996 for the Tb^{All}. Moreover, the robustness and predictive ability of the model was emphasized by two error types *(RMSE* and *SEP)*; the result was very encouraging 2.0168 K for the *RMSE*, and 0.4723% for the *SEP*.

The model encompasses specific families such as alkanes as well as all families of hydrocarbons. Furthermore, our model is able to distinguish even between the isomers of position with high precision. The QSPR-ANN model developed in this study, can be used in the field of industry and petroleum engineering, and also in other fields related to the pure hydrocarbons to estimate their *Tb*, only from its molecular structure using simple descriptors.

ACKNOWLEDGMENT

The authors appreciate the efforts of the (LSPN) and (LBMPT) laboratory team for their encouragement throughout this project. The authors are also thankful to the organizing committee of SNPCM1'2018, ENSET-SKIKDA, for their efforts.

REFERENCES

[1] Riazi, M. R. (2005). Characterization and properties of petroleum fractions (Vol. 50). ASTM international.

[2] Bose, B. K. (2000). Energy, environment, and advances in power electronics. *IEEE Transactions on Power Electronics*, *15*(4), 688-701.

[3] Leprince, P. (2001). *Petroleum refining. Vol. 3 conversion processes* (Vol. 3). Editions Technip.

[4] Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, *69*(1), 17-20.

[5] Chan, P. Y., Tong, C. M., & Durrant, M. C. (2011). Estimation of boiling points using density functional theory with polarized continuum model solvent corrections. *Journal of Molecular Graphics and Modelling*, *30*, 120-128.

[6] Duchowicz, P. R., & Castro, E. A. (2013). The Importance of the QSAR-QSPR Methodology to the Theoretical Study of Pesticides. *International Journal of Chemical Modeling*, *5*(1), 35.

[7] Mauri, A., Consonni, V., & Todeschini, R. (2016). Molecular descriptors. *Handbook of Computational Chemistry*, 1-29.

[8] Karelson, M., Lobanov, V. S., & Katritzky, A. R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical reviews*, *96*(3), 1027-1044.

[9] Yousefinejad, S., & Hemmateenejad, B. (2015). Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, *149*, 177-204.

[10] Visco Jr, D. P., Pophale, R. S., Rintoul, M. D., & Faulon, J. L. (2002). Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *Journal of Molecular Graphics and Modelling*, *20*(6), 429-438.

[11] Todeschini, R., & Consonni, V. (2008). *Handbook of Molecular Descriptors* (Vol. 11). John Wiley & Sons, New York.

[12] Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., & Dobchev, D. A. (2010). Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chemical reviews*, *110*(10), 5714-5789.

[13] Wessel, M. D., & Jurs, P. C. (1995). Prediction of normal boiling points of hydrocarbons from molecular structure. *Journal of chemical information and computer sciences*, *35*(1), 68-76.

[14] Pan, Y., Jiang, J., & Wang, Z. (2007). Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network. *Journal of hazardous materials*, *147*(1-2), 424-430.

[15] Hamadache, M., Amrane, A., Hanini, S., & Benkortbi, O. (2018). Multilayer Perceptron Model for Predicting Acute Toxicity of Fungicides on Rats. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, *3*(1), 100-118.

[16] Lahiouel, Y. (1995). Industrial experimental dataset on petroleum fractions and pure hydrocarbons, *Refinery of Skikda*, Skikda, ALGERIA.

[17] KDB. (2017). *Korean Thermophysical Properties Data Bank–Cheric (KDB)* [Online]. Available: http://www.cheric.org/research/kdb/, (Accessed 02/12/2017).

[18] Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L., & Moussa, C. S. (2016a). A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *Journal of hazardous materials*, *303*, 28-40.

[19] Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

[20] Garson, G. D. (1991). Interpreting neural-network connection weights. *AI expert*, *6*(4), 46-51.

[21] Goh, A. T. C. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3), 143-151.

[22] Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *Molecular Informatics*, *26*(5), 694-701.

[23] Zhou, L., Wang, B., Jiang, J., Pan, Y., & Wang, Q. (2017). Predicting the gas-liquid critical temperature of binary mixtures based on the quantitative structure property relationship. *Chemometrics and Intelligent Laboratory Systems*, *167*, 190-195.

[24] Kraim, K., Khatmi, D., Saihi, Y., Ferkous, F., & Brahimi, M. (2009). Quantitative structure activity relationship for the computational prediction of α-glucosidase inhibitory. *Chemometrics and Intelligent Laboratory Systems*, *97*(2), 118-126.

[25] Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives*, *111*(10), 1361.

[26] Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2014). Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective. *Current computer-aided drug design*, *10*(2), 137-147.

[27] Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791-4810.

[28] Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set descriptor space: a review. *ATLA-NOTTINGHAM-*, *33*(5), 44

APPENDICES (SUPPLEMENTARY DATA).

Appendix A. Part of dataset used in this study [13,16,17].

N°	*UPAC Name*	*Chemical formula*	$Tb^{exp}(K)$	$Tb^{cal}(K)$	*Tb Splitting*
1	Ethane	C2H6	184,6[b]	183,6967	1st Test
2	Propane	C3H8	231,1[b]	231,1680	1st Train
3	n-butane	C4H10	272,65[a]	272,2369	2nd Test
4	i-butane	C4H10	261,43[a]	260,6658	2nd Train
5	n-pentane	C5H12	309,22[a]	308,4504	3rd Train
28	n-octacosane	C28H58	704,8[b]	707,7089	22nd Train
29	n-dotriacontane	C32H66	738,8[b]	739,3861	7th Test
30	n-hexatriacontane	C36H74	770,1[b]	758,6414	23rd Train
31	2,2-dimethylpropane (neopentane)	C5H12	282,6[b]	281,3343	8th Test
32	2-methylpentane	C6H14	333,41[a]	333,4679	24th Train
77	1-cis,2-trans,4-trimethylcyclopentane	C8H16	389,91[a]	388,8706	63rd Train
78	1-trans,2-cis,3-trimethylcyclopentane	C8H16	383,3[a]	387,6239	64th Train
79	2,2,3-trimethylhexane	C9H20	406,8[c]	406,3322	65th Train
80	2,2,4-trimethylhexane	C9H20	399,69[a]	399,8705	15th Test
81	2,2,5-trimethylhexane	C9H20	397,24[c]	395,7781	66th Train
98	2,2,3,3-tetramethylhexane	C10H22	433,48[c]	434,2614	82nd Train
99	Propylene	C3H6	225,4[b]	225,2432	17th Test
100	1-butene	C4H8	266,9[b]	268,5800	83rd Train
119	trans-2-butene	C4H8	274[b]	275,4693	100th Train
120	i-butene	C4H8	266,2[b]	261,1068	20th Test
133	3,3-dimethyl-1-butene	C6H12	314,4[b]	319,0478	109th Train
134	2-methyl-2-pentene	C6H12	340,4[b]	341,6988	25th Test
135	1,5-hexadiene	C6H10	332,6[b]	333,8126	110th Train
136	Cyclopropane	C3H6	240,4[b]	240,6259	111th Train
137	Cyclobutene	C4H8	285,7[b]	284,9319	112th Train
138	Cyclopentane	C5H10	322,4[a]	323,7218	113th Train
161	Tridecylcyclopentane	C18H36	598,6[c]	594,1797	130th Train
162	Tetradecylcyclopentane	C19H38	599[c]	606,7687	131st Train
163	Pentadecylcyclopentane	C20H40	625[c]	618,1861	132nd Train
164	1-methyl-4-isopropylcyclohexane	C10H20	443,87[a]	441,6034	133rd Train
165	1,1-dimethylcyclopentane	C7H14	360,96[a]	360,9700	32nd Test
204	1-trans-methyl-2-ethylcyclohexane	C9H18	424,87[a]	423,2367	163rd Train
205	1-ethyl-2,3-dimethylcyclohexane	C10H20	424,87[a]	432,6397	164th Train
206	1-methyl-ethyl-benzene	C9H12	425,561[a]	426,1248	165th Train
219	sec-butylbenzene	C10H14	446,48[a]	453,6651	175th Train
220	p-diethylbenzene	C10H14	456,9[b]	457,1811	176th Train
221	diphenyl (biphenyl)	C12H10	528,1[b]	526,2607	177th Train
222	Naphthalene	C10H8	491,1[b]	493,7525	178th Train
223	2,2,3,3-tetramethylbutane	C8H18	379,4[b]	380,2639	179th Train

[a]from experimental data base of Lahiouel [16]; [b]from study Wessel &Jurs [13]; [c]from KDB [17].

Appendix B. Weights and bias of the best QSPR-ANN model (16-12-1).

Input – Hidden Layers Connections								**Hidden – Output Layers Connection**	
Wight							Bias	Wight	Bias
$\mathbf{W}^{I}_{j1}$	$\mathbf{W}^{I}_{j2}$	$\mathbf{W}^{I}_{j3}$	……	$\mathbf{W}^{I}_{j14}$	$\mathbf{W}^{I}_{j15}$	$\mathbf{W}^{I}_{j16}$	$\mathbf{b}^{H}_{j}$	$\mathbf{W}^{H}_{1j}$	$\mathbf{b}^{O}_{1}$
0,009864	0,131273	-0,159378	……	0,072699	-0,034833	0,124340	0,025964	-0,019756	
0,069998	0,075598	-0,021941	……	-0,186341	0,223640	0,197385	-0,033483	0,241887	
⋮	⋮	⋮	……	⋮	⋮	⋮	⋮	⋮	-0,238553
0,124394	0,019825	0,064614	……	0,261647	0,290865	0,071172	0,103426	-0,514280	
0,060392	0,007065	-0,060827	……	0,093913	-0,070512	-0,034277	-0,031219	0,430837	

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Faculté des Sciences et de la Technologie
Département de Génie des Procédés
Université 08 mai 1945 – Guelma

Laboratoire des Silicates, polymères et des Nanocomposites

Ecole Normale Supérieure d'Enseignement Technologique ENSET – SKIKDA

1er Séminaire National de Physique et Chimie des Matériaux (SNPCM1'2018) -ENSET – SKIKDA, 28,29 Novembre 2018.

Robust QSPR-ANN Estimation Model for Normal Boiling Point Temperature of Pure Hydrocarbons and Petroleum Fractions

Presented by : ➢ **Mr. ROUBEHIE FISSA. Mohamed[1,*]**

Co-Auteurs : ➢ **Pr. LAHIOUEL. Yasmina[1], Dr. KHAOUANE. Latifa[2] et Pr. HANINI. Salah[2]**

➢ *Authors Affiliation :*

[1]Laboratoire des Silicates, Polymères et des Nanocomposites (LSPN), Université 8 mai 1945 Guelma, BP 401, Guelma 24000, Algérie.
[2]Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT), Université Yahia Farès de Médéa, 26000 Médéa, Algérie.

➢ *Corresponding e-mail adresses: roubehiefissa.mohamed@univ-guelma.dz; mohfismasbioprophar@mail.com*

Attestation de comité d'organisation d'un séminaire nationale avec comité de lecture



جامعة الدكتور يحي فارس بالمدية
UNIVERSITE Dr.YAHIA FARES DE MEDEA

SNEPT 2020

PREMIER SEMINAIRE NATIONAL
DE L'ENVIRONNEMENT ET DES PHENOMENES DE TRANSPORT
03-04 Mars 2020

جامعة يحيى فارس - المدية

ATTESTATION DE PARTICIPATION

Ce certificat est fièrement remis au

ROUBEHIE FISSA Mohamed

En reconnaissance de sa participation distinguée au comité d'organisation du premier séminaire national de l'Environnement et des phénomènes de transport SNEPT20 à l'université de Médéa.

Dr. FERHAT Samira
La présidente du SNEPT20

General Chair of SNEPT 20
Dr. FERHAT Samira

Université Yahia Fares de Médea
SNEPT20
03.04 Mars 2020
Faculté de Technologie

LBMPT
Laboratoire de Biomatériaux et Phénomènes de Transport

كلية التكنولوجيا
Faculté de Technologie

LME
Laboratoire Matériaux et Environnement