THE JOURNAL OF PHYSICAL CHEMISTRY A

Article

Cite This: J. Phys. Chem. A 2019, 123, 8305−8313

pubs.acs.org/JPCA

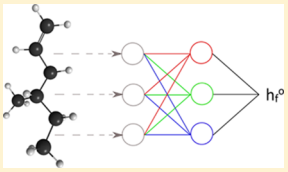
Machine Learning To Predict Standard Enthalpy of Formation of Hydrocarbons

Kiran K. Yalamanchi,*[,†] Vincent C. O. van Oudenhoven,[†] Francesco Tutino,[†] M. Monge-Palacios,[†] Abdulelah Alshehri,[†] Xin Gao,[‡] and S. Mani Sarathy*[,†]

[†]Clean Combustion Research Center, Physical Sciences and Engineering Division and [‡]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

Supporting Information

ABSTRACT: Thermodynamic properites of molecules are used widely in the study of reactive processes. Such properties are typically measured via experiments or calculated by a variety of computational chemistry methods. In this work, machine learning (ML) models for estimation of standard enthalpy of formation at 298.15 K are developed for three classes of acyclic and closed-shell hydrocarbons, viz. alkanes, alkenes, and alkynes. Initially, an extensive literature survey is performed to collect standard enthalpy data for training ML models. A commercial software (Dragon) is used to obtain a wide set of molecular descriptors by providing SMILES strings. The molecular descriptors are used as input features for the ML models. Support vector regression (SVR) and artificial neural networks are used with a two-level K-fold cross-validation (K-fold CV) workflow. The first level is for estimation of accuracy of both the ML models, and the second level is for generation of the final models. The SVR model is selected as the best model based on error estimates over 10-fold CV. The final SVR model is compared against conventional Benson's group additivity for a set of octene isomers from the database, illustrating the advantages of the proposed ML modeling approach.



INTRODUCTION

Accurate thermodynamic data of species are essential for chemical kinetic models to predict the outcome of complex chemical processes. With the need for accurate chemical kinetic mechanisms that include a larger number of components, the species included in the mechanisms are increasing continuously.[1] Quantum chemistry calculations and experiments can be used to determine all thermodynamic properties of interest; however, these are computationally intensive and require specialized skillsets. In this work, we propose a machine learning framework for quantitative structure−property relationships (QSPRs) to predict an important thermodynamic property, standard enthalpy of formation, herein referred to as "enthalpy". While the methodology developed in this work is general and could be applied, in principle, to any class of chemical compounds, we restrict our investigation to three classes of hydrocarbons, viz. alkanes, alkenes, and alkynes, which are the most commonly included species in kinetic models, especially in combustion studies.

Machine learning has been used in recent years for predicting molecular properties. Rupp et al.[2] used kernel ridge regression to predict atomization energies of organic molecules based on the norm between Coulomb matrices. By training a single time on a finite subset of known solutions, the need to explicitly solve the Schrodinger equation is bypassed. Hansen et al.[3] developed on this work to predict atomization energies. They selected a number of established machine learning techniques and investigated the influence of the molecular representation on the method's performance. In particular, they obtained molecular representations resulting from three variations on the Coulomb matrix. In addition, Hu et al.,[4] Wu and Xu,[5] and Sun et al.[6] used hybrid techniques of combining ML with quantum chemistry methods for predicting heats of formation of organic molecules. Their study is limited in scope as it is built upon a small set of compounds. This is because of a significantly large amount of computational work required to generate a large number of quantum mechanics-based training samples. In any case, the computational complexity of these hybrid techniques limits the ability to generate thermodynamic data at a large scale.

Direct predictions of other physical properties for complex mixtures are also reported in the literature. Abdul Jameel et al.[7] used artificial neural networks (ANNs) to predict research and motor octane numbers using functional groups supplemented with molecular weight and branching index as inputs. Similarly, de Oliveira et al.[8] used ANNs to predict flash point, cetane index, and sulfur content (S1800) of diesel blends using distillation curves as inputs. Saldana et al.[9] developed models for the prediction of two fuel properties, flash points and cetane numbers, using 150 molecular descriptors based on minimized geometries generated from Materials Studio software[10] and functional groups. This work was extended by Saldana et al.[11] to predict the melting point and heat of combustion using the same input representation. The enthalpy of combustion database for this work includes a wide variety of

Received: May 20, 2019
Revised: August 29, 2019
Published: August 29, 2019

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

hydrocarbons and was taken from the DIPPR database[12] and *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*.[13] Sennott et al.[14] used ANNs to predict the cetane number of biofuel candidates based on the molecular structure. They selected a set of 32 molecular descriptors to be inputs of their ANN by an iterative reduction of an initial extensive group generated by Dragon,[15] as described by Todeschini and Consonni.[16] Coley et al.[17] employed convolutional neural networks (CNNs) for the prediction of aqueous solubility, octanol solubility, melting point, and toxicity. Coley et al.'s work[17] uses a graph-based CNNs for molecular embedding in contrast to the previous studies. Machine learning has been used in all these cases because of the complexity related to using first principle-based calculations.

Group contribution methods are an alternative approach for estimating thermodynamic properties. Several group contribution methods exist, and these can be used to estimate thermodynamic and other properties from molecular structures. These include the simple Joback method[18] that sums up group contributions and the UNIFAC method[19] that includes group interaction parameters. The most widely used method for estimating enthalpy, heat capacity, and entropy is Benson's group additivity, hereafter referred as GA. GA was proposed in 1958,[20] and several other works[21−24] have contributed to facilitate widespread use of this method by adding new group values and revising existing ones for improving accuracy. A computer program THERM was made available by Ritter[25] to facilitate the calculation of thermodynamic properties of interest to combustion and reaction modeling using GA. A cloud-based platform for calculating thermodynamic properties is also availabe on the CloudFlame website (https://cloudflame.kaust.edu.sa/).[26] Benson's GA uses second-order group values and correction factors such as steric effects and intermolecular interactions to account for neighboring effects. GA requires users to identify internal and external symmetry of molecules for estimation of entropy and to identify gauche interactions for enthalpy. The reliance of GA on the user's chemical intuition makes it prone to errors. In addition, a previous work[23] has shown that the effects of secondary interactions are difficult to capture in group additivity schemes because the nominal corrections to enthalpy vary between molecules. In summary, while GA has obtained widespread acceptance, its lack of accuracy and need for chemical intuition motivates the use for an artificial intelligence methdology.

Support vector regression (SVR) and ANNs are utilized in this study as these ML algorithms can capture the complex relationship between input and output vectors that are required to map largely convoluted molecular representations into thermodynamics properties. The final trained models can be used as an effective tool to predict the enthalpy of any molecule belonging to the aforementioned classes with no theoretical/experimental data in the literature. For any ML application, the first part is to curate an exhaustive database from the literature that can be used to train the algorithm. In this study, data are collected from various databases,[27−32] which include experimental and theoretical values, as described in the next section. Various molecular descriptors calculated from Dragon[15] primarily consisting of functions that are able to extract the topology of the chemical structures are analyzed for the feature selection to provide inputs for the ML models. Molecular descriptors and the feature selection are explained in the next sections, together with brief description on the used ML algorithms and the training procedure. The performances of the final models are presented together with their respective sensitivity analyses. Finally, a comparison with GA is made.

METHODOLOGIES

Data Curation. The training dataset used in this work is made up of 310 noncyclic hydrocarbon compounds. In particular, this set of hydrocarbons consists of alkane, alkene, and alkyne species with one or more unsaturations and up to 18 carbon atoms, linear and branched. The dataset is composed of constitutional isomers as well as *Z* and *E* stereoisomers, making the developed ML model capable of discerning between different kinds of isomers. An extensive search for enthalpy data was conducted, and data from multiples sources were gathered and merged. The complete dataset containing all the features and respective enthalpy is available in the Supporting Information. Table 1 shows the distributions of hydrocarbons in the dataset. Several compounds had multiple reported values. For the sake of consistency in model development, a priority was given to experimental values over theoretical ones; among the experimental measurements, priority was given to the most recent ones.

Table 1. Distribution of Hydrocarbons in the Final Dataset

type of hydrocarbon	no. of compounds	max. chain length
alkanes	66	18
alkenes	195	16
alkynes	49	9

All of the experimental values are collected from databases by Ghahremanpour et al.[27] and Pedley et al.[28] The former has experimental values collected from refs 12 13, and 30 along with theoretical values calculated at the CBS-QB3,[33,34] G2,[35] G3,[36] G4,[37] W1BD,[38] and W1U[38] composite ab initio levels of theory. The latter consists of enthalpy values derived from experiments along with values predicted based on groups. In the final dataset of 310 species, 286 species are from Ghahremanpour et al.,[27] and 18 are from Pedley et al.[28]

Data for species hexa-1,5-diyne, penta-1,3-diyne, and 1,2,3-pentatriene were taken from G4 theoretical calculations of Ghahremanpour et al.[27] Among the theoretical values reported by Ghahremanpour et al.,[27] the G4 composite technique was prioritized over the others. This is done because the root-mean-squared deviation (RMSD) from the reported experimental values shows values of 3.06, 3.23, 4.18, and 4.76 kcal/mol for the G4, G3, G2, and CBS-QB3 techniques, respectively. In general, the G4 composite technique represents a good method for estimating the enthalpies. This composite method was developed as a further improvement to G3, and the differences between both techniques can be found elsewhere.[36,37] Here, we highlight three of them. First, the G4 and G3 methods use the B3LYP/6-31G(2df,p) and MP2(full)/6-31G(d) levels for the geometry optimization, respectively; the former uses the same ab initio level for the frequency calculation, while the latter uses the lower level HF/6-31G(d). In addition, the single-point energy calculations performed by the G4 method lead to more accurate and robust calculations by adding extra d-polarization functions to the basis set used by the G3 technique and also by replacing the QCISD(T)/6-31G(d) calculation by the CCSD(T)/6-31G(d) level of theory.

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

For 1,2,3-butatriene species, enthalpy was taken from Goldsmith et al. dataset,[29] who used the RQCISD(T)/cc-pv∞QZ//B3LYP/6-311++G(d,p) single-point ab initio level. 2,3,5-trimethyl-hexane and 1-dodecene enthalpy values were taken from CRC database of thermodynamic properties.[30]

Data Processing. The use of molecular information for training ML algorithms has been approached in various ways in the literature. Coulomb matrices, molecular descriptors, and convolutions have all been used to provide molecular information, with largely different levels of precision.[2,3,8,9,11,14,17] Each approach represents chemical structures to correlate one or more specific microscopic properties with a macroscopic physical quantity that one wants to predict. Coulomb matrices and convolution methods do not require information other than a molecular fingerprint, which requires additional data to accurately capture all interrelations between the plain molecular structure and the property of interest. Molecular descriptors have been used in various QSPR models for machine learning. A molecular descriptor is a numerical value that is defined as a function of the molecular structure. This method requires that every molecule in the domain of a given function be mapped to a single numerical value. Saldana et al.[9,11] used molecular descriptors based on minimized geometries generated from the Materials Studio software. However, geometry optimization costs increase with increasing molecular complexity. Sennott et al.[14] used descriptors generated from Dragon,[15] which are directly calculated from the simplified molecular-input line-entry system (SMILES) of a molecule without optimization of geometry. Dragon is a commercial software that can be used to calculate a wide set of descriptors by providing SMILES of molecules. Todeschini and Consonni[16] have reviewed and reported the formulae used to calculate the comprehensive molecular descriptors available in Dragon. Dragon has as many as 5255 molecular descriptors in 30 categories. For this study, 3177 descriptors from 24 applicable categories are calculated in the dataset considered. These categories include constitutional and topological indices, among others, shown in Table 2 along with the number of descriptors in each of those categories. Since methane is a single-carbon compound, it does not have most of these descriptors, so we removed methane and reduced the number of molecules in the dataset to 309.

The primary objective of this work is to obtain a machine learning algorithm with a level of accuracy comparable with QM calculations. This requires that the number of input features be maximized. However, many of the above-mentioned 3177 descriptors are either incomplete or redundant. When a descriptor is not known for all samples, the descriptor was removed to maintain the size of the sample dataset. Furthermore, redundant descriptors increase training noise and reduce the final model's generalization. Therefore, the number of descriptors was filtered to produce a more robust ML model. Duplicate descriptors with identical values were removed, reducing the number of descriptors to 2536. For example, functional group count descriptors describing groups consisting more than C and H atoms are zero for the dataset in the study, so they could be neglected. Next, descriptors that are not applicable for some molecules in the dataset were removed also, reducing the number of descriptors to 2439. Finally, the number of descriptors was further reduced by removing highly correlated descriptors having an absolute value of Pearson's correlation coefficient[39] greater than 0.8. For all the combinations of descriptors, the correlation coefficient was calculated, and a final set of descriptors with the correlation coefficient value between any pair less than 0.8 was selected, thereby forming a final set with 261 descriptors. The reduction of features in the last step is a trade-off between having the features necessary to distinguish two different molecules (e.g., isomers) and having a minimum set of distinctive features.

Table 2. Descriptor Categories Used in This Study

category of descriptors	number of descriptors
constitutional indices	47
topological indices	75
walk and path counts	46
connectivity indices	37
information indices	50
2D matrix-based descriptors	607
2D autocorrelations	213
burden eigen values	96
P_VSA-like descriptors	55
ETA indices	23
edge adjaceny indices	324
geometrical descriptors	38
3D matrix-based descriptors	99
3D autocorrelations	80
RDF descriptors	210
3D-MoRSE descriptors	224
GETAWAY descriptors	273
Randic molecular profiles	41
functional group counts	154
atom-centered fragments	115
atom-type E-state indices	172
CATS 2D	150
molecular properties	20
drug-like indices	28
total	3177

A common practice in ML applications is to normalize input features (i.e., descriptors) to facilitate model training. Min−max normalization was adopted in this work as there is a high variance in the magnitude of various feature vectors. This normalization procedure transforms all of the features to limits within the 0−1 interval, thereby avoiding steep gradients in one direction during optimization.

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{min}} - x_{\text{max}}}$$

Machine Learning Models. In this study, two popular machine learning models, viz. support vector regression (SVR) and artificial neural networks (ANN), were used. The details of each of these are explained in the following sections. The codes for development of these ML models were written in Python using Keras[40] libraries with TensorFlow[41] as backend. The codes are available as an open source on the GitHub repository (http://github.com/vcovo/enthalpy-prediction). The workflow followed for error estimation and final model generation for these two models is very similar. It consists of a two-level K-fold cross-validation (K-fold CV), (i) one of which is used to determine errors over the entire dataset and thus the performance of the method, and (ii) the second is used to validate the search over a predefined hyperparameter grid.

In the first part of error determination, the workflow consists of two loops, as shown in Figure 1. The outer loop consists of the division of the complete dataset into k folds. The folds are

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

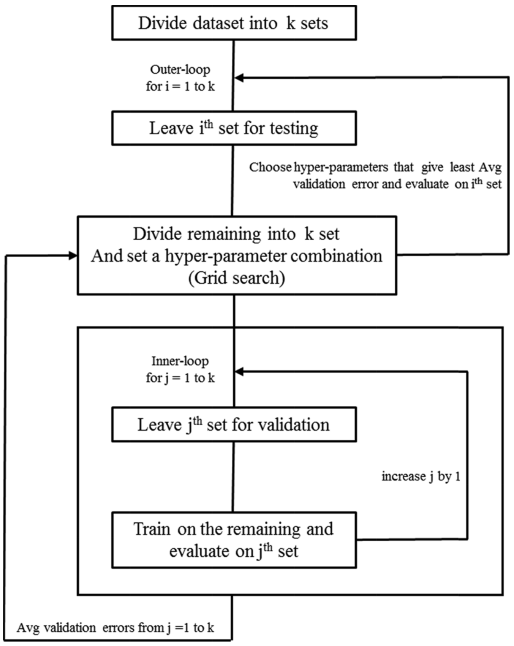


Figure 1. Workflow for error estimation of an ML model with K-fold CV.

iterated over until each fold has been used as the test set for a single time. The inner loop is used to determine the best hyperparameters for each particular split of the outer loop using a method formally known as "grid search". A search space is defined, and for each combination of parameters, the remaining folds are divided into k folds, and like before, they are iterated over until each set has been used as a validation set. The average over the folds is used as a metric for the performance of that particular hyperparameter combination. The hyperparameter combination that leads to the minimum average validation error is used to define the model for that particular outer split. It is retrained on the k-1 folds used as the training set and tested on the test set. The performance on this test set is the representation of the performance of the model.

In summary, each of the k sets is set aside as a test set, and a model is trained with the remaining data with no information from the test set passing onto the training set. Finally, this model is used to estimate accuracy on the left-out test set. When repeating this procedure for k sets, error metrics are obtained for k sets, which provides insights into ML model performance. Although this method has higher computational cost compared to normal division of dataset into training, validation, and test set, it gives a more unbiased picture of ML model accuracy since each data point is taken into the test set once.

The final model is generated by following the inner loop in the aforementioned workflow without leaving anything for the test set. This makes sure that all the collected data are used for the predictions using the final model. The common practice in the ML community is to use the 5-fold, 10-fold, or leave-one-out cross-validation (LOOCV) method (extreme case of K-fold CV where K = number of data points). In the present study, 10-fold CV is used since the dataset is small enough to afford the computation time but too large to use the LOOCV method.

RESULTS AND DISCUSSION

Support Vector Regression. Support vector machines (SVMs)[42] are a type of supervised learning algorithm widely used for classification problems in the field of machine learning. Using the same idea of SVM, support vector regression was introduced by Drucker et al.[43] SVMs perform the classification task by representing *n*-features in an *n*-dimensional space and finding a hyperplane that is able to linearly separate the initial data space in two subsets with the largest margin between the two classes. If the linear separation cannot be performed solely with SVM, then it is complemented by kernel functions. The use of kernel functions projects the dataset into a higher dimensional space and reduces nonlinear boundary classifications to a linear one. SVRs function in the same way and fit a nonlinear function by means of high-dimensional kernel-induced feature space.

A radial basis function (RBF) kernel was adopted for the SVR here, considering that the number of features and data is comparable.[44] The hyperparameters (i.e., penalty coefficient and epsilon) were tuned for better performance of the SVR. These parameters were optimized using the grid search method, as explained in the previous section, to achieve better accuracy against the validation sets. Epsilon defines the permissible error, allowing the objective function to not penalize predictions that fall within these bounds. The dataset (provided in the Supporting Information) has an inherent uncertainty that is estimated to be around 0.5 kcal. Therefore, epsilon was parameterized between 0.1 and 0.5. An epsilon value of less than 0.1 would decrease the generality of the model, while increasing it beyond 0.5 would allow for too much freedom. The penalty coefficient C is a regularization parameter that controls overfitting. More concretely, it influences the size of the margin of the hyperplane. Large values of C will force the SVR to choose a smaller margin and thereby achieve higher training accuracy. Conversely, a smaller value of C will lead to a larger margin at the cost of having a lower training accuracy. Tuning this parameter finds a good balance between the overall training accuracy and the margin size; the latter is typically proportional to the model's generalization capabilties. The best set of hyperparameters is chosen based on the results of the grid search. These are used to train the model on the dataset fed to the inner loop and to predict enthalpy values in the test set. The mean absolute error (MAE) and R^2 scores for the SVR model using the 10-fold CV method for error estimation are given in Table 3 for each of 10-folds divided randomly.

Artificial Neural Networks. Artificial neural networks (ANNs) have been proposed as universal nonlinear approximators. A detailed explanation of ANNs is outside of the scope of this paper, so we focus on the architecture of the present model and how it was determined. The largest decision to be made when desigining an ANN is the number of hidden layers. These are the layers between the input layer, which simply consists of the features, and the output layer, which, for our case, is a single number: enthalpy. These layers are made up of a variable number of nodes. The nodes are essentially new features that are created from the original inputs through a combination function, summation in our case, and then passed through an activation function. The activation function is what allows for nonlinearity in ANNs; the rectified linear unit (ReLu) was chosen for this study. It was empirically determined that two hidden layers allow the model to capture

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

Table 3. Mean Absolute Errors and R^2 Scores for the 10-Fold CV Method Using SVR and ANN

	SVR		ANN	
fold number	MAE (kcal/mol)	R^2 score	MAE (kcal/mol)	R^2 score
1	1.024	0.997	1.635	0.990
2	0.989	0.997	1.668	0.994
3	0.821	0.999	1.256	0.995
4	1.879	0.996	2.170	0.994
5	1.555	0.996	1.602	0.996
6	1.457	0.991	1.621	0.992
7	1.555	0.996	1.899	0.994
8	1.627	0.996	2.669	0.979
9	0.858	0.999	1.575	0.996
10	1.865	0.988	2.188	0.993
average	1.363	0.995	1.828	0.992

the patterns of our dataset and was thus the basic architecture used for the rest of the optimization.

The mean absolute error was used as the loss function, and dropout was used for regularization. Considering its suitability for ReLu activation function, He-uniform[45] initialization was used. He uniform samples initial weights of ANN from a uniform distribution within a limit set by the number of weights between layers in ANN. The optimizer of choice is Adam.[46] To further optimize the ANN, a grid search was performed on the number of nodes in each hidden layer and the respective dropout coefficients, in addition to the batch size and the number of epochs. An epoch is a complete training cycle: the training examples are fed through the network, and the respective error is "back-propagated". The batch size is the number of examples that pass through the ANN, affecting which patterns are picked up by the ANN. Last, the dropout coefficients determine how many nodes per layer are "dropped out" per epoch, which is proportional to the degree of regularization. Similar to SVR, the hyperparameter combinations were validated using 10-fold CV. Table 4 shows the hyperparameter search space for the ANNs.

Table 4. Hyperparameter Search Space for ANN

parameter	possible values
number of units	$\{10*2\hat{}x \in N \mid 1 \leq x \leq 4\}$
number of epochs	$\{1000x \in N \mid 1 \leq x \leq 5\}$
batch size	$\{2\hat{}x \in N \mid 2 \leq x \leq 9\}$
dropout rate	$\{x/10 \in N \mid 1 \leq x \leq 4\}$

The MAE and R^2 scores for each ANN model are given in Table 3. The random state was fixed for both SVR and ANN, so the test set for each fold is the same to enable a direct comparison of both models. From Table 3, it is clear that the SVR model performs better than the ANN, but the difference is small with average MAE varying by 0.465 kcal. The better performance of the SVR model is due to more extensive search and tuning of hyperparameters. For the ANN model, a higher granularity of the layer size and dropout coefficients were not considered due to computational cost. To put that in perspective, training the final ANN model (1-fold, one set of hyperparameters) takes around 52 s compared to 1.4 s for the SVR model. This was timed using a 2.80 GHz Intel Xeon (E) processor. The actual times will vary on different systems. However, the ratio of the time taken for the training of these two kinds of models should stay the same. However, it can be noted that both ML models converge to a similar accuracy, although both are theoretically different models. For these reasons, the final model is chosen to be the SVR model, and this will be used for comparison with group additivity and sensitivity analysis in the next two sections.

Comparison with Group Additivity. In comparison with the ML model presented in this study, Benson's group additivity (GA) uses only group counts as descriptors along with correction factors. While group counts are the main constituent in a GA scheme, other complex descriptors and their effect on enthalpy cannot be captured. The ML-based model used in the present study uses a wide set of descriptors that may capture complex interactions better than that in GA. Furthermore, GA uses a linear regression model to determine group values for specific groups present in a series of molecules. Linear regression models train well only for datasets that are homogeneous for the types of descriptors considered. When this is not the case, GA tends to lose out some information from the training dataset. For example, consider the pairs ethylene/propene and allene/1,2-butadiene. The difference in groups between these two pairs is the same, that is, one allylic carbon atom. This means that the difference in enthalpy between the ethylene/propene pair should be the same as that between the allene/1,2-butadiene pair, which is 7.87 kcal/mol according to GA values taken from RMG.[47] However, these differences are 7.74 and 6.76, respectively, in the experimental values taken from Ghahremanpour et al.[27] and the NIST[31] database. The difference between 7.74 and 6.76 is 0.98 kcal/mol, which is higher than the the sum of experimental uncertainities of these four molecules, which is 0.26 kcal/mol (individual uncertainities are provided in the Supporting Information). This illustrates the problem GA suffers from in these situations. As new data from experiments and theoretical calculations are generated, it would be difficult to fit them within the GA scheme. Despite these drawbacks, GA is a good approximation, considering its simplicity and the extent to which there are linear dependencies of enthalpy on group counts (7.74−6.76 in the example discussed). The present study takes into account both drawbacks of GA by fitting a nonlinear function (SVR/ANN) with a wide variety of molecular descriptors.

To exemplify the advantage of the ML model over GA, both models should be trained on the same dataset and compared. However, it would be a tedious task to train group values according to the database considered in this study. For this reason, GA calculations taken from RMG[47] for a set of octene isomers are compared with SVR results. RMG was selected for comparison because it only requires species notation input (similar to our ML models) to estimate thermodynamic properties, thereby removing any human errors associated with assigning incorrect groups, symmetry numbers, gauche interactions, etc. Note that the SVR model was not trained on this dataset and has not seen it before. Figure 2 presents the comparison of GA, SVR, and experimental values for 3-ethyl-4-methylpent-1-ene, 3,3,4-trimethylpent-1-ene, 3,4,4-trimethylpent-1-ene, 2-ethyl-3,3-dimethylbut-1-ene, (2Z)-3-ethyl-4-methylpent-2-ene and (2E)-3-ethyl-4-methylpent-2-ene, and (2Z)-3,4,4-trimethylpent-2-ene and (2E)-3,4,4-trimethylpent-2-ene. Good agreement is shown between the SVR model results and experimental data taken from Yaws' handbook,[13] with an average error of 0.96 kcal/mol. The experimental data from Yaws' handbook[13] have a mean uncertainity value of 0.61 kcal/mol. Considerable differences are observed between GA

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

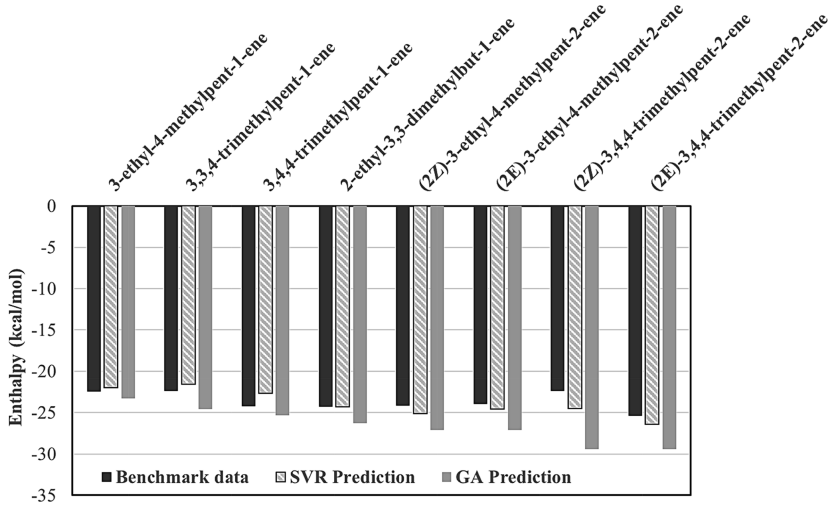


Figure 2. Comparison of SVR and GA models for octene isomers with benchmark data taken from Yaws' handbook.[13]

and experimental values, with an average error of 2.89 kcal/mol and a maximum of 4.01 kcal/mol for (2E)-3,4,4-trimethylpent-2-ene. The difference between the SVR model/experimental value and GA could possibly be due to several reasons: different training datasets, possible difficulty in fitting group values, and incorrect group contribution parameters. If the former is the case, then updating group values could make GA more accurate. However, as the size of the dataset increases, even this would become difficult, as explained in the previous section with examples of *n*-decane, *n*-undecane, and *n*-dodecane. To gauge the potential corrections that could be made by using the SVR model over GA, a comparison of predicted values of both these models is made for all the nonane isomers and is shown in Figure 3. Note that the dataset used for training in this study consists of only 8 of these 35 isomers, as indicated in Figure 3. A maximum of 4.74 kcal/mol difference is observed for 2-methyl-3-ethylhexane with an average difference of 1.13 kcal/mol. Both the comparisons between SVR and GA models made in this section illustrate the potential improvement that can be made in predicting enthalpy.

Sensitivity Analysis. Sensitivity analysis can determine the relative contribution of input factors on the calculated output parameters. A review and comparison of sensitivity analysis methods have been discussed by Gevrey et al.[48] Although these were discussed for ANN, most of them are applicable for any general regression model. In this work, the "perturb" method is used. This method assesses the effect of small changes in each input feature on the SVR output. An input feature is increased by 10% of its standard deviation over its mean, while others are assigned their mean values. The output from the modified SVR model is generated and compared with the output with all the features assigned with their mean values. This difference is then normalized by the amount a particular feature was changed, as shown in eq 1. The sensitivity values are compared, and the 10 most sensitive features are shown in Table 5. One thing to note here is that the descriptors are not completely independent. This means that a descriptor alone cannot be changed by changing the molecular structure.

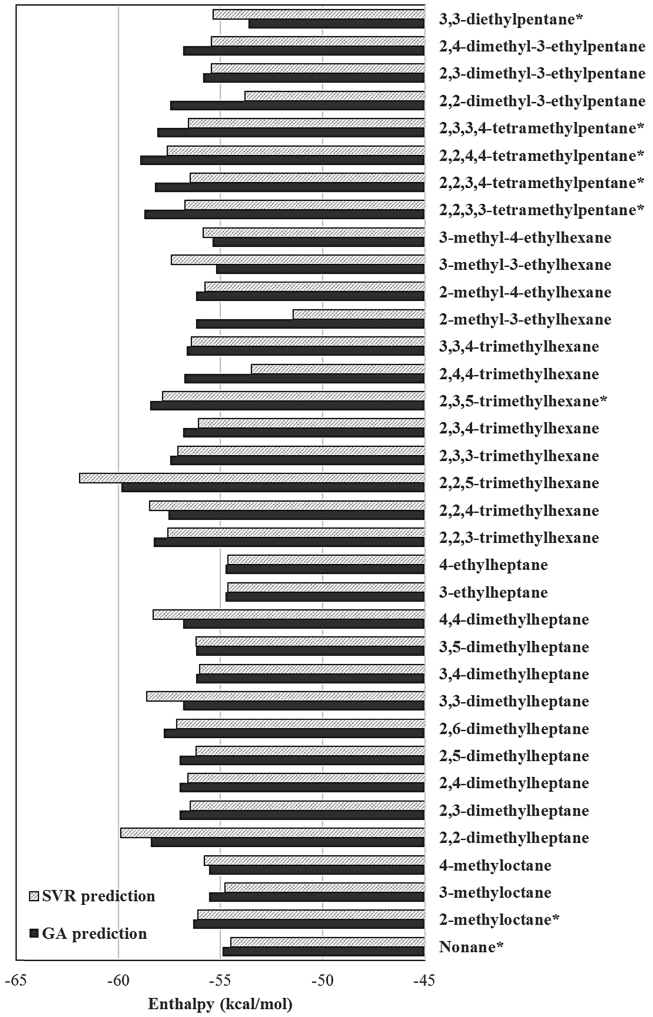


Figure 3. Comparison of SVR and GA models for nonane isomers (data with asterisks are from the training dataset).

$$\text{sensitivity}_n = \frac{\text{SVR (feature}_n \text{ changed by 10\% SD)} - \text{SVR (all features at their mean)}}{\text{10\% SD of feature}_n} \quad (1)$$

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

Table 5. Ten Most Sensitive Features for SVR Using the Perturb Method

feature	description
P_VSA_MR_7	P_VSA-like on molar refractivity, bin 7
nBM	number of multiple bonds
nTB	number of triple bonds
P_VSA_MR_6	P_VSA-like on molar refractivity, bin 6
ZM1V	first Zagreb index by valence vertex degrees
n=C=	number of allene groups
SpMax2_Bh(s)	largest eigenvalue no. 2 of Burden matrix weighted by I-state
AMW	average molecular weight
P_VSA_s_4	P_VSA-like on I-state, bin 4
GATS2m	Geary autocorrelation of lag 2 weighted by mass

The most sensitive descriptor is a P_VSA-like descriptor; in addition, two more P_VSA-like descriptors are among the top 10, indicating their importance in our calculations. P_VSA descriptors are calculated as the sum of atomic contributions of atoms falling in a specific bin/range for a property to the van der Walls surface area (VSA).[49] Labute et al.[49] defined VSA descriptors based on three properties including molar refractivity, therefore, polarizability and pointed out that the descriptors are weakly correlated with one another. Furthermore, the work uses exclusively VSA descriptors for predicting various properties, suggesting their capability to embed the molecular information in descriptors. These descriptors were further extended to other properties in Dragon,[15] including the I-state (intrinsic state). The I-state of an atom is the ratio of π and lonepair electrons to the count of the σ bonds for the considered atom. Therefore, the P_VSA descriptors highlighted in our work as the most sensitive ones clearly show the link between properties such as molar polarizability/electronic structure and enthalpy of formation.

nBM and nTB are constitutional descriptors defining number of multiple and triple bonds, respectively. The high sensitivity of these descriptors stems from the fact that adding multiple bonds changes enthalpy significantly; for example, we observe enthalpy values of −20.05, 12.52, and 54.37 kcal/mol for ethane, ethylene, and acetylene, respectively. n=C= is a functional group count for allene groups, which also affects the enthalpy of a molecule appreciably; for example, the difference in enthalpy between 1,2-butadiene and 1,3-butadiene is 12.55 kcal/mol. Other senstive features include the following: SpMax2_Bh(s), the second largest eigenvalue of Burden matrix constructed from I-states of atoms; AWM, which is the average molecular weight; GATS2m, defined as Geary correlation[50] and calculated from atomic masses with a topological distance of 2; and ZM1V, that is, the sum of squares of the degrees of the vertices in a hydrogen-depleted molecular graph. More details of these descriptors can be found elsewhere.[16] From the physical standpoint, it may be difficult to understand how exactly these descriptors affect the enthalpy of formation; however, these other sensitive descriptors found in our work are related to the molecular bond and electronic structure and therefore can be expected to exert a pronounced effect on the enthalpy of formation, explaining its sensitivitity. Group counts, such as those used in GA schemes, are included in the molecular descriptors considered in this study. However, the absence of direct group counts (except n=C=) in the list of sensitive features suggests that there are superior descriptors for predicting enthalpy.

To get insights into the performance of the 10 most sensitive descriptor set (Table 5) we identified, a SVR model for enthalpy using these top 10 descriptors was developed and the model showed a reasonable low mean absolute error of 2.34 kca/mol. We therefore conclude that these kinds of descriptors might be playing a special role in determining the thermochemical properties of molecules and may have to be included in QSPR models to achieve a good performance.

Prospective Improvements. Although machine learning models are better compared to presently employed group additivity schemes, improvement is needed to achieve better predictions. A frequency plot in Figure 4 shows the distribution of data over the range of enthalpy values present. In the K-fold error estimation values in Table 3, the highest error values come from a test set that consists of extreme values with little data available, as shown in the frequency plot. Errors can be reduced by generating additional data that would add data points to these low-frequency enthalpy values. This could be achieved by performing either experiments or high-accuracy quantum chemical calculations. Furthermore, the workflow employed in this study is general and can be expanded to the other categories of hydrocarbons and also to other thermodynamic properties, for example, entropy and heat capacities.

SUMMARY AND CONCLUSIONS

This study complements efforts to generate robust chemical kinetic mechnaims by providing a more accurate means of predicting thermodynamic data. A data science approach was

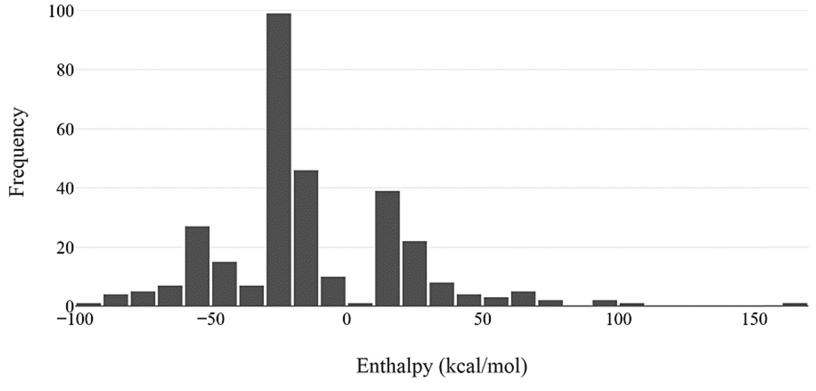


Figure 4. Enthalpy frequency histogram.

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

followed to estimate enthalpy for a particular class of hydrocarbons (i.e., alkanes, alkenes, and alkynes). SVR and ANN models were trained on an extensive dataset collected from the literature. The final SVR and ANN models are available in the Supporting Information with details of the procedure followed for training and instructions on how to use them. The input to these models can be generated from Dragon by providing SMILES strings for the compounds. Estimated errors for both the models suggest that SVR gives a better accuracy because more extensive hyperparameter tuning was possible with less computational expense.

A comparison of the estimated enthalpy of the SVR model with traditional group additivity suggests that there is large scope of improvement that can be achieved by using machine learning models. Supplementing this, a sensitivity analysis pointed out that enthalpy is correlated with descriptors that are not direct group counts. The advantages of machine learning models over the traditional group additivity method can be summarised as follows:

1. A wide set of molecular descriptors can be considered, including functional groups.
2. A nonlinear function can relate input features with calculated outputs, in contrast to a simple linear relation.
3. A more robust methodology for realistic error estimation is possible with ML models, as well as a simpler means of retraining and improving models as new data become available.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpca.9b04771.

Enthalpy database used in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: kiran.yalamanchi@kaust.edu.sa (K.K.Y.).
*E-mail: mani.sarathy@kaust.edu.sa (S.M.S.).

ORCID

Kiran K. Yalamanchi: 0000-0002-9990-0046
M. Monge-Palacios: 0000-0003-1199-5026
S. Mani Sarathy: 0000-0002-3975-6206

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work at King Abdullah University of Science and Technology (KAUST) was supported by the KAUST Clean Fuels Consortium (KCFC) and its member companies.

■ REFERENCES

(1) Lu, T.; Law, C. K. Toward Accommodating Realistic Fuel Chemistry in Large-Scale Computations. *Prog. Energy Combust. Sci.* **2009**, *35*, 192−215.
(2) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(3) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(4) Hu, L.; Wang, X.; Wong, L.; Chen, G. Combined First-Principles Calculation and Neural-Network Correction Approach for Heat of Formation. *J. Chem. Phys.* **2003**, *119*, 11501−11507.
(5) Wu, J.; Xu, X. The X1 Method for Accurate and Efficient Prediction of Heats of Formation. *J. Chem. Phys.* **2007**, *127*, 214105.
(6) Sun, J.; Wu, J.; Song, T.; Hu, L.; Shan, K.; Chen, G. Alternative Approach to Chemical Accuracy: A Neural Networks-Based First-Principles Method for Heat of Formation of Molecules Made of H, C, N, O, F, S, and Cl. *J. Phys. Chem. A* **2014**, *118*, 9120−9131.
(7) Abdul Jameel, A. G.; Van Oudenhoven, V.; Emwas, A.-H.; Sarathy, S. M. Predicting Octane Number Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks. *Energy Fuels* **2018**, *32*, 6309−6329.
(8) de Oliveira, F. M.; de Carvalho, L. S.; Teixeira, L. S. G.; Fontes, C. H.; Lima, K. M. G.; Câmara, A. B. F.; Araújo, H. O. M.; Sales, R. V. Predicting Cetane Index, Flash Point, and Content Sulfur of Diesel−Biodiesel Blend Using an Artificial Neural Network Model. *Energy Fuels* **2017**, *31*, 3913−3920.
(9) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Pidol, L.; Jeuland, N.; Creton, B. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. *Energy Fuels* **2011**, *25*, 3900−3908.
(10) *Materials Studio*; Accelrys Software Inc., San Diego, CA: 2009.
(11) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. On the Rational Formulation of Alternative Fuels: Melting Point and Net Heat of Combustion Predictions for Fuel Compounds Using Machine Learning Methods. *SAR QSAR Environ. Res.* **2013**, *24*, 259−277.
(12) Rowley, R. L.; Wilding, W. V.; Oscarson, J. L.; Yang, Y.; Zundel, N. A.; Daubert, T. E.; Danner, R. P. *DIPPR® Data Compilation of Pure Compound Properties, Design Institute for Physical Properties*; AIChE: New York, 2003.
(13) Yaws, C. L.; *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel: Norwich, NY, 2003.
(14) Sennott, T.; Gotianun, C.; Serres, R.; Ziabasharhagh, M.; Mack, J. H.; Dibble, R. Artificial Neural Network for Predicting Cetane Number of Biofuel Candidates Based on Molecular Structure. In *ASME 2013 Internal Combustion Engine Division Fall Technical Conference*; American Society of Mechanical Engineers, 2013.
(15) *Dragon 7.0*; Kode Chemoinformatics srl, 2017, https://chm.kode-solutions.net/.
(16) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008; Vol. *11*.
(17) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.
(18) Joback, K. G.; Reid, R. C. Estimation of pure-component properties from Group-Contributions. *Chem. Eng. Commun.* **1987**, *57*, 233−243.
(19) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086−1099.
(20) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546−572.
(21) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.
(22) Cohen, N.; Benson, S. W. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419−2438.
(23) Sabbe, M. K.; Saeys, M.; Reyniers, M.-F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. Group Additive Values for the Gas Phase Standard Enthalpy of Formation of Hydrocarbons and Hydrocarbon Radicals. *J. Phys. Chem. A* **2005**, *109*, 7466−7480.
(24) Burke, S. M.; Simmie, J. M.; Curran, H. J. Critical Evaluation of Thermochemical Properties of C1−C4 Species: Updated Group-Contributions to Estimate Thermochemical Properties. *J. Phys. Chem. Ref. Data* **2015**, *44*, No. 013101.

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313

(25) Ritter, E. R. THERM: A Computer Code for Estimating Thermodynamic Properties for Species Important to Combustion and Reaction Modeling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 400−408.
(26) Goteng, G. L.; Nettyam, N.; Sarathy, S. M. CloudFlame: Cyberinfrastructure for Combustion Research. In *2013 International Conference on Information Science and Cloud Computing Companion*; IEEE, 2013; pp 294−299.
(27) Ghahremanpour, M. M.; van Maaren, P. J.; Ditz, J. C.; Lindh, R.; van der Spoel, D. Large-Scale Calculations of Gas Phase Thermochemistry: Enthalpy of Formation, Standard Entropy, and Heat Capacity. *J. Chem. Phys.* **2016**, *145*, 114305.
(28) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*, 2nd ed., Chapman & Hall, London, 1986.
(29) Goldsmith, C. F.; Magoon, G. R.; Green, W. H. Database of Small Molecule Thermochemistry for Combustion. *J. Phys. Chem. A* **2012**, *116*, 9033−9057.
(30) Lide, D. R., Ed. *CRC Handbook of Chemistry and Physics 90th edition*; CRC Press, 2009.
(31) Linstrom, P. J.; Mallard, W. G. *NIST Chemistry WebBook*, National Institute of Standards and Technology, Gaithersburg MD, 20899. http://webbook.nist.gov
(32) Ruscic, B.; Pinzon, R. E.; von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoy, D.; Wagner, A. F. Active Thermochemical Tables: Thermochemistry for the 21st Century. *J. Phys. Conf. Ser.* **2005**, *16*, 561−570.
(33) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A Complete Basis Set Model Chemistry. VI. Use of Density Functional Geometries and Frequencies. *J. Chem. Phys.* **1999**, *110*, 2822−2827.
(34) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A Complete Basis Set Model Chemistry. VII. Use of the Minimum Population Localization Method. *J. Chem. Phys.* **2000**, *112*, 6532−6542.
(35) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 Theory for Molecular Energies of First- and Second-row Compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.
(36) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764−7776.
(37) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126*, 084108.
(38) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A., Jr.; Frisch, M. J.; Martin, J. M. L. Unrestricted Coupled Cluster and Brueckner Doubles Variations of W1 Theory. *J. Chem. Theory Comput.* **2009**, *5*, 2687−2693.
(39) Pearson's Correlation Coefficient. In *Encyclopedia of Public Health*; Kirch, W., Ed.; Springer Netherlands: Dordrecht, 2008; pp 1090−1091.
(40) Chollet, F. *Keras*. 2015, https://keras.io.
(41) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. Tensorflow: A System for Large-Scale Machine Learning. In *12th Symposium on Operating Systems Design and Implementation*; 2016; pp 265−283.
(42) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(43) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; Vapnik, V. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9*; MIT Press, 1997; pp 155−161.
(44) Vert, J. P.; Tsuda, K. A. Primer on Kernel Methods. *Kernel Methods in Computational Biology*; MIT Press, 2004; *47*, 35−70.
(45) He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision*; IEEE, 2015.
(46) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv1412.6980. arXiv.org e-Print archive, https://arxiv.org/abs/1412.6980.
(47) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212−225.
(48) Gevrey, M.; Dimopoulos, I.; Lek, S. Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models. *Ecol. Modell.* **2003**, *160*, 249−264.
(49) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(50) Geary, R. C. The Contiguity Ratio and Statistical Mapping. *Inc. Stat.* **1954**, *5*, 115−146.

DOI: 10.1021/acs.jpca.9b04771
J. Phys. Chem. A 2019, 123, 8305−8313