THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

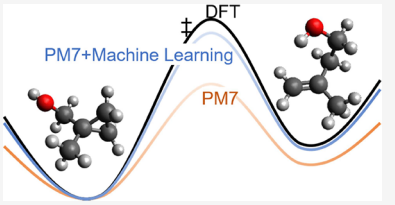
pubs.acs.org/JPCA
Article

Barrier Height Prediction by Machine Learning Correction of Semiempirical Calculations

Xabier García-Andrade, Pablo García Tahoces, Jesús Pérez-Ríos, and Emilio Martínez Núñez*

Cite This: J. Phys. Chem. A 2023, 127, 2274−2283
Read Online

ABSTRACT: Different machine learning (ML) models are proposed in the present work to predict density functional theory-quality barrier heights (BHs) from semiempirical quantum mechanical (SQM) calculations. The ML models include a multitask deep neural network, gradient-boosted trees by means of the XGBoost interface, and Gaussian process regression. The obtained mean absolute errors are similar to those of previous models considering the same number of data points. The ML corrections proposed in this paper could be useful for rapid screening of the large reaction networks that appear in combustion chemistry or in astrochemistry. Finally, our results show that 70% of the features with the highest impact on model output are bespoke predictors. This custom-made set of predictors could be employed by future Δ-ML models to improve the quantitative prediction of other reaction properties.



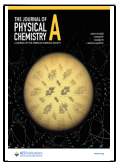
1. INTRODUCTION

Transition state theory (TST) provides a useful means to study the kinetics of elementary chemical reactions.[1] Depending on the specific version, TST requires a more or less exhaustive knowledge of the potential energy surface of the system.[2] In the absence of strong tunneling effects, the value of the Gibbs energy of activation $\Delta G^{\ddagger}$ [Gibbs energy difference between the transition state (TS) and the reactant(s)] is sufficient to predict the rate of reaction. At 0 K, $\Delta G^{\ddagger}$ is just the electronic energy difference between the TS and reactant including their zero-point vibrational energies (ZPEs), called the barrier height (BH). Although the BH does not include the thermal correction to enthalpy and the entropic contribution, sometimes it is employed as a proxy for the true Gibbs energy of activation. Nevertheless, predicting highly accurate BHs (of sub-kcal/mol accuracy) requires the use of expensive ab initio methods, such as the gold standard coupled cluster including single and double excitations with perturbative triple excitations [CCSD(T)].[3] Fortunately, today's state-of-the-art density functionals predict BHs that are rather close to the accurate CCSD(T),[4] thus being the method of choice for modeling large systems. However, even density functional theory (DFT) becomes prohibitive for biochemical systems or for complex reaction networks of medium-size systems.

With the surge of large computational and experimental data sets, machine learning (ML) is shifting the paradigm to data-driven predictive modeling. This approach has been pursued to predict activation energies and BHs in previous studies.[5−15] By way of example, Choi et al. developed different ML models to predict activation energies of gas-phase reactions, with the tree-boosting method showing the best performance.[5] More recently, Green and co-workers have demonstrated that it is possible to predict accurate BHs using a deep learning (DL) model given only reactant and product graphs.[6,8] Green's DL model was trained on a gas-phase organic chemistry (GPOC) data set of 12,000 chemical reactions involving carbon, hydrogen, nitrogen, and oxygen. The calculations were carried out at the DFT ωB97X-D3/def2-TZVP quantum chemistry level, which has been shown to predict BHs with a mean absolute error (MAE) of 3.5 kcal/mol against a CCSD(T)-F12 reference.[16] An updated version of the GPOC is available,[16] with BHs calculated at the CCSD(T)-F12 level of theory; in a follow-up work our models will be improved using the newest data set. Green and co-workers recently improved their model using fewer parameters and proper data splits to estimate performance on unseen reactions.[8] In addition, Habershon and co-workers employed this basis set to predict rates of chemical reactions.[17] Alexandrova and co-workers have also shown that topological descriptors of the quantum mechanical charge density in the reactant state can be used to predict BHs for Diels−Alder reactions.[9] Hybrid models combining traditional TS modeling and ML are also employed to predict BHs for nucleophilic aromatic substitution reactions in solution.[10]

Semiempirical quantum mechanical (SQM) methods are significantly faster than DFT and provide results with sufficient accuracy when applied to molecules of the same type as those of the training set.[18] However, except when the interest is in a specific reaction,[19−23] training sets do not usually include data of TSs, which results in inaccurate BH predictions. In an attempt to model the reactivity of organic reactions with useful

Received: November 28, 2022
Revised: February 19, 2023
Published: March 6, 2023



https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

accuracy, Stewart developed the SQM method called PM7-TS.[18] Using a training set of 97 BHs obtained from collections of high-level calculations, the MAE using PM7-TS was 3.8 kcal/mol, as compared with the MAEs for PM7 of 11.0 kcal/mol and for PM6 of 12.2 kcal/mol.[18] However, Jensen and co-workers benchmarked PM7-TS using BHs for five model enzymes and found an MAE of 19 kcal/mol, while the MAEs for PM6 and PM7 were around 12−15 kcal/mol.[24] Iron and Janes[25] have also shown that SQM methods perform very poorly in predicting transition metal BHs: using a new data set with high-accurate energies, the MAEs of PM6, PM7, and PM7-TS are 21.6, 106.4, and 68.2 kcal/mol, respectively. In his PM7 paper, Stewart already acknowledged that the predictive power of PM7-TS was unknown at the time and suggested parameter re-optimization as more BHs became available.[18]

An alternative to parameter optimization is to develop analytical[26] or ML corrections of the SQM calculations. The latter are usually termed Δ-ML because the model predicts the difference between the benchmark and the approximate baseline calculation (SQM in this case).[27] There are some examples in the literature of the successful use of ML to improve the accuracy of both DFT[28−30] and SQM calculations.[31,32]

In this work, we leverage ML to predict BHs with DFT accuracy at the cost of SQM calculations. The model employs multitask deep neural network (DNN), gradient-boosted trees by means of the XGBoost (XGB) interface,[33] and Gaussian process (GP) regression trained on a curated version of the GPOC data set.[7] Gradient boosting regression has been successfully applied to predict BHs in Diels−Alder reactions,[9] and the reactivity of transition metal complexes.[15] Similarly, GP has shown a great performance in complex potential energy surface fittings,[34−37] predicting spectroscopic constants of diatomic molecules[38,39] and second virial coefficients of organic and inorganic compounds.[40] The selected SQM model was PM7,[18] which is overall the most accurate method implemented in MOPAC2016.[41] To fully exploit the SQM calculations, several input features are constructed from the electronic and structural properties of reactant, TSs, and products. Moreover, the model makes different predictions for cases where two TSs exist for the same rearrangement.[42] A similar synergistic SQM/ML approach to predict activation energies for a diverse class of C−C bond-forming nitro-Michael additions has been recently proposed.[43]

The ML correction proposed in this paper could be employed in conjunction with SQM-based methods for automated reaction mechanism prediction like AutoMeKin.[44−47]

2. METHODS

2.1. Performance of PM7-TS on the GPOC Data Set.

Since the accuracy of PM7-TS is uncertain (*vide supra*), its performance was evaluated on the GPOC data set of BHs.[7] Figure 1 shows the correlation between the ωB97X-D3/def2-TZVP BHs and the values predicted by PM7-TS. In general, PM7-TS significantly underestimates the BHs with an MAE of 22.5 kcal/mol, which is in line with the deviation obtained by Jensen and co-workers on a data set of model enzymes[24] and much greater than the reported error of 3.8 kcal/mol on the training set employed to optimize the PM7-TS parameters.[18] These results call for an alternative method to predict accurate SQM-based BHs. The proposal of the present work is to employ ML models to correct the SQM values.

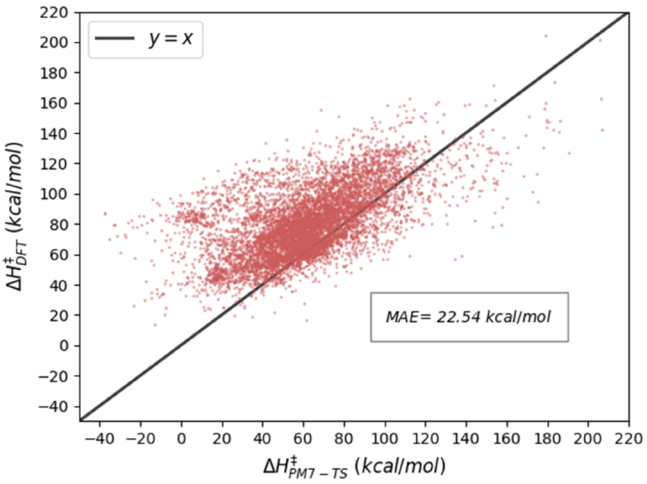


Figure 1. Performance of PM7-TS on the GPOC data set.

2.2. Data Set Curation.

The target for our ML models is the difference $BH^{DFT} - BH^{PM7}$, where BH^{DFT} and BH^{PM7} are the BHs obtained at the benchmark (DFT) and PM7 levels, respectively. The GPOC data set developed by Green and co-workers is employed here.[7] It contains 11,960 reactions, with energies for reactant, TSs, and the product obtained at the ωB97X-D3/def2-TZVP level of DFT. The DFT BHs were directly obtained by subtracting the reactant energy from the TS energy including their ZPEs. Obtaining BHs at the PM7 SQM level entails a more involved process, as several sanity checks are required. All SQM calculations were carried out with MOPAC2016[41] and the settings employed in the different PM7 calculations are detailed in the Supporting Information (SI). A flow chart diagram explaining how the PM7 BHs were obtained is shown in Figure 2. The first step, labeled as TS optimization in the figure, consists of optimizing the TSs at the PM7 level using as initial guesses the geometries optimized at the DFT level. Some structures could not be optimized at the PM7 level and were discarded. Then, for each successfully optimized TS structure, an IRC calculation[48] is carried out in each direction (IRC = 1 and IRC = −1 in the figure). The IRC end points are compared with the reactant and product present in the data set. For such a comparison, the eigenvalues of the corresponding adjacency matrices (with their diagonals representing the atomic numbers) were employed.[49] Obtaining identical eigenvalues ensures that the connectivity of each structure (reactant and product) is the same at both levels of theory (DFT and SQM). When the connectivity differs for either reactant or product, the reaction is discarded. Otherwise, both the IRC end point and the structure from the data set are optimized at the PM7 level and compared to ensure they present the same conformation. For this last comparison that involves 3D structures, the eigenvalues of a weighted adjacency matrix are employed.[49]

From the initial 11,860 reactions, 8355 survived this screening process, meaning that roughly 70% of the samples could be utilized in our model. This means that our approach is limited to situations where a TS can be optimized. The use of reverse BHs did not lead to a major improvement during the training of the model but increased the computational cost, so this form of data augmentation was discarded.

An exploratory data analysis (EDA) of the curated GPOC data set was then carried out. The detailed results of our EDA are collected in the SI. Reactions in the data set contain up to

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

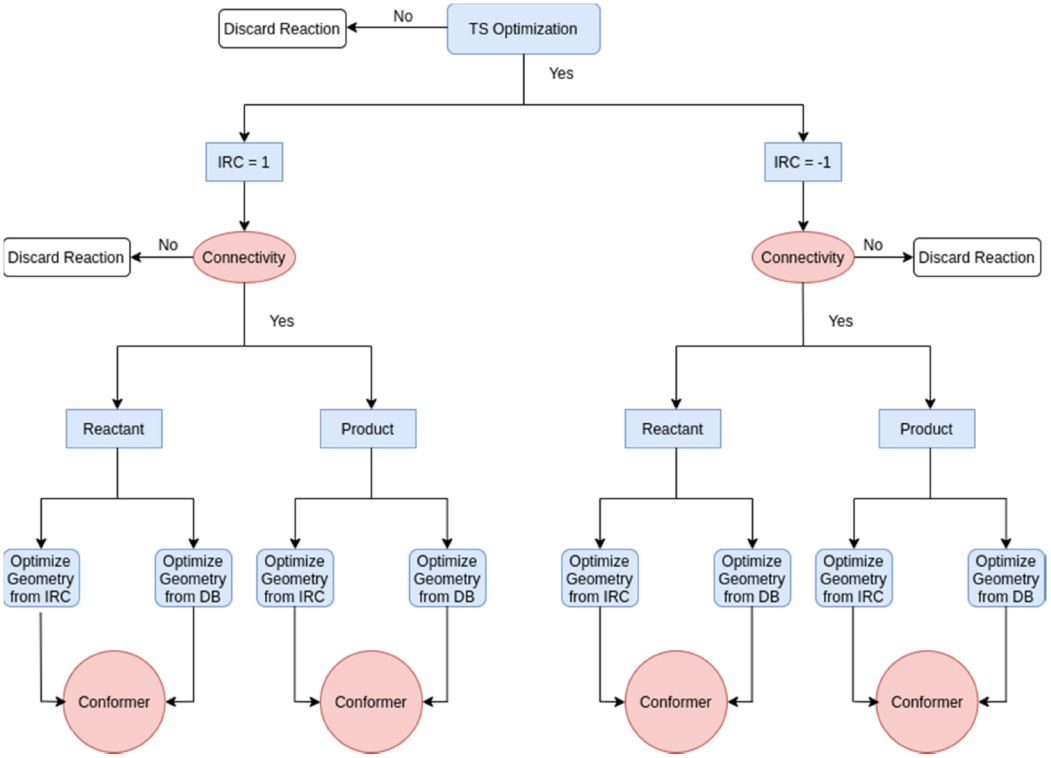


Figure 2. SQM data set generation flow diagram.

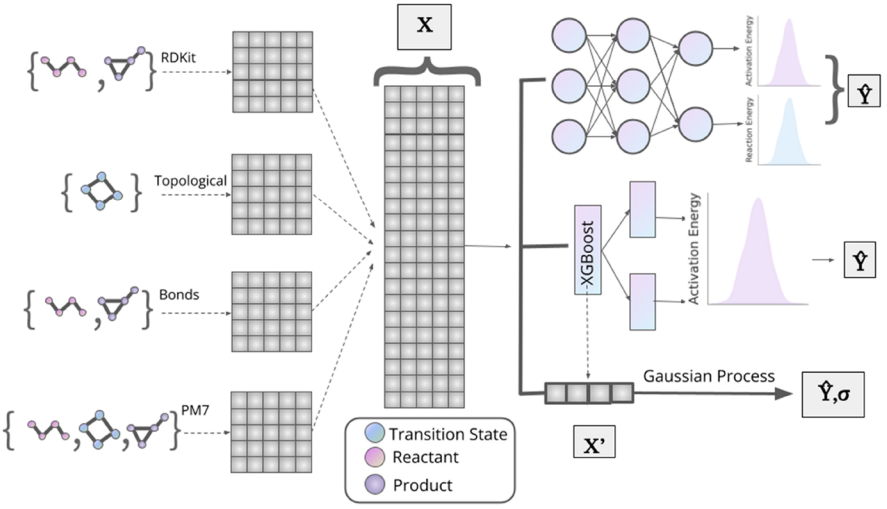


Figure 3. Workflow for the prediction of barrier heights using three machine learning models to correct SQM barrier heights. Two types of descriptors are employed: standard RDKit-based and our own custom set that comprises three subtypes. These features $\boldsymbol{X}$ are input to DNN and XGB regressors, whereas the input features for the GP are labeled by $\boldsymbol{X}'$, which is a subset of $\boldsymbol{X}$ informed by the feature importance results from the XGB model. The DNN model predicts the BH and the energy difference between the reactant and product. The XGB and GP models predict the BH only.

seven heavy atoms (C, N, or O) per molecule and consist of unimolecular reactions leading to one or more products (although most reactions are isomerizations).

2.3. Machine Learning Models. Figure 3 shows the workflow for the three ML models employed in this study to correct SQM BHs. A crucial step of the models is the calculation of a set of descriptors (or input features) that encode the most useful information present in every reaction. Our models employ two types of descriptors: (a) standard RDKit-based descriptors and (b) a custom set based on the SQM calculations and chemical intuition. Figure 3 also shows how every species in the reaction (namely, TS, reactant, or product) contributes to each set of descriptors.

The first set of descriptors X_{RDKit} is obtained from the cheminformatics library RDKit.[50]

Each descriptor of this type $X_{\mathrm{RDKit},\,i}$ is calculated as follows:

$$X_{\mathrm{RDKit},i} = X^{\mathrm{P}}_{\mathrm{RDKit},i} - X^{\mathrm{R}}_{\mathrm{RDKit},i} \tag{1}$$

where $X^{\mathrm{R}}_{\mathrm{RDKit},\,i}$ and $X^{\mathrm{P}}_{\mathrm{RDKit},\,i}$ refer to the ith RDKit descriptor of the reactant and product, respectively. If a specific descriptor remains invariant in the reaction (like the molecular weight),

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

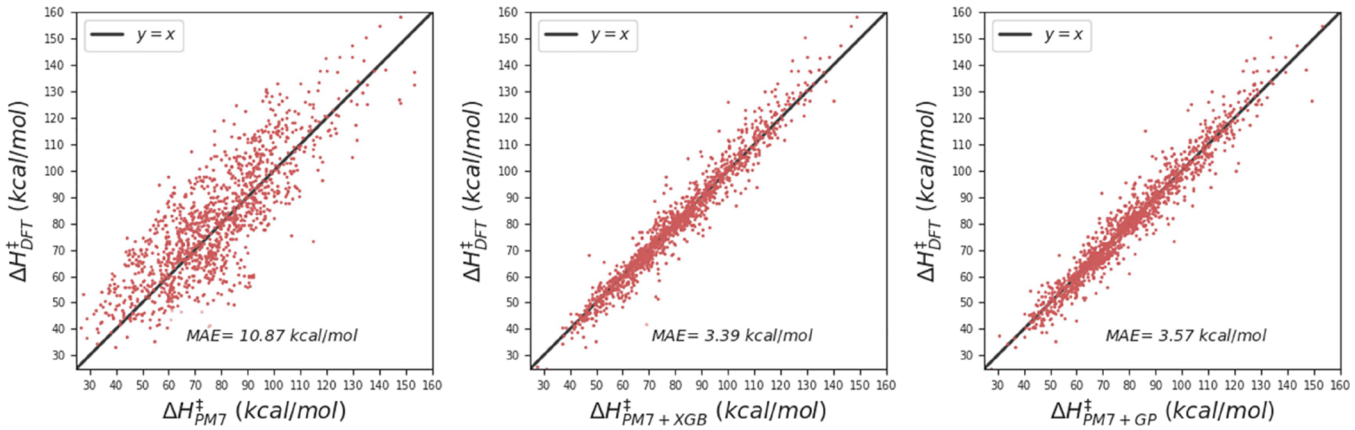


Figure 4. Barrier height predictions at DFT, PM7, PM7 + XGB, and PM7 + GP levels.

the raw value is employed instead of eq 1. The X_{RDKit} set contains 132 descriptors (see the SI for details).

Besides the above standard set of descriptors, a custom set is also employed in this work. This set is specifically tailored to extract the most relevant features of chemical reactions. It comprises information on the topology of the TS, the number of bonds that change in the reaction, and results from the PM7 calculations.

An advantage of our model is that the approximate TS structures calculated at the PM7 level of theory can be employed as input features. Specifically, the 3D geometries are converted into molecular graphs, represented in the form of adjacency matrices, using the definitions employed in AutoMeKin.[47] From the molecular graphs, some topological descriptors X_{topol} can be constructed. These include Randic's connectivity index,[51] the spectral gap (or lowest nonzero eigenvalue of the Laplacian matrix, λ_1^{TS}), the Estrada index,[52] or the Zagreb index;[53] the full list of topological descriptors can be found in the SI. The Laplacian matrix defined as $D - A$ (with D and A being the degree and adjacency matrices, respectively) is calculated from a weighted adjacency matrix to account for 3D structures of the TSs.[47] Topological descriptors provide a measure of the extent of branching or the tightness of the TS structure.

The subset X_{bonds} includes the number of broken and formed bonds of each type, i.e., all pairings of H, C, N, and O atoms. For instance, this set includes the descriptors +CO and −CH, which refer to the number of formed CO bonds and a number of broken CH bonds, respectively.

The last subset of descriptors X_{PM7} capitalizes on the PM7 calculations. This set includes the BH, a rough proxy for the rate constant $e^{-\mathrm{BH}}$, the imaginary frequency at the TS ν_1^{TS}, and differences between ZPEs of reactant, product, and TS: $\mathrm{ZPE}^{\mathrm{R}}$, $\mathrm{ZPE}^{\mathrm{P}}$, $\mathrm{ZPE}^{\mathrm{TS}}$, respectively. The subset also comprises electronic descriptors like the eigenvalues of the bond order matrix calculated at the TS, the global "hardness",[54] and Mulliken's electronegativity[55] at the TS (η^{TS} and α^{TS}, respectively), and differences between the self-polarizability[56] of reactant $\pi_{\mathrm{R}}^{\mathrm{s}}$ and product $\pi_{\mathrm{P}}^{\mathrm{s}}$. While some of these descriptors are readily available from a frequency calculation at the TS, others are obtained using the keyword SUPER in MOPAC.

Having defined the input features, a correlation matrix was built where each entry represents the Pearson coefficient r for every pair of descriptors. A threshold was established such that if the correlation coefficient exceeds this value, one of the descriptors is dropped from the input features. The threshold was optimized by cross-validation and set to $r = 0.9$.

These stacked descriptors are input to the three models depicted in Figure 3: DNN, XGB, and GP. DNN works in a multitask approach, where the output includes, besides the BH, the energy difference between the reactant and product. This approach has been shown to enhance predictions and generalization power, even if our interest is only in the BHs.[6,57] The architecture of the DNN model (number of hidden layers and number of neurons) as well as other hyperparameters were fine-tuned in a fivefold cross-validation fashion using a grid search, considering some hyperparameters to be orthogonal.

Nevertheless, since our set of descriptors consists of heterogeneous tabular data and the amount of data is limited by DL standards, we decided to use two alternative approaches, that perform better in this case, XGB and GP.[58] In particular, we chose XGB[33] implementation of gradient boosting techniques, which achieves state-of-the-art results and provides sparsity-aware algorithms particularly suited for our data set. In this case, hyperparameters were optimized by means of the Bayesian optimization library Optuna,[59] using fivefold cross-validation as well. Furthermore, considering that gradient boosting techniques that rely on decision trees as weak learners assign higher importance to descriptors that will be more relevant for other models, we find a more succinct descriptor X' containing only 49 features to feed in the GP model. The GP model, after being exposed to the training data, generates a multivariate Gaussian prior distribution that by means of Bayesian inference leads to a posterior distribution for the test set. Thus, leading to a prediction with a confidence interval based on the inherent Bayesian nature of the model.

Following common practices in the ML literature as well as considering the size of the data set, the data was split into 85% training, 5% validation, and 10% testing. The first data set partitioning was made prior to any hyperparameter optimization phase, relying on random splits. For the validation set, we relied on cross-validation.

TensorFlow,[60] XGB,[33] and MATLAB[61] were used for the DNN, XGB, and GP models, respectively. The specifics of the models can be looked up in the provided repository or notebook.

3. RESULTS AND DISCUSSION

3.1. Performance of the Machine Learning Models.

The MAEs obtained in this work using ML models DNN,

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

XGB, and GP are 3.69, 3.39, and 3.57 kcal/mol, respectively. Figure 4 shows the correlation between the reference (DFT) vs the predicted values of the BHs obtained with the XGB and GP models for the test set. In the interest of simplicity, the figures only display results for our best models (XGB and GP).

The performance of our models is comparable to Green's for the same number of training points[6] and markedly better than either PM7 or PM7-TS. The MAEs vs the number of training data points for our models are shown in Figure 5, where MAEs

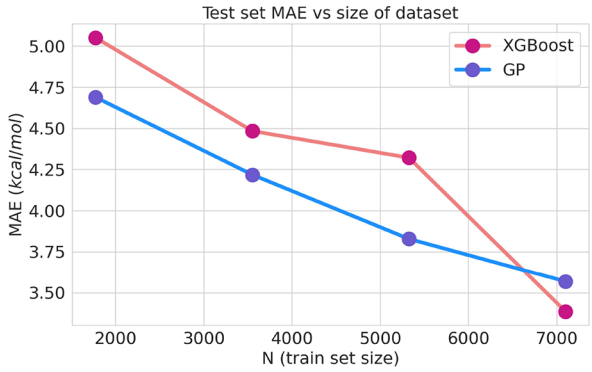


Figure 5. MAEs vs train set size for the GP and XGBoost models.

decrease as new training data points are included. In comparing both models, GP is more data efficient, as it performs better for a smaller number of data points. Nevertheless, GP seems to converge as new training data points are considered, which is not the case for XGB. The latter outperforms GP when the total data set is used becoming the preferable model for larger data sets. For 7100−7500 data points, the MAEs obtained in this work are 3.57 and 3.39 kcal/mol for GP and XGB, respectively, which can be compared with a value greater than 3.6 kcal/mol obtained by Green and co-workers.[6] It should be noted here that the procedure employed in this work cannot exploit the sort of data augmentation employed in Green's work by including reverse reactions because many of our features refer to the TSs structures, which are common to both direct and reverse reactions.

Figure 6 shows the error distribution on the target variable for XGB and GP in comparison with the results obtained with MOPAC's PM7 calculations. For the XGB and GP models, most reactions show an error smaller than 10 kcal/mol, in stark contrast with PM7-MOPAC predictions, thus showing a superior accuracy of the ML models with respect to PM7. As mentioned in the Introduction, a new data set is available,[16] with BHs calculated at the CCSD(T)-F12 level of theory. The performance of our models against a CCSD(T) reference could be worse than the one obtained here. The most recent and accurate data set[16] will be employed in a separate study to improve on our current models.

3.2. Interpretability. Models can be interpreted in terms of their feature importances, i.e., how much a certain feature contributes to the prediction. Feature importances are obtained in present work from the SHAP values,[62] which resort to game-theoretic approaches to measure the contribution to the model output by each descriptor. The underlying principle is to measure the expected change in output when using different combinations of descriptors.

Figure 7 shows a SHAP (SHapley Additive exPlanations) summary plot, which displays the magnitude and direction of a

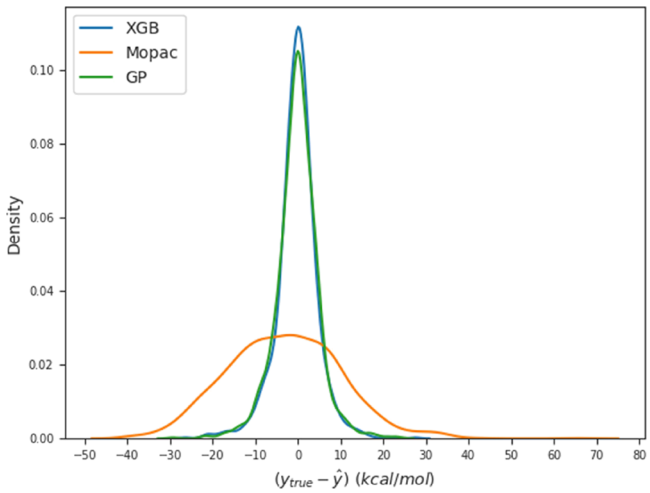


Figure 6. Error distribution on the target variable for the ML models (XGB and GP) in comparison with the one obtained directly from the MOPAC calculations.

feature's effect. Interestingly, 70% (14/20) of the most important features belong to the custom set. As expected, the features with the greatest impact on the model output are the values of the PM7 barrier height BH_{PM7} and the proxy for the rate constant $e^{-BH_{PM7}}$. The MAEs obtained with XGB without BH_{PM7} and $e^{-BH_{PM7}}$ are 4.10 and 3.47 kcal/mol, respectively. While they encode the same information, and gradient boosting (or any other model relying on decision trees as weak learners) are, in principle, invariant with respect to monotonic transformations, in this case, we included both the transformed and original descriptor. This can lead to collinearity, but XGB can handle these situations, and based on both feature importance and the increase in model performance, we decided to keep both descriptors.

Figure 7 also shows that PM7 tends to underestimate high BHs and vice versa, which is reflected by the positive impact on the model output for high BHs. Our result is in agreement with a recent ML model to predict activation energies from DFT calculations, where the DFT-computed activation energy was also the most important feature.[10]

The "hardness" η^{TS} and Mulliken's electronegativity α^{TS} calculated at the TS also rank very high on the global feature importance plot. Using Koopman's theorem,[63] they can be approximated as $\eta = (\varepsilon_{LUMO} - \varepsilon_{HOMO})/2$ and $\alpha = -(\varepsilon_{LUMO} + \varepsilon_{HOMO})/2$, where ε_{LUMO} and ε_{HOMO} are the energies of the lowest unoccupied molecular orbital (LUMO) and of the highest occupied molecular orbital (HOMO), respectively. These descriptors have been employed as an index to predict the chemical behavior and reactivity[64−70] and even to locate TSs.[71] The value of η decreases as the molecule departs from its equilibrium position, attaining a minimum at the TS. The LUMO/HOMO energies have also been employed to predict activation energies in Diels−Alder reactions.[11]

With similar impacts on the model output, the absolute value of the imaginary frequency ω_1^{TS} and the lowest nonzero eigenvalue of the Laplacian λ_1^{TS} (or spectral gap) at the TS are also among the most important descriptors according to Figure 7. Both provide a measure for the tightness of the TS structure, with the imaginary frequency also containing information on the mass of the atoms involved in the reaction coordinate.

The number of formed CH and CO bonds (+CH and +CO, respectively), the number of broken CH bonds (−CH), ZPE

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

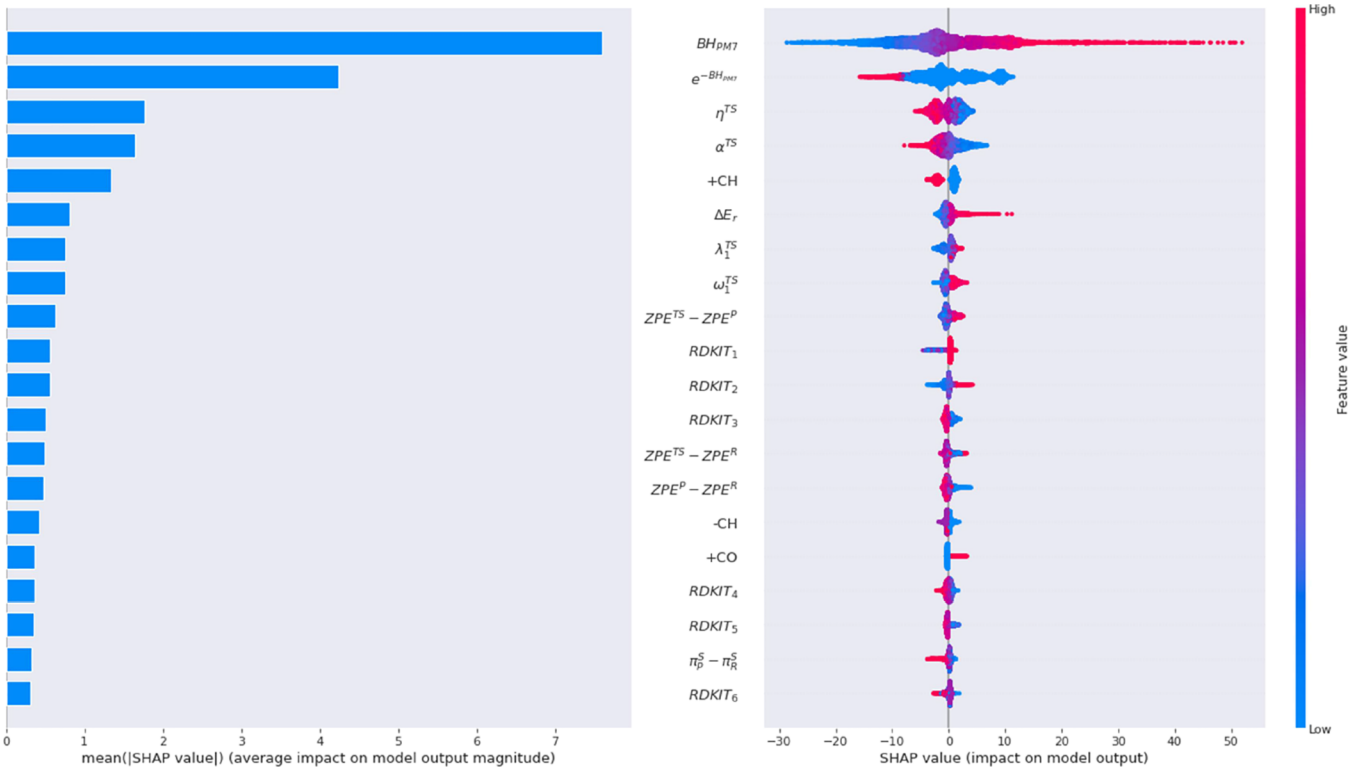


Figure 7. SHAP values for the top 20 most relevant descriptors and their impact on model output.

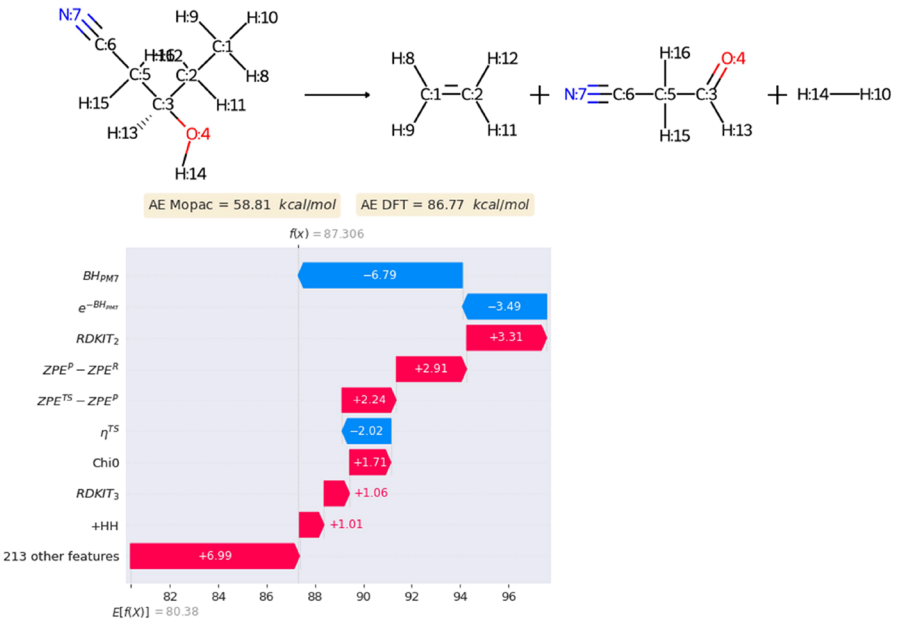


Figure 8. Model output interpretation for a single reaction.

differences among reactant, TS, and product, the PM7 reaction enthalpy (ΔE_r), or the self-polarizabilities (π_R^s and π_P^s) also contribute among the most important features. The importance of the number of formed/broken bonds of different types in the model output can be explained by the accuracy of SQM methods predicting bond energies, which strongly depends on the bond type.[72]

RDKit descriptors considered important include SMR_VSA (RDKIT 1), LabuteASA (RDKIT 2),[73] and VSA_EState2 (RDKIT 7). These descriptors grant a measure of the approximate accessible van der Waals surface area per atom. Other relevant descriptors are as follows: Balaban J (RDKIT 3), referring to the connectivity distance of the molecular graph,[74] Chi0_v[75] (RDKIT 6), which is also a topological-based descriptor, and MolLogP[76] (RDKIT 5), which refers to atom-based partition coefficients.

Figure 8 showcases how SHAP values can be used for interpretation of a single reaction. It shows the descriptors that contribute the most to shift the prediction of the model from its average (expected) prediction. Not surprisingly, BH_{PM7} and

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

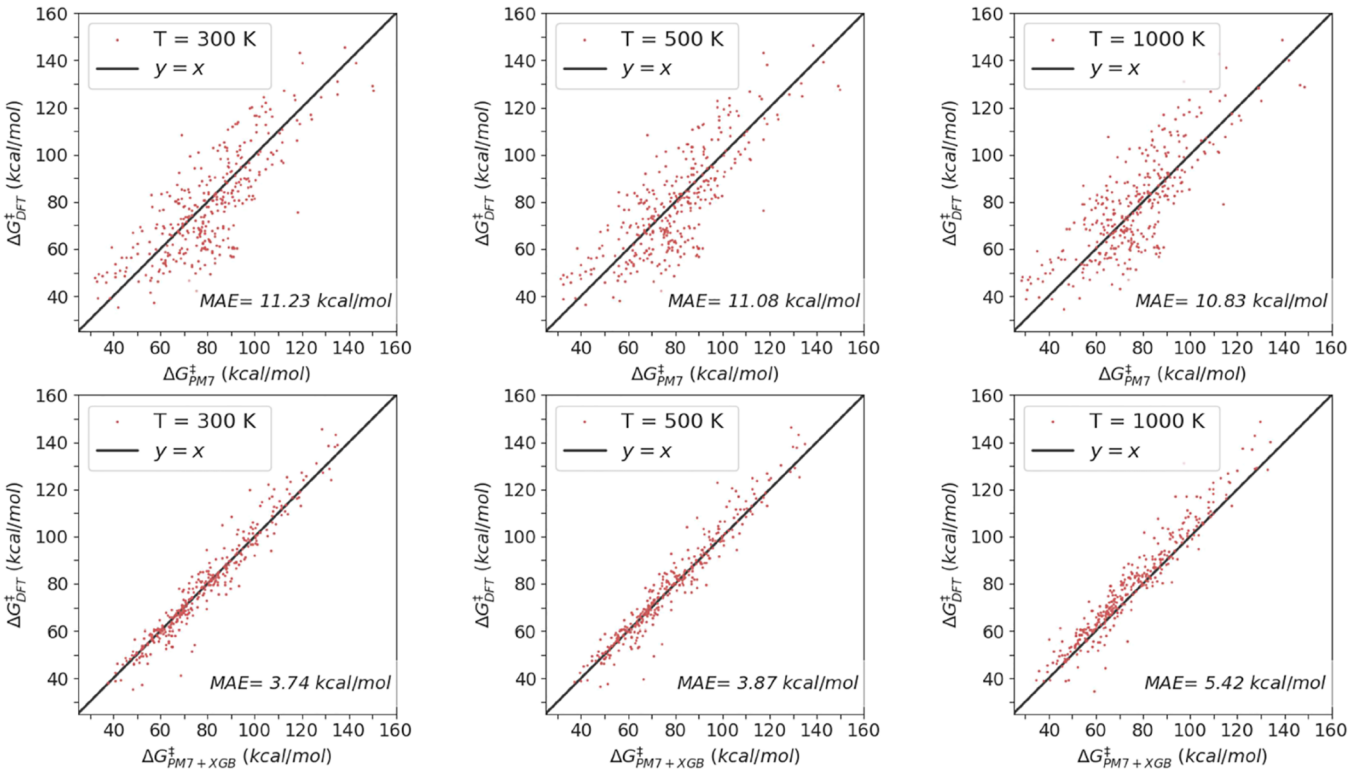


Figure 9. Correlation of the DFT, PM7, and PM7 + XGB values for the Gibbs energy difference between the reactant and transition state $\Delta G^{\ddagger}$ at T = 300, 500, and 1000 K.

$e^{-BH_{PM7}}$as well as other descriptors of Figure 7 contribute significantly also for this particular reaction. Additionally, since this reaction involves the formation of molecular hydrogen, the number of formed H−H bonds (+HH) is also an important descriptor.

3.3. Entropic Effects. In mechanistic and kinetics studies of chemical reactions the quantity of interest is the Gibbs energy of activation $\Delta G^{\ddagger}$, rather than the BH. The reason is that the former includes enthalpic and entropic corrections to the electronic and ZPE energies. Reaction channels that are not very competitive at low temperatures/energies might become predominant at high temperatures/energies because of entropic factors.[77] Therefore, the prediction of $\Delta G^{\ddagger}$ is crucial when the interest is the kinetics and the determination of the predominant mechanism.

The calculation of $\Delta G^{\ddagger}$ is straightforward when the geometries and vibrational frequencies of the reactant and TS are available. The values of $\Delta G^{\ddagger}$ have been obtained in this work at different temperatures using the thermochemistry module of AutoMeKin[47] for the reference and SQM calculations using the rigid rotor/harmonic oscillator approximation. In the absence of a scaling factor for the ωB97X-D3/def2-TZVP vibrational frequencies, the value of 0.9914 was employed; this is the recommended value for the related ωB97X-D/def2-TZVP model chemistry.[78] Furthermore, the PM7 vibrational frequencies were corrected using the recommended scaling factors.[79]

Figures 9 and S11 display the correlation between the reference (DFT), the PM7, PM7 + XGB, and PM7 + GP predictions for $\Delta G^{\ddagger}$ at three different temperatures: 300, 500, and 1000 K, respectively. At the two lowest temperatures, $\Delta G^{\ddagger}$ predictions are roughly of the same accuracy as those for the BH. However, for the highest temperature of 1000 K, the ML predictions start to deteriorate and the MAE at this temperature is 5.30 kcal/mol for the GP model. A clear improvement to the model would be to use a multitask ML model to correct the SQM vibrational frequencies. Nevertheless, the current accuracy of our models significantly improves the PM7 accuracy, and it may suffice for fast screening of reaction networks.

4. CONCLUSIONS

The main conclusions of this work are summarized below:

a) Cheap SQM calculations can be leveraged to obtain DFT-quality BHs by means of ML.

b) The MAEs of our ML models (multitask DNN, gradient-boosted trees by means of the XGB interface, and GP regression) are of the same magnitude as those obtained in previous work.

c) The analysis of the models shows that the custom-made descriptors obtained from the MOPAC calculations are, in general, considered more important than those obtained from standard cheminformatics libraries.

d) Our MOPAC-based descriptors could be widely adopted in future quantitative predictions of reaction properties.

e) Our ML models could be used for screening large reaction networks, or they could be implemented in automated reaction mechanism programs based on SQM calculations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.2c08340.

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283

Exploratory data analysis; details of the hyperparameter optimization; descriptor explanation; links to the data and code employed in this work; and free energies of activation obtained with the GP model (PDF)

AUTHOR INFORMATION

Corresponding Author

Emilio Martínez Núñez − *Department of Physical Chemistry, University of Santiago de Compostela, Santiago de Compostela 15782, Spain;* orcid.org/0000-0001-6221-4977; Email: emilio.nunez@usc.es

Authors

[#]Xabier García-Andrade − *AWS Networking Science, Dublin D04 HH21, Ireland*

Pablo García Tahoces − *Department of Electronics and Computer Science, University of Santiago de Compostela, Santiago de Compostela 15782, Spain*

Jesús Pérez-Ríos − *Department of Physics, Stony Brook University, Stony Brook, New York 11794, United States; Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794-3800, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.2c08340

Notes

The authors declare no competing financial interest.
[#]Work done prior to joining AWS.

ACKNOWLEDGMENTS

This work was partially supported by Consellería de Cultura, Educación e Ordenación Universitaria (Grupo de referencia competitiva ED431C 2021/40) and by Ministerio de Ciencia e Innovación through Grant #PID2019-107307RB-I00. J.P.-R. acknowledges the support of the Simons Foundation.

REFERENCES

(1) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. Current Status of Transition-State Theory. *J. Phys. Chem.* **1996**, *100*, 12771−12800.
(2) Bao, J. L.; Truhlar, D. G. Variational transition state theory: theoretical framework and recent developments. *Chem. Soc. Rev.* **2017**, *46*, 7548−7596.
(3) Zhang, J.; Valeev, E. F. Prediction of Reaction Barriers and Thermochemical Properties with Explicitly Correlated Coupled-Cluster Methods: A Basis Set Assessment. *J. Chem. Theor. Comput.* **2012**, *8*, 3175−3186.
(4) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(5) Choi, S.; Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chem. − Eur. J.* **2018**, *24*, 12354−12358.
(6) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992−2997.
(7) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **2020**, *7*, 137.
(8) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976−3986.
(9) Vargas, S.; Hennefarth, M. R.; Liu, Z.; Alexandrova, A. N. Machine Learning to Predict Diels−Alder Reaction Barriers from the Reactant State Electron Density. *J. Chem. Theor. Comput.* **2021**, *17*, 6203−6213.
(10) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **2021**, *12*, 1163−1175.
(11) Ravasco, J. M. J. M.; Coelho, J. A. S. Predictive Multivariate Models for Bioorthogonal Inverse-Electron Demand Diels−Alder Reactions. *J. Am. Chem. Soc.* **2020**, *142*, 4235−4241.
(12) Glavatskikh, M.; Madzhidov, T.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Malakhova, D.; Marcou, G.; Varnek, A. Predictive Models for Kinetic Parameters of Cycloaddition Reactions. *Mol. Inf.* **2019**, *38*, No. e1800077.
(13) Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis. *Mol. Inf.* **2019**, *38*, No. 1800104.
(14) Madzhidov, T. I.; Gimadiev, T. R.; Malakhova, D. A.; Nugmanov, R. I.; Baskin, I. I.; Antipin, I. S.; Varnek, A. A. Structure−reactivity relationship in Diels−Alder reactions obtained using the condensed reaction graph approach. *J. Struct. Chem.* **2017**, *58*, 650−656.
(15) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584−4601.
(16) Spiekermann, K.; Pattanaik, L.; Green, W. H. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Sci. Data* **2022**, *9*, 417.
(17) Ismail, I.; Robertson, C.; Habershon, S. Successes and challenges in using machine-learned activation energies in kinetic simulations. *J. Chem. Phys.* **2022**, *157*, No. 014109.
(18) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1−32.
(19) Martinez-Nunez, E.; Vazquez, S. A. Three-center vs. four-center HF elimination from vinyl fluoride: a direct dynamics study. *Chem. Phys. Lett.* **2000**, *332*, 583−590.
(20) Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. Direct dynamics calculations with NDDO (neglect of diatomic differential overlap) molecular orbital theory with specific reaction parameters. *J. Phys. Chem.* **1991**, *95*, 4618−4627.
(21) Martinez-Nunez, E.; Estevez, C. M.; Flores, J. R.; Vazquez, S. A. Product energy distributions for the four-center HF elimination from 1,1-difluoroethylene. A direct dynamics study. *Chem. Phys. Lett.* **2001**, *348*, 81−88.
(22) Gonzalez-Vazquez, J.; Fernandez-Ramos, A.; Martinez-Nunez, E.; Vazquez, S. A. Dissociation of difluoroethylenes. I Global potential energy surface, RRKM, and VTST calculations. *J. Phys. Chem. A* **2003**, *107*, 1389−1397.
(23) Gonzalez-Vazquez, J.; Martinez-Nunez, E.; Fernandez-Ramos, A.; Vazquez, S. A. Dissociation of difluoroethylenes. II Direct Classical Trajectory Study of the HF elimination from 1,2-difluoroethylene. *J. Phys. Chem. A* **2003**, *107*, 1398−1404.
(24) Kromann, J. C.; Christensen, A. S.; Cui, Q.; Jensen, J. H. Towards a barrier height benchmark set for biologically relevant systems. *PeerJ* **2016**, *4*, No. e1994.
(25) Iron, M. A.; Janes, T. Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange−Correlation Functionals: The MOBH35 Benchmark Database. *J. Phys. Chem. A* **2019**, *123*, 3761−3781.
(26) Pérez-Tabero, S.; Fernández, B.; Cabaleiro-Lago, E. M.; Martínez-Núñez, E.; Vázquez, S. A. New Approach for Correcting Noncovalent Interactions in Semiempirical Quantum Mechanical Methods: The Importance of Multiple-Orientation Sampling. *J. Chem. Theor. Comput.* **2021**, *17*, 5556−5567.
(27) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-

Machine Learning Approach. *J. Chem. Theor. Comput.* **2015**, *11*, 2087−2096.
(28) Plehiers, P. P.; Lengyel, I.; West, D. H.; Marin, G. B.; Stevens, C. V.; Van Geem, K. M. Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations. *Chem. Eng. J.* **2021**, *426*, No. 131304.
(29) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.
(30) Gao, T.; Li, H.; Li, W.; Li, L.; Fang, C.; Li, H.; Hu, L.; Lu, Y.; Su, Z.-M. A machine learning correction for DFT non-covalent interactions based on the S22, S66 and X40 benchmark databases. *J. Cheminform.* **2016**, *8*, 24.
(31) Wan, Z.; Wang, Q.-D.; Liang, J. Accurate prediction of standard enthalpy of formation based on semiempirical quantum chemistry methods with artificial neural network and molecular descriptors. *Int. J. Quantum Chem.* **2021**, *121*, No. e26441.
(32) Zhu, J.; Vuong, V. Q.; Sumpter, B. G.; Irle, S. Artificial neural network correction for density-functional tight-binding molecular dynamics simulations. *MRS Commun.* **2019**, *9*, 867−873.
(33) Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, California, USA, 2016; 785−794.
(34) Cui, J.; Krems, R. V. Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with Gaussian processes. *J. Phys. B At. Mol. Opt. Phys.* **2016**, *49*, No. 224001.
(35) Christianen, A.; Karman, T.; Vargas-Hernández, R. A.; Groenenboom, G. C.; Krems, R. V. Six-dimensional potential energy surface for NaK−NaK collisions: Gaussian process representation with correct asymptotic form. *J. Chem. Phys.* **2019**, *150*, No. 064106.
(36) Dai, J.; Krems, R. V. Interpolation and Extrapolation of Global Potential Energy Surfaces for Polyatomic Systems by Gaussian Processes with Composite Kernels. *J. Chem. Theor. Comput.* **2020**, *16*, 1386−1395.
(37) Sugisawa, H.; Ida, T.; Krems, R. V. Gaussian process model of 51-dimensional potential energy surface for protonated imidazole dimer. *J. Chem. Phys.* **2020**, *153*, 114101.
(38) Liu, X.; Meijer, G.; Pérez-Ríos, J. On the relationship between spectroscopic constants of diatomic molecules: a machine learning approach. *RSC Adv.* **2021**, *11*, 14552−14561.
(39) Liu, X.; Meijer, G.; Pérez-Ríos, J. A data-driven approach to determine dipole moments of diatomic molecules. *Phys. Chem. Chem. Phys.* **2020**, *22*, 24191−24200.
(40) Cretu, M. T.; Pérez-Ríos, J. Predicting second virial coefficients of organic and inorganic compounds using Gaussian process regression. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2891−2898.
(41) Stewart, J. J. P. *MOPAC2016, Stewart Computational Chemistry*: Colorado Springs, CO, USA, 2016, HTTP://OpenMOPAC.net (accessed July 01, 2022).
(42) Carpenter, B. K.; Ellison, G. B.; Nimlos, M. R.; Scheer, A. M. A Conical Intersection Influences the Ground State Rearrangement of Fulvene to Benzene. *J. Phys. Chem. A* **2022**, *126*, 1429−1447.
(43) Farrar, E. H. E.; Grayson, M. N. Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction. *Chem. Sci.* **2022**, *13*, 7594−7603.
(44) Martínez-Núñez, E. An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 222−234.
(45) Martínez-Núñez, E. An automated transition state search using classical trajectories initialized at multiple minima. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14912−14921.
(46) Varela, J. A.; Vazquez, S. A.; Martinez-Nunez, E. An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis. *Chem. Sci.* **2017**, *8*, 3843−3851.
(47) Martínez-Núñez, E.; Barnes, G. L.; Glowacki, D. R.; Kopec, S.; Peláez, D.; Rodríguez, A.; Rodríguez-Fernández, R.; Shannon, R. J.; Stewart, J. J. P.; Tahoces, P. G.; Vazquez, S. A. AutoMeKin2021: An open-source program for automated reaction discovery. *J. Comput. Chem.* **2021**, *42*, 2036−2048.
(48) Taketsugu, T.; Gordon, M. S. Dynamic reaction path analysis based on an intrinsic reaction coordinate. *J. Chem. Phys.* **1995**, *103*, 10042−10049.
(49) Vazquez, S. A.; Otero, X. L.; Martinez-Nunez, E. A Trajectory-Based Method to Explore Reaction Mechanisms. *Molecules* **2018**, *23*, 3156.
(50) Landrum, G. *RDKit: Open-source cheminformatics* (2016). https://www.rdkit.org (accessed July 01, 2022).
(51) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(52) Estrada, E. Characterization of the folding degree of proteins. *Bioinformatics* **2002**, *18*, 697−704.
(53) Gutman, I.; Trinajstić, N. Graph theory and molecular orbitals. Total φ-electron energy of alternant hydrocarbons. *Chem. Phys. Lett.* **1972**, *17*, 535−538.
(54) Parr, R. G.; Pearson, R. G. Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512−7516.
(55) Mulliken, R. S. A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *J. Chem. Phys.* **1934**, *2*, 782−793.
(56) Coulson, C. A.; Longuet-Higgins, H. C.; Bell, R. P. The electronic structure of conjugated systems II. Unsaturated hydrocarbons and their hetero-derivatives. *Proc. R. Soc. Lond. A* **1947**, *192*, 16−32.
(57) Ye, Z.; Yang, Y.; Li, X.; Cao, D.; Ouyang, D. An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol. Pharmaceutics* **2019**, *16*, 533−541.
(58) Popov, S.; Morozov, S.; Babenko, A., Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations*; Addis Ababa, Ethiopia, 2020.
(59) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: Anchorage, AK, USA, 2019; 2623−2631.
(60) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, Software available from tensorflow.org.
(61) *MATLAB, R2022a*; The MathWorks Inc.: Natick, Massachusetts, 2022.
(62) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56−67.
(63) Koopmans, T. Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica* **1934**, *1*, 104−113.
(64) Datta, D. "Hardness profile" of a reaction path. *J. Phys. Chem.* **1992**, *96*, 2409−2410.
(65) Ordon, P.; Tachibana, A. Nuclear reactivity indices within regional density functional theory. *J. Mol. Model.* **2005**, *11*, 312−316.
(66) Chandra, A. K.; Nguyen, M. T. Density Functional Approach to Regiochemistry, Activation Energy, and Hardness Profile in 1,3-Dipolar Cycloadditions. *J. Phys. Chem. A* **1998**, *102*, 6181−6185.
(67) Zhan, C.-G.; Nichols, J. A.; Dixon, D. A. Ionization Potential, Electron Affinity, Electronegativity, Hardness, and Electron Excitation

Energy: Molecular Properties from Density Functional Theory Orbital Energies. *J. Phys. Chem. A* **2003**, *107*, 4184−4195.
(68) Alfrey, T., Jr.; Price, C. C. Relative reactivities in vinyl copolymerization. *J. Polym. Sci.* **1947**, *2*, 101−106.
(69) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793−1874.
(70) De Proft, F.; Geerlings, P. Conceptual and Computational DFT in the Study of Aromaticity. *Chem. Rev.* **2001**, *101*, 1451−1464.
(71) Beg, H.; De, S. P.; Ash, S.; Misra, A. Use of polarizability and chemical hardness to locate the transition state and the potential energy curve for double proton transfer reaction: A DFT based study. *Comput. Theor. Chem.* **2012**, *984*, 13−18.
(72) Qu, X.; Latino, D. A. R. S.; Aires-de-Sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *J. Cheminform.* **2013**, *5*, 34.
(73) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(74) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(75) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In. *Rev. Comput. Chem.* 2007, 367 −422.
(76) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(77) Vazquez, S. A.; Martinez-Nunez, E. HCN elimination from vinyl cyanide: product energy partitioning, the role of hydrogen-deuterium exchange reactions and a new pathway. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6948−6955.
(78) Kesharwani, M. K.; Brauer, B.; Martin, J. M. L. Frequency and Zero-Point Vibrational Energy Scale Factors for Double-Hybrid Density Functionals (and Other Selected Methods): Can Anharmonic Force Fields Be Avoided? *J. Phys. Chem. A* **2015**, *119*, 1701−1714.
(79) Rozanska, X.; Stewart, J. J. P.; Ungerer, P.; Leblanc, B.; Freeman, C.; Saxe, P.; Wimmer, E. High-Throughput Calculations of Molecular Properties in the MedeA Environment: Accuracy of PM7 in Predicting Vibrational Frequencies, Ideal Gas Entropies, Heat Capacities, and Gibbs Free Energies of Organic Molecules. *J. Chem. Eng. Data* **2014**, *59*, 3136−3143.

https://doi.org/10.1021/acs.jpca.2c08340
J. Phys. Chem. A 2023, 127, 2274−2283