JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

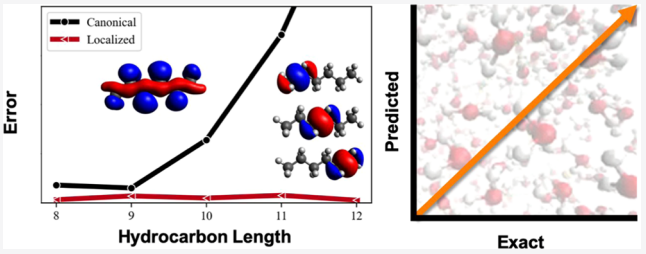
Transferable MP2-Based Machine Learning for Accurate Coupled-Cluster Energies

Jacob Townsend and Konstantinos D. Vogiatzis*

Cite This: J. Chem. Theory Comput. 2020, 16, 7453−7461 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning methods have enabled the low-cost evaluation of molecular properties such as energy at an unprecedented scale. While many of such applications have focused on molecular input based on geometry, few studies consider representations based on the underlying electronic structure. Directing the attention to the electronic structure offers a unique challenge that allows for a more detailed representation of the underlying physics and how they affect molecular properties. The target of this work is to efficiently encode a lower-cost correlated wave function derived from MP2 to predict a higher-cost coupled-cluster singles-and-doubles (CCSD) wave function based on correlation-pair energies and the contributing electron promotions (excitations) and integrals. The new molecular representation explores the short-range behavior of electron correlation and utilizes distinct models that differentiate between two-electron promotions from the same molecular orbital or from two different orbitals. We present a re-engineered set of input features that provide an intuitive description of the orbital properties involved in electron correlation. The overall models are found to be highly transferable and size extensive, necessitating very few training instances to approach the chemical accuracy of a broad spectrum of organic molecules. The efficiency and transferability of the novel representation are demonstrated on a series of linear hydrocarbons, the potential energy surface of the water dimer, and on the GDB-9 database. For the GDB-9 database, we found that data from only 140 randomly selected molecules are adequate to achieve chemical accuracy for more than 133 000 organic molecules.



1. INTRODUCTION

Machine learning (ML) methods continue to have an increasingly prominent role in many scientific fields including chemistry. Such techniques have been widely applied for evaluating potential energy surfaces (PES),[1−14] drug discovery,[15−21] and material and molecular design.[22−32] The application of ML to the prediction of quantum-mechanical energies or properties provides massive computational savings by encoding a molecular structure via a fingerprint or molecular representation,[6,33−40] where a popular strategy is to use the structural properties to predict energetics, typically at the density functional theory (DFT) level.[26,39]

While most ML applications in chemistry have been directed toward connecting physical structure to desired properties, a significant effort toward developing ML methods that directly improve the electronic structure description of molecules and materials has been explored.[41] For example, ML has been adopted to improve density functionals or bypass the Kohn−Sham equations altogether.[42−48] Lei and Medford have used convolutional neural networks to map Maxwell−Cartesian spherical harmonic kernels for functional construction.[49] Chandrasekaran et al. have introduced a representation to map atomic environments of grid points to generate electron charge densities, and therefore bypassing the Kohn−Sham equations altogether.[47] Similarly, wave function-based ML approaches have been developed, which largely focus on the reduction of their substantial computational cost. Coe has developed an ML method to learn a selective configuration interaction (CI) expansion more efficiently than the conventional perturbative or Monte Carlo-based sampling,[50] a method that has been applied to potential energy curves.[51] Additionally, Yang and co-workers have developed an artificial neural network (ANN)-based ansatz to solve complete active space (CAS)-CI wave functions.[52] Schütt and co-workers have also shown a remarkable technique to predict quantum-mechanical wave functions using a localized basis of atomic orbitals based on atoms and atom pairs, which can be used to accelerate the convergence of the self-consistent field procedure.[53]

In particular, efforts have been made to model the coupled-cluster methods, as the coupled-cluster singles-and-doubles with perturbative triples [CCSD(T)] has been denoted the

Received: September 7, 2020
Published: November 3, 2020

"gold standard" in computational chemistry for its accuracy in systems well-described by a single reference state.[54] Nudejima and co-workers performed an ML-based energy density analysis and were able to predict complete-basis-limit-extrapolated CCSD(T) energetics with Hartree−Fock (HF) densities obtained from small basis sets. McGibbon et al. predicted CCSD(T) interaction energies using Møller−Plesset perturbation theory (MP2) and symmetry adapted perturbation theory (SAPT0) as features for an artificial neural network. Margraf and Reuter developed a method to predict CCSD correlation energies utilizing a representation based on the MP2 amplitudes.[55] Miller et al. have developed a transferable molecular-orbital-based approach for the calculation of CCSD(T) energetics using Fock, Coulomb, and exchange matrices obtained from HF theory using Gaussian Process Regression (GPR).[56,57] More recently, they have demonstrated that regression clustering can be used to significantly reduce the training times of GPR by producing an ensemble of smaller models.[58] Lastly, there have been efforts to extend the approach of using HF orbitals to predict correlated energies using density matrices.[59,60]

In our previous work, we introduced the prediction of CCSD wave function parameters, the two-electron amplitudes, based on properties of the MP2 wave function.[61] In this work, we expand on the previous study to provide a new method based on the MP2 wave function and accurately predict CCSD energies, with a particular focus on the role of localization of the electron correlation and its performance on machine-learned energetics. With traditional post-HF methods, the effective correlation space grows with molecular size, which is cost prohibitive for methods that incur harsh computational scaling such as coupled-cluster.[62] However, the short-range locality of electron correlation has been targeted to reduce computational expense since its introduction to correlated methods in the 1980s,[63] in realization that the electron promotions (excitations) associated with correlation energy should not grow with molecular size. Localized correlation methods omit unimportant electronic configurations from the correlated domain or approximate them with a lower level of theory to capture correlation at a fraction of the computational expense.[64−68] However, a perhaps underutilized property of localized orbitals is the transferability of correlation contributions,[69−72] which has been utilized within machine learning of correlation,[56−58,60] but its role in the success of such methods has not been fully explored. Our goal is to introduce a novel representation to accurately predict correlation energies within chemical accuracy of the respective method (1 kcal mol^{-1}) that are chemically transferable with respect to changes in molecular structure and system size based on the underlying MP2 wave function. In Section 2, the construction of the representation is explained, and its properties are explored in Section 4.1. The transferability with respect to size and system are examined in Sections 4.2 and 4.5, respectively. The examination of the performance of the model when the underlying MP2 performs poorly is examined on the carbon dimer in Section 4.3. The potential energy surface of water dimer computed by the new methodology is discussed in Section 4.4, while basis set effects are presented in Section 4.6. Finally, the conclusions from this study are summarized in Section 5.

2. THEORY AND IMPLEMENTATION

In CCSD theory, the projected CC equations are solved to determine the converged cluster amplitudes, which subsequently determine the energy

$$E_{\text{corr}}^{\text{CCSD}} = \sum_{\substack{i>j \\ a>b}} \langle ij||ab\rangle (t_i^a t_j^b - t_i^b t_j^a + t_{ij}^{ab}) \tag{1}$$

where $\{ij\}$ and $\{ab\}$ correspond to indices of the occupied and virtual molecular orbitals, respectively, and t_i^a,t_{ij}^{ab} correspond to the one- and two-electron cluster amplitudes, respectively. An initial guess to two-electron amplitudes t_{ij}^{ab} is usually provided by the MP2 promotion (excitation) amplitudes. Therefore, to circumvent the solution of the CC equations, one needs to generalize a representation that can effectively map between the two-electron promotions of MP2 to the converged cluster amplitudes. One way to accomplish this is to rewrite eq 1 in terms of Nesbet's theorem as

$$E_{\text{corr}} = \sum_{i>j} \varepsilon_{ij} \tag{2}$$

where each ε_{ij} is the total pair-correlation energy corresponding to the occupied orbital-pair ij. Each ε_{ij} is composed of a square matrix containing elements from eq 1, or more explicitly

$$\varepsilon_{ij} = \sum_{a>b} e_{ij}^{ab} \tag{3}$$

$$e_{ij}^{ab} = \langle ij \,||\, ab\rangle (t_i^a t_j^b - t_i^b t_j^a + t_{ij}^{ab}) \tag{4}$$

where $\{ab\}$ run over the indices of the number of virtual orbitals. In principle, eq 2 would correspond to the CCSD correlation energy if the exact coefficients comprising the two electrons coefficients were known (t_i^a, t_j^b, t_{ij}^{ab}). While the basis of our approach to predict pair energies was inspired by the work of Miller et al.,[56] our approach is to focus on using the properties of the underlying MP2 wave function. By exploiting the locality and transferability of electron correlation, we aim to find a systematic, learnable connection between MP2 and more accurate correlated methods. We hypothesize that it is possible to learn the connection between the MP2 and CC wave functions and provide a consistent methodology for the accurate prediction of CC energies by circumventing the solution of the projected CC equations.

In this paragraph, we turn our attention to the development of the new representation for correlation-pair energies ε_{ij}. Based on the notion that electron correlation is local (short-range) and transferable, we introduce a Δ-ML implementation,[73] where the MP2 pair-correlation energies, $\varepsilon_{ij}^{\text{MP2}}$, and their associated promotions and integrals were used to predict the CCSD pair-correlation energies that are summed to produce the total CCSD electronic energy. Therefore, the goal is to learn the systematic connection between MP2 and CCSD correlation. With these aims in mind, the resulting representation aims to capture as much of the full MP2 correlation energy as possible while keeping a compact feature matrix for computational tractability. In our implementation, diagonal and off-diagonal ε_{ij} elements were predicted on two independent models (dual model), an approach that enhances model flexibility and increased accuracy. The rationale behind this selection is that the correlation energy term ε_{ii} of two electrons promoted from the same molecular orbital i has a higher contribution to the total CCSD energy than the off-

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461

diagonal terms ε_{ij}. Overall, the CC correlation-pair energies, ε_{ij}, were predicted from the following features

- $\varepsilon_{ij\{\mathrm{MP2}\}}$
- $\langle ii||jj\rangle$
- $e_{ij\ \{\mathrm{MP2}\}}^{ab}$ matrix
- $\langle ij||ab\rangle$ matrix
- $\langle ii||aa\rangle$ matrix
- $\langle jj||bb\rangle$ matrix
- $\langle aa||bb\rangle$ matrix
- MP2 t_2 amplitude matrix
- Missing $\varepsilon_{ij\{\mathrm{MP2}\}}$ correlation in representation (*vide infra*)

where all of the matrices have been sorted with respect to the energy contributions of $e_{ij\{\mathrm{MP2}\}}^{ab}$. The two-electron integrals were added in to indirectly provide additional information about the energy contributions $e_{ij\{\mathrm{MP2}\}}^{ab}$ and their corresponding orbitals. For the diagonal elements, ε_{ii}, the $\langle ii||jj\rangle$ were excluded (since $i = j$), and thus, their corresponding model has one less feature. In the spirit of localized correlation methods, each of these matrices was truncated to a fixed number of elements. Specifically, the most positive and most negative contributions were included in the representation to ensure that the representation does not contain a correlation exceeding the MP2 value. This truncation parameter is chosen based on the chemical application (*vide infra*) but can be considered a hyperparameter; however, a general consideration is that larger basis sets will require more elements from these matrices as the electron correlation becomes more delocalized over a greater number of virtual orbitals. This study uses 20 two-electron promotions, since larger feature spaces did not lead to increased accuracy and utilized more computational resources in model training. The final feature, the missing $\varepsilon_{ij\{\mathrm{MP2}\}}$ correlation, contains the truncated MP2 energy not contained in the 20 $e_{ij\{\mathrm{MP2}\}}^{ab}$ values. In the construction of a model based solely on electronic structure, the produced model is invariant with respect to molecular rotations, translations, and permutations.

3. COMPUTATIONAL DETAILS

All HF calculations were performed in the Psi4 program.[74] MP2 and CCSD calculations for training used a modified Psi4Numpy[75] spin-factored CCSD implementation.[76] The Psi4 program package was used for the computation of CCSD energies for evaluating the accuracy of the trained models. MP2 and CCSD calculations utilize the frozen-core approximation. Orbital localization was performed with a threshold of 10^{-12} on both the occupied and virtual orbital sub-blocks. In this study, the Boys[77] and Pipek–Mezey[78] (PM) orbital localization schemes were used. All calculations were performed with the 6-31G[79] basis set except where the basis set effects were explored, where STO-3G[80] and cc-pVDZ[81] were used as representative smaller and larger basis sets than the 6-31G, respectively.

All features were scaled using MinMaxScaler in the Sci-Kit Learn[82] module, and the XGBoostRegressor in the xgboost package was used for regression due its scalability, efficiency, and accuracy.[83] While kernel-based methods such as kernel ridge regression and Gaussian process regression were also tested and often provided superior performance for smaller training sets, their N^3 scaling with respect to training instances was impractical for larger training sets. For consistency, we have chosen to use xgboost throughout this study, which follows $N \times \log(N)$ scaling. Neural networks were also tested but did not produce greater accuracy for this application. Hyperparameters were selected via a threefold training cross validation among training examples to optimize the depth and child weight, and L2 regularization for the models. The optimized hyperparameters are located in the Supporting Information.

4. RESULTS AND DISCUSSION

4.1. Properties of the New Representation. As previously discussed, in the generation of the representation for each pair energy, ε_{ij}, the corresponding orbital space comprising e_{ij}^{ab} was truncated to a fixed number of electron promotions. Thus, the model only considers a fraction of the promotions containing the full $\varepsilon_{ij}^{\mathrm{MP2}}$. In principle, localizing the orbitals should concentrate the e_{ij}^{ab} terms such that the electron correlation can be captured with significantly fewer terms. We have chosen to examine the behavior of the representation on hexane because its size is large enough so that the benefits of localization can be assessed, while it sets the stage for the exploration of transferability toward other larger hydrocarbons (Section 4.2). To evaluate the behavior of orbital localization on the aforementioned feature space, the fraction of the full MP2 energy was plotted with respect to the number of promotions included in the representation of each pair energy, which is shown for hexane in Figure 1a. Canonical orbitals include the least amount of correlation within the representation at approximately 20% of the total MP2 correlation. Both

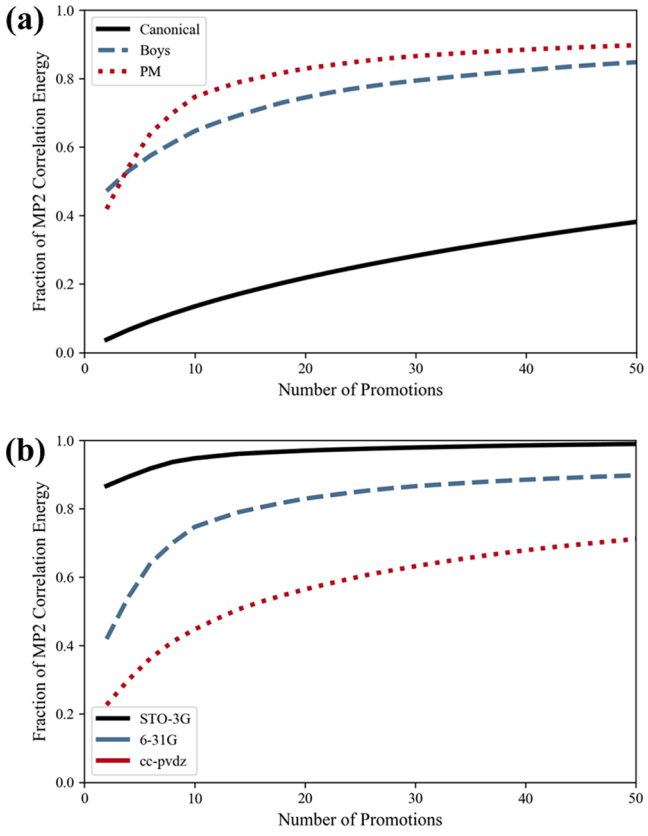


Figure 1. (a) Fraction of the total MP2 correlation energy included in the representation of hexane versus the number of promotions from $e_{ij\{\mathrm{MP2}\}}^{ab}$ (eq 4) in the 6-31G basis with canonical, Boys, and Pipek–Mezey orbitals. (b) Fraction of MP2 correlation energy included in the representation of hexane with PM-localized orbitals versus the number of promotions with the STO-3G, 6-31G, and cc-pVDZ basis.

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453–7461

Boys and PM orbitals significantly increased the total MP2 energy contained in the representation, capturing more than 70% of the total MP2 correlation energy using only 20 promotions for each ϵ_{ij}. PM orbitals outperform Boys with respect to capturing MP2 correlation in fewer promotions.

All data shown on Figure 1a were obtained with the 6-31G basis set, but as the basis set is expanded, the e_{ij}^{ab} matrix grows substantially, and the correlation contained within the representation is subsequently diminished. This is shown for hexane with PM orbitals for STO-3G, 6-31G, and cc-pVDZ, corresponding to 44, 82, and 154 basis functions, respectively (Figure 1b). STO-3G quickly captures the complete MP2 correlation, due to the small number of virtual orbitals. Subsequently, 6-31G and cc-pVDZ representations contain substantially less of the overall MP2 correlation, as the increasing virtual orbital space creates a sparser e_{ij}^{ab} matrix. The same behavior can be seen for the canonical and Boys orbitals. Therefore, this suggests larger basis sets may be increasingly difficult as much of the correlation is not contained in the representation. However, as the basis set size increases, $\epsilon_{ij}^{\mathrm{MP2}}$ is a better approximation to $\epsilon_{ij}^{\mathrm{CCSD}}$, which may offset this challenge. The results found in Figure 1 provide intuition about the representation and should be considered as a guide for a reasonable selection for the number of promotions to be included. The accuracy of the method is not significantly affected if a reasonable number is selected, as it is demonstrated in the next sections. The performance impact of increasing the basis set size is further investigated in Section 4.6.

4.2. Hydrocarbon Series and Size Extensivity. An important consideration for a useful application of ab initio-based ML applications is the extrapolation of system size, i.e., where systems can be trained on smaller systems and have transferable accuracy to larger systems. Such extrapolations are essential due to the poor scaling with respect to a system size of electronic structure theory methods, and therefore benefit from training on small systems to predict the properties of large systems and have shown promising results for the ML-based force fields and previous ML-based ab initio studies.

In order for system size extrapolation to be effective, the representation must show systematic treatment for both small and large molecules. By considering that larger molecules have more pair-correlation terms, the method should be extensive in its treatment toward molecule size. A lack of size extensivity has been highlighted as a weakness for molecular representations that feature global feature sets for a molecule.[49,59] However, given that the representation is centered on the MP2 pair-correlation energies and sorted contributions, one must determine whether the representation will have a learnable behavior as the correlation about the molecule becomes more delocalized. Therefore, the properties and performance of the new representation with increasing system size were evaluated on saturated linear chain hydrocarbons.

The fraction of the total MP2 correlation captured within the truncated promotion space for ethane to decane with 20 promotions that were used throughout this study is shown in Figure 2. The evolution of captured correlation throughout the promotion space is monitored in Figures S3−S5 in the Supporting Information. With canonical orbitals, the amount of correlation contained within the representation decreases with system size. Ethane contains approximately 66% of the total MP2 correlation energy, whereas decane contains only 13%. Using either Boys or PM localization schemes, the

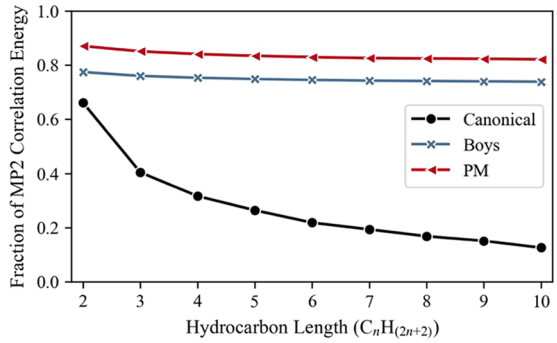


Figure 2. Fraction of MP2 correlation energy contained within the representation of hydrocarbons ethane through decane considering 20 pair promotions.

correlation contained within the representations remains relatively static. For both sets of localized orbitals, moving from pentane to decane results in an approximately 1% reduction in the total MP2 correlation within the representation. PM orbitals showed the highest efficiency, with 82% of the correlation captured for decane at 20 promotions, followed by Boys with 74%, and lastly canonical with only 13%.

Using a single conformation of propane through heptane hydrocarbons (C_nH_{2n+2}, n = 3−7), a model was made to predict the CCSD energies of the lengthier octane through octadecane (n = 8−18) for canonical, Boys-localized, and PM-localized orbitals, respectively (Figure 3). For the canonical

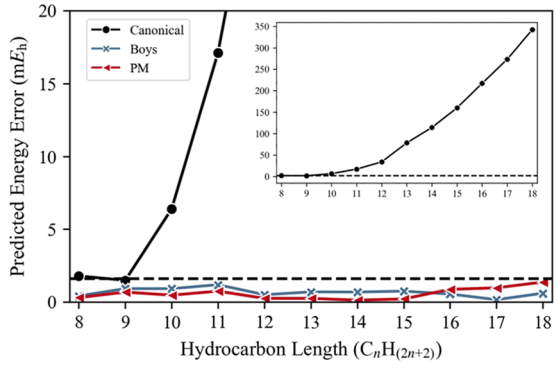


Figure 3. Predicted CCSD energy error for hydrocarbons octane through octadecane (n = 8−18) based on the model trained on single configurations of propane through heptane (n = 3−7). In the inset, the CCSD energy errors computed with canonical errors are shown in greater detail.

orbitals, the error was near chemical accuracy relative to the reference for octane and nonane. However, for the larger molecules, the errors increase significantly due to the decreasing correlation contained in the representation. For both Boys and PM-localized orbitals, the models achieved accuracy within the 1 kcal mol^{-1} threshold for the entire hydrocarbon series, with mean absolute errors (MAE) of 0.67 and 0.57 mE_h, respectively. This result suggests that the PM orbital localization scheme is the best choice when the aim is molecular size extrapolation.

4.3. C_2 Dissociation. The behavior of the new model on cases with challenging electronic structure is discussed in this section. For that purpose, we investigate the dissociation of the carbon dimer (C_2), which remains a challenge for many theoretical methods as it contains many near-degenerate

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461

electronic configurations even close to the equilibrium geometry.[84] C_2 presents an unusual ground-state bonding configuration since it contains two π bonds with little or no σ bonding.[85] While C_2 has been extensively studied using full CI and a plethora of multireference methods,[84−90] it has been selected herein because MP2 fails to describe its dissociation while CCSD performs reasonably well.[87] By selecting a case where MP2 is a poor description for CCSD, it evaluates whether the model is robust even when the initial representation is a poor approximation to the target CCSD. The C_2 dissociation was modeled by uniform sampling 20 points from 0.8 to 3.7 Å for training and to reproduce the full dissociation, which is shown in Figure 4. CCSD and MP2 are

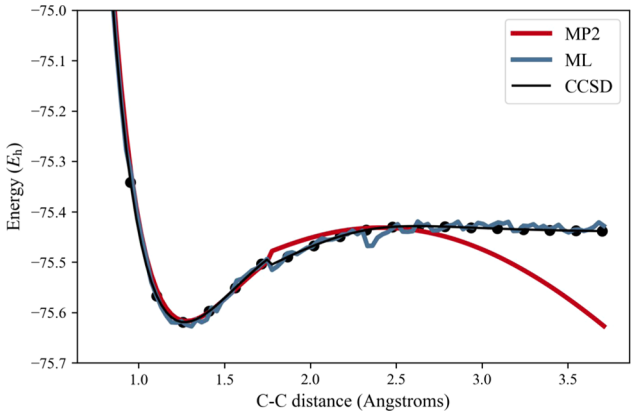


Figure 4. C_2 dimer dissociation using MP2 (red), CCSD (black), and a Boys-localized ML model (blue) utilizing the 6-31G basis set. The black points represent training points for the ML model.

in relative agreement until approximately 2.5 Å, when MP2 begins to fail considerably and the correlation energy becomes unphysical. Despite this failure, the ML model is still able to qualitatively regenerate the CCSD dissociation. From the region of 3.0−3.7 Å, MP2 shows an absolute average deviation from CCSD of 108 mE_h, whereas the ML model was 6.7 mE_h. While the results are not quantitatively accurate, they demonstrate that the model is able to qualitatively reproduce the CCSD result on a challenging electronic structure when the underlying MP2 representation is a poor description of CCSD.

4.4. Potential Energy Surface of Water 2510. As a second example, we have examined the performance of the method in predicting energies on a complex data set of interacting systems to evaluate a full potential energy surface (PES). To demonstrate this capability, the Water 2510 data set[91,92] was utilized, which contains 2510 water dimers at a vast array of distances and conformations. Although the database focuses only on water dimers, it contains a diverse set of noncovalent interactions, including highly repulsive short-range electrostatic interactions (up to +150 kcal mol^{-1} interaction energy).[93] More specifically, this application is particularly challenging due to the broad distribution of interaction energies, which is shown in Figure S6. For this application, 10% of the dimers were used for training and 90% for testing. The results for Boys and PM-localized orbitals are found in Figure 5, which shows a mean absolute error (MAE) of 9.7 × 10^{-5} E_h and 1.1 × 10^{-4} E_h, respectively, whereas canonical orbitals had a substantial MAE of 7.67 × 10^{-4} E_h (Figure S7). The mean absolute percentage error (MAPE) for

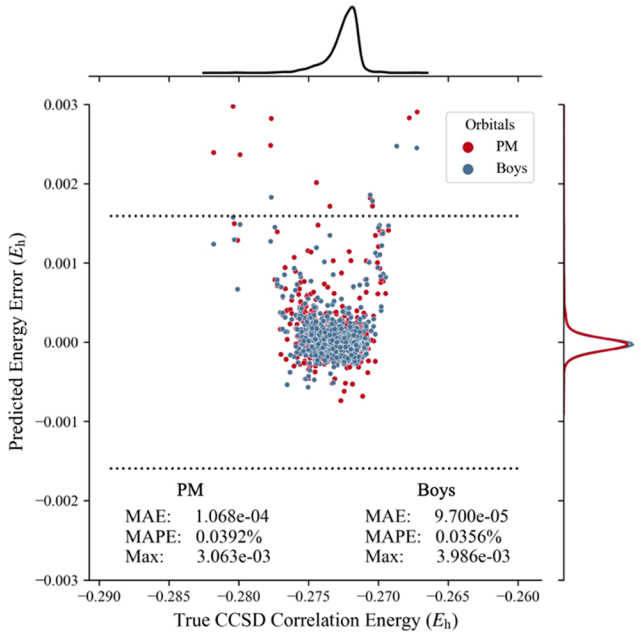


Figure 5. Predicted CCSD energy error versus the true correlation energy with their respective distributions for Boys and PM-localized orbitals using the 6-31G basis set. The dotted black lines represent the 1 kcal mol^{-1} limit.

Boys and PM is 0.0392 and 0.0457%, respectively. For the Boys-localized orbitals, of the 2259 molecular dimers in the testing set, only seven had an error exceeding 1.0 kcal mol^{-1}. Figure 5 shows the systems predicted most poorly were those that are at the edges of the distribution of correlation energy. This error can be ameliorated with a more sophisticated strategy such as active learning.[39,94,95]

4.5. Examination of Transferability. With the previous models having demonstrated promising applicability with respect to accurate CCSD energies for PES mapping that is scalable to larger systems, we considered the transferability between different chemical systems. For this, we utilize the GDB-9 database,[96] which contains 133 885 small organic molecules optimized at the DFT level. The performance of CCSD energy predictions was evaluated using 0.01, 0.1, 1, and 10% of the database, respectively, to test the remaining approximately 121 000 molecules (90%) and is shown for each orbital localization scheme in Figure 6.

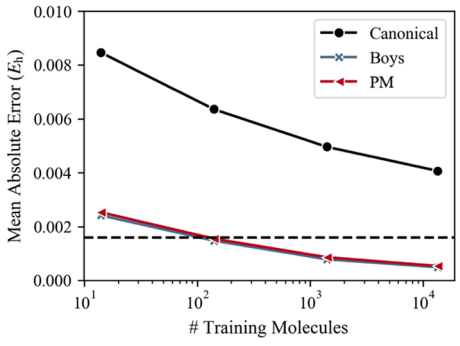


Figure 6. Learning curves for the models based on canonical, Boys-localized, and Pipek−Mezey-localized orbitals in the evaluation of CCSD energetics on the GDB-9 database using the 6-31G basis set. The horizontal black dashed line represents the 1 kcal mol^{-1} limit.

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461

In agreement with the previous applications, canonical orbital model accuracy is poorer than those trained with localized orbitals. In the GDB-9 screening, the Boys localization has lower error across all training set sizes. Both Boys and PM model mean absolute errors are below 1 kcal mol^{-1} using only 140 molecules for training. The Boys localization model had MAEs of 0.78 and 0.49 mE_h for 1 and 10% of the database as training, respectively, while the MAEs from PM are 0.86 and 0.53 mE_h, respectively. This is a substantial improvement over MP2, which averaged a 56.8 mE_h difference from CCSD, as it is shown in Figure S8. The predicted and actual correlation energies for the model using Boys-localized orbitals of 10% of the GDB-9 (i.e., 13 388 molecules used as input data) for training are shown in Figure 7. The correlation between predicted and exact data is perfectly linear, with a slope of 1.0012 and a Pearson's correlation coefficient (R^2) of 0.9998.

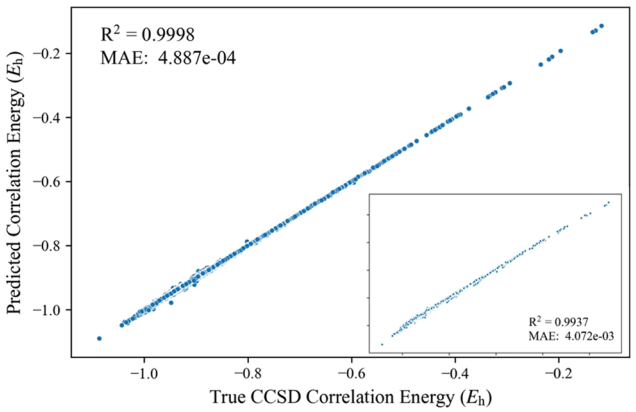


Figure 7. Predicted CCSD correlation energy versus the true CCSD correlation energy for 120 493 molecules in the GDB-9 database using the Boys localization method (data from the 10% of the GDB-9 database were used for training). Results obtained from the canonical model are shown in the inset for comparison. The R^2 and MAE (in E_h) values are provided for both models using the 6-31G basis set.

A PM7-based geometry-featurized method to compute coupled-cluster energetics needed approximately 20k training instances from the GDB-9 data set to reduce the MAE to 1 kcal mol^{-1}, which is over 100 times the number of training instances of our method to achieve such accuracy (0.1% or 140 molecules for our model).[97] However, using DFT energies and a geometry-based representation, sub 1 kcal mol^{-1} error was achieved in just 500 training points. An electronic structure-theory-based representation using HF features, named the MOB-ML model,[57] predicts coupled-cluster energies on a 350 K thermally sampled QM7b data set with mean absolute errors of approximately 1.25 kcal mol^{-1} with 140 training structures, which is impressive given the small training set size and small basis set (cc-pVDZ), as well as the addition of perturbative triples $E_{(T)}$ correction. However, extrapolating to the larger molecules of the GDB-13 data set, the errors were increased to 3.88 mE_h (2.4 kcal mol^{-1}).

4.6. Basis Set Effects. Lastly, the model was examined to determine the performance deprecation in the expansion of the basis set. As shown in Figure 1b, the electron correlation becomes more delocalized over a larger number of orbital promotions as the virtual space grows, leading to a sparser representation. Therefore, less of the total MP2 energy is contained within the truncated promotion space, which comprises our models. To examine the implications of expanding the basis set, we have generated models for predictions on a subset of the GDB-9 data set. Each model was trained on the same set of molecules that spanned 1% of the full data set and used to predict the energies of 6100 randomly selected molecules using models trained and tested on the STO-3G, 6-31G, and cc-pVDZ basis sets, respectively, and are shown in Figure 8. For each subset of orbital types,

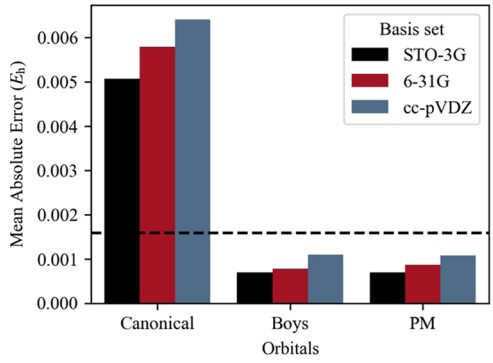


Figure 8. Mean absolute errors on a subset of the GDB-9 data set utilizing different basis set sizes for canonical and the Boys- and PM-localized orbitals. The black horizontal dotted line represents an error of 1.0 kcal mol^{-1}.

there is a systematic decrease in model performance upon expansion of the basis set. For the Boys-localized models in increasing basis set size from STO-3G to 6-31G the MAE goes from 0.69 to 0.77 mE_h. Further increase of the basis set size to the substantially larger cc-pVDZ increases the MAE to 1.09 mE_h, which is still below the 1 kcal mol^{-1} threshold. Similar relative increases in error are shown for the canonical and PM-localized orbitals. While the performance remains reasonable for these basis sets, it is clear that further model development may be necessary to utilize routinely applied basis sets for coupled-cluster methods. These improvements may be found by further investigating the locality of the electron correlation or providing a more global representation of each orbital through the use of persistent images[38,98,99] and is a target for future studies.

5. CONCLUSIONS

This study introduced a new representation of the MP2 wave function for ML that allowed for accurate prediction of CCSD energies using localized orbitals. The generated representations require no additional calculations, and since they are based on the electronic structure, they are invariant with respect to translation, permutation, and rotation. A re-engineered set of input features that provide an intuitive description of the orbital properties involved in electron correlation were introduced. The pair energy representation contains a subset of the largest contributions to the MP2 energies and properties of those respective promotions. A combination of this approach together with orbital localization and the dual training model showed that the representation essentially converges its contained correlation with respect to size, and therefore was able to predict CCSD energies within chemical accuracy of larger linear chain hydrocarbons based on smaller ones. The new method also shows promising results on evaluating potential energy surfaces and was able to map a complex data set containing many unique conformations of

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461

water dimers, yielding a MAE of 0.06 kcal mol^{-1}. The chemical transferability was also examined through the evaluation of the GDB-9 database. The localized models were able to accurately provide CCSD correlation energies with mean absolute errors below 1.0 kcal mol^{-1} using as few as 140 molecules for training. The Boys localization model using approximately 10% of the GDB-9 database for training provided a mean absolute error of 0.31 kcal mol^{-1}. We believe that this work is a step toward building a general-purpose and transferable model to accurately predict coupled-cluster energies across the periodic table. While methods that circumvent the MP2 step have a significantly reduced cost at the time of testing, ultimately, timings will also depend on the amount and size of molecules included for training. Finally, we envision that the new representation, together with recent advances in localized equation-of-motion coupled cluster,[100] can be used for accelerating the computation of excited states with machine learning.[101−103] Future works will expand the applicability with the inclusion of the perturbative triples term (T) and an unrestricted wave function implementation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00927.

Additional details about learner choice, hyperparameters, and accuracy of the model (PDF)

Townsend20_CC_results (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Konstantinos D. Vogiatzis − *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, United States;* orcid.org/0000-0002-7439-3850; Email: kvogiatz@utk.edu

Author

Jacob Townsend − *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00927

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the National Science Foundation (CHE-1800237) and the University of Tennessee (start-up grant) for financial support of this work and the Advanced Computer Facility (ACF) of the University of Tennessee for computational resources.

■ REFERENCES

(1) Lorenz, S.; Scheffler, M.; Gross, A. Descriptions of Surface Chemical Reactions Using a Neural Network Representation of the Potential-Energy Surface. *Phys. Rev. B* **2006**, *73*, 1−13.

(2) Morawietz, T.; Behler, J. A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van Der Waals Corrections. *J. Phys. Chem. A* **2013**, *117*, 7356−7366.

(3) Pukrittayakamee, A.; Malshe, M.; Hagan, M.; Raff, L. M.; Narulkar, R.; Bukkapatnum, S.; Komanduri, R. Simultaneous Fitting of a Potential-Energy Surface and Its Corresponding Force Fields Using Feedforward Neural Networks. *J. Chem. Phys.* **2009**, *130*, 134101.

(4) Jose, K. V. J.; Artrith, N.; Behler, J. Construction of High-Dimensional Neural Network Potentials Using Environment-Dependent Atom Pairs. *J. Chem. Phys.* **2012**, *136*, 194111.

(5) Handley, C. M.; Popelier, P. L. A. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.

(6) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(7) Dawes, R.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: A Strategy for Efficient Automatic Data Point Placement in High Dimensions. *J. Chem. Phys.* **2008**, *128*, 084107.

(8) Witkoskie, J. B.; Doren, D. J. Neural Network Models of Potential Energy Surfaces: Prototypical Examples. *J. Chem. Theory Comput.* **2005**, *1*, 14−23.

(9) Behler, J.; Lorenz, S.; Reuter, K. Representing Molecule-Surface Interactions with Symmetry-Adapted Neural Networks. *J. Chem. Phys.* **2007**, *127*, 014705.

(10) Guo, Y.; Tokmakov, I.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Improving Efficiency via Local Approximants. *J. Chem. Phys.* **2007**, *127*, 184108.

(11) Manzhos, S.; Carrington, T. Using Neural Networks to Represent Potential Surfaces as Sums of Products. *J. Chem. Phys.* **2006**, *125*, 194105.

(12) Botu, V.; Ramprasad, R. Learning Scheme to Predict Atomic Forces and Accelerate Materials Simulations. *Phys. Rev. B* **2015**, *92*, 094306.

(13) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(14) Lorenz, S.; Groß, A.; Scheffler, M. Representing High-Dimensional Potential-Energy Surfaces for Reactions at Surfaces by Neural Networks. *Chem. Phys. Lett.* **2004**, *395*, 210−215.

(15) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(16) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X. Q. S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* **2018**, *20*, 1−10.

(17) Thomford, N. E.; Senthebane, D. A.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural Products for Drug Discovery in the 21st Century: Innovations for Novel Drug Discovery. *Int. J. Mol. Sci.* **2018**, *19*, 1578.

(18) Lavecchia, A. Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects. *Drug Discovery Today* **2019**, *24*, 2017−2032.

(19) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science.* **2019**, *365*, eaax1566.

(20) Gawehn, E.; Hiss, J. A.; Brown, J. B.; Schneider, G. Advancing Drug Discovery via GPU-Based Deep Learning. *Expert Opin. Drug Discovery* **2018**, *13*, 579−582.

(21) Schneider, G. Automating Drug Discovery. *Nat. Rev. Drug Discovery* **2018**, *17*, 97−113.

(22) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.

(23) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442−452.

(24) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal

Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513−524.
(25) Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137−5152.
(26) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069−7077.
(27) Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64*, 2311−2323.
(28) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939−8954.
(29) Janet, J. P.; Gani, T. Z. H.; Steeves, A. H.; Ioannidis, E. I.; Kulik, H. J. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* **2017**, *56*, 4898−4910.
(30) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595−6612.
(31) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825−24836.
(32) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354−2377.
(33) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 1−19.
(34) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of Compounds for Machine-Learning Prediction of Physical Properties. *Phys. Rev. B* **2017**, *95*, 144110.
(35) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(36) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.
(37) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B* **2014**, *89*, 205118.
(38) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem., Int. Ed.* **2018**, *57*, 4164−4169.
(39) Townsend, J.; Micucci, C. P.; Hymel, J.; Maroulas, V.; Vogiatzis, K. Representation of Molecular Structures with Persistent Homology for Machine Learning Applications in Chemistry. *Nat. Commun.* **2020**, *11*, 3230.
(40) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148*, 241718.
(41) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336−2347.
(42) Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to Machine Learning: Recent Approaches to Materials Science−a Review. *J. Phys. Mater.* **2019**, *2*, 032001.
(43) Seino, J.; Kageyama, R.; Fujinami, M.; Ikabata, Y.; Nakai, H. Semi-Local Machine-Learned Kinetic Energy Density Functional with Third-Order Gradients of Electron Density. *J. Chem. Phys.* **2018**, *149*, 241705.
(44) Li, L.; Baker, T. E.; White, S. R.; Burke, K. Pure Density Functional for Strong Correlation and the Thermodynamic Limit from Machine Learning. *Phys. Rev. B* **2016**, *94*, 245129.
(45) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.
(46) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K. R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.
(47) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the Electronic Structure Problem with Machine Learning. *npj Comput. Mater.* **2019**, *5*, 22.
(48) Lei, X.; Medford, A. J. Design and Analysis of Machine Learning Exchange-Correlation Functionals via Rotationally Invariant Convolutional Descriptors. *Phys. Rev. Mater.* **2019**, *3*, 063801.
(49) Jung, H.; Stocker, S.; Kunkel, C.; Oberhofer, H.; Han, B.; Reuter, K.; Margraf, J. T. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem* **2020**, *2*, e1900052.
(50) Coe, J. P. Machine Learning Configuration Interaction. *J. Chem. Theory Comput.* **2018**, *14*, 5739−5749.
(51) Coe, J. P. Machine Learning Configuration Interaction for Ab Initio Potential Energy Curves. *J. Chem. Theory Comput.* **2019**, *15*, 6179−6189.
(52) Yang, P. J.; Sugiyama, M.; Tsuda, K.; Yanai, T. Artificial Neural Networks Applied as Molecular Wave Function Solvers. *J. Chem. Theory Comput.* **2020**, *16*, 3513−3529.
(53) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.
(54) Helgaker, T.; Coriani, S.; Jørgensen, P.; Kristensen, K.; Olsen, J.; Ruud, K. Recent Advances in Wave Function-Based Methods of Molecular-Property Calculations. *Chem. Rev.* **2012**, *112*, 543−631.
(55) Margraf, J. T.; Reuter, K. Making the Coupled Cluster Correlation Energy Machine-Learnable. *J. Phys. Chem. A* **2018**, *122*, 6343−6348.
(56) Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772−4779.
(57) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A Universal Density Matrix Functional from Molecular Orbital-Based Machine Learning: Transferability across Organic Molecules. *J. Chem. Phys.* **2019**, *150*, 131103.
(58) Cheng, L.; Kovachki, N. B.; Welborn, M.; Miller, T. F. Regression Clustering for Improved Accuracy and Training Costs with Molecular-Orbital-Based Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6668−6677.
(59) Peyton, B. G.; Briggs, C.; D'Cunha, R.; Margraf, J. T.; Crawford, T. D. Machine-Learning Coupled Cluster Properties through a Density Tensor Representation. *J. Phys. Chem. A* **2020**, *124*, 4861−4871.
(60) Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground State Energy Functional with Hartree−Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155−7165.
(61) Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129−4135.
(62) Townsend, J.; Kirkland, J. K.; Vogiatzis, K. D. In *Post-Hartree-Fock Methods: Configuration Interaction, Many-Body Perturbation Theory, Coupled-Cluster Theory*. S.M., Blinder; J.E., House, Ed.; Elsevier, 2019, pp 63−117.
(63) Pulay, P. Localizability of Dynamic Electron Correlation. *Chem. Phys. Lett.* **1983**, *100*, 151−154.
(64) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.
(65) Sparta, M.; Neese, F. Chemical Applications Carried out by Local Pair Natural Orbital Based Coupled-Cluster Methods. *Chem. Soc. Rev.* **2014**, *43*, 5032−5041.
(66) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.
(67) Nagy, P. R.; Kállay, M. Approaching the Basis Set Limit of CCSD(T) Energies for Large Molecules with Local Natural Orbital

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461

Coupled-Cluster Methods. *J. Chem. Theory Comput.* **2019**, *15*, 5275−5298.
(68) Nagy, P. R.; Samu, G.; Kállay, M. Optimization of the Linear-Scaling Local Natural Orbital CCSD(T) Method: Improved Algorithm and Benchmark Applications. *J. Chem. Theory Comput.* **2018**, *14*, 4193−4215.
(69) Silva, A. F.; Vincent, M. A.; McDonagh, J. L.; Popelier, P. L. A. The Transferability of Topologically Partitioned Electron Correlation Energies in Water Clusters. *ChemPhysChem* **2017**, *18*, 3360−3368.
(70) Flocke, N.; Bartlett, R. J. Localized Correlation Treatment Using Natural Bond Orbitals. *Chem. Phys. Lett.* **2003**, *367*, 80−89.
(71) Saebo, S.; Pulay, P. Local Treatment of Electron Correlation. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213−236.
(72) Carpenter, J. E.; Weinhold, F. Transferability of Natural Bond Orbitals. *J. Am. Chem. Soc.* **1988**, *110*, 368−372.
(73) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(74) Smith, D. G. A.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A.; et al. Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry. *J. Chem. Phys.* **2020**, *152*, 184108.
(75) Smith, D. G. A.; Burns, L. A.; Sirianni, D. A.; Nascimento, D. R.; Kumar, A.; James, A. M.; Schriber, J. B.; Zhang, T.; Zhang, B.; Abbott, A. S.; et al. PSi4NUmPy: An Interactive Quantum Chemistry Programming Environment for Reference Implementations and Rapid Development. *J. Chem. Theory Comput.* **2018**, *14*, 3504−3511.
(76) Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J. A Direct Product Decomposition Approach for Symmetry Exploitation in Many-Body Methods. I. Energy Calculations. *J. Chem. Phys.* **1991**, *94*, 4334−4345.
(77) Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300−302.
(78) Pipek, J.; Mezey, P. G. A Fast Intrinsic Localization Procedure Applicable for Ab Initio and Semiempirical Linear Combination of Atomic Orbital Wave Functions. *J. Chem. Phys.* **1989**, *90*, 4916−4926.
(79) Hehre, W. J.; Ditchfield, K.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.
(80) Hehre, W. J.; Ditchfield, R.; Stewart, R. F.; Pople, J. A. Self-Consistent Molecular Orbital Methods. IV. Use of Gaussian Expansions of Slater-Type Orbitals. Extension to Second-Row Molecules. *J. Chem. Phys.* **1970**, *52*, 2769−2773.
(81) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(82) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Varoquaux, G.; Gramfort, A.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(83) Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; pp 785−794.
(84) Sherrill, C. D.; Piecuch, P. The X $\sum$g+1, B Δg1, and B′ $\sum$g+1 States of C2: A Comparison of Renormalized Coupled-Cluster and Multireference Methods with Full Configuration Interaction Benchmarks. *J. Chem. Phys.* **2005**, *122*, 124104.
(85) Zou, W.; Cremer, D. C2 in a Box: Determining Its Intrinsic Bond Strength for the X1$\sum$g + Ground State. *Chem. - Eur. J.* **2016**, *22*, 4087−4099.
(86) Sharma, S. A General Non-Abelian Density Matrix Renormalization Group Algorithm with Application to the C2 Dimer. *J. Chem. Phys.* **2015**, *142*, 024107.
(87) Abrams, M. L.; Sherrill, C. D. Full Configuration Interaction Potential Energy Curves for the X 1$\sum$ G+,B 1Δ g, and B′ 1$\sum$ G+ States of C 2: A Challenge for Approximate Methods. *J. Chem. Phys.* **2004**, *121*, 9211−9219.
(88) Booth, G. H.; Cleland, D.; Thom, A. J. W.; Alavi, A. Breaking the Carbon Dimer: The Challenges of Multiple Bond Dissociation with Full Configuration Interaction Quantum Monte Carlo Methods. *J. Chem. Phys.* **2011**, *135*, 084104.
(89) Varandas, A. J. C. Extrapolation to the Complete-Basis-Set Limit and the Implications of Avoided Crossings: The X1 $\sum$g+, B 1Δg, and B′1 σ G+ States of C2. *J. Chem. Phys.* **2008**, *129*, No. 234103.
(90) Jiang, W.; Wilson, A. K. Multireference Composite Approaches for the Accurate Study of Ground and Excited Electronic States: C2, N2, and O2. *J. Chem. Phys.* **2011**, *134*, 034101.
(91) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Polarizable Interaction Potential for Water from Coupled Cluster Calculations. I. Analysis of Dimer Potential Energy Surface. *J. Chem. Phys.* **2008**, *128*, 094313.
(92) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van Der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315*, 1249−1252.
(93) Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7*, 2197−2203.
(94) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(95) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(96) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.
(97) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.
(98) Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushtanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L. Persistence Images: A Stable Vector Representation of Persistent Homology. *J. Mach. Learn. Res.* **2017**, *18*, 1−35.
(99) Maroulas, V.; Nasrin, F.; Oballe, C. A Bayesian Framework for Persistent Homology. *SIAM J. Math. Data Sci.* **2020**, *2*, 48−74.
(100) Izsák, R. Single-Reference Coupled Cluster Methods for Computing Excitation Energies in Large Molecules: The Efficiency and Accuracy of Approximations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1445.
(101) Kiyohara, S.; Tsubaki, M.; Mizoguchi, T. Learning Excited States from Ground States by Using an Artificial Neural Network. *npj Comput. Mater.* **2020**, *6*, 68.
(102) Chen, W. K.; Liu, X. Y.; Fang, W. H.; Dral, P. O.; Cui, G. Deep Learning for Nonadiabatic Excited-State Dynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 6702−6708.
(103) Dral, P. O.; Barbatti, M.; Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 5660−5663.

https://dx.doi.org/10.1021/acs.jctc.0c00927
J. Chem. Theory Comput. 2020, 16, 7453−7461