THE JOURNAL OF PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA Article

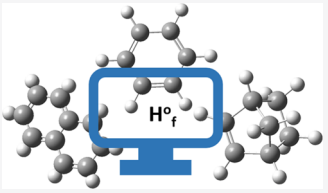
Data Science Approach to Estimate Enthalpy of Formation of Cyclic Hydrocarbons

Kiran K. Yalamanchi,* M. Monge-Palacios, Vincent C. O. van Oudenhoven, Xin Gao, and S. Mani Sarathy*

Cite This: J. Phys. Chem. A 2020, 124, 6270–6276 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In spite of increasing importance of cyclic hydrocarbons in various chemical systems, studies on the fundamental properties of these compounds, such as enthalpy of formation, are still scarce. One of the reasons for this is the fact that the estimation of the thermodynamic properties of cyclic hydrocarbon species via cost-effective computational approaches, such as group additivity (GA), has several limitations and challenges. In this study, a machine learning (ML) approach is proposed using a support vector regression (SVR) algorithm to predict the standard enthalpy of formation of cyclic hydrocarbon species. The model is developed based on a thoroughly selected dataset of accurate experimental values of 192 species collected from the literature. The molecular descriptors used as input to the SVR are calculated via alvaDesc software, which computes in total 5255 features classified into 30 categories. The developed SVR model has an average error of approximately 10 kJ/mol. In comparison, the SVR model outperforms the GA approach for complex molecules and can be therefore proposed as a novel data-driven approach to estimate enthalpy values for complex cyclic species. A sensitivity analysis is also conducted to examine the relevant features that play a role in affecting the standard enthalpy of formation of cyclic species. Our species dataset is expected to be updated and expanded as new data are available to develop a more accurate SVR model with broader applicability.



1. INTRODUCTION

With the development of alternative fuels coming from different sources, as well as new additives from petroleum, cyclic hydrocarbons have become important components of current and future fuels.[1,2] Furthermore, cyclic hydrocarbons, such as polycyclic aromatic hydrocarbons (PAH), are common intermediates in flames that lead to soot formation. Cyclic hydrocarbons are important not only in combustion chemistry but also in other fields; cyclic unsaturated hydrocarbons can lead to the formation of Criegee intermediates[3] and highly oxidized organic compounds,[4] with implications in pollutant formation and climate. Therefore, knowledge of their molecular properties can help build models for atmospheric and combustion modeling. Despite their importance, lesser is known about the oxidation process of cyclic hydrocarbons compared to their aliphatic counterparts, as more theoretical and experimental studies have been concerned with the latter.[5] Nevertheless, additional data for complex cyclic hydrocarbons, saturated and unsaturated, have been derived in the recent years.[6,7]

Chemical kinetics and molecular thermochemistry studies are of vital importance for the development of kinetic models, which are useful to gain knowledge of the oxidation properties of fuels. Accurate thermochemical properties are critical for combustion modeling, as has been recently proved by vom Lehn et al.,[8] who observed that ignition delay of diethyl ether mixtures is more sensitive to the enthalpy of formation of certain species than to kinetic parameters. Different tools can be used to estimate thermochemical properties. Quantum chemistry calculations can yield accurate predictions but become impractical for large molecules with many heavy atoms. A more computationally feasible alternative, although less accurate, is the group additivity method,[9] which calculates the enthalpy of formation from so-called group contribution values. However, this approach assumes that each group is independent, and the contributions of those groups are additive; this may result in a poor description of the thermochemistry of cyclic species, as ring strain is influenced by more than just immediate neighboring atoms considered when defining group values. To cope with this limitation, ring corrections are often added; however, this alternative group additivity method only yields accurate predictions for cyclic species closely related to ones that are included in the training database, thereby limiting its applicability. Some other difficulties in the implementation of ring corrections have been reported elsewhere.[9,10]

Alternative techniques to estimate thermochemical properties are therefore necessary to promote further studies on the oxidation properties of cyclic hydrocarbons. Machine learning

Received: March 29, 2020
Revised: July 9, 2020
Published: July 10, 2020

(ML) has become a popular tool to predict molecular properties in recent years. ML models for estimating several chemical properties, viz., octane number,[11] flash point,[12,13] cetane number,[12,13] melting point,[14] solubility, and toxicity,[15] have been developed using different input representations. Saldana et al.[14] used ML compounded with geometry optimization for estimation of heat of combustion using the data taken from the DIPPR database[16] and the Yaws' handbook of thermodynamic and physical properties of chemical compounds.[17] Although their dataset included a diverse set of compounds, most of them were aliphatic and thus showed highest errors for cyclic species, with an overall error of 52 kJ/mol. Li et al.[18] used "active learning" to develop a self-training and continuously evolving model for enthalpy of polycyclic species calculated from the low-level density functional theory (DFT) B3LYP/6-31G(2df,p). Since the training dataset was from low-level theory calculations, error in their work could be attributed to both error from the ML model and error from the DFT calculations, compared with experimental values or calculations at a higher level of theory. However, their large initial dataset and self-evolving algorithm provide a platform for predicting enthalpy for a wide spectrum of molecules.

In the present study, we focus on estimating the standard enthalpy of formation at 298 K and 1 atm (hereafter referred to as enthalpy) of cyclic species using ML techniques based on a dataset consisting exclusively of accurate experimental data. Our goal is to develop an ML model that yields accurate enthalpies of cyclic species, which are scarce, to facilitate further studies on the oxidation of cyclic hydrocarbons. Section 2 provides details on the dataset used in this study along with details of the input features used for the ML model.

2. DATA CURATION

The database used in this work consists of experimental measurements reported by Ghahremanpour et al.,[19] CRC,[20] and Minenkov et al.[21] The dataset consists of 192 cyclic hydrocarbon species including both saturated and unsaturated species with up to 14 carbon atoms. Among those 192 species, both constitutional isomers and *R* and *S* enantiomers (species with chiral centers) are included, making our ML model able to discern between different kinds of isomers.

Due to the lack of data reported for cyclic hydrocarbons beyond C_{14} species, we restricted our dataset to species with up to 14 carbon atoms. Larger species with several rings can assume different structures and conformers that alter their thermochemistry, so training an ML model for such species requires a large and comprehensive dataset. This is the case for C_{15}−C_{18} hydrocarbons, for which accurate enthalpies are only available for 10 species. In addition, three-membered ring species were also excluded due to their lower importance in combustion and their skewed values of enthalpy due to high cyclic strain energies. Thus, the final dataset consists of cyclic hydrocarbon species with a minimum of four-membered ring and a maximum of 14 carbon atoms. Figure 1 shows the distribution of species present in our dataset with respect to their number of carbon atoms. It should be noted that the present approach allows us to update the training dataset as new data become available to improve the ML model for a wider range of cyclic hydrocarbons. The dataset with the species considered for our ML model, along with the excluded ones, is provided in the Supporting Material.

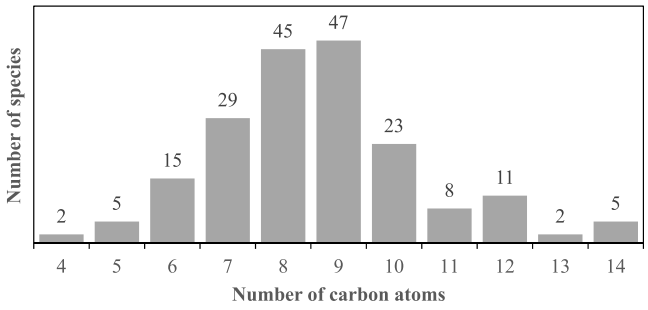


Figure 1. Distribution of species with respect to their number of carbon atoms.

Several input representation methods have been used in the past for predicting molecular properties, such as Coloumb matrices,[22] various sets of molecular descriptors,[12,14] and convolutions.[15] Coloumb matrices and convolution methods require a molecular fingerprint; training such methods requires a relatively large dataset to find relationships between plain molecular structure and a macroscopic property. On the contrary, molecular descriptors are formulations derived from the molecular structure and can be used for small to large datasets with a varying number of descriptors. Todeschini and Consonni[23] compiled a comprehensive set of molecular descriptors that can be calculated from simplified molecular input line entry system (SMILES) in alvaDesc.[24] Out of this comprehensive set of 5255 molecular descriptors sorted in 30 categories, a relevant subset of 5072 descriptors from 29 categories are used in this study. Duplicate descriptors with identical values, such as alcohol functional group counts, which are zero for all of the species in our dataset, are removed to reduce the number of descriptors to 2478. A process to further downselect among these descriptors is explained in Section 3 along with the ML workflow. The complete dataset used for the ML model containing all of the molecular descriptors is provided in the Supporting Material.

3. MACHINE LEARNING FRAMEWORK

The ML model development consists of a workflow and model training. The ML workflow comprises division of the dataset for error estimation and final model development. The training consists of using the ML algorithm to train on the dataset following the workflow and then fine-tuning the hyperparameters associated with the algorithm. Sections 3.1 and 3.2 discuss each part in detail. All of the scripts for this study are written in Python using scikit-learn library[25] with TensorFlow[26] backend. All scripts are available in the GitHub repository (github.com/kiranyalamanchi/enthalpy-cyclic-species).

3.1. Workflow. Two commonly used workflows in ML are as follows: (i) divide the dataset into training/validation/test sets used for training the model, then fine-tune the hyperparameters, and estimate the model performance, and (ii) divide the dataset into *k* sets and use a *k*-fold cross-validation (*k*-fold CV) method to estimate the accuracy of the ML algorithm on the dataset, and then train the final model on the entire dataset. The *k*-fold CV method is computationally demanding but gives a better representation of model accuracy. This method is adopted herein because of the small dataset size. In the accuracy estimation part, the workflow consists of two loops, an outer loop and an inner loop. First, the entire dataset is divided into *k*-sets, $i = 1$ to k, as shown in Figure 2.

https://dx.doi.org/10.1021/acs.jpca.0c02785
J. Phys. Chem. A 2020, 124, 6270−6276

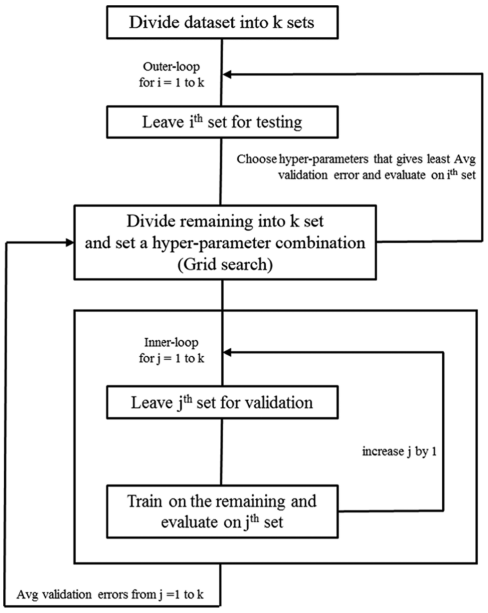


Figure 2. Workflow for error estimation of an ML model with *k*-fold CV.

The outer loop constitutes of $i = 1$ to *k*-folds iterated over until each fold has been used as the test set. For each of *i*th set in the outer loop, the remaining data points expect those in *i*th set are mixed and again divided into $j = 1$ to k sets for the inner loop, as shown in Figure 2. The inner loop is used to determine the best hyperparameters for each particular split of the outer loop from a defined search space. For each combination of hyperparameters, all of the $j = 1$ to *k*-folds in the inner loop are iterated over until each set has been used as a validation set. The accuracy of the inner $j = 1$ to *k*-folds is averaged and used as a metric for the performance of that particular hyperparameter combination. The hyperparameter combination that has the minimum average validation error is used to train a model for that particular *i*th outer split. This model is trained on the remaining $k - 1$ folds (leaving *i*th set from $i = 1$ to *k*-sets) as the training set and then tested on the test set (*i*th set). Since the entire dataset is divided into *k* sets for the outer loop, we have *k* errors, which represents the model's performance. Finally, a model is trained on the entire dataset to obtain the final model, which can be used for finding enthalpy of a new species that is not present in the original dataset. A similar cross-validation pipeline was followed in our previous work for the development of ML models for enthalpy of linear species,[27] and more details on adopted *k*-fold workflow can be found therein. It should be noted that this methodology follows best practice in ML and ensures that the final model is not simply an overfitting of the dataset.

3.2. Model Development. Support vector regression (SVR) and artificial neural network (ANN) are widely used ML algorithms for regression tasks. While ANN performs better for tasks with large datasets, SVR can be very efficient in dealing with smaller ones;[27] hence, the latter is selected for this study. SVR[28] is a regression counterpart developed from support vector machines (SVM),[29] which are large margin classifiers widely used for classification problems. The idea of SVM is to transform a feature space using kernels to create a hyperplane that separates different classes with a large margin. SVR is based on a similar concept except that the hyperplane aims to fit the data. The function of SVR can be mathematically described by eq 1, where ε and ξ_i, ξ_i^* are positive numbers that describe the allowable error (ε) and additional error above ε for a data point i in the training set, respectively. Equation 2 describes the radial basis function (RBF) kernel. Therefore, constraint terms describe the distance of Y_i (output of data point *i*) from the predicted value by the transformed hyperplane using a kernel with parameters ω. This distance is the prediction error and is permitted to be within ε, while the additional ξ_i, ξ_i^* has to be minimized. The term $a\|\omega a\|^2$ is the regularization term used in the cost function to be minimized to avoid overfitting. The parameters *C*, ε, and γ are known as hyperparameters, which are adjusted using the validation set, i.e., in the inner loop for the *k*-fold CV method, as explained in Section 3.1.

$$\text{minimize}_{\omega,b,\xi_i,\xi_i^*} \frac{1}{2}\left\|\omega\right\|^2 + C\sum_{n}^{i=1}\xi_i + \xi_i^* \tag{1}$$

$$\text{such that } Y_i - \langle\omega, X_i\rangle - b \le \varepsilon + \xi_i$$

$$\langle\omega, X_i\rangle + b - Y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0 \ \forall\ i$$

$$\langle\omega, X_i\rangle = \exp\left(\frac{-\|\omega - X_i\|^2}{2\gamma^2}\right) \tag{2}$$

The processed dataset consists of 192 species with 2478 features, which is a large amount of features compared to the data points present. Using this entire feature set could cause overfitting and noise. To reduce the number of features, the dataset is divided randomly into 90/10% training/test, and the test set accuracy is used to select the number of features. Note that this split method is just used for feature selection and not for error estimation, for which the *k*-fold CV workflow is used. Pearson's correlation coefficient[30] is a metric used to measure correlation between vectors and for reduction of features. For a given correlation coefficient threshold, input features with an absolute correlation coefficient exceeding the threshold are removed, and then hyperparameters are tuned on the training set. The model is then used to predict on the test set such that the mean absolute error (MAE) can be obtained, as shown in Figure 3. The plot consists of errors for five different random training/test splits (each shown in different color), which are considered to avoid any bias due to a particular split. All of the test set errors for different splits decrease for a correlation

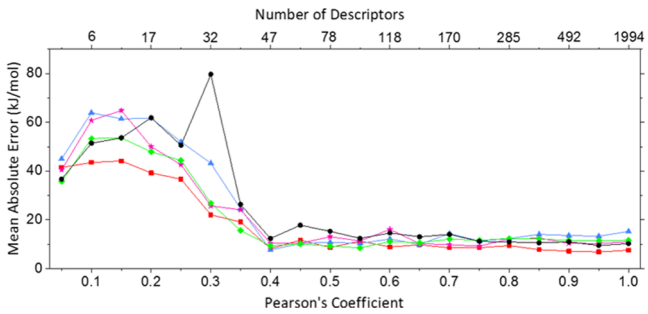


Figure 3. Variation of mean absolute error of various test sets with respect to Pearson's coefficient. This is used to reduce the number of features for training the model. Different colors correspond to different random states (training/test splits) used for dividing the dataset.

https://dx.doi.org/10.1021/acs.jpca.0c02785
J. Phys. Chem. A 2020, 124, 6270−6276

coefficient value of 0.4. A lower correlation coefficient value increases error due to an insufficient number of descriptors available to carry all of the molecular information. A higher value gives approximately similar error, but some noise is observed because the model is tuned based on different descriptors that are correlated. A correlation coefficient of 0.4 corresponds to a reduced set of 47 descriptors.

Table 1 shows the number of descriptors present from each category in the final reduced set of descriptors, the individual list of which is provided in the Supporting Material. These 47 descriptors are used with 192 data points for error estimation using the *k*-fold CV method, and the results are discussed in Section 4.

Table 1. Descriptors of Each Category in the Reduced Set

category of descriptors	number of descriptors
constitutional indices	5
ring descriptors	6
topological indices	3
2D autocorrelations	7
edge adjacency indices	11
3D MoRSE descriptors	4
WHIM descriptors	3
other minor descriptors	8
total	47

4. RESULTS AND DISCUSSION

Using the reduced set of descriptors, the entire dataset was split randomly into *k* sets, and the *k*-fold CV method was applied. It is a common practice to set the value of *k* to 5, 10, or the size of dataset, which are called 5-fold, 10-fold, and leave one out CV (LOOCV), respectively. Since the dataset in this study is small, using 5-fold would leave out too much data for the model to adequately train. LOOCV is a computationally expensive workflow for very small datasets. Therefore, *k* is set to 10 for the present study such that there are 10 test sets in the outer loop of the *k*-fold CV, resulting in 10 sets of errors. The regression (R2) score and MAE for these 10 sets are given in Table 2. An average MAE of 9.71 kJ/mol is achieved using SVR with the 10-fold CV workflow. The highest errors observed for fold numbers 6 and 7 arise from the species with a high prediction error, azulene and cyclohepta-1,3,5-triene, respectively. This is due to their unique structures compared to other species in the dataset, i.e., a seven-membered ring with resonant double bonds.

Table 2. Regression (R2) Score and MAE for SVR Model Using *k*-Fold CV Workflow Applied to the Dataset in This Study

fold number	R2 score	MAE (kJ/mol)
1	0.986	7.32
2	0.992	8.40
3	0.992	7.70
4	0.994	7.14
5	0.992	10.67
6	0.957	13.02
7	0.986	14.35
8	0.990	9.47
9	0.990	10.59
10	0.992	8.45
average	0.987	9.71

It is also worth mentioning that the ML model prediction accuracy is more sensitive to the availability of data for species with similar structures than for species with the same number of carbon atoms. This can be explained by C_{13} species having lower average error than C_{10} species, as shown in Figure 4,

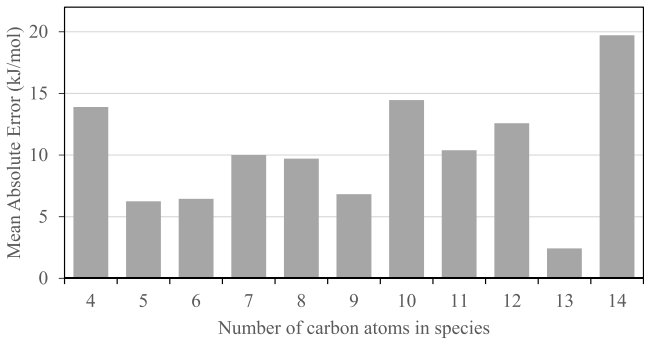


Figure 4. Variation of accuracy with respect to number of carbon atoms in species (averaged).

despite the fact that the number of C_{10} and C_{13} species are 23 and 2, respectively. The two C_{13} species, 9H-fluorene and diphenylmethane, are planar and comprise mainly benzene rings, for which there are much data available. In contrast, the C_{10} species include some unique structures, such as azulene, for which it is more difficult to predict its enthalpy due to a few similar species in the database.

4.1. Comparison with Group Additivity. Bensons group additivity (GA)[9] is an easy and fast method for calculating thermochemical properties. Han et al.[31] refined the GA approach by including ring corrections to account for the complexity of ring structures to improve the performance of GA for cyclic species. Despite this improvement, and due to limitations in the underlying assumptions of GA, this method is only effective for estimating the molecular thermochemistry of relatively simple organic molecules, and its accuracy rapidly decreases as the species become more complex, as was reported by Li et al.[18]

A direct estimation of the improvement achieved by our SVR model over the GA approach needs both models to be trained on the same data, but this is beyond the scope of this work. Therefore, in Figure 5, GA values taken from RMG[32] and present SVR predictions are compared to experimental

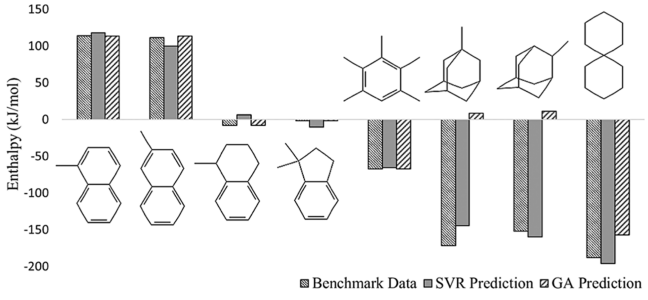


Figure 5. Comparison of GA and SVR predictions against benchmark experimental data used in this study. Species from left to right: 1-methylnaphthalene, 2-methylnaphthalene, 1-methyl-1,2,3,4-tetrahydronaphthalene, 1,1-dimethylindan, pentamethyl-benzene, 1-methyladamantane, 2-methyladamantane, and spiro[5.5]undecane.

https://dx.doi.org/10.1021/acs.jpca.0c02785
J. Phys. Chem. A 2020, 124, 6270−6276

values for the eight C_{11} species in the dataset. It is important to mention that the SVR predictions in these comparisons span over all of the folds shown in Table 2. Therefore, every prediction is from a model that was trained on a dataset that does not contain the corresponding predicted species; the species was used for external validation in a test set in the *k*-fold workflow. Both GA (5.77 kJ/mol MAE) and SVR (7.82 kJ/mol MAE) achieve good accuracy on predictions of enthalpy values for molecules with relatively simple structures such as 1-methylnaphthalene, 2-methylnaphthalene, 1-methyl-1,2,3,4-tetrahydronaphthalene, 1,1-dimethylindan, pentamethyl-benzene, and spiro[5.5]undecane. The slightly better performance of GA could be attributed to the fact that the SVR values are from a model that was not trained on these particular species, whereas this may not be the case for GA. SVR outperforms the GA method for molecules with more complex structures, such as 1-methyladamantane and 2-methyladamantane. The average error for 1-methyladamantane and 2-methyladamantane is greater for GA (171.54 kJ/mol MAE) and smaller for SVR (17.57 kJ/mol MAE). This is a promising result, indicating that the SVR method yields more accurate predictions for complex cyclic molecules for which the GA approach usually fails, as has been noted by Li et al.[18] Cyclic molecules with multiple rings connected together assume complex and unique geometry that contributes to enthalpy via ring strain, making it difficult for the GA approach to capture complexity mere ring corrections.

4.2. Sensitivity Analysis. To get more insights into the molecular descriptors affecting enthalpy, a sensitivity analysis was performed. The SVR model trained on the whole dataset is used for sensitivity analysis in this section. The perturb method, as discussed in Gevrey et al.,[33] was used to find the sensitivity of enthalpy to each of the input features. All of the input features are first assigned their mean, and then one of the input features is increased by 10% of its standard deviation. The output from the SVR model using the modified input, when one of the feature values was increased, is compared to the output with all of the features assigned with their mean values. This difference is then normalized by the 10% of the standard deviation of the feature that is perturbed. Finally, sensitivity of a feature is determined by taking the absolute value of the resulting normalized values. These sensitivity values are used to sort the features, and the top 10 sensitive features are listed in Table 3.

Among the 10 features listed in Table 3, four (nCsp3, nDB, MW, and nTA) are simple constitutional descriptors describing chemical composition of a compound without any information about its molecular geometry or atom connectivity. Out of the remaining six, four are ring descriptors (nR03, nR04, Rbrid, and nR05) which describe the number and size of rings as well as the number of ring bridges; these descriptors are crucial to discern among different cycles and their arrangement within the dataset. MAXDN is the maximum of the field effects on the atoms in a molecule due to the perturbation of all other atoms.[34] Mor26u is a 3D MoRSE descriptor calculated from the whole molecule structure and is one of the few descriptors that could discern between species with chiral centers. This descriptor can be used to establish relationships between enthalpy and chirality. Although the Morse descriptor is difficult to interpret, many studies have found them to be useful in quantitative structure−property relationship (QSPR) studies.[35] MAXDN and Mor26U are the only sensitive descriptors that are related to the geometry of molecules, while the remaining are constitutional and ring descriptors.

These findings are in contrast to those of the noncyclic compounds previously studied[27] by us, wherein enthalpy predictions were found to be highly sensitive to P_VSA descriptors, which are related to the polarizability and electronic structure of the species. This indicates that enthalpy of cyclic compounds is highly influenced by the number and size of their cycles, as well as by the arrangement these cycles adopt, that is, by ring descriptors. These conclusions are helpful for the development of further ML models on cyclic and/or linear hydrocarbons, which can benefit from our sensitivity analysis that ranks the importance of different descriptors.

Table 3. List of the 10 Most Sensitive Features

feature	description
$nCsp^3$	number of sp^3-hybridized carbon atoms
nDB	number of double bonds
MW	molecular weight
nR03	number of three-membered rings
nTA	number of all atoms
nR04	number of four-membered rings
Rbrid	ring bridge count
nR05	number of five-membered rings
MAXDN	maximal electro topological negative variation
Mor26u	signal 26/unweighted

5. CONCLUSIONS

A data-driven approach based on the SVR algorithm was developed to predict enthalpy values for cyclic hydrocarbons with a dataset of 192 species collected from Ghahremanpour et al.,[19] CRC,[20] and Minenkov et al.[21] Molecular descriptors from alvaDesc[24] were used as input features that are generated from the output of SMILES, which are chemical formulas encoded as text strings. A *k*-fold workflow with an SVR algorithm was used to determine the accuracy with which our ML model could predict standard enthalpy of formation of cyclic hydrocarbons. In comparison to the group additivity method, our ML model performs better for species with complex structures. Sensitivity analysis reveals that simple molecular descriptors reflecting the ring nature and overall size of species play a more important role compared to more complex descriptors involving the shape of a species. Our ML model represents an accurate and computationally practical alternative to well-established GA and quantum chemistry methods for the prediction of enthalpy data of cyclic species, which are scarce in the literature despite their importance in combustion.

Access to more data can greatly enhance the accuracy and predictive capability of the presented ML model. Therefore, it is expected that this ML model will be improved as new and accurate data for enthalpy of cyclic hydrocarbons become available. However, calculating accurate enthalpies for cyclic and large hydrocarbons can be computationally expensive, and methods like active learning used by Li et al.[18] for better design of experiments (DoE) should be considered for this task. Furthermore, it can be useful to train the existing model with large datasets consisting of uniform inputs, as opposed to widely varying features. In this sense, it would be also beneficial to identify a subset of cyclic species specifically relevant to

https://dx.doi.org/10.1021/acs.jpca.0c02785
J. Phys. Chem. A 2020, 124, 6270−6276

combustion and fine-tune active learning model with the modified subset.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c02785.

Enthalpy database used in this study (.csv); molecular descriptors used in this study for all of the species used for developing ML model (.csv); final reduced set of descriptors (.csv); all of the supporting information compiled (.xlsx) (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Kiran K. Yalamanchi − *Physical Sciences and Engineering Division, Clean Combustion Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia;* orcid.org/0000-0002-9990-0046; Email: kiran.yalamanchi@kaust.edu.sa

S. Mani Sarathy − *Physical Sciences and Engineering Division, Clean Combustion Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia;* orcid.org/0000-0002-3975-6206; Email: mani.sarathy@kaust.edu.sa

Authors

M. Monge-Palacios − *Physical Sciences and Engineering Division, Clean Combustion Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia;* orcid.org/0000-0003-1199-5026

Vincent C. O. van Oudenhoven − *Physical Sciences and Engineering Division, Clean Combustion Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*

Xin Gao − *Computer, Electrical and Mathematical Sciences and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.0c02785

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research under the award number OSR-2019-CRG7-4077, and the KAUST Clean Fuels Consortium (KCFC) and its member companies.

■ REFERENCES

(1) Yang, Y.; Boehman, A. L. Oxidation Chemistry of Cyclic Hydrocarbons in a Motored Engine: Methylcyclopentane, Tetralin, and Decalin. *Combust. Flame* **2010**, *157*, 495−505.

(2) Chen, F.; Li, N.; Li, S.; Li, G.; Wang, A.; Cong, Y.; Wang, X.; Zhang, T. Synthesis of Jet Fuel Range Cycloalkanes with Diacetone Alcohol from Lignocellulose. *Green Chem.* **2016**, *18*, 5751−5755.

(3) Monge-Palacios, M.; Rissanen, M. P.; Wang, Z.; Sarathy, S. M. Theoretical Kinetic Study of the Formic Acid Catalyzed Criegee Intermediate Isomerization: Multistructural Anharmonicity and Atmospheric Implications. *Phys. Chem. Chem. Phys.* **2018**, *20*, 10806−10814.

(4) Rissanen, M. P.; Kurtén, T.; Sipilä, M.; Thornton, J. A.; Kangasluoma, J.; Sarnela, N.; Junninen, H.; Jørgensen, S.; Schallhart, S.; Kajos, M. K.; et al. The Formation of Highly Oxidized Multifunctional Products in the Ozonolysis of Cyclohexene. *J. Am. Chem. Soc.* **2014**, *136*, 15596−15606.

(5) Simmie, J. M. Detailed chemical kinetic models for the combustion of hydrocarbon fuels. *Prog. Energy Combust. Sci.* **2003**, *29*, 599−634.

(6) Zhang, I. Y.; Wu, J.; Xu, X. Accurate Heats of Formation of Polycyclic Saturated Hydrocarbons Predicted by Using the XYG3 Type of Doubly Hybrid Functionals. *J. Comput. Chem.* **2019**, *40*, 1113−1122.

(7) Gao, C. W.; Vandeputte, A. G.; Yee, N. W.; Green, W. H.; Bonomi, R. E.; Magoon, G. R.; Wong, H.-W.; Oluwole, O. O.; Lewis, D. K.; Vandewiele, N. M.; et al. JP-10 Combustion Studied with Shock Tube Experiments and Modeled with Automatic Reaction Mechanism Generation. *Combust. Flame* **2015**, *162*, 3115−3129.

(8) vom Lehn, F.; Cai, L.; Pitsch, H. Impact of Thermochemistry on Optimized Kinetic Model Predictions: Auto-Ignition of Diethyl Ether. *Combust. Flame* **2019**, *210*, 454−466.

(9) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.

(10) Yu, J.; Sumathi, R.; Green, W. H. Accurate and Efficient Method for Predicting Thermochemistry of Polycyclic Aromatic Hydrocarbons- Bond-Centered Group Additivity. *J. Am. Chem. Soc.* **2004**, *126*, 12685−12700.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) Yaws, C. L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel: Norwich, NY, 2003.

(18) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123*, 2142−2152.

(19) 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*, No. 114305.

(20) *CRC Handbook of Chemistry and Physics*, 90th Edition; Lide, D. R., Ed.; CRC Press: Boca Raton, Florida, 2009

(21) Minenkov, Y.; Wang, H.; Wang, Z.; Sarathy, S. M.; Cavallo, L. Heats of Formation of Medium-Sized Organic Compounds from Contemporary Electronic Structure Methods. *J. Chem. Theory Comput.* **2017**, *13*, 3537−3560.

(22) 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.
(23) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008; Vol. *11*.
(24) Alvascience Inc. alvaDesc. https://www.alvascience.com/alvadesc/.
(25) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J. et al. In *API Design for Machine Learning Software: Experiences from the Scikit-Learn Project*, ECML PKDD Workshop: Languages for Data Mining and Machine Learning; 2013; pp 108−122.
(26) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *Tensorflow: A System for Large-Scale Machine Learning*, 12th Symposium on Operating Systems Design and Implementation; 2016; pp 265−283.
(27) Yalamanchi, K. K.; van Oudenhoven, V. C. O.; Tutino, F.; Monge-Palacios, M.; Alshehri, A.; Gao, X.; Sarathy, S. M. Machine Learning To Predict Standard Enthalpy of Formation of Hydrocarbons. *J. Phys. Chem. A* **2019**, *123*, 8305−8313.
(28) 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.
(29) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(30) Pearson's Correlation Coefficient. In *Encyclopedia of Public Health*; Kirch, W., Ed.; Springer Netherlands: Dordrecht, 2008; pp 1090−1091.
(31) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation. *Int. J. Chem. Kinet.* **2018**, *50*, 294−303.
(32) 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.
(33) 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.
(34) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Molecular Descriptors. *Chemosphere* **2000**, *41*, 763−777.
(35) Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE Descriptors Explained. *J. Mol. Graphics Modell.* **2014**, *54*, 194−203.

https://dx.doi.org/10.1021/acs.jpca.0c02785
J. Phys. Chem. A 2020, 124, 6270−6276