Machine learning based models for accessing thermal conductivity of liquids at different temperature conditions

Rosa Moreno Jimenez, Benoît Creton, Samuel Marre

Machine learning based models for accessing thermal conductivity of liquids at different temperature conditions

Rosa Moreno Jimenez[a,b], Benoit Creton[a]*, Samuel Marre[b]*

[a] *IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France.*

[b] *CNRS, Univ. Bordeaux, ICMCB, UMR 5026, F-33600 Pessac, France.*

*Corresponding authors: benoit.creton@ifpen.fr and samuel.marre@icmcb.cnrs.fr

Abstract

Combating the global warming-related climate change demands prompt actions to reduce greenhouse gas emissions, particularly carbon dioxide. Biomass-based biofuels represent a promising alternative fossil energy source. To convert biomass into energy, numerous conversion processes are performed at high pressure and temperature conditions and the design and dimensioning of such processes requires thermophysical property data, particularly thermal conductivity, which are not always available in the literature. In this paper, we proposed the application of Chemoinformatics methodologies to investigate the prediction of thermal conductivity for hydrocarbons and oxygenated compounds. A compilation of experimental data, followed by a careful data curation were performed to establish a database. The support vector machine algorithm has been applied to the database leading to models with good predictive abilities. The SVR model has then been applied to an external set of compounds, *i.e.* not considered during the training of models. It showed that our SVR model can be used for the prediction of thermal conductivity values for temperatures and/or compounds that are not covered experimentally in the literature.

Keywords: QSPR; thermal conductivity; temperature; oxygenated compounds, hydrocarbons

List of acronyms

CCC	Concordance Correlation Coefficient
EoS	Equation of State
FGCD	Functional Group Count Descriptors
GHGs	Greenhouse gases
MAE	Mean Absolute Error
ML	Machine Learning
MM	Molar Mass
n-CV	n-fold Cross-Validation
PCA	Principal Component Analysis
QSPR	Quantitative Structure-Property Relationship
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification
SQA	Sequential Quadratic Approximation
SVM	Support Vector Machines
SVR	Support Vector Regression

Introduction

In the context of reducing the concentration of greenhouse gases (GHGs) emissions in the atmosphere, the use of biomass-derived energy represents a promising alternative to fossil resources. To convert biomass into useful forms of energy (*e.g.* biofuels), a variety of conversion processes are necessary, and their design and dimensioning (from a chemical engineer's point of view) needs as support, the knowledge of thermophysical properties. For example, transport properties involving mass and heat transfers are widely considered in industrial applications such as separators, reactors, or heat exchangers. More specifically focusing on aircrafts, the fuel itself represents the main heat transfer fluid within the thermal management systems. The thermal conductivity (λ) of fluids therefore represents an essential thermodynamic data to consider for the design and the development of processes leading to alternative fuels.

Several experimental techniques have been developed to measure the thermal conductivity, *e.g.* the transient hot-wire method – consisting in monitoring temperature changes of a thin metallic wire – which is considered to be one of the standard macroscopic techniques for fluids [1,2]. Due to the phenomena of convection and radiation, which result in heat loss, it is difficult to accurately determine thermal conductivity using these methods within large-scale equipment. Recently, by downsizing experiments to reach scales of milli- or microfluidic techniques, Oudebrouckx *et al.* [3] and Moreno *et al.* [4] succeeded in predicting thermal conductivity values with few microlitres of products. For instance, Moreno *et al.* proposed new experimental data for 2,5-dimethylfuran for temperatures ranging from 293.15 to 333.15K [4].

The use of alternative methods is however essential, whether to supplement experimental data or even to feed process simulators. Indeed, data on thermal conductivity for oxygenated compounds, particularly under extreme temperature and

pressure conditions, remain limited or even unavailable in the literature. Existing models have been used in some scenarios for the prediction of thermal conductivity of liquids [5–7]. However, these types of models have not often been applied to a wide range of polar substances, such as alcohols, ketones, organic acids, among others. Accurate prediction involving equation of state (EoS) remains difficult due to the scarcity of experimental data, which directly affects the accuracy of the parameters required for EoS. Hence, another alternative for modeling thermal conductivity is the use of machine learning based approaches, as proposed by Malatesta *et al.* for mixtures of hydrocarbons used as aviation turbine fuel surrogates [8]. Lu *et al.* recently reported the development of a quantitative structure-property relationship (QSPR) based model to predict the thermal conductivity of diverse organic compounds in liquid phase [9]. Their model was trained on a set of chemical compounds – similar to our target – including chemical families such as alcohols, ethers, aldehydes, ketones, acids, and esters. The 6-descriptor linear model was developed by Lu *et al.* combining a genetic function approximation and a variety of molecular descriptors derived from geometries optimized using density functional theory. However, this model approximates the temperature dependency of thermal conductivity as a constant identical value for all compounds, which experimental data do not seem to show. Moreover, it seems that the Training/Test sets splitting randomly performed led to compounds representation in both sets for different temperature values [9].

In this work, we propose to revisit the development of a QSPR based model for the prediction of thermal conductivity using a ML algorithm leading to a ‘non-linear’ approach combined to a compilation of accurate reference data to obtain a coherent database and performing an appropriate Training/Test set splitting. After presenting the data collection and curation methods and the strategy followed to develop new QSPR

based models, we expose the predictive performances of models and discuss their utilization for external data prediction, before concluding.

Materials and Methods

These last years, we have devoted large efforts in the application of machine learning on chemical databases to derive QSPR based models for the prediction of various property values [10,11]. These approaches aim at identifying non-obvious correlations between property values of the matter and some features encoding information about the matter. Reviews already exist containing detailed elements regarding the developments and applications of QSPR based models, and best practices in developing such models [12,13].

Database

The accuracy of predictive ML-based models is mainly related to the quality of data used as support to develop these models, and thus one of the keystones of such a work success stands in the database itself. The available experimental data on thermal conductivity was collected from existing databases, mainly DIPPR [14,15] but also the DETHERM [16], NIST and data from the recent literature, including data for 2,5-dimethylfuran recently measured by our group using Microfluidics [4]. It should be emphasized that no data originating from models or even from data fittings has been considered. Hydrocarbons and oxygenated compounds were the focus of the study, and Figure 1 shows the distributions of these two classes of compounds in the database, as well as distributions of considered subfamilies. It shows that hydrocarbons and oxygenates represent roughly 40% and 60% of chemicals, respectively. Hydrocarbons may then be discretized in terms of alkanes, alkenes, and cyclic molecules such as naphthenics and aromatics. In the database, cyclic compounds represent about 31% of hydrocarbons and the remaining ones

are for 61% saturated paraffins (n- and i-alkanes) and 8% alkenes. Regarding oxygenated compounds, the database includes, in decreasing order of occurrence: alcohols, esters, carboxylic acids, ethers, ketones, formats, aldehydes, and epoxides. It should be noted that 6 oxygenates are polyfunctionals, *i.e.* they are constituted of at least two of the latter chemical characteristics.

For each compound, as many data points as possible were collected to account for the temperature dependence of λ. In some cases, experimental thermal conductivity values reported for one compound at a specific temperature can vary depending on the sources of data, and only one value has been retained for each molecule and temperature condition. The priority was given to values reported as 'Accepted' by the DIPPR staff reviewers, and when necessary, averages were taken. Then, parameters of the expression proposed by Jamieson – equation (1) – were regressed using data points within the database:

$$\lambda(T) = A\left(1 + B\tau^{1/3} + C\tau^{2/3} + D\tau\right) \quad (1)$$

with $\tau = 1 - {}^{T}/_{T_c}$, where T_c is the critical temperature in K [17]. The as-obtained A to D parameters were subsequently used together with equation (1) to generate, for each compound, 10 data points by varying the temperature from the lowest of collected data up to T_c. Although the direct use of experimental data is a common practice in QSPR development, the use of pseudo-experimental data as proposed here enables, when modelling a property dependence as a function of a characteristic (here, the temperature), to reduce any over-representation or unbalance effect of compounds as well as a reduction of the noise in reference data.

From conclusions drawn in previous studies [18–20], solely descriptors derived from the chemical and structural formulae were considered and hereafter labelled as functional group count descriptors (FGCD). In this latter family of molecular descriptors

are included counts of atoms and groups of atoms identified as relevant from chemical aspects. Such a simple representation of compounds has been shown to provide relevant descriptors usable in QSPR procedure [21]. FGCD under consideration in this study, labelled from X1 to X42, are listed in Table 1. As an example, the FGCD labelled X4 denotes the number of CH_3 groups. The molar mass (MM) of neat compounds was also computed and used as an additional descriptor (labelled X42). Simplified molecular input line entry specification (SMILES) codes were assigned to each neat compound within the database. FGCD were automatically calculated using the RDKit [22] and SMILES arbitrary target specification (SMARTS) matching functionalities [23], SMARTS codes corresponding to considered FGCD are given in Table 1.

The complete database containing names, SMILES, investigated temperatures and pseudo-experimental λ values for 157 hydrocarbons and oxygenated compounds is available in the supporting information.

Chemical space representation

The information contained within our database was pre-processed by applying a principal component analysis (PCA) on molecular descriptor values. The three first principal components resulting from the PCA were used as an approximated graphical representation of the chemical space for our database. Figure 2 represents the projections of compounds within this chemical space. As seen, some molecules are isolated from all others, located at the borders of the domain, this is typically the case for molecules such as butyl octadecenoate, bis(2-ethylhexyl) benzene-1,2-dicarboxylate, Dihexyl hexanedioate, or even 2,4,6-Trimethyl-1,3,5-trioxane. These latter data points and many others – 32 compounds in total – thus appear as structural outliers, and their presence in external sets may induce violations of the applicability domain of models. These 32

molecules were therefore systematically used during the learning processes, as detailed hereafter.

Machine learning modelling

During last decades of QSPR model developments, the use of external validation has been shown necessary to ensure their ability to be applied to new fluids, *i.e.* not considered within the data set used to train the model [24]. Its popular version is the n-fold cross-validation (n-CV) in which the data set is randomly divided in approximately equal *n* portions. An aggregate of (*n-1*) portions forms a Training set − used to train models, and the remaining portion constitutes an external set or Test set − used to evaluate model's performances. We emphasize that no data point belonging to external sets was used to derived models. This procedure is repeated *n* times choosing for each a new portion of data as an external set. The subject of external validation for QSPR analysis has been addressed by Muratov *et al.* [25]. Considering our database content with ten temperatures – ten thermal conductivity values – for each molecule, a 'compound out' strategy was applied in this study, meaning that a molecule belongs to only one fold. To avoid any strong violation of the applicability domain of models during the cross-validation procedure, we fixed 32 molecules in a specific fold always used to form Training sets. A 5-CV was applied to the 125 remaining molecules, and the Training and Test sets thus represent 84% and 16% of the database, respectively.

We demonstrated in a number of previous works that the combination of molecular descriptors such as FGCD and Support Vector Machines (SVM) provides accurate solutions in terms of property modelling [20,21,26]. The Support Vector Regression (SVR) as implemented within the LibSVM library [27] was employed, with both linear and radial basis function kernels, and with an epsilon insensitive zone [28].

According to this method, three parameter values need to be optimized: cost, epsilon, and gamma. We followed the approach previously proposed by Gantzer et al. [29], and the Sequential Quadratic Approximation (SQA) method implemented in our in-house program [30] was used to optimize the SVR parameter values within a 5-CV procedure. Finally, a model was developed using the set of optimized parameters and considering the 1570 data points of the database.

Models are evaluated according to their ability to predict reference thermal conductivity values. Predicted values are compared to reference pseudo-experimental data, and the performances of models are evaluated by means of metrics such as Mean Absolute Error (MAE, equation (2)), Root Mean Squared Error (RMSE, equation (3)), coefficient of determination (R^2, equation (4)), or Concordance Correlation Coefficient (CCC, equation (5)), defined respectively as:

$$MAE = \frac{1}{N}\sum_{1}^{N}|y_i - x_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_i - x_i)^2} \tag{3}$$

$$R^2 = 1 - {\sum_1^N (y_i - x_i)^2}\Big/{\sum_1^N (x_i - \bar{x})^2} \tag{4}$$

$$CCC = \frac{2\sum_1^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^N (x_i - \bar{x})^2 + \sum_1^N (y_i - \bar{y})^2 + N(\bar{x} - \bar{y})^2} \tag{5}$$

with y_i the predicted value, x_i the experimental value, $\bar{x}$ the average of experimental thermal conductivity values, $\bar{y}$ the average of predicted thermal conductivity values, and N is the number of data points in the considered set. Chirico *et al.* have shown that the use of CCC is advocated considering various scenarios such as location shifts, scale shifts, and location plus scale shifts [31,32]. MAE and RMSE values have the unit of the property under consideration, while R^2 and CCC are unitless.

Results and discussion

The entire collection of experimental data, curated as detailed in the previous section, was used to derive predictive models. From the performed regressions of experimental data points using equation (1), we generated a set of pseudo-experimental data subsequently used to train a machine learning algorithm. Figure 3a) presents evolutions of thermal conductivity values as a function of the temperature for alkan-1-ols. No clear trend appears for this subfamily of alcohols, with for example slope values in between -0.0002 to -0.0003 $W.m^{-1}.K^{-2}$, for pentan-1-ol and octan-1-ol, respectively. A similar study was carried out for n-alkanes and Figure 3b) presents the evolution of thermal conductivity values as a function of temperature for this family of compounds. The figure clearly shows an evolution of slopes when moving from light linear alkanes to higher numbers of carbon atoms, from -0.0006 to -0.0002 $W.m^{-1}.K^{-2}$. It is interesting to note that this latter value is in line with that reported by Lu *et al.* in their model [9]. The observed variation in slopes from one compound to another justifies our decision to use an algorithm that establishes non-linear correlations between the descriptors and the reference thermal conductivity values, such as SVM. Because of its obvious effect on thermal conductivity values – λ values are negatively correlated with T – the temperature (expressed in Kelvin) was considered as an additional descriptor. This approach is commonly employed for temperature dependent properties [33].

A 5-CV was applied resulting in a splitting of the database into 5 folds, plus one additional containing the molecules fixed in training to avoid violation of the applicability domain. Note that for each compound, the dependency of λ on temperature is ensured by ten data points, which avoids any over-representation or unbalance effect of compounds within the database. The cost, epsilon, and gamma parameters of the SVR were optimized considering the six folds. Table 2 presents RMSE and R^2 values calculated on the Training

and Test sets for the five ephemeral models generated during the 5-CV. It shows that values of metrics are roughly stable from one decomposition to another with for instance, mean RMSE and R^2 values of 0.0025 $W.m^{-1}.K^{-1}$ and 0.984 on Training sets, respectively. A final model was then trained using the set of optimized parameters (cost = 537.930000; epsilon = 2.305940 ; gamma = 0.419818) and considering all reference data within the Training set. Performances of the obtained SVM-based model were evaluated using metrics such as MAE, RMSE, R^2, and CCC, which values are presented in Table 3. When applied to Training data, metrics indicate that the developed model accurately reproduces reference data, with for instance a R^2 of 0.992. The good agreement between reference and predicted values using the SVR model is illustrated in Figure 4. This latter presents the scatterplots of reference *vs.* predicted thermal conductivity values using the SVM-based model. All data points are not too scattered on both sides of the bisector, indicating that predicted values are in good agreement with reference data. However, several data points appear to have been overestimated by the model, more precisely, they correspond to the 10 data related to dodecan-1-ol for which absolute relative deviations of about 11% are observed with respect to pseudo-experimental data. Analysing the information contained in Figure 3a), data points corresponding to dodecan-1-ol (represented by dark blue circles) seem to be abnormally low roughly considering the trend with the number of carbon atoms in alkan-1-ols, while emphasising that we have taken into account the values quoted as 'Accepted' by the DIPPR [14,15]. It should be noted that, in the database used by Lu *et al.* [9], reference thermal conductivity values for dodecan-1-ol range from 0.148 $W.m^{-1}.K^{-1}$ (at 440 K) to 0.166 (at 310 K) $W.m^{-1}.K^{-1}$, in reasonable agreement with our predicted values (an absolute relative deviations of about 5%). For their 6-descriptor linear model, Lu *et al.* reported a R^2 value of 0.914 and a RMSE of 0.0067 $W.m^{-1}.K^{-1}$. However, authors did not perform the Training/Test splitting following a ‘compound-out’

approach, which represents a too favourable scenario and avoid any conclusion regarding predictive ability of the model.

During the data curation, many chemical compounds were discarded due to an insufficient number of experimental data to regress parameters of equation (1). These data points were set aside to be considered as a Validation set. Our SVM-based model was applied to compounds and temperature conditions within the Validation set. Predictive performances of our SVR model were evaluated on the validation set using metrics such as MAE, RMSE, R^2, and CCC, which values are presented in Table 3. Metrics suggested that the developed model has a good predictive ability and accurately reproduces reference data, with for instance a R^2 value of 0.836 and a RMSE of 0.0080 $W.m^{-1}.K^{-1}$. As for the Training set, the good agreement observed between reference and predicted values using the SVR model is illustrated in Figure 4. For compounds belonging to the Validation set, the SVR model was used to predict thermal conductivity values for temperatures up to critical temperatures. A to D parameters of equation (1) were then regressed on the basis of predicted values. The as-obtained parameter values are reported in Table 4. As an example, Figure 5 presents the evolution of thermal conductivity values as a function of temperature for diethyl oxalate – with SMILES formula CCOC(=O)C(=O)OCC –, and proposes a comparison between predictions using the SVR model and equation (1) fed with regressed parameters displayed in Table 4, and the experimental data reported by Riedel [34]. It shows the good agreement between the three sources of data at 293.15 K. Figure 5 illustrates the possible use of our SVR model: the prediction of thermal conductivity values for temperatures and/or compounds that are not covered experimentally in the literature.

Conclusions

We proposed here the application of Chemoinformatics methodologies to investigate the thermal conductivity prediction for hydrocarbons and oxygenated compounds in liquid phase. A compilation of experimental data and then a careful data curation were performed to establish a database of reference values. A pre-processing of the data has demonstrated that the use of algorithm leading to linear models would be a too strong approximation in terms of modelling for predicting the effects of temperature. Thus, a SVM algorithm was applied to the database to generate predictive models, within a 5-CV procedure and we emphasize the 'compound out' strategy followed for the splitting of data into Training and Test sets. Comparisons performed with respect to reference data have demonstrated that our SVR model has good predictive reliability and robustness, with a MAE of 0.0018 $W.m^{-1}.K^{-1}$ calculated on the entire database.

The SVR model has been used to predict the evolution of the thermal conductivity as a function of the temperature for a series of compounds for which there were too few experimental data in the literature to be included in the learning process. Performed comparisons showed the good agreement between predicted values and available experimental data. This confirms that our SVR model can be used for the prediction of thermal conductivity values for temperatures and/or compounds that are not covered experimentally in the literature.

Acknowledgements

The authors would like to thank Dr Claire Marlière, Lionel Teulé-Gay, and Olivier Nguyen for discussions that preceded this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors reported there is no funding associated with the work featured in this article.

List of figures:

References

[1] J.J. Healy, J.J. de Groot, and J. Kestin, *The theory of the transient hot-wire method for measuring thermal conductivity,* Physica B+C 82 (1976), pp. 392–408.

[2] Y. Nagasaka and A. Nagashima, *Simultaneous measurement of the thermal conductivity and the thermal diffusivity of liquids by the transient hot-wire method,* Review of Scientific Instruments 52 (1981), pp. 229–232.

[3] G. Oudebrouckx, T. Vandenryt, S. Bormans, P.H. Wagner, and R. Thoelen, *Measuring Thermal Conductivity in a Microfluidic Device With the Transient Thermal Offset (TTO) Method,* IEEE Sensors J. 21 (2021), pp. 7298–7307.

[4] R. Moreno Jimenez, B. Creton, C. Marliere, L. Teule-Gay, O. Nguyen, and S. Marre, *A microfluidic strategy for accessing the thermal conductivity of liquids at different temperatures,* Microchemical Journal 193 (2023), p. 109030.

[5] X.-Q. Guo, C.-Y. Sun, S.-X. Rong, G.-J. Chen, and T.-M. Guo, *Equation of state analog correlations for the viscosity and thermal conductivity of hydrocarbons and reservoir fluids,* Journal of Petroleum Science and Engineering 30 (2001), pp. 15–27.

[6] S. Khosharay, K. Khosharay, G. Di Nicola, and M. Pierantozzi, *Modelling investigation on the thermal conductivity of pure liquid, vapour, and supercritical refrigerants and their mixtures by using Heyen EOS,* Physics and Chemistry of Liquids 56 (2017), pp. 124–140.

[7] L.F. Cardona, L.A. Forero, and J.A. Velásquez, *A group contribution method to model the thermal conductivity of pure substances,* Fluid Phase Equilibria 564 (2023), p. 113592.

[8] W.A. Malatesta and B. Yang, *Aviation Turbine Fuel Thermal Conductivity: a Predictive Approach Using Entropy Scaling-Guided Machine Learning with Experimental Validation,* ACS omega 6 (2021), pp. 28579–28586.

[9] H. Lu, W. Liu, F. Yang, H. Zhou, F. Liu, H. Yuan, G. Chen, and Y. Jiao, *Thermal Conductivity Estimation of Diverse Liquid Aliphatic Oxygen-Containing Organic Compounds Using the Quantitative Structure-Property Relationship Method,* ACS omega 5 (2020), pp. 8534–8542.

[10] B. Creton, *Chemoinformatics at IFP Energies Nouvelles: Applications in the Fields of Energy, Transport, and Environment,* Molecular informatics 36 (2017), p. 1700028.

[11] C. Hall, B. Creton, B. Rauch, U. Bauder, and M. Aigner, *Probabilistic Mean Quantitative Structure–Property Relationship Modeling of Jet Fuel Properties,* Energy Fuels 36 (2022), pp. 463–479.

[12] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. de Hemptinne, P. Ungerer, B. Rousseau, and C. Adamo, *A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes,* Chemical reviews 115 (2015), pp. 13093–13164.

[13] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, and D.A. Dobchev, *Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction,* Chemical reviews 110 (2010), pp. 5714–5789.

[14] J.C. Bloxham, M.E. Redd, N.F. Giles, T.A. Knotts, and W.V. Wilding, *Proper Use of the DIPPR 801 Database for Creation of Models, Methods, and Processes,* J. Chem. Eng. Data 66 (2021), pp. 3–10.

[15] R.L. Rowley, W.V. Wilding, A. Congote, and N.F. Giles, *The Use of Database Influence Factors to Maintain Currency in an Evaluated Chemical Database,* Int J Thermophys 31 (2010), pp. 860–874.

[16] D.F. Ilten, *DETHERM: Thermophysical property data for the optimization of heat-transfer equipment,* J. Chem. Inf. Comput. Sci. 31 (1991), pp. 160–167.

[17] D.T. Jamieson, *Thermal conductivity of liquids,* J. Chem. Eng. Data 24 (1979), pp. 244–246.

[18] N. Villanueva, B. Flaconnèche, and B. Creton, *Prediction of Alternative Gasoline Sorption in a Semicrystalline Poly(Ethylene),* ACS combinatorial science 17 (2015), pp. 631–640.

[19] D.A. Saldana, B. Creton, P. Mougin, N. Jeuland, B. Rousseau, and L. Starck, *Rational Formulation of Alternative Fuels using QSPR Methods: Application to Jet Fuels,* Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles 68 (2012), pp. 651–662.

[20] B. Creton, B. Veyrat, and M.-H. Klopffer, *Fuel sorption into polymers: Experimental and machine learning studies,* Fluid Phase Equilibria 556 (2022), p. 113403.

[21] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, and B. Creton, *Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods,* Energy Fuels 25 (2011), pp. 3900–3908.

[22] *RDKit: Open-Source Cheminformatics Software, Accessed in 2023. URL: http://www.rdkit.org/.*

[23] *SMARTS - a language for describing molecular patterns; daylight chemical information systems inc.: Laguna niguel, ca, Accessed in 2023. URL: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.*

[24] P. Gramatica, *Principles of QSAR models validation: internal and external,* QSAR Comb. Sci. 26 (2006), pp. 694–701.

[25] E.N. Muratov, E.V. Varlamova, A.G. Artemenko, P.G. Polishchuk, and V.E. Kuz'min, *Existing and Developing Approaches for QSAR Analysis of Mixtures*, Molecular informatics 31 (2011), pp. 202–221.

[26] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, and B. Creton, *On the Rational Formulation of Alternative Fuels: Melting Point and Net Heat of Combustion Predictions for Fuel Compounds Using Machine Learning Methods*, SAR and QSAR in environmental research 24 (2013), pp. 259–277.

[27] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, ACM Trans. Intell. Syst. Technol. 2 (2011), pp. 1–27.

[28] L.M. Hiot, Y.S. Ong, Y. Tenne, and C.-K. Goh, *Computational Intelligence in Expensive Optimization Problems*, Adaptation, learning, and optimization v. 2. Springer, Berlin, Heidelberg, 2010.

[29] P. Gantzer, B. Creton, and C. Nieto-Draghi, *Comparisons of Molecular Structure Generation Methods Based on Fragment Assemblies and Genetic Graphs*, Journal of chemical information and modeling 61 (2021), pp. 4245–4258.

[30] D. Sinoquet, H. Langouët, and S. Da Veiga, *A new spring for geoscience: EAGE conference and exhibition, Barcelona 14-17 June 2010, 72, Conference proceedings & exhibitors' catalogue*. DB Houten, NL, 2010.

[31] L.I.-K. Lin, *A Concordance Correlation Coefficient to Evaluate Reproducibility*, Biometrics 45 (1989), p. 255.

[32] N. Chirico and P. Gramatica, *Real External Predictivity of QSAR Models: How to Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient*, Journal of chemical information and modeling 51 (2011), pp. 2320–2335.

[33] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, N. Ferrando, and B. Creton, *Prediction of Density and Viscosity of Biofuel Compounds Using Machine Learning Methods*, Energy Fuels 26 (2012), pp. 2416–2426.

[34] L. Riedel, *Thermal Conductivity of Liquids (in German),* Mitt. Kaltetech. Inst. Karlsruhe 2 (1948).

Table 1. Molecular descriptors selected for the development of models.

Label	SMARTS	Label	SMARTS
X1	[H]	X22	[CX4H2][CX3H0](=[O])[OX2H0]
X2	[C,c]	X23	[CX3H1](=[O])[OX2H0]
X3	[O,o]	X24	[CX4H3][OX2H0][#6]
X4	[CX4H3]	X25	[CX4H2][OX2H0][#6]
X5	[CX4H2]	X26	[CX4H1][OX2H0][#6]
X6	[CX4H1]	X27	[CX3H0](=[O])[OX2H1]
X7	[CX4H0]	X28	[CX3H1](=[O])[OX2H1]
X8	[CX3H2]=[CX3H1]	X29	[OX2H0][CX3H0](=[O])
X9	[CX3H1]=[CX3H1]	X30	[OX2H1][CX4H2][CX4H2][OX2H0][#6]
X10	[CX3H2]=[CX3H0]	X31	[OX2H0][CX4H1][OX2H0]
X11	[cH1]	X32	[cX3H0][O][cX3H0]
X12	[cH0]	X33	[#6][CX3H1](=O)
X13	[c][CX4H3]	X34	[#6][OX2H0][#6]
X14	[c][CX4H2]	X35	[#6][CX3H0](=[O])[#6]
X15	[OX2H1]	X36	[#6][CX4H2][OX2H1]
X16	[CX4H3][OX2H1]	X37	[#6][CX4H1]([#6])[OX2H1]
X17	[c][OX2H1]	X38	[CX4H3][CX3H0](=[O])[OX2H1]
X18	[CX4H3][CX3H0](=[O])[#6]	X39	[#6][CX3H0](=[O])[#8]
X19	[CX4H2][CX3H0](=[O])[#6]	X40	[O,o;R]
X20	[CX3H1](=[O])	X41	[C;R]
X21	[CX4H3][CX3H0](=[O])[OX2H0]	X42	MM

Table 2. RMSE (in $W.m^{-1}.K^{-1}$) and R^2 values calculated on Training and Test sets for the ephemeral SVR based models generated during the 5-CV.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Training					
RMSE	0.0021	0.0025	0.0025	0.0026	0.0027
R^2	0.988	0.983	0.983	0.982	0.982
Test					
RMSE	0.0050	0.0057	0.0060	0.0054	0.0056
R^2	0.944	0.918	0.915	0.940	0.917

Table 3. Performance characteristics (statistical metrics) of the SVR based model calculated on the Training set (1570 data points) and Validation set (31 data points).

	Training	Validation
MAE ($W.m^{-1}.K^{-1}$)	0.0018	0.0060
RMSE ($W.m^{-1}.K^{-1}$)	0.0024	0.0080
R^2	0.992	0.836
CCC	0.996	0.920

Table 4. Values of the parameters in equation (1) regressed on predictions using the SVR model.

IUPAC name	A	B	C	D	Tc (K)
octan-2-one	0.076	0.386	-1.790	3.050	632.7
pentan-2-one	0.088	0.207	-0.987	2.203	561.1
prop-2-en-1-ol	0.097	0.191	-0.921	2.093	545.1
octanoic acid	0.086	0.116	-0.639	1.764	693.0
1-propoxypropane	0.059	0.333	-1.625	4.243	530.6
2-methylpropanal	0.066	0.504	-2.419	4.684	544.0
4-methylphenol	0.108	0.286	-1.265	1.802	704.6
octadec-9-enoic acid	0.114	0.017	-0.231	0.943	781.0
diethyl oxalate	0.100	0.118	-0.599	1.562	618.0
dibenzofuran	0.103	0.376	-1.665	2.019	824.0
bis(2-ethylhexyl) hexanedioate	0.100	0.312	-1.486	2.088	845.0
prop-2-enoic acid	0.106	0.123	-0.605	1.466	615.0
1,3-diacetyloxypropan-2-yl acetate	0.163	0.007	-0.034	0.049	701.4
dec-1-ene	0.063	0.461	-2.169	4.284	617.0
2,2,5-trimethylhexane	0.046	0.442	-2.208	4.601	569.8
2-methylbutane	0.054	-0.103	-0.062	3.023	460.4
butyl propanoate	0.060	0.359	-1.733	3.999	594.5
methyl benzoate	0.065	0.652	-2.988	4.758	702.0
propanal	0.104	0.208	-1.030	2.696	505.0
1,2-bis(2-methoxyethoxy)ethane	0.101	0.176	-0.868	1.885	651.0
1-methyl-2-propan-2-ylbenzene	0.056	0.687	-3.118	4.913	657.0