J. Chem. Inf. Model. 2023
https://doi.org/10.1021/acs.jcim.3c00231

Temperature-Dependent Density and Viscosity Prediction for Hydrocarbons: Machine Learning and Molecular Dynamics Simulations

Pawan Panwar[1], Quanpeng Yang[1], and Ashlie Martini[1+]

[1]Department of Mechanical Engineering, University of California Merced,
5200 North Lake Road, Merced, CA 95343, USA
E-mail: amartini@ucmerced.edu

ORCID IDs:

- Pawan Panwar: https://orcid.org/0000-0001-5859-3927
- Quanpeng Yang: https://orcid.org/0000-0002-4395-9866
- Ashlie Martini: https://orcid.org/0000-0003-2017-6081

Abstract

Machine learning-based predictive models allow rapid and reliable prediction of material properties and facilitate innovative materials design. Base oils used in the formulation of lubricant products are complex hydrocarbons of varying size and structure. This study developed Gaussian process regression-based models to accurately predict the temperature-dependent density and dynamic viscosity of 305 complex hydrocarbons. In our approach, strongly correlated/collinear predictors were trimmed, important predictors were selected by least absolute shrinkage and selection operator (LASSO) regularization and prior domain knowledge, hyperparameters were systematically optimized by Bayesian optimization, and the models were interpreted. The approach provided versatile and quantitative structure−property relationship (QSPR) models with relatively simple predictors for determining the dynamic viscosity and density of complex hydrocarbons at any temperature. In addition, we developed molecular dynamics simulation-based descriptors and evaluated the feasibility and versatility of dynamic descriptors from simulations for predicting material properties. It was found that the models developed using a comparably smaller pool of dynamic descriptors performed similarly in predicting density and viscosity to models based on many more static descriptors. The best models were shown to predict density and dynamic viscosity with coefficient of determination (R^2) values of 99.6% and 97.7%, respectively, for all datasets, including a test dataset of 45 molecules. Finally, partial dependency plots (PDPs), individual conditional expectation (ICE) plots, local interpretable model-agnostic explanation (LIME) values, and trimmed model R^2 values were used to identify the most important static and dynamic predictors of density and viscosity.

Introduction

Lubricants play a crucial role in enhancing the performance of mechanical systems by improving their friction and wear characteristics. The properties of lubricants are determined by blending base oils with additives. The base oil, which constitutes up to 98% of a lubricant, can be derived from crude oil, biological sources, or produced synthetically [1,2]. It primarily consists of complex hydrocarbons, including paraffins, isoparaffins, aromatics, and naphthenic molecules with varying carbon numbers [3]. The molecular composition of the base oil directly influences its properties and, consequently, the performance of the lubricant [3].

Viscosity and density are two crucial properties of base oils that significantly impact the hydrodynamics of lubrication [4]. These properties directly influence friction, wear, and the overall lifespan of mechanical systems [5,6]. Kinematic viscosity, ratio of dynamic viscosity (referred to as viscosity in the manuscript) and density at 40 and 100°C enable calculation of viscosity index [7] which is most an industrial standard parameter to quantify the broader temperature performance capability of lubricants. Therefore, selecting an appropriate base oil with viscosity and density that meet the specific requirements of an application is crucial [3,6]. However, the knowledge of physical, chemical, and thermodynamic properties is often unavailable for compounds that have not been characterized or synthesized yet. Additionally, experimental measurements for molecules generated *in silico* can be expensive and time-consuming [8,9]. To overcome these limitations, quantitative structure-property relationships (QSPR) models have been developed to establish quantitative relationships between molecular features of hydrocarbons and their density and viscosity under application conditions. QSPR is a molecular descriptor-based modeling approach that correlates measured physical or chemical properties with descriptor [11–15]. It has been extensively used to predict biological, toxicological, and physicochemical endpoints [16], especially in the pharmaceutical industry [9,16–20].

Previous studies have employed various multivariate statistical tools, such as multiple linear regression (MLR), polynomial regression (PR), cluster analysis, principal component analysis (PCA), and partial least-squares regression (PLS), to develop QSPR models [21–28]. These models have provided reasonably accurate predictions, but they often rely on experimentally determined values as descriptors, limiting their applicability for discovering new materials and predicting their properties. Recent advances in machine learning (ML) have significantly improved the accuracy of QSPR models for estimating lubricant properties [23,24,28-32]. Machine learning algorithms, such as artificial neural networks (ANN) and Gaussian process regression (GPR) have been utilized to develop robust models. These machine learning (ML)-based models have showcased enhanced and efficient predictive capacities compared to traditional experimental and molecular dynamics (MD) simulation methods [23,24,28-32]. This is primarily because, once an accurate model is formulated, it serves as a more time-efficient and cost-effective alternative.

Several studies have employed machine learning to predict the viscosity of different systems, including biofuels, ionic liquids, alkanes, lubricants, Lennard-Jones fluids, and binary mixtures [33-39]. Notably, some of these studies [37,38] have also utilized MD simulations to generate viscosity data. For instance, the predictions of kinematic viscosity for alkanes achieved impressive R^2 values of 0.998 and 0.899 using artificial neural network and free volume theory,

respectively [36]. However, this study developed a viscosity model for a relatively smaller number of alkanes with up to 20 carbon atoms. These studies demonstrate the efficacy of machine learning and molecular dynamics simulations in predicting viscosity for various systems.

Most previous QSPR models of viscosity used static descriptors, those derived directly from chemical formulas. Most studies qualitatively correlated viscosity whereas only a few studies quantitively correlated viscosity with dynamics descriptors obtained from MD simulations [48-55]. Although some recent examples demonstrate the power of molecular dynamics-based descriptors in predicting material properties, such as supramolecular gelation [56], no previous studies have combined static and dynamic descriptors in QSPR models and compared their performance side-by-side.

In this study, we propose a QSPR approach to develop temperature-dependent viscosity and density models for complex hydrocarbons. Our models utilize both static and dynamic molecular descriptors and employ GPR as a simple and interpretable machine learning algorithm. Various parameters are used to assess the model quality during training, validation, and testing, and the models' predictive capabilities are evaluated on independent subsets of molecules. Additionally, we employ model-agnostic interpretation techniques to determine the impact of each descriptor on the model predictions. This detailed interpretation, along with the significance of model terms, can aid in the selection of existing hydrocarbons or the design of new molecules with desired viscosity and density. Furthermore, the approach presented in this study can be extended to predict other important material properties for a range of applications.

Methods and Materials

This section outlines the design and training process of the descriptor-based ML models. The flowchart in Figure 1 shows the overall workflow of the ML approach. Step 1 is the collection of experimental data used to train, validate, and test the models. Step 2 is calculation of a large set of molecular descriptors or model predictors (predictors = descriptors + operating conditions). Step 3 is selection of the significant predictors from the set of predictors using the least absolute shrinkage and selection operator (LASSO) regularization, *F*-test, and elimination of the strongly correlated predictors using correlation/collinearity analyses. Step 4 is development of models with all possible combinations of significant predictors using GPR. Step 5 is the optimization of hyperparameters and then selection of the best models using model assessment parameters: coefficient of determination (R^2), root mean squared error (RMSE), and the variance inflation factor (VIF) values. Step 6 is model-agnostic interpretation of the best models by partial dependency plots (PDP), individual conditional expectation (ICE) plots, average local-interpretable model-agnostic explanations (LIME), and relative decrease in R^2 values due to trimming a predictor. Lastly, Step 7 is evaluation of the final models using test data.

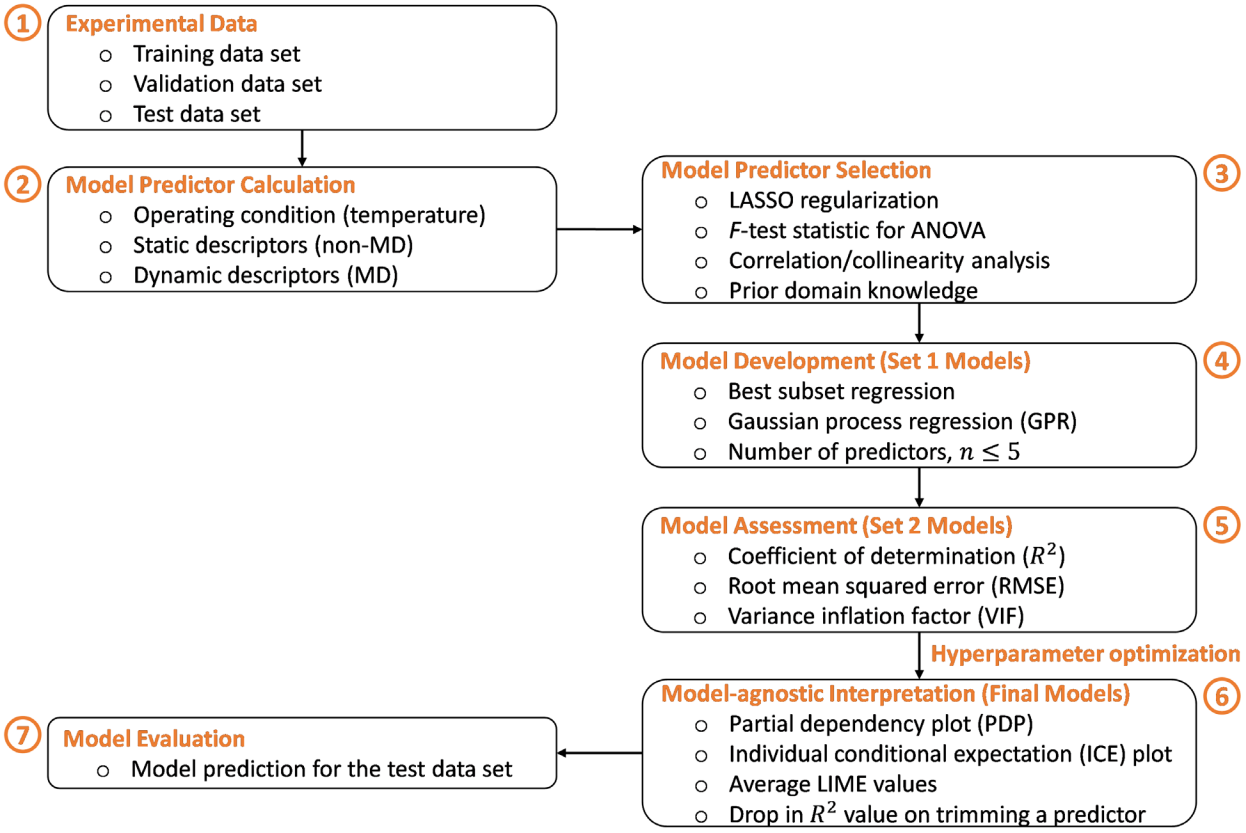


Figure 1: The overall workflow of the ML approach to design, train, and evaluate the models.

Experimental Data

The dynamic viscosity and density of 305 pure hydrocarbons (C_8 to C_{50}) at a wide range of temperatures were obtained from the American Petroleum Institute (API) Research Project 42 [57]. These 305 hydrocarbons include n-paraffins, branched-paraffins, 1-olefin, branched-olefins, non-fused ring naphthene, fused ring naphthene, non-fused ring aromatic, and fused ring aromatics. Schematics of some of the hydrocarbons are shown in Figure 2 to illustrate the diversity of molecule structures. The molecular weights of the 305 hydrocarbons range from 110.20 to 703.30 g/mol. The viscosities range from 0.29 cP to 2.00×10^4 cP, and the densities range from 0.67 g/cc to 1.12 g/cc. The viscosities and densities of these molecules, along with their molecular formulas and simplified molecular input line entry system (SMILES) [58] codes can be found in the Supporting Information (Tables S1 and S2). In addition, schematics of all the molecules can be found in the Supporting Information. Viscosity and density data at atmospheric pressure and temperatures ranging from 0 °C to 135 °C were used to develop the models. The API Research Report provided both density and viscosity for most hydrocarbons. However, for some hydrocarbons or temperatures, only density or viscosity was reported. In total, 1292 viscosity data points and 1474 density data points were included in the model development.

Due to the large dataset, we used the holdout cross-validation technique. Data were divided randomly into three partitions to develop and assess the models: training, validation, and test

datasets. First, 70% of the 305 molecules (215 molecules) were used to train the models, and then 15% of the 305 molecules (45 molecules) were used to validate the models during development. Lastly, the remaining 15% of the 305 molecules (45 molecules) that were not in the training and validation datasets were selected to assess the accuracy of the predictions from the developed models. All molecules are listed in Table S1, each identified with a partition ID, either 1, 2, or 3, to indicate if it was in the training, validation, or test dataset. All the experimental data can be found in the Supporting Information.

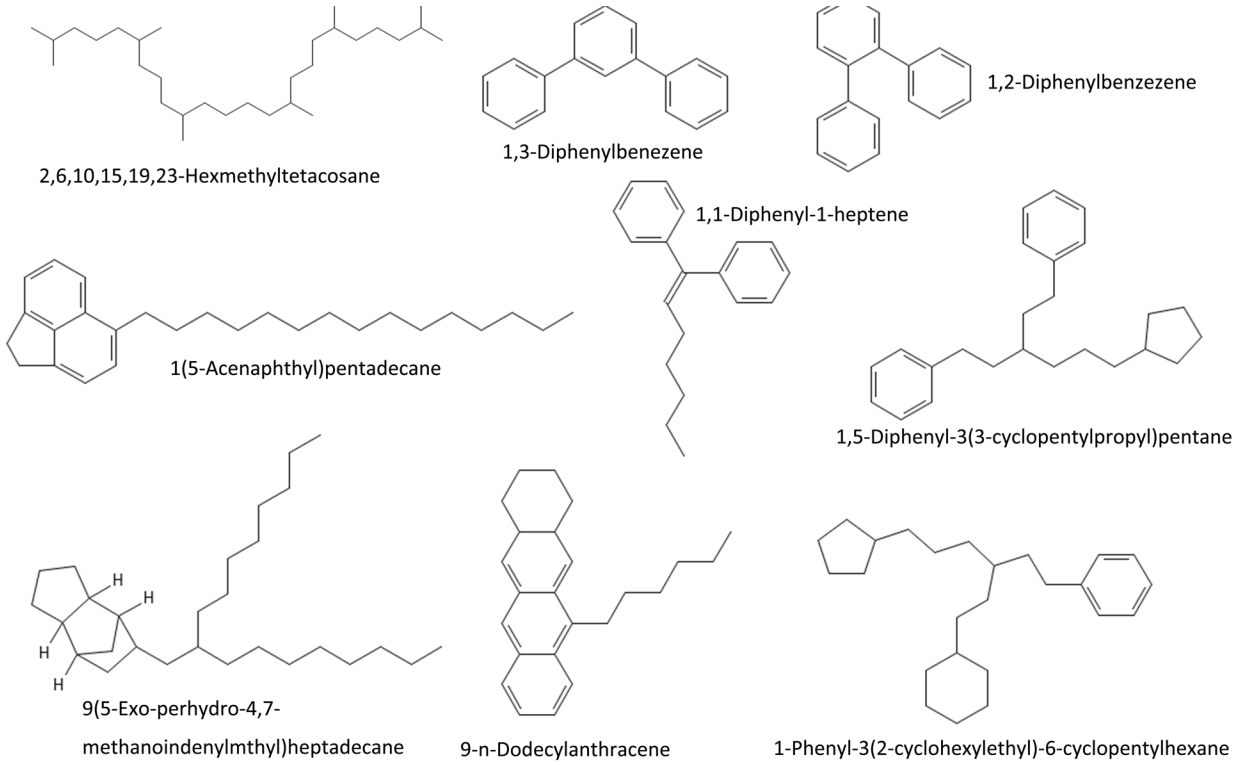


Figure 2: Structures of representative hydrocarbons used in the model training, validation, or test datasets.

Model Predictors

The model predictors used in this study were temperature and molecular descriptors. Temperature was included as an operating condition predictor since viscosity and density are both inversely related to temperature [59–62]. Molecular descriptors are mathematical representations of the physical and chemical properties of molecules [46]. In this study, we classified molecular descriptors into two categories based on the level of molecular representation required for calculating them, either static descriptors or dynamic descriptors. Static molecular descriptors are one-dimensional (1D) and two-dimensional (2D) descriptors that do not require three-dimensional (3D) coordinates of the atoms in a molecule for calculation. Dynamic molecular descriptors are 3D descriptors, commonly known as geometric descriptors, that require the 3D coordinates of atoms for calculation. Dynamic descriptors are more robust and better able to capture the molecular conformations under different operating conditions such as temperature, pressure, and speed. However, dynamic descriptors require a higher computational cost to calculate. Since all predictors, except temperature, were molecular descriptors, the terms “predictors” and “descriptors” can be effectively used interchangeably.

A total of 1444 static descriptors were obtained using an open-source software, PaDEL [63] by providing SMILES codes [58] of the molecules in Table S1. The details of all 1444 static descriptors can be found in the Supporting Information. In addition, 156 dynamic descriptors were determined using MD simulations. Of these, 57 dynamic descriptors were directly obtained from the simulations and 99 dynamic descriptors were calculated by postprocessing the atomic trajectories from the simulations. The 57 dynamic descriptors obtained directly from MD simulations include stress tensor, energies, density, volume, and dipole moment. The remaining dynamic descriptors were calculated via our open-source Python package, PyL3dMD [51], which utilizes the 3D coordinates, connectivity, and charge of atoms obtained at each timestep of the MD simulation. As it is not possible to provide the physical and chemical significance of all descriptors used in this study, only the significance of highly correlated predictors is provided along with the model interpretation. The definition of each descriptor used in this study can be found in the Supporting Information.

For each molecule, a cubic model system with periodic boundaries containing around 5000 atoms and 5.0 nm^3 was created using an open-source software called Packmol [64]. The model volume of the systems was large enough to minimize finite-size effects across the periodic boundary [65] and to reduce pressure and stress fluctuations [66], enabling accurate and reliable simulations. All atomic interactions were described using the All Atom Optimized Potentials for Liquid Simulations (OPLS) [67], which is one of the most popular and accurate potentials for calculating transport properties of hydrocarbons. The OPLS parameters for the hydrocarbons were obtained using the LigParGen [68] and BOSS [69] software packages, which are open-source. We used the CM1A-LBCC [70] charge model to assign charges to the hydrocarbon atoms. Finally, dynamic simulations were run using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software with a time step of 1.0 fs [71].

The following is the protocol for the MD simulations from which the dynamic descriptors were calculated. First, a previously developed, robust equilibration molecular dynamics (EMD) simulation approach was followed to optimize the geometries of the molecules at atmospheric pressure and the same temperatures as given in the API Research Report [72]. Energy minimization of the system was performed using the conjugate gradient algorithm. Second, the system was heated to 1000 K for 0.25 ns in the canonical (NVT) ensemble to achieve homogeneity. Third, the system density was equilibrated at 1.0 atm and a target temperature for 1.0 ns in the isothermal–isobaric (NPT) ensemble using the Nosé–Hoover thermostat and barostat [73,74], with damping coefficients of 100 and 25 fs, respectively. Fourth, while maintaining a constant temperature for 0.5 ns in the NVT ensemble, the simulation box was deformed until the density of the fluid reached the average density computed from the previous NPT simulations. Finally, the system was equilibrated using the final configuration from NVT as the initial configuration for 0.25 ns in the microcanonical (NVE) ensemble.

After equilibration, two independent production runs of 1.0 ns and 3.0 ns were carried out in the NPT and NVE ensembles, respectively, for calculating dynamic descriptors. The dynamic descriptors for the density models were calculated from the NPT ensemble, since models that allow volume fluctuation are commonly used to calculate density in MD. For the viscosity models,

dynamic descriptors were calculated from the NVE ensemble, to be consistent with typical simulation methods used to calculate viscosity from MD simulations with the Green-Kubo formula [48,55]. The density, *rho*, calculated from the NPT ensemble, was used as a descriptor for both the density and viscosity models. The trajectories of atoms were stored every 1000 fs in each production run. The dynamic descriptors were calculated based on each stored frame of the trajectories and averaged over the last 50% of time in the production simulation. The LAMMPS scripts and data files of all molecules and the calculated static and dynamic descriptors are provided in the Supporting Information.

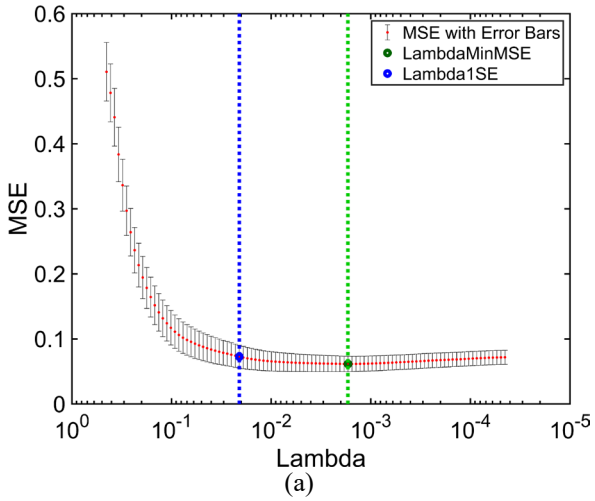
It is worth noting that a key difference between directly calculating density or viscosity from MD simulations and using simulations in the same ensembles for calculating descriptors is that the latter requires significantly less computational time. For example, here, we used only 3 ns of NVE simulation data to predict viscosity using molecular descriptors from the simulation whereas, in our previous studies [48,55], we had to run simulations for 400 ns, to accurately calculate the viscosity of a lubricant.

Model Predictor Selection

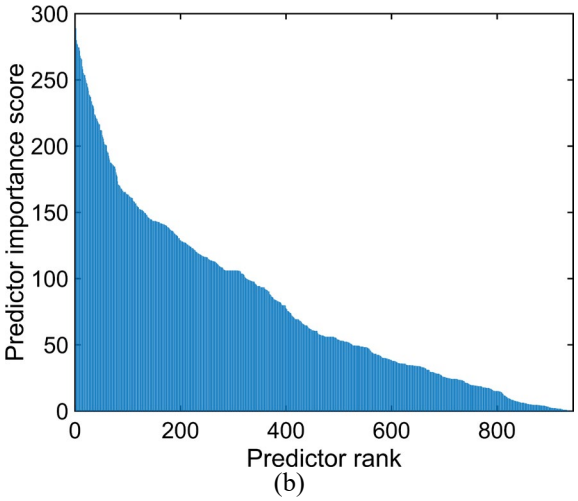
Developing models of all possible combinations of a large set of predictors is inefficient and only feasible using supercomputing resources. Predictor selection reduces the dimensionality of data by selecting only a subset of predictors to create a model that accurately predicts measured responses. The primary objectives of predictor selection techniques are to improve prediction performance, provide faster and more cost-effective predictors, and improve model interpretability [75]. Therefore, after gathering a large complex set of potential predictors, the next step was to remove redundant, unimportant, and strongly correlated predictors to avoid unreliable and unstable estimates from the regression models. Therefore, LASSO regularization [76,77], *F*-test [78], correlation/collinearity analysis [79,80], and prior domain knowledge were used to remove redundant and strongly correlated predictors and select the most important predictors of the response variable (density/viscosity).

Here, we explain the predictor selection approach for static descriptors used to develop viscosity models. We started with 1444 static descriptors. After removing descriptors with any missing values, infinite values, or the identical values for all molecules, we were left with 944 static descriptors. Next, a LASSO fit with 10-fold cross-validation was performed, and the descriptors in the sparsest model within one standard error of the minimum mean squared error (MSE) (*LambdaMinMSE*, as shown in Figure 3a were selected. In LASSO regularization, the coefficients of covariates that were strongly correlated with one another or were less relevant to the response variable (in this case, viscosity) were eliminated from the pool of predictors. After performing LASSO regularization, we were left with only 37 identified important predictors. Secondly, using an *F*-Test, we ranked the importance of all 944 static descriptors for the response variable, as shown in Figure 3b. The *F*-test is the statistic used for analysis of variance (ANOVA) to examine the importance of each predictor individually. The *p*-value, also known as probability value, is a statistical measurement used to validate a hypothesis against observed data. A small *p*-value or a large negative $log(p)$ value of the test statistic indicates the importance of the corresponding predictor. In this context, the negative logarithmic *p*-value serves as the predictor's

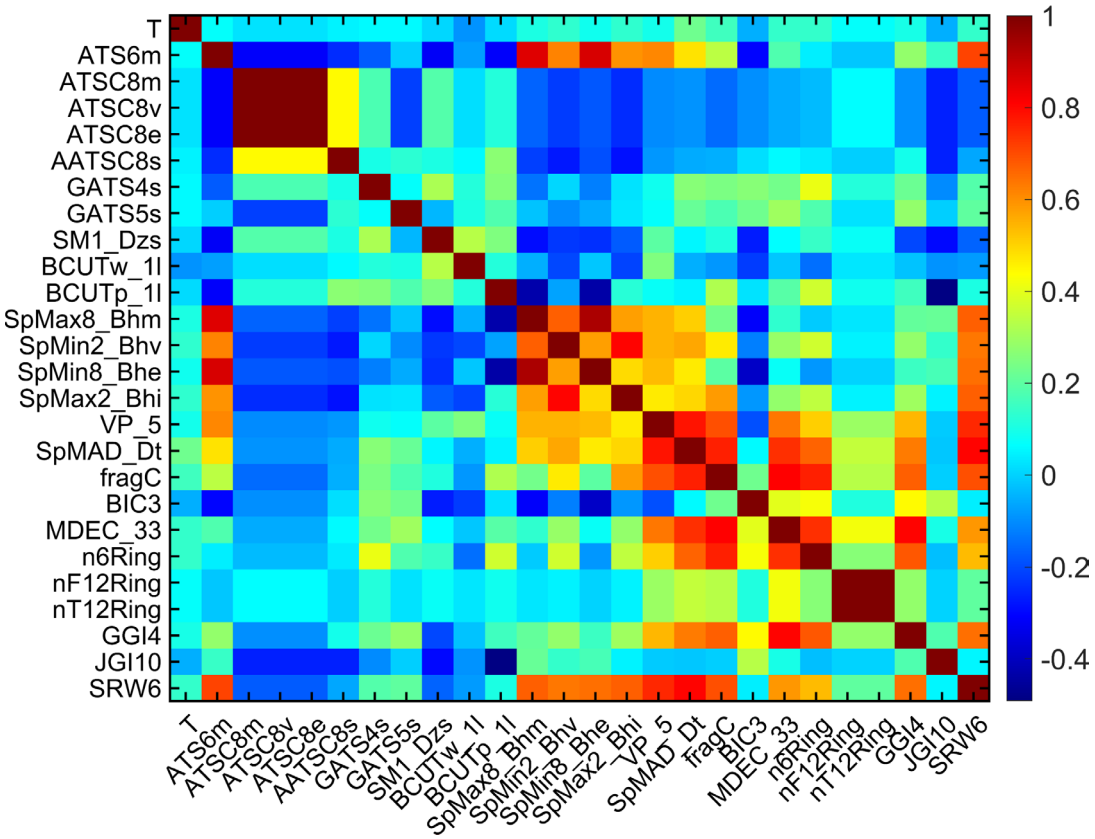
score, indicating its importance of the corresponding predictor. Using the scores of the *F*-tests and prior domain knowledge of important predictors, we included highly important predictors that LASSO regularization might have removed from the pool. Thirdly, a correlation matrix was used to assess the cross-correlation of the predictors and remove strongly correlated predictors, as shown in Figure 3c. The correlation matrix is a standard measure of the strength of pairwise linear relationships. Finally, predictors that were not eliminated at this modeling step were considered the most significant and were used for developing the viscosity models. The same predictor selection approach was used for the static descriptors of the density models and the dynamic descriptors of the density and viscosity models.



(a)



(b)



(c)

Figure 3: (a) MSE of the LASSO fit using 10-fold cross-validation. The lambda that results in the lowest MSE is the green dotted line whereas the blue dotted line is the lambda that is within one standard error of the lowest MSE. (b) Score of all predictors using the *F*-test where, in this example, the most important predictor is given rank 1 and the least important predictor is given rank 944. (c) Pairwise linear correlation coefficients of the descriptors where the dark red and dark blue represent highly positively and negatively correlated predictors, respectively.

Model Development and Assessment

After selecting important predictors and randomly dividing the experimental data into three datasets for training, validation, and testing, ML algorithm-based models were developed. GPR was chosen for its flexibility and tractability. GPR models are nonparametric kernel-based probabilistic models [81]. GPR was combined with the best subset regression approach to develop GPR-based models using each possible combination of predictors from the pool of selected important predictors. In this approach, all possible models were developed with up to five predictors or until a significant increase of the R^2 [82] was observed by increasing the number of predictors. The models generated at this stage were designated as the first set of models. As described in the workflow of the model development in Figure 4, the best subset regression approach was applied. The goal of this approach was to choose a subset that maximizes model performance while minimizing complexity, which could prevent overfitting by using the minimal number of predictors that is necessary for the model but no more [83].

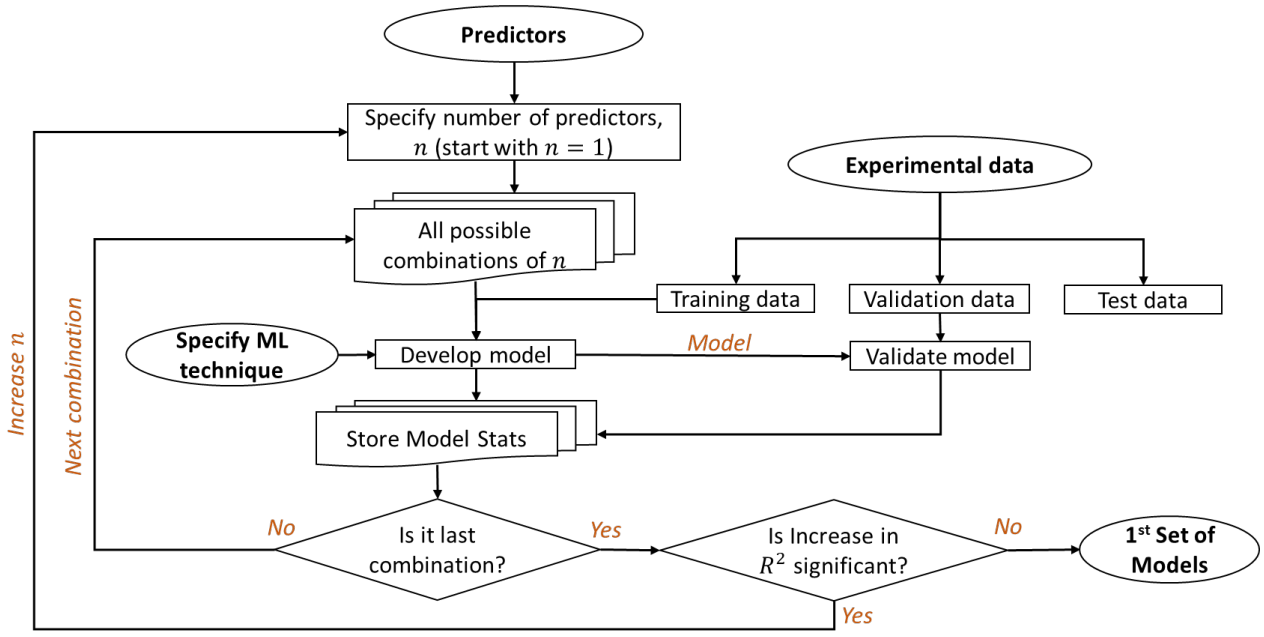


Figure 4: Workflow of the model development step.

In this step of model development, a holdout cross-validation technique was implemented to validate the trained models, and the statistics of the training and validation sets were recorded to assess the quality of the models. These statistics include R^2 and RMSE values for the training and validation datasets, as well as the VIF of each predictor. The predictive performance of the models was assessed based on the R^2 and RMSE values. R^2 is a statistical measure of fit that quantifies the variation of the response variable that can be predicted by the predictor(s) in a regression model. R^2 was calculated using Equation 1.

$$R^2 = \frac{\sum_{i=1}^{N}\left(y_i^{exp} - \bar{y}\right)^2 - \sum_{i=1}^{N}\left(y_i^{exp} - y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(y_i^{exp} - \bar{y}\right)^2} \quad (1)$$

where, N is the total number of data points, i is the i^{th} data point, y_i^{exp} is the experimental value of the response variable, y_i^{pred} is the model predicted value of the response variable, and $\bar{y}$ is the mean experimental value of the response variable. The RMSE, which is a measure of the difference between values predicted by a model and values observed by experiment, was calculated using Equation 2.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{pred}\right)^2} \quad (2)$$

VIF was used to evaluate the multicollinearity between three or more predictors with the response variable and with each other. VIF was calculated using Equation 3 [80].

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} = diag(\mathbf{C}^{-1}) \quad (3)$$

where, $\mathbf{C}$ is the correlation matrix or matrix of the correlation coefficient, and R_j^2 is the R^2 of predictor j on the remaining predictors. When the variation/trend of a predictor j is nearly a linear combination of the other predictors, then R_j^2 is close to 1 and the *VIF* for that predictor is correspondingly large. If R_j^2 is 0 (no collinearity), then VIF is 1, which is the lowest possible value of VIF. We used VIF > 5 as a benchmark for the presence of multicollinearity [84] and discarded models with VIFs higher than 5.

The top 100 models with the highest R^2 value, lowest RMSE value, and VIF values less than or equal to 5 for all predictors were selected as the second set of the models. These models exhibit sensitivity to the numerous hyperparameters in the GPR algorithm, impacting their predictive performance. Therefore, hyperparameters of the models in the second set were optimized using a 5-fold cross-validation technique. The hyperparameters were tuned by exploring the multidimensional combinatorial hyperparameter space using the Bayesian optimization algorithm [85,86]. Bayesian optimization was chosen because, unlike other optimization techniques, it utilizes information from past function evaluations and does not solely rely on local gradient and Hessian approximations [87]. This enables the optimization search to rapidly reach the minimum, even for nonconvex functions [88].

After retraining the top 100 models with optimized hyperparameters, we chose the best model for each response variable as the model with the highest R^2 value, the lowest RMSE value, VIFs ≤ 5, and for which the trends of PDPs [89,90] were consistent with the expected physical behavior. For example, it is commonly known that the viscosity and density of liquids decrease as temperature increases. If a model does not match the expected trend, it is not correct, even if it has a perfect R^2 of 1.0 (i.e., 100% accurate). PDP depicts the marginal effect of a predictor on the outcome of a model, and the extent of change in the response variable to a change in a predictor indicates the global importance of that predictor. In addition, for two models with similar statistics, preference was given to models with simple descriptors that are easy to understand and calculate, when two models had similar statistics.

Model Interpretation and Evaluation

To better understand the predictions, we systematically interpreted our best models and their predictors. We conducted model-agnostic interpretation using PDPs [89,90], ICE [90], and LIME [91]. PDP is a tool to investigate global importance, which represents the contribution of a predictor to the overall prediction of data, whereas ICE and LIME are tools used to investigate local importance, which represents the contribution of a predictor to the prediction of each data point. The PDPs do not reveal hidden dependencies because they only show averaged relationships between a predictor and response variable. ICE plots can be used to identify interactions among model variables and detect unusual subgroups in the datasets [92]. Therefore, to investigate heterogeneities in partial dependence originating from different observations, ICE plots were generated for each predictor in the best models. As the name suggests, the LIME value or weight

represents local importance, however, the mean of the LIME values for all data points can be used as a global representation of predictor importance. A positive (negative) mean LIME value implies a positive (negative) relationship of the predictor with respect to the response variable. The importance of predictors was analyzed by trimming predictors one at a time from the best model and observing the performance (R^2) of the trimmed model. Finally, the best models for viscosity and density were evaluated with a new set of hydrocarbons (test dataset) to verify their predictive performance across a wide range of temperatures.

Results and Discussion

Density Models

We were able to achieve good temperature-dependent density models with three or fewer predictors. The three best static descriptor-based models for temperature-dependent density with one, two, and three static predictors are Equations 4-6, called Model I, II, and III. Model I is only a function of Broto-Moreau autocorrelation-lag 2/weighted by Sanderson electronegativities ($AATS2e$). $AATS2e$ is a spatial autocorrelation calculated from molecular graph, that is connectivity of atoms of a molecule [93] where e in $AATS2e$ is the Sanderson electronegativity [94] of atoms in a molecule, whereas 2 in $AATS2e$ is the lag or the topical distance between two connected atoms in a molecule. Therefore, it is a measure of molecular connectivity and complexity. Example calculations of these static descriptors can be found elsewhere [46]. Model II is a function of temperature (T) in °F and $AATS2e$. Model III is a function of temperature, conventional bond order ID number of order 3 ($piPC3$), and fraction of rotatable bonds, including terminal bonds ($RotBtFrac$). The conventional bond order ID number is a molecular weighted path number calculated from weighting graph edges (bonds) with conventional bond order, which is defined as 1, 2, 3, or 1.5 for single, double, triple, or aromatic bonds, respectively. $piPC3$ is a conventional bond order weighted measure of molecular connectivity and complexity [46]. The conventional bond orders for single, double, triple, and aromatic bonds are 1, 2, 3, and 1.5 [46]. $RotBtFrac$ is the fraction of rotatable bonds over the total number of bonds in a molecule [46]. Rotatable bonds are bonds that meet the three following criteria: (a) single bond connected by heavy atoms with the heavy atoms connected to at least one atom (including hydrogen atom), (b) the external bond by which the heavy atom is connected must not a triple bond unless the triple bonded atom is connected to another atom, and (c) the bond must not be part of a ring [46].

Model I: $$\rho(T) = f(AATS2e) \quad (4)$$

Model II: $$\rho(T) = f(T, AATS2e) \quad (5)$$

Model III: $$\rho(T) = f(T, piPC3, RotBtFrac) \quad (6)$$

The best three dynamic descriptor-based models for density with one, two, and three dynamic predictors are Equations 7-9. Model I is only a function of simulation-calculated density (rho) from the NPT ensemble. Model II is a function of rho and the radius of gyration (R_g) of the molecule which quantifies molecular size. Model III is a function of rho, R_g, and energy due to

van der Waals interactions ($evdwl$). Note that *rho* is different from ρ, although both are density: *rho* is a dynamic descriptor calculated from the simulations and ρ is the experimentally measured fluid density predicted by the ML model. We include *rho* as a descriptor because it can be calculated from a very short simulation and is part of the PyL3dMD python package [51]. The density models with dynamic descriptors excluding the simulation-calculated density are also provided in the Supporting Information.

Model I: $$\rho(T) = f(rho) \quad (7)$$

Model II: $$\rho(T) = f(rho, R_g) \quad (8)$$

Model III: $$\rho(T) = f(rho, R_g, evdwl) \quad (9)$$

Table 1: Model assessment parameters for the density models with static and dynamic descriptors.

Parameter		**Static Descriptors**			**Dynamic Descriptors**		
		Model I	Model II	Model III	Model I	Model II	Model III
Training	R^2	0.893	0.989	0.999	0.983	0.994	1.000
	RMSE	0.025	0.008	0.003	0.010	0.006	0.001
Validation	R^2	0.910	0.993	1.000	0.985	0.994	1.000
	RMSE	0.024	0.007	0.001	0.010	0.006	0.000
Test	R^2	0.874	0.935	0.981	0.971	0.977	0.988
	RMSE	0.029	0.021	0.011	0.014	0.012	0.009
Average R^2		0.892	0.972	0.993	0.980	0.988	0.996
Maximum VIF		1.000	1.003	1.316	1.000	1.203	4.113

Table 1 lists the R^2 and RMSE values for the density models with static and dynamic descriptors for the training, validation, and test datasets. The average R^2 values for the training, validation, and test datasets are also reported to enable comparison of the models. The maximum VIF for each predictor is also given to indicate the degree of multicollinearity. Due to the single predictor in Model I with both static and dynamic descriptors, the VIF value is a perfect 1, but the VIF increased as the number of predictors increased in Models II and III. The statistics in Table 1 show that the model (Equation 4) with a single static descriptor $AATS2e$ was able to reach an accuracy of 89.2% in predicting the density of the hydrocarbons. When a temperature term was added to the model, that is, in Model II (Equation 5), the R^2 for the test dataset increased to 97.2%. The best model with the static descriptors, Model III (Equation 6), has R^2 values of 99.9%, 100.0%, and 98.1% for the training, validation, and test datasets.

From the statistics in Table 1, the density model with a single dynamic descriptor (Equation 7), rho calculated from only 1.0 ns of simulation time, was able to achieve an accuracy of 98.0% in predicting density. This is higher than any single static descriptor and any combination of two static descriptors. When other simulation-calculated descriptors were included, i.e., Models II and

III in Equations 8 and 9, the R^2 values increased to around 99%. The best density model with dynamic descriptors, Model III (Equation 9), had R^2 values of 100.0%, 100.0%, and 98.8% for the training, validation, and test datasets. The perfect R^2 value on the training dataset indicates that the complexity of the model was able to describe the relationship between the descriptors and the densities.

Figures 5a and 5b show experimental density and density predicted by the best model (Model III) with static and dynamic descriptors for all datasets over a wide temperature range. The blue dashed lines represent ideal predictions. The model-predicted density for the training, validation, and test datasets is shown as black circles, red squares, and green triangles, respectively. From the statistics in Table 1 and Figure 5, we can see both models performed exceptionally well with only three descriptors. Furthermore, the model with three dynamic descriptors performed slightly better than the model with three static descriptors.

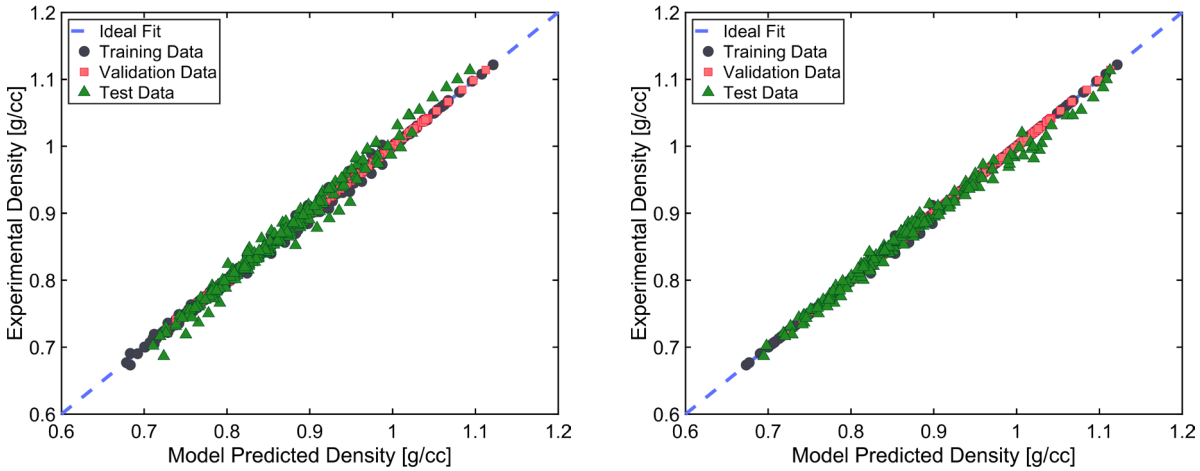


Figure 5: Experimental density vs. density predicted by the best models (Model III) with (a) static and (b) dynamic descriptors for the training (black circles), validation (red squares), and test data (green triangles) sets. The blue dashed lines represent the ideal prediction.

We can visualize the relationships between each model predictor in a trained regression model and model-predicted responses using the PDPs and ICE plots. In Figure 6, the circle symbols show the predicted response for each data point. The PDP (red line) shows the averaged relationship, whereas ICE plots (gray lines) show an individual dependence for each observation [90], resulting in one line per observation. The PDPs are offset such that the y-axis starts at zero to illustrate cumulative effects and the importance of each predictor in the model. Figures 6a and 6b show the relationship between the model-predicted density and each static predictor in Equation 6 and each dynamic descriptor in Equation 9, respectively.

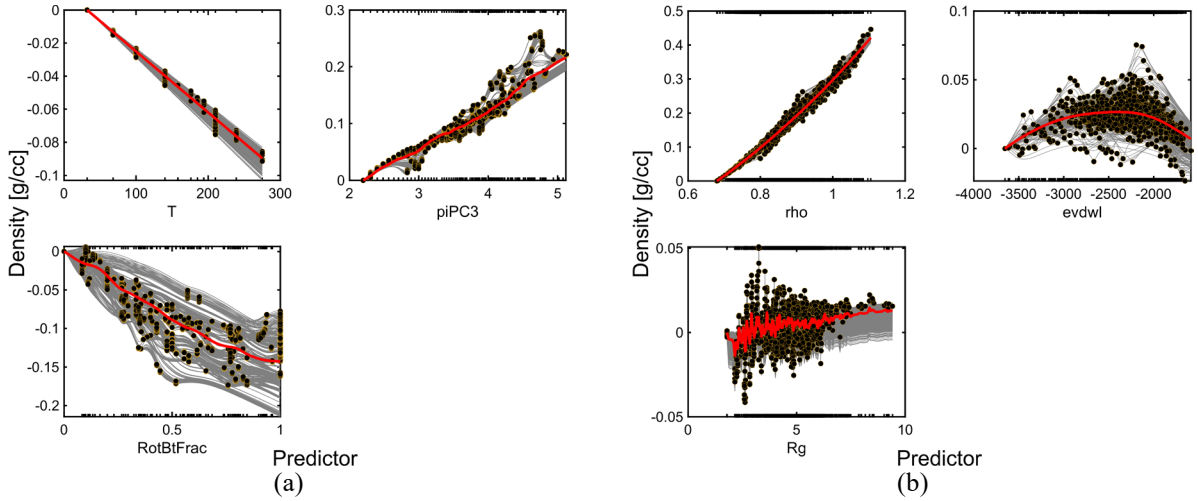


Figure 6: The partial dependence (red curve) and the individual conditional expectation (grey curves) of each predictor in the best density models (Model III) with (a) static and (b) dynamic descriptors. The scattered circular symbols represent the relationship between a predictor variable and density for each observation. The data are offset so that the density starts from zero to better illustrate the cumulative effect of a predictor on density.

PDPs and ICE plots show that the density decreases with increasing temperature (T) in °F and fraction of rotational bonds in the molecules ($RotBtFrac$) but increases with increasing $piPC3$. A larger value of $RotBtFrac$ indicates greater ease of rotation of the backbone in the molecules (i.e., chain flexibility) [95]. In Figure 6b, it must be noted that the negative sign of $evdwl$ suggests that the interaction is driven by an attractive force. Therefore, a higher negative value of $evdwl$ means stronger van der Waals interactions. Figure 6b shows that the experimental density and *rho* are directly and strongly correlated, as expected. However, the trends for radius of gyration (R_g) and van der Waals interaction energy ($evdwl$) are not definitive from these plots. As a complementary analysis, Figure 7 shows the average LIME values over all observations for the models with static and dynamic descriptors. The sign of the LIME values reveals that, on average, the density of hydrocarbons decreases with increasing T, $RotBtFrac$, R_g, and $evdwl$, but increases with increasing $piPC3$ and rho.

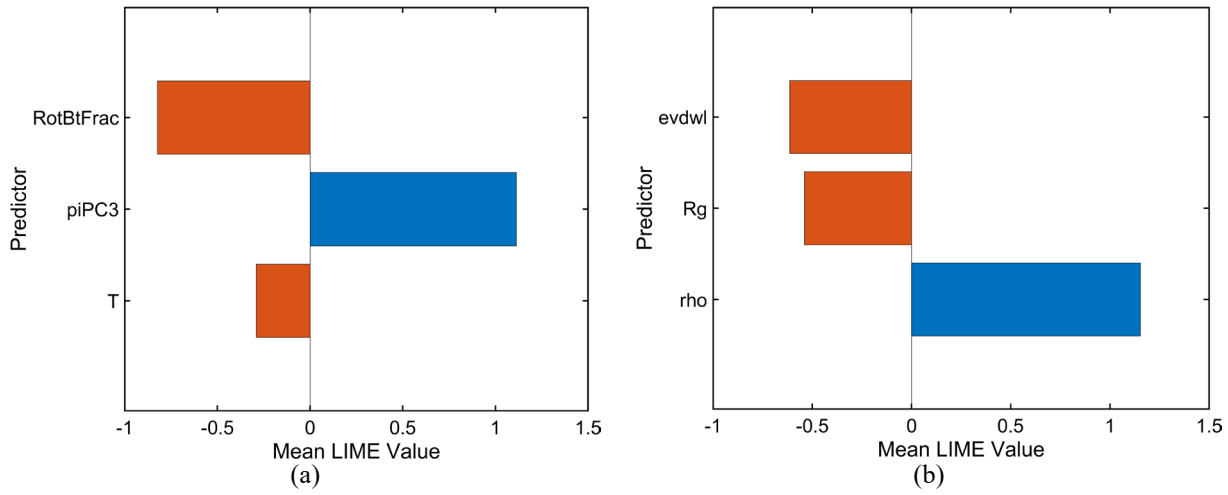


Figure 7: The average LIME value for each predictor in the best density models (Model III) with (a) static and (b) dynamic descriptors. The orange and blue colors represent negative and positive relationships between a predictor and the response variable. The size of a bar represents the overall importance of a predictor.

To rank the importance of each predictor in the models, we removed each predictor one at a time and observed the performance of the model with a dropped predictor. Table 2 gives the average R^2 of the training, validation, and test datasets for each predictor when it was trimmed from the best density model. For instance, when the T term was dropped from the best density model with static descriptors, the average R^2 value decreased from 99.3% to 90.0%, that is, by 9.3%. A larger decrease in R^2 indicates more importance. The decrease in performance of the best model was 9.3%, 18.3%, or 21.4% when the T, $RotBtFra$, or $piPC3$ term was dropped from the best density models. Therefore, the order of importance of the predictors in the best density model with static descriptors is $piPC3 > RotBtFrac > T$. This is consistent with the magnitudes in the LIME plots in Fig. 7a. Similarly, for the best density model with dynamic descriptors, the order of predictor importance is $rho > R_g \approx evdwl$. Both $evdwl$ and R_g had only small effects on the density model compared to the simulation-calculated density. Analysis of both static and dynamic models indicates that the most important predictors of density are T, $RotBtFrac$, $piPC3$, and rho.

Table 2: Performance of the best density models when a predictor was removed from the models.

Static Descriptors (Equation 6)			**Dynamic Descriptors** (Equation 9)		
Term Dropped	Avg. R^2	Drop in Avg. R^2	Term Dropped	Avg. R^2	Drop in Avg. R^2
None	99.3%	0.0%	None	99.6%	0.0%
T	90.0%	9.3%	$evdwl$	98.8%	0.8%
$RotBtFrac$	81.0%	18.3%	R_g	98.1%	1.5%
$piPC3$	77.9%	21.4%	rho	82.0%	17.6%

We also developed density models with combined static and dynamic descriptors. The combined models performed slightly better (R^2 of 0.997) than the models with only static (R^2 of 0.993) and only dynamic descriptors (R^2 of 0.996). Therefore, they were not analyzed further but are provided in the Supporting Information.

Viscosity Models

Instead of training models directly for viscosity (η), we trained the models for logarithmic viscosity ($\log \eta$), based on the knowledge the viscosity decreases exponentially with temperature. Unlike density, we were not able to achieve good temperature-dependent viscosity models with one or two predictors. The best three models with two, three, and four static predictors for temperature-dependent viscosity are Equations 10-12. Model I is a function of temperature (T) in °F and the first kappa shape index ($Kier1$) [96]. Model II is a function of T, $RotBtFrac$, and the molecular weight (MW). Model III has the same terms as Model II plus the $PetitjeanNumber$ [43]. The terms $RotBtFrac$ in Model II and $RotBFrac$ in Model III are the fraction of rotatable bonds over the total number of bonds, including ($RotBtFrac$) and excluding ($RotBFrac$) terminal bonds, respectively. $Kier1$ is a connectivity descriptor which quantifies the complexity in connectivity of the molecule. $PetitjeanNumber$ is a topological anisometry descriptor which

quantifies the molecular shape [43]. It is calculated from the generalized radius and diameter of the molecule [43]. Example calculations of these static descriptors can be found elsewhere [43,46].

Model I: $$\log\eta(T) = f(T, Kier1) \quad (10)$$

Model II: $$\log\eta(T) = f(T, RotBtFrac, MW) \quad (11)$$

Model III: $$\log\eta(T) = f(T, RotBFrac, MW, PetitjeanNumber) \quad (12)$$

The best three models for viscosity with two, three, and four dynamic predictors are Equations 13-15. Model I is a function of rho and kinetic energy (ke). Model II is a function of rho, energy due to improper interactions ($eimp$), and acylindricity (c). $eimp$ quantifies the stiffness of the molecule, proportional to the inverse of rotatable bonds in the molecule. c is a measure of cylindricity in the distribution of atoms in a molecule. c is zero when a molecule is cylindrically symmetric and increases as the molecule deviates from this shape [88]. Model III is a function of rho, $eimp$, T, and a diagonal component of the moment of inertia tensor (I, i.e., size or distribution of atomic mass from the center of mass of a molecule). It was found that any diagonal component (I_{xx}, I_{yy}, I_{zz}) of the moment of inertia tensor resulted in similar predicting performance when used in Equation 15, likely because the dynamic descriptors were calculated from equilibration MD simulations with a cubic simulation box.

Model I: $$\log\eta(T) = f(rho, ke) \quad (13)$$

Model II: $$\log\eta(T) = f(rho, eimp, c) \quad (14)$$

Model III: $$\log\eta(T) = f(T, rho, eimp, I_{zz}) \quad (15)$$

Table 3: Model assessment parameters for the dynamic viscosity models with static and dynamic descriptors.

Parameter		Static Descriptors			Dynamic Descriptors		
		Model I	Model II	Model III	Model I	Model II	Model III
Training	R^2	0.949	0.989	0.991	0.942	0.994	0.999
	RMSE	0.161	0.076	0.070	0.173	0.010	0.020
Validation	R^2	0.970	0.988	0.989	0.95	0.987	0.998
	RMSE	0.129	0.082	0.078	0.167	0.003	0.031
Test	R^2	0.834	0.936	0.952	0.809	0.876	0.932
	RMSE	0.275	0.171	0.148	0.324	0.293	0.176
Average R^2		0.918	0.971	0.977	0.900	0.952	0.977
Maximum VIF		1.007	1.229	1.409	1.021	1.115	1.874

Table 3 gives the R^2 and RMSE values of the viscosity models with static and dynamic descriptors for the training, validation, and test datasets. The statistics in Table 3 show that the R^2

values for the best model with static descriptors are 99.1%, 98.9%, and 95.2% for training, validation, and test datasets. Similarly, the R^2 values for the best model with dynamic descriptors are 99.9%, 99.8%, and 93.2% for training, validation, and test datasets. The average R^2 value of 97.7% is the same for the models with static or dynamic descriptors. The very high R^2 value on the training dataset indicates that the complexity of the model was able to describe the relationship between the descriptors and the viscosities.

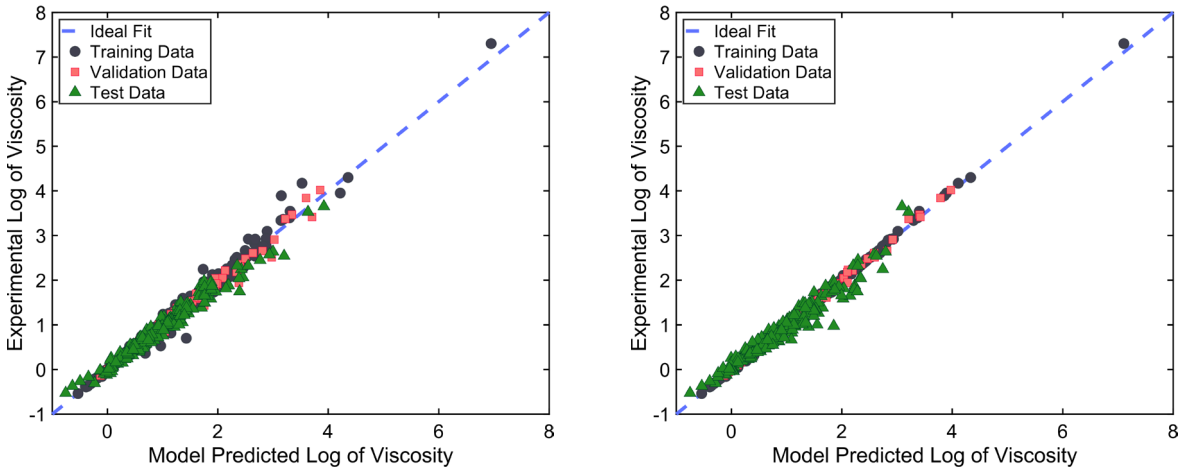


Figure 8: Model predicted viscosity obtained from the best models (Model III) with (a) static (b) dynamic descriptors for the training (black circles), validation (red squares), and test data (green triangles) sets. The blue dashed lines represent the ideal predictions.

Figure 8 shows experimental viscosity and viscosity predicted using the best models with static and dynamic descriptors for all datasets over a wide range of temperatures. Note that the y-axis is $\log \eta(T)$. From the statistics in Table 3 and Figure 9, we can see that both models performed exceptionally well with only three or four predictors, including temperature. The model with three static descriptors and the model with three dynamic descriptors performed the same, with an average R^2 value of 97.7%.

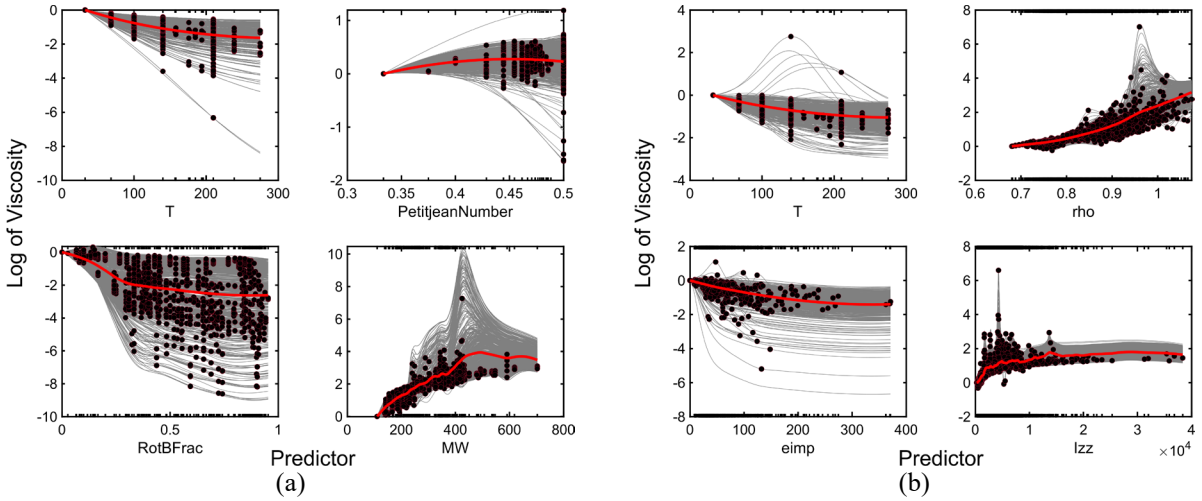


Figure 9: The partial dependency (red curves) and the individual conditional expectation (grey curves) of each predictor in the best viscosity models (Model III) with (a) static (b) dynamic descriptors. The scattered circular symbols represent the relationship between a predictor variable and viscosity for each observation. The plots are offset so that the viscosity starts from zero better to illustrate the cumulative effect of each predictor.

Like density, Figure 9 shows ICE plots (gray lines) and a PDP plot (red line) for each predictor. Figures 9a and 9b show the PDP and ICE plots of predictors in the best models with static (Equation 12) and dynamic (Equation 15) descriptors. The circle symbols are the predicted response by the predictor for each data point. PDPs of static descriptors show viscosity decreasing with increasing T and $RotBFrac$, increasing with MW. PDPs of dynamic descriptors show that viscosity decreases with increasing T and $eimp$, but increases with the *rho* and I_{zz}. Similar finding were reported in a recent QSPR study [30] where $PetitjeanNumber$, $RotBFrac$, and MW were found to be correlated with viscosity. Many previous experimental and simulation studies reported a power-law relationship between MW and viscosity [55,98], consistent with the trend in Figure 9b. Figure 10 shows the average LIME values over all observations for the viscosity models with static and dynamic descriptors. The direction of the bars in the LIME plots indicate the same trend of the predictor with viscosity as the PDPs and ICE plots in Figure 9, i.e., viscosity decreases with increasing $RotBFrac$, T, but increases with increasing $PetitjeanNumber$ and MW. The consistency between all three interpretation tools validates the various methods used to interpret the developed GPR models for establishing predictors relationship and importance to the response variable (density/viscosity).

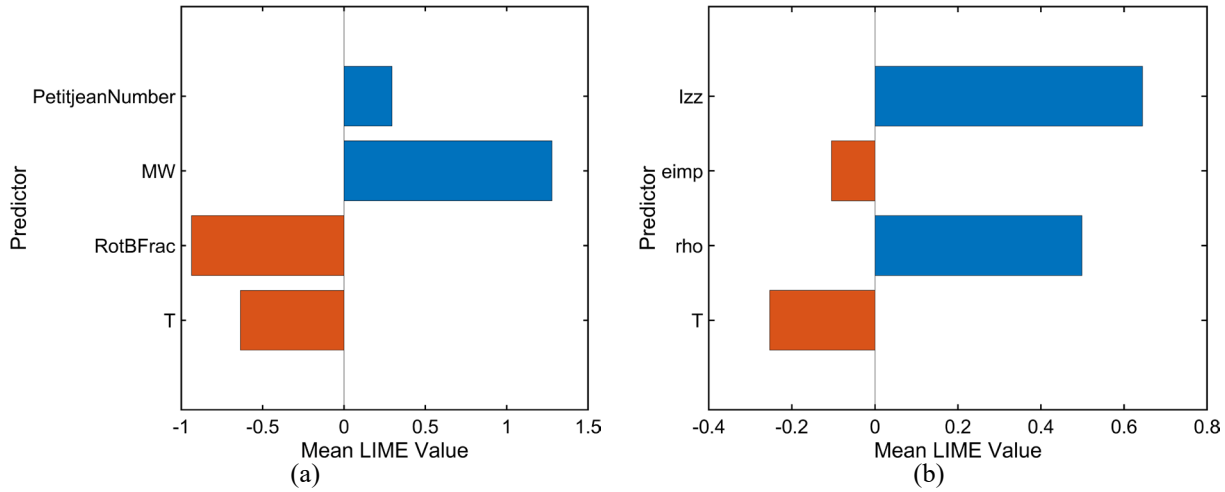


Figure 10: The average LIME value or coefficient for each predictor in best viscosity models (Model III) with the (a) static (b) dynamic descriptors. The orange and blue colors represent the negative and positive relationship between a predictor and the response variable. The size of a bar represents the overall importance of a predictor of a model.

Table 4 gives the R^2 values of viscosity models for the training, validation, and test datasets when a predictor was dropped from the best models with static and dynamic descriptors. The R^2 decreased by 0.6%, 12.3%, 29.1%, or 54.4% when $PetitjeanNumber$, $RotBFrac$, MW, or T term was dropped from the best viscosity model with static descriptors. This indicates that the least important predictor is the $PetitjeanNumber$, consistent with the results of the LIME analysis in Fig. 10a. Similarly, the last important dynamic descriptor is $eimp$, again as shown by the LIME analysis. Therefore, the most important predictors of viscosity are T, MW, $RotBFrac$, $eimp$, I_{zz}, and rho. This finding contradicts a previous study where the $PetitjeanNumber$ was found to be highly correlated with viscosity [30,99], but may be explained by the fact that our models were developed for only pure hydrocarbons.

Table 4: Performance of the best viscosity models when a predictor was removed from the models.

Static Descriptors (Equation 12)			**Dynamic Descriptors** (Equation 15)		
Term Dropped	Avg. R^2	Drop in Avg. R^2	Term Dropped	Avg. R^2	Drop in Avg. R^2
None	97.7%	0.0%	None	97.7%	0.0%
$PetitjeanNumber$	97.1%	0.6%	$eimp$	90.2%	7.5%
$RotBFrac$	85.4%	12.3%	T	77.7%	20.0%
MW	68.6%	29.1%	I_{zz}	60.3%	37.4%
T	43.4%	54.4%	rho	53.7%	44.0%

We developed viscosity models with combined both static and dynamic descriptors. The combined models performed only slightly better (R^2 of 0.982) than the models with only static (R^2 of 0.977) or only dynamic (R^2 of 0.977) descriptors. These models were not analyzed further but are provided in the Supporting Information.

The details of all models can be found in the Supporting Information. The MATLAB code for the best density and viscosity models with static and dynamic descriptors are also provided in the Supporting Information.

Conclusions

A GPR-based model was trained with Bayesian optimization to accurately predict the dynamic viscosity and density of complex hydrocarbons over a wide range of temperatures. We presented a top-down systematic approach to developing simple models using various robust ML algorithms. Our approach (1) removed redundant and strongly correlated predictor, (2) assessed the risk of overfitting and underfitting in models, (3) ensured that important predictors were included in the model, (4) assessed the quality of the model predictions, and (5) included model-agnostic interpretation. The best subset regression approach evaluated all combinations of significant predictors, which ensured that the minimum number of predictors was used and prevented over fitting. Notably, although the developed models involved very few (less than or equal to five) and relatively simple predictors but showed high accuracy in the prediction of experimental dynamic viscosity and density as a function of temperature for a variety of hydrocarbons. Multiple model-agnostic interpretations methods consistently showed that the T, $RotBtFrac$, $piPC3$, and rho of the molecules were the most important predictors for density, while T, MW, $RotBFrac$, $eimp$, I_{zz}, and rho were the most important predictors for viscosity.

We evaluated the feasibility and versatility of using dynamic and static molecular descriptors to predict density and viscosity of hydrocarbons. Since dynamic descriptor-based models involve the relative positions of the atoms in 3D space, it is commonly expected dynamic descriptors to provide more information and discrimination power for similar molecular structures and molecule conformations than static descriptors [100]. Further, sometimes the same SMILES string can represent different molecules; for example, the SMILES string CC(C)=O represents both Acetone and Dimethyl Ether, but their molecular structures and properties are different. However, static descriptors calculated based on their SMILES string will be indistinguishable between these two molecules. In addition, the relative positions of the atoms in a molecule change with the condition (e.g., temperature, pressure), which cannot be captured by static descriptors. Since dynamic descriptors are calculated based on the state of a molecule at a given condition, they contain more accurate information than static descriptors. Based on this, dynamic descriptors should have advantages over static descriptors for predicting viscosity and density, especially at different temperatures as done in this study. And, indeed, we found that models with dynamic descriptors performed as well as or better than models with static descriptors, even though the pool of dynamic descriptors (157) was significantly smaller than that of the static descriptors (1444). To further assess the benefit of using dynamic descriptors, models for viscosity, density, or other properties as a function of multiple operating conditions, such as temperature, pressure, and shear rate, should be developed.

Importantly, the ML-based predictive models developed in this study, which can be used to quickly predict the viscosity and density of hydrocarbons at given temperatures, could enable the design of novel hydrocarbon molecules with tunable properties. Moreover, although our

models in this study were only trained for density and viscosity of hydrocarbons, it provides a method that can be extended to other properties of a wider range of materials.

Associated Content

Data Availability Statement

The data underlying this study are openly available in the GitHub repository "`panwarp`" at https://github.com/panwarp/SupplementaryMaterials

Supporting Information

The Supporting Information is available free of charge on the GitHub at https://github.com/panwarp/SupplementaryMaterials

- Schematics of all molecules.
- Definition of the molecular descriptors.
- All experimental data with the static and dynamic descriptors of all molecules.
- LAMMPS data files of the molecules with forcefield parameters and initially built atomic positions. LAMMPS input files to run the MD simulations.
- MATLAB files of the best models to predict temperature-dependent density and viscosity.
- Density and viscosity models with combined static and dynamic descriptors.
- Density models with dynamic descriptors excluding simulation-calculated density.

Nomenclature/Abbreviations

ANN	artificial neural network
ANOVA	analysis of variance
API	American petroleum institute
EMD	equilibrium molecular dynamics
GPR	gaussian process regression
ICE	individual conditional expectation
LAMMPS	large-scale atomic/molecular massively parallel simulator
LASSO	least absolute shrinkage and selection operator
LIME	local interpretable model-agnostic explanation
MD	molecular dynamics
ML	machine learning
MLR	multiple linear regression
NPT	isothermal–isobaric ensemble
NVE	canonical ensemble
NVT	microcanonical ensemble
OPLS	optimized potentials for liquid simulations
PCA	principal component analysis
PDP	partial dependency plot
PLS	partial least-squares regression
PyL3dMD	python LAMMPS 3D molecular descriptors package
QSPR	quantitative structure-property relationships
RMSE	root mean squared error
SMILES	simplified molecular input line entry system
VIF	variance inflation factor

$AATS2e$	Broto-Moreau autocorrelation-lag 2/weighted by Sanderson electronegativities
c	acylindricity
$\mathbf{C}$	correlation matrix or matrix of the correlation coefficient
$eimp$	energy due to improper interaction
$evdwl$	energy due to van der Waals interactions
i	i^{th} data point
I_{zz}	diagonal component of moment of inertia tensor
ke	kinetic energy
$Kier1$	first kappa shape index
$\log \eta(T)$	logarithmic of temperature-dependent dynamic viscosity
MW	molecular weight of molecule
N	total number of data points
$PetitjeanNumber$	Petitjean number
$piPC3$	conventional bond order ID number of order 3
rho	MD simulation-calculated density from NPT ensemble
R_g	radius of gyration
R^2	coefficient of determination or R-squared
R_j^2	R^2 of predictor j on the remaining predictors
$RotBFrac$	fraction of rotatable bonds, excluding terminal bonds
$RotBtFrac$	fraction of rotatable bonds, including terminal bonds
T	temperature in °F
$\bar{y}$	mean experimental value of the response variable
y_i^{exp}	experimental value of the response variable
y_i^{pred}	model predicted value of the response variable
$\eta(T)$	temperature-dependent dynamic viscosity
$\rho(T)$	temperature-dependent density

Acknowledgements

The authors thank Dr. Jack Zakarian for identifying the API 42 Project resource as well as Julian Gonzalez and Shaun Flannigan for transferring the data into electronic form. We thank Dr. Simon Sung for guiding in some of aspect of machine learning.

References

[1] Rizvi, S. Q. A. *A Comprehensive Review of Lubricant Chemistry, Technology, Selection, and Design*, ASTM International, Conshohocken. **2009**.

[2] Erhan, S. Z.; Asadauskas, S. Lubricant Basestocks from Vegetable Oils. Ind. Crops Prod. 11, **2000**, pp. 277–282.

[3] Jameel, A. G. A.; Sarathy, S. M. Lube Products: MolecularCharacterization of Base Oils. Encycl. Anal. Chem. Appl. Theory Instrum. **2006**, pp. 1–14.

[4] Ewen, J. P.; Gattinoni C.; Thakkar, F. M.; Morgan, N.; Spikes, H. A.; Dini, D.; A Comparison of Classical Force-Fields for Molecular Dynamics Simulations of Lubricants. Materials. 9, **2016**, pp. 1–17.

[5] Stachowiak, G; Batchelor, A. *Engineering Tribology*, Burlington. **2006**.

[6] Höglund, E. Influence of Lubricant Properties on Elastohydrodynamic Lubrication. Wear. 232, **1999**, pp. 176–184.

[7] ASTM D2270. Standard practice for calculating viscosity index from kinematic viscosity at 40 and 100° C. Annual Book of Standards, **2003**.

[8] Moity, L.; Molinier, V.; Benazzouz, A.; Barone, R; Marion, P; Aubry, J. M. In Silico Design of Bio-Based Commodity Chemicals: Application to Itaconic Acid Based Solvents. Green Chem. 16, **2014**, pp. 146–160.

[9] Dearden, John; Worth, Andrew. *In Silico Prediction of Physicochemical Properties*. Luxembourg. **2007**.

[10] Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; De Hemptinne, J. C.; Ungerer, P.; Rousseau, B.; Adamo, C. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. Chem. Rev. 115, **2015**, pp. 13093–13164.

[11] Roy, K.; Kar, S.; Das, R. N. A Primer on QSAR/QSPR Modeling: Fundamental Concepts. SpringerBriefs Mol. Sci. 1, **2015**, pp. 1–121.

[12] Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A Practical Overview of Quantitative Structure-Activity Relationship. EXCLI J. 8, **2009**, pp. 74–88.

[13] Ajmani, S.; Jadhav, K.; Kulkarni, S. A.; Group-Based QSAR (G-QSAR): Mitigating Interpretation Challenges in QSAR. QSAR Comb. Sci. 28, **2009**, pp. 36–51.

[14] Oskarsdottir, G.; Bhan, A.; Venkatasubramanian, V.; Thomson, K. T.; Snively, C. M.; Katare, S.; Lauterbach, J. A.; Caruthers, J. M. Catalyst Design: Knowledge Extraction from High-Throughput Experimentation. J. Catal. 216, **2003**, pp. 98–109.

[15] Bernazzani, L.; Duce, C.; Micheli, A; Mollica, V; Sperduti, A; Starita, A; Tiné, M. R. Predicting Physical-Chemical Properties of Compounds from Molecular Structures by Recursive Neural Networks. J. Chem. Inf. Model. 46, **2006**, pp. 2030–2042.

[16] Selassie, C. D; Garg, R; Kapur, S; Kurup, A; Verma, R. P; Mekapati, S. B; Hansch, C; Comparative QSAR and the Radical Toxicity of Various Functional Groups. Chem. Rev. 102, **2002**, pp. 2585–2605.

[17] Grover, M; Singh, B; Bakshi, M; Singh, S. Quantitative Structure-Property Relationships in Pharmaceutical Research - Part 2. Pharm. Sci. Technol. Today, 3, **2000**, pp. 50–57.

[18] Grover, M; Singh, B; Bakshi, M; Singh, S. Quantitative Structure-Property Relationships in Pharmaceutical Research - Part 1. Pharm. Sci. Technol. Today, 3, **2000**, pp. 50–57.

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

[20] Dearden, J. C; Rotureau, P; Fayet, G. QSPR Prediction of Physico-Chemical Properties for REACH. SAR QSAR Environ. Res. 24, **2013**, pp. 279–318.

[21] Ivanciuc, O; Ivanciuc, T; Filip, P. A; Cabrol-Bass, D. Estimation of the Liquid Viscosity of

Organic Compounds with a Quantitative Structure-Property Model. J. Chem. Inf. Comput. Sci. 39, **1999**, pp. 515–524.

[22] Suzuki, T; Ohtaguchi, K; Koide, K. Computer-Assisted Approach to Develop a New Prediction Method of Liquid Viscosity of Organic Compounds. Comput. Chem. Eng. 20, **1996**, pp. 161–173.

[23] Suzuki, T; Ebert, R. U; Schüürmann, G. Development of Both Linear and Nonlinear Methods to Predict the Liquid Viscosity at 20 °C of Organic Compounds. J. Chem. Inf. Comput. Sci. 37, **1997**, pp. 1122–1128.

[24] Suzuki, T; Ebert, R. U; Schüürmann, G. Application of Neural Networks to Modeling and Estimating Temperature-Dependent Liquid Viscosity of Organic Compounds. J. Chem. Inf. Comput. Sci. 41, **2001**, pp. 776–790.

[25] Katritzky, A. R; Chen, K; Wang, Y; Karelson, M; Lucic, B; Trinajstic, N; Suzuki, T; Schüürmann, G. Prediction of Liquid Viscosity for Organic Compounds by a Quantitative Structure-Property Relationship. J. Phys. Org. Chem; 13, **2000**, pp. 80–86.

[26] Lučić, B; Bašic, I; Nadramija, D; Miličević, A; Trinajstić, N; Suzuki, T; Petrukhin, R; Karelson, M; Katritzky, A. R. Correlation of Liquid Viscosity with Molecular Structure for Organiccompounds Using Different Variable Selection Methods. Arkivoc. **2002**, pp. 45–59.

[27] Cocchi, M; De Benedetti, P. G; Seeber, R; Tassi, L; Ulrici, A. Development of Quantitative Structure-Property Relationships Using Calculated Descriptors for the Prediction of the Physicochemical Properties (ND, ρ, Bp, ∈, η) of a Series of Organic Solvents. J. Chem. Inf. Comput. Sci. 39, **1999**, pp. 1190–1203.

[28] Kauffman, G. W; Jurs, P. C. Prediction of Surface Tension, Viscosity, and Thermal Conductivity for Common Organic Solvents Using Quantitative Structure-Property Relationships. J. Chem. Inf. Comput. Sci. 41, **2001**, pp. 408–418.

[29] Koutsoukos, S; Philippi, F; Malaret, F; Welton, T. A Review on Machine Learning Algorithms for the Ionic Liquid Chemical Space. Chem. Sci. 12, **2021**, pp. 6820–6843.

[30] Loh, G. C; Lee, H. C; Tee, X. Y; Chow, P. S; Zheng, J. W. Viscosity Prediction of Lubricants by a General Feed-Forward Neural Network. J. Chem. Inf. Model. 60, **2020**, pp. 1224–1234.

[31] Goussard, V; Duprat, F; Ploix, J. L; Dreyfus, G; Nardello-Rataj, V; Aubry, J. M. A New Machine-Learning Tool for Fast Estimation of Liquid Viscosity. Application to Cosmetic Oils. J. Chem. Inf. Model. 60, **2020**, pp. 2012–2023.

[32] Nashawi, I. S; Elgibaly, A. A. Prediction of Liquid Viscosity of Pure Organic Compounds via Artificial Neural Networks. Pet. Sci. Technol. 17, **1999**, pp. 1107–1144.

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

[34] Paduszyński, K.; Domańska, U. Viscosity of Ionic Liquids: An Extensive Database and a

New Group Contribution Model Based on a Feed-Forward Artificial Neural Network. J Chem Inf Model. **2014**, 54, 1311–1324.

[35] Beckner, W.; Mao, C. M.; Pfaendtner, J. Statistical Models Are Able to Predict Ionic Liquid Viscosity across a Wide Range of Chemical Functionalities and Experimental Conditions. Mol Syst Des Eng. **2018**, 3, 253–263.

[36] Santak, P.; Conduit, G. Predicting Physical Properties of Alkanes with Neural Networks. Fluid Phase Equilib. **2019**, 501, 112259.

[37] Kajita, S.; Kinjo, T.; Nishi, T. Autonomous Molecular Design by Monte-Carlo Tree Search and Rapid Evaluations Using Molecular Dynamics Simulations. Commun Phys. **2020**, 3, 77.

[38] Avula, N. V. S.; Veesam, S. K.; Behera, S.; Balasubramanian, S. Building Robust Machine Learning Models for Small Chemical Science Data: The Case of Shear Viscosity of Fluids. Mach Learn Sci Technol. **2022**, 3, 045032.

[39] Bilodeau, C.; Kazakov, A.; Mukhopadhyay, S.; Emerson, J.; Kalantar, T.; Muzny, C.; Jensen, K. Machine Learning for Predicting the Viscosity of Binary Liquid Mixtures. Chemical Engineering Journal. **2023**, 142454.

[40] Heintz, J; Belaud, J. P; Pandya, N; Teles Dos Santos, M; Gerbaud, V. Computer Aided Product Design Tool for Sustainable Product Development. Comput. Chem. Eng. 71, **2014**, pp. 362–376.

[41] Dreyfus, G. *Neural Networks: Methodology and Applications*. **2005**.

[42] Goulon, A; Picot, T; Duprat, A; Dreyfus, G. Predicting Activities without Computing Descriptors: Graph Machines for QSAR. SAR QSAR Environ. Res. 18, **2007**, pp. 141–153.

[43] Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. J. Chem. Inf. Comput. Sci. 32, **1992**, pp. 331–337.

[44] Dehmer, M; Varmuza, K; Bonchev, D. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. **2012**.

[45] Todeschini, R; Consonni, V. *Handbook of Molecular Descriptors*. **2000**.

[46] Todeschini, R; Consonni, V. *Molecular Descriptors for Chemoinformatics*. **2010**.

[47] Puzyn, T; Gajewicz, A; Leszczynska, D; Leszczynski, J. Nanomaterials–the next great challenge for QSAR modelers. Recent Advances in QSAR Studies: Methods and Applications. **2010**, 383-409.

[48] Panwar, P; Michael, P; Devlin, M; Martini, A. Critical Shear Rate of Polymer-Enhanced Hydraulic Fluids. Lubricants. 8, **2020**, pp. 1–15.

[49] Kioupis, L. I; Maginn, E. J. Molecular Simulation of Poly-α-Olefin Synthetic Lubricants: Impact of Molecular Architecture on Performance Properties. J. Phys. Chem. B. 103, **1999**, pp. 10781–10790.

[50] Gordon, P. A. Characterizing Isoparaffin Transport Properties with Stokes-Einstein Relationships. Ind. Eng. Chem. Res. 42, **2003**, pp. 7025–7036.

[51] Panwar, P; Yang, Q; Martini, A. PyL3dMD: Python LAMMPS 3D Molecular Descriptors Package. J. Cheminformatics. **2023**. https://github.com/panwarp/PyL3dMD

[52] Martini, A; Ramasamy, U. S; Len, M. Review of Viscosity Modifier Lubricant Additives. Tribol. Lett. 66. **2018**.

[53] Len, M; Ramasamy, U. S; Lichter, S; Martini, A. Thickening Mechanisms of Polyisobutylene in Polyalphaolefin. Tribol. Lett. 65. **2018**.

[54] Ramasamy, U. S; Len, M; Martini, A. Correlating Molecular Structure to the Behavior of Linear Styrene–Butadiene Viscosity Modifiers. Tribol. Lett. 65. **2017**.

[55] Panwar, P; Schweissinger, E; Maier, S; Hilf, S; Sirak, S; Martini, A. Effect of Polymer Structure and Chemistry on Viscosity Index, Thickening Efficiency, and Traction Coefficient of Lubricants. J. Mol. Liq. 359. **2022**.

[56] Van Lommel, R.; Zhao, J.; De Borggraeve, W. M.; De Proft, F.; Alonso, M. Molecular Dynamics Based Descriptors for Predicting Supramolecular Gelation. Chem Sci. **2020**, 11, 4226–4238.

[57] American Petroleum Institute. *Properties of Hydrocarbons of High Molecular Weight Synthesized by Research Project 42 of the American Petroleum Institute*, New York. **1967**.

[58] Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. 28, **1988**, pp. 31–36.

[59] Mary, C; Philippon, D; Lafarge, L; Laurent, D; Rondelez, F; Bair, S; Vergne, P. New Insight into the Relationship between Molecular Effects and the Rheological Behavior of Polymer-Thickened Lubricants under High Pressure. Tribol. Lett. 52, **2013**, pp. 357–369.

[60] Larsson, R; Kassfeldt, E; Byheden, Å; Norrby, T. Base Fluid Parameters for Elastohydrodynamic Lubrication and Friction Calculations and Their Influence on Lubrication Capability. J. Synth. Lubr. 18, **2001**, pp. 183–198.

[61] Pensado, A. S; Comuñas, M. J. P; Fernández, J. The Pressure-Viscosity Coefficient of Several Ionic Liquids. Tribol. Lett. 31, **2008**, pp. 107–118.

[62] Ramasamy, U. S; Bair, S; Martini, A. Predicting Pressure-Viscosity Behavior from Ambient Viscosity and Compressibility: Challenges and Opportunities. Tribol. Lett. 57. **2015**.

[63] Yap C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. J. Comput. Chem. 32, **2010**, pp. 1466–1474.

[64] Martinez, L; Andrade, R; Birgin, E. G; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. J. Comput. Chem. 30, **2009**, pp. 2157–2164.

[65] Gartner, T. E; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. Macromolecules. 52, **2019**, pp. 755–786.

[66] Landau, L. D; Lifshitz, E. M; Reichl, L. E. Statistical Physics, Part 1. Phys. Today. 31. **1981**.

[67] Jorgensen, W. L; Maxwell, D. S; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. Am. Chem. Soc. 118, **1996**, pp. 11225–11236.

[68] Dodda, L. S; De Vaca, I. C; Tirado-Rives, J; Jorgensen, W. L. LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands. Nucleic Acids Res. 45, **2017**, pp. W331–W336.

[69] Jorgensen, W. L. BOSS - Biochemical and Organic Simulation System. Encycl. Comput. Chem. 3, **1998**, pp. 3281–3285.

[70] Dodda, L. S; Vilseck, J. Z; Tirado-Rives, J; Jorgensen, W. L. 1.14∗CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations. J. Phys. Chem. B. 121, **2017**, pp. 3864–3870.

[71] Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. J. Comput. Phys. 117, **1995**, pp. 1–19.

[72] Maginn, E. J; Messerly, R. A; Carlson, D. J; Roe, D. R; Elliott, J. R. Best Practices for Computing Transport Properties 1. Self-Diffusivity and Viscosity from Equilibrium Molecular Dynamics [Article v1.0]. Living J. Comput. Mol. Sci. 1, **2019**, pp. 1–20.

[73] NosÉ, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. Mol. Phys. 100, **2002**, pp. 191–198.

[74] Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. Phys. Rev. A, 31, **1985**, pp. 1695–1697.

[75] Elisseeff, A; Guyon, I. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3. **2003**, pp. 1157–1182.

[76] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B, 58, **1996**, pp. 267–288.

[77] Zou, H; Hastie, T. Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, **2005**, pp. 301–320.

[78] Miller, R. E. Analysis of Variance. Chem. Eng. (New York), 92, **1985**, pp. 173–178.

[79] Belsley, D. A; E. Kuh, R. E. W. *Regression Diagnostics*, John Wiley & Sons, Inc. New York, NY. **1980**.

[80] Gogtay, N. J; Thatte, U. M. Principles of Correlation Analysis. J. Assoc. Physicians India, 65, **2017**, pp. 78–81.

[81] Rasmussen, C. E; C. K. I. W. *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Massachusetts. **2006**.

[82] Hastie, T; Tibshirani, R; Friedman, J. *The Elements of Statistical Learning*, Springer New York, New York, NY. **2001**.

[83] Hawkins DM. The problem of overfitting. Journal of chemical information and computer

sciences. 2004 Jan 26;44(1):1-2.

[84] O'Brien, R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual. Quant. 41, **2007**, pp. 673–690.

[85] Snoek, J; Larochelle, H; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. Adv. Neural Inf. Process. Syst. 4, **2012**, pp. 2951–2959.

[86] Gelbart, M. A; Snoek, J; Adams, R. P. Bayesian Optimization with Unknown Constraints. Uncertain. Artif. Intell. - Proc. 30th Conf. UAI 2014. **2014**, pp. 250–259.

[87] Bull, A. D. Convergence Rates of Efficient Global Optimization Algorithms. J. Mach. Learn. Res. 12, **2011**. pp. 2879–2904.

[88] Shahriari, B; Swersky, K; Wang, Z; Adams, R. P; De Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. Proc. IEEE, 104, **2016**, pp. 148–175.

[89] Jerome H; Friedman. Greedy Function Approximation a Gradient Boosting Machine. Ann. Stat. **2001**, pp. 1189–1232.

[90] Goldstein, A; Kapelner, A; Bleich, J; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. J. Comput. Graph. Stat. 24, **2015**, pp. 44–65.

[91] Ribeiro, M. T; Singh, S; Guestrin, C. 'Why Should i Trust You?' Explaining the Predictions of Any Classifier. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, **2016**, pp. 1135–1144.

[92] Wright, R. Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots. Explor. SAS ® Enterp. Min. Spec. Collect. **2018**, pp. 1950–2018.

[93] Broto,, P; Devilers, J. Autocorrelation of Properties Distributed on Molecular Graphs. **1990**.

[94] Sanderson, R. T. Electronegativity and Bond Energy. J. Am. Chem. Soc. 105, **1983**, pp. 2259–2261.

[95] Wang, S. Q. On Chain Statistics and Entanglement of Flexible Linear Polymer Melts. Macromolecules, 40, **2007**, pp. 8684–8694.

[96] Hall, L. H; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. Rev. Comput. Chem. 2, **2007**, pp. 367–422.

[97] Theodorou, D. N; Suter, U. W. Shape of Unperturbed Linear Polymers: Polypropylene. Macromolecules. 18, **1985**, pp. 1206–1214.

[98] Colby, R. H; Fetters, L. J; Graessley, W. W. Melt Viscosity-Molecular Weight Relationship for Linear Polymers. Macromolecules, 20, **1987**, pp. 2226–2237.

[99] Morgan, M. J; Mukwembi, S; Swart, H. C. A Lower Bound on the Eccentric Connectivity Index of a Graph. Discret. Appl. Math. 160, **2012**, pp. 248–258.

[100] Yasuda, I.; Kobayashi, Y.; Endo, K.; Hayakawa, Y.; Fujiwara, K.; Yajima, K.; Arai, N.; Yasuoka, K. Combining Molecular Dynamics and Machine Learning to Analyze Shear

Thinning for Alkane and Globular Lubricants in the Low Shear Regime. ACS Appl Mater Interfaces. **2023**, 15, 8567–8578.