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

pubs.acs.org/JPCA Article

# Exploration of the Two-Electron Excitation Space with Data-Driven Coupled Cluster

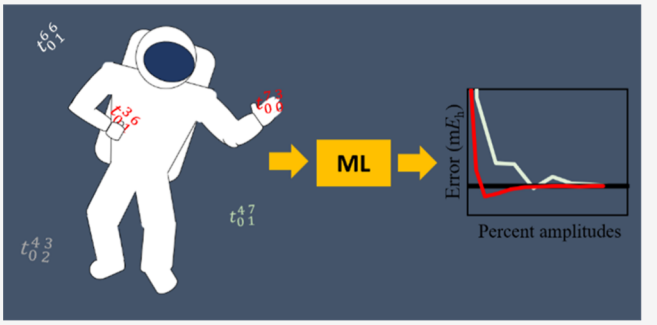
*Published as part of The Journal of Physical Chemistry A virtual special issue "Machine Learning in Physical Chemistry Volume 2".*

P. D. Varuna S. Pathirage, Justin T. Phillips, and Konstantinos D. Vogiatzis*

Cite This: J. Phys. Chem. A 2024, 128, 1938–1947 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Computational cost limits the applicability of post-Hartree–Fock methods such as coupled-cluster on larger molecular systems. The data-driven coupled-cluster (DDCC) method applies machine learning to predict the coupled-cluster two-electron amplitudes ($t_2$) using data from second-order perturbation theory (MP2). One major limitation of the DDCC models is the size of training sets that increases exponentially with the system size. Effective sampling of the amplitude space can resolve this issue. Five different amplitude selection techniques that reduce the amount of data used for training were evaluated, an approach that also prevents model overfitting and increases the portability of data-driven coupled-cluster singles and doubles to more complex molecules or larger basis sets. In combination with a localized orbital formalism to predict the CCSD $t_2$ amplitudes, we have achieved a 10-fold error reduction for energy calculations.



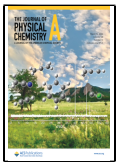

## 1. INTRODUCTION

Electron correlation is the heart of molecular electronic structure theory.[1] Hartree–Fock (HF) theory includes Fermi correlation in an exact sense and electron repulsion in an average, mean-field manner since it fails to describe the instantaneous Coulombic repulsion between electron pairs. Post-HF methods such as configuration interaction, many-body perturbation theory, and coupled-cluster (CC) theory introduce the missing electron correlation by considering a systematic expansion of the *N*-electron basis formed from all possible electronic configurations within a given orbital basis.[2] Among post-HF methods, coupled-cluster singles and doubles with perturbative triples [CCSD(T)][3] provide a balance between accuracy and efficiency for molecular systems without degeneracies or near-degeneracies. A variety of different approaches have been introduced that aim to accelerate the convergence of the CCSD(T) and increase its applicability, including explicitly correlated methods,[4,5] linear scaling methods,[6] pair-natural orbital expansions,[7−11] fragmentation schemes,[12,13] and high-performance computing.[14]

Machine learning (ML) has been recently proposed as an alternative to traditional methodologies for further acceleration of CCSD(T). We can separate these methods into two groups based on the feature space (representation) used as input to ML models.[15] In the first group, the feature space depends on atomic positions or atomic functions that are able to encode the local environment of atoms in molecules and materials.[16,17] The "Accurate NeurAl networK engINe for Molecular Energies" (ANI) family of methods[18−21] is a representative example where a vector that contains specific radial and angular chemical information on an individual atom environment is introduced as input. These models are trained with density functional theory or, partially, with CC data and offer significant speedup for the reliable calculation of energies and forces.

The second group of methods utilizes a descriptor space that is based on quantum chemical information obtained from a low-level method such as HF or second-order perturbation theory (MP2) and aims to predict results from a higher-level method [e.g., CCSD or CCSD(T)]. This type of molecular representation is indirectly correlated to the atomic positions, and it can offer direct transferability within molecular environments since they encode the underlying electronic structure properties.[22−30] Along these lines, we have developed a data-driven methodology for the prediction of CCSD two-

Received: October 3, 2023
Revised: January 16, 2024
Accepted: January 22, 2024
Published: February 29, 2024



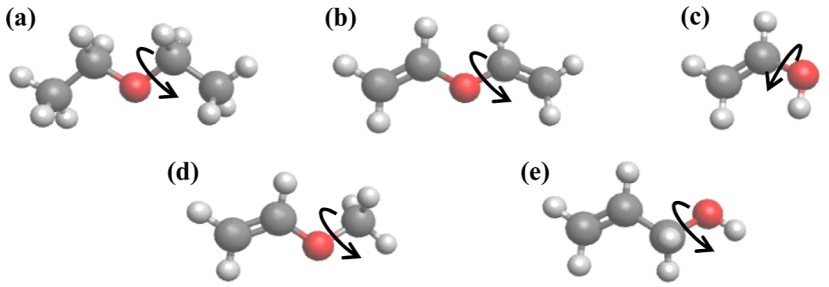



**Figure 1.** Optimized structures of five organic molecules considered in this study: (a) diethyl ether, (b) divinyl ether, (c) vinyl alcohol, (d) methyl vinyl ether, and (e) allyl alcohol. Molecular conformers for training and testing of the DDCC models were generated by rotating the dihedral angle indicated by curved arrows.

electron excitation ($t_2$) amplitudes using MP2-level electronic structure data.[31,32] The data-driven coupled-cluster singles-and-doubles (DDCCSD) scheme has some attractive features, such as transferability between systems of different sizes. However, as the number of basis functions and the system size increase, the number of $t_2$ amplitudes increases, which eventually introduces a computational bottleneck.

In this work, we extended DDCCSD by introducing two new key features that aim to expand the applicability of the method to larger molecular systems. First, we have applied a localized orbital formalism into the data-driven coupled cluster (DDCC) methodology, and second, we have developed five different approaches for the data-point selection from the amplitude space [score to bins (SB), clustering to bins (CB), large amplitudes (LA), electronic correlation (EC), and probabilistic selection (PS)]. The amplitude selection schemes reduce the computational burden of the training and testing process and allow us to apply DDCC on larger molecules and larger basis sets, which was previously unfeasible. Another important aspect of the systematic selection of training data points is the increased accuracy of the ML models by preventing overfitting.

This article is organized as follows: a brief theoretical background of DDCCSD together with a new approach to the calculation of the feature weights is presented in Section 2. The computational details related to the data generation used in this work are given in Section 3. Section 4 introduces the definitions, rationale, and technical aspects of the five amplitude selection schemes that are examined in this study. The evaluation of the performance of different selection schemes is discussed in Section 5, and finally, concluding remarks are presented in Section 6.

## 2. THEORETICAL BACKGROUND

**2.1. Data-Driven Coupled-Cluster Singles and Doubles.** In CC theory, the correlation energy ($E_{\text{corr}}^{\text{CCSD}}$) is computed by using the following equation.

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

The indices $i$ and $j$ correspond to occupied orbitals, $a$ and $b$ to virtual orbitals, $t_i^a$ and $t_{ij}^{ab}$ are the one- and two-electron amplitudes, respectively, and $\langle ij||ab\rangle$ is the two-electron integral between orbitals $i$, $j$, $a$, and $b$. The variational computation of the $t_i^a$ and $t_{ij}^{ab}$ amplitudes is a nontrivial process, and thus, the amplitudes are computed by solving the projected CC equations iteratively

$$\langle \mu|\exp(-\hat{T})\hat{H}\exp(\hat{T})|\Psi_0\rangle = 0 \tag{2}$$

In eq 2, $\psi_0$ is the HF reference wave function, $\hat{T}$ is the cluster operator, and $\mu$ the determinantal excitation manifold (in CCSD, $\mu$ are either the singly or doubly excited determinants). Typically, the MP2 amplitudes $t_{ij(\text{MP2})}^{ab}$ are used as an initial guess for CC $t_{ij}^{ab}$ amplitudes

$$t_{ij(\text{MP2})}^{ab} = \frac{\langle ab||ij\rangle}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{3}$$

where $\varepsilon_p$ is the energy of the orbital $p$. In the DDCCSD approach, the initial CC $t_{ij}^{ab}$ amplitudes are predicted by ML using 30 input features from MP2. Since MP2 amplitudes are the initial guess for the iterative solution of the projected CC equations, they are selected as one of the features used in the DDCCSD method. Other features of the DDCCSD are the numerator and denominator terms of the MP2 amplitude equation (eq 3). The denominator is further broken into individual terms, and the occupied and virtual orbital energies are also introduced in the input vector. The individual orbital energy is broken into the one-electron, Coulomb, and exchange contributions. Another feature of the DDCCSD method is whether the excited electrons are promoted to the same virtual orbital (binary). For the full feature list, see Supporting Information, Section S-I.

The predicted amplitudes ($t_{ij(\text{DDCC})}^{ab}$) are then introduced into the singles ($\mu = \psi_i^a$) projected equations of eq 2 and a single step is performed that updates the $t_i^a$ amplitudes ($t_{i(1)}^a$ or $t_{1(1)}$). The DDCCSD energy is then computed as

$$E_{\text{corr}}^{\text{CCSD}} = \sum_{\substack{a<b \\ i<j}} \langle ij||ab\rangle t_{ij(\text{DDCC})}^{ab} + \sum_{\substack{a<b \\ i<j}} \langle ij||ab\rangle (t_{i(1)}^a t_{j(1)}^b - t_{i(1)}^b t_{j(1)}^a) \tag{4}$$

Note that in the next paragraphs, we will refer to a generic $t_{ij}^{ab}$ amplitude as $t_2$. Alternatively, we can introduce the predicted $t_2$ amplitudes to the solver of CCSD and iteratively obtain the exact CCSD energies in less computational effort, as we have shown previously.[31] In this work, we solely consider the first approach, where the predicted amplitudes from ML are used for the computation of an approximate CCSD energy.

**2.2. Feature Weights.** In our previous work on the DDCCSD method,[31] a weight $w_i$ was assigned to each feature, which was optimized via a grid search for a data set of water molecules in different conformations close to the equilibrium geometry. In order to extend the transferability of DDCC, we have applied an alternative approach for weight assignment to features. We have calculated the Pearson correlation coefficient



https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

($r$) between each of the features with the CCSD $t_2$ amplitude value for the training data

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (5)$$

Here, $x_i$ is the value of feature $x$, $y_i$ is the value of the $i$th $t_2$ amplitude, $\bar{x}$ is the mean of feature $x$, and $\bar{y}$ is the mean of the value of CCSD $t_2$ amplitudes. This correlation coefficient value is used as the weight for each feature (see Supporting Information, Section S-II) and is used for the comparison of canonical molecular orbitals (CMOs) and localized molecular orbitals (LMOs) and for the comparison of different amplitude selection schemes.

## 3. COMPUTATIONAL DETAILS

All quantum chemical calculations, feature extraction, and output data collection ($t_2$ amplitudes) were performed using the Psi4NumPy software[33] as described previously.[31] For the purposes of this study, we used five molecules for the calibration and performance evaluation of the new models; those are diethyl ether, divinyl ether, vinyl alcohol, methyl vinyl ether, and allyl alcohol molecules (Figure 1). Initial ground-state structures were optimized with the B3LYP[34,35] density functional and the 6-31G[36,37] basis set, with the Psi4[38] quantum chemical program package.

In a previous study on machine-learned CCSD pair energies, we demonstrated that models trained with LMOs are more transferable than the CMOs and that the Foster−Boys (Boys) localization scheme provided slightly better accuracy than the Pipek−Mezey localization scheme.[39] In this work, the Boys localization scheme is introduced in DDCC and its performance is compared with canonical orbitals. For the comparison of the performance of the CMOs and LMOs, we used six training sets. The first five training sets consisted of 30 conformers of an individual molecule, while the sixth training set consisted of a mixture of six conformers from each molecule. Similarly, five test sets, each consisting of 50 conformers of a molecule, were used for testing. Conformers for training were generated by rotating the C−O bond of each of the five molecules shown in Figure 1. Conformers for testing were generated by rotating the same bond. For this analysis, the STO-3G[40,41] basis set was used to calculate the CCSD $t_2$ amplitudes.

For the evaluation of the amplitude selection schemes, we used the mixed training set since it better represents the transferability of the DDCCSD models. For testing, we used five conformers from each molecule. Initially, we used the STO-3G basis set with both LMO and CMO formalisms and then expanded our study to the cc-pVDZ and aug-cc-pVDZ basis sets with LMOs. All the conformers used for training and testing are included in the Supporting Information (SI_Conformers.txt document).

Three basis sets were considered in this study. Initially, we used the STO-3G basis set for the comparison of the CMO and LMO models and for determining the best set of hyperparameters for each scheme for further studies. A set of 1,180,056 data points (amplitudes) was available in total for training from the conformers when the STO-3G basis set was used. In order to study the performance of the amplitude selection schemes with increasing number of data points, we further used cc-pVDZ[42] and aug-cc-pVDZ[43] basis sets. A total of 57,357,372 data points were available for training with the cc-pVDZ basis set and 197,707,512 data points were available with the aug-cc-pVDZ basis set (see Supporting Information, Section S-III, for the CCSD $t_2$ amplitude distribution for all basis sets considered in this study). Figure 2 shows a log plot of the distribution of CCSD $t_2$ amplitudes for the aug-cc-pVDZ basis set.

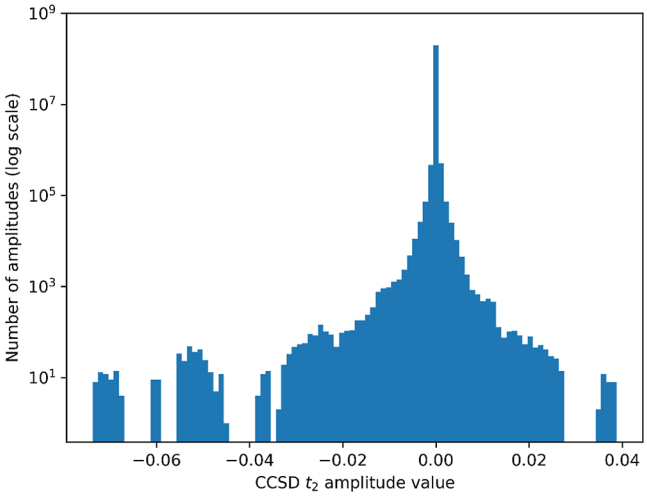



**Figure 2.** Distribution of CCSD $t_2$ amplitudes (on a log scale) computed with the aug-cc-pVDZ basis set.

All DDCCSD models were trained with the random forest ML algorithm[44] using the scikit-learn package.[45] During training, random forest models create a set of decision trees; the number of decision trees is treated as a hyperparameter that must be defined. The prediction for a new data point is the mean of the output from each decision tree. The first step of our study was a hyperparameter search with respect to the number of estimators (decision trees), the choice of the loss function, as well as the minimum number of data points. For the hyperparameter optimization, models were trained with 30 different water conformers and tested with 100 different water conformers with the STO-3G basis set. From this analysis, we found that the optimum set of hyperparameters was 350 estimators with squared error loss function and a minimum sample split of 2 (see Supporting Information, Section S-IV). These hyperparameters were used throughout this study.

For the assessment of the DDCC models trained with truncated amplitude spaces, the mean absolute error (MAE) was calculated for each training and test set combination where $n$ is the number of test conformers

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |E_{\mathrm{CCSD},i} - E_{\mathrm{DDCCSD},i}|}{n} \quad (6)$$

All the features were scaled using MinMaxScaler[45] as implemented in the scikit-learn package.[45]

## 4. AMPLITUDE SELECTION SCHEMES

The main objective of this study is the development and assessment of data selection techniques for the reduction of the amplitude space needed for the DDCC model training. Two aspects were considered for decreasing the number of training data points. The first aspect is computational cost and time reduction. For example, the random forest ML algorithm has a training time complexity of $O(n \times \log(n) \times d \times k)$ where $n$ is the number of training data points, $d$ is the number of features,


https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

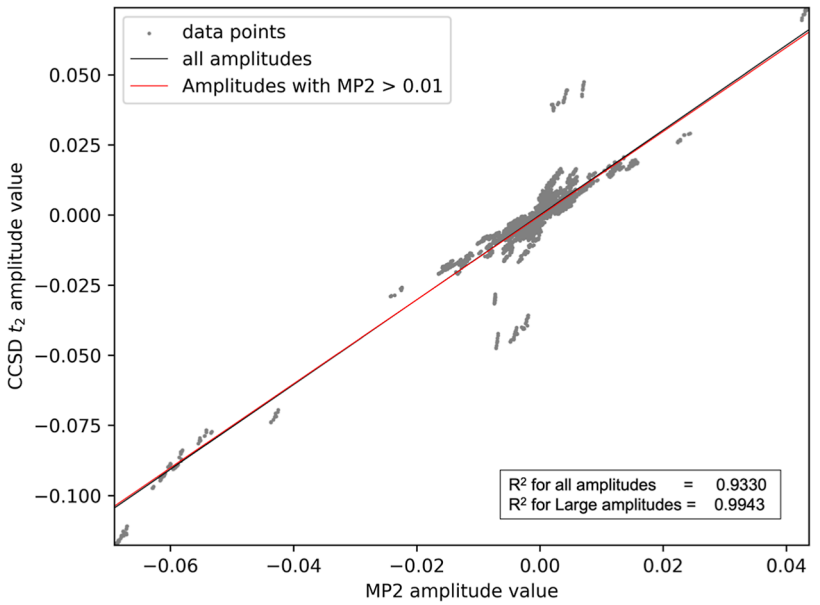



**Figure 3.** CCSD $t_2$ amplitude value vs MP2 amplitude value for all the amplitudes calculated with the STO-3G basis set and LMO formalism.

and $k$ is the number of decision trees. Thus, it becomes evident that the number of input data directly affects the training time. The second consideration is the prevention of model overfitting since a significant number of amplitudes for a given molecule have values close to zero. The number of amplitudes with values closer to zero is also size-dependent. As a solution, we propose the selection of an amplitude subspace that is used for DDCC training without affecting the model accuracy. In this study, we discuss five selection approaches. The first method is SB, where a value (score) is assigned to each amplitude based on their scaled feature scores. The second method involves ML clustering for sorting data into bins (CB). The work of Maitra et al.[46−49] formed the basis for the LA scheme. In their study, they used a set of amplitudes that are greater than a cutoff value to train models that would predict amplitudes that are less than the cutoff value. Here, we are selecting only the amplitudes that are larger than a particular cutoff value for training and testing, and we will refer to it as a LA scheme. The fourth method considers the MP2 correlation energy of two-electron excitations $ij \rightarrow ab$ for selecting amplitudes for training and testing (EC scheme), and the fifth method involves a probabilistic selection (PS scheme) of the initial amplitude space based on feature values. These five schemes are presented in detail in the following paragraphs.

**4.1. Score to Bins.** In the SB approach, a score $S_i$ is assigned to each amplitude based on the values of the scaled features

$$S_i = \sqrt{\sum_{n=1}^{30} (w_n f_n^i)^2} \tag{7}$$

Here, $w_n$ is the weight of feature $n$, and $f_n^i$ is the value of the $n$th feature of the $i$th amplitude. The score $S_i$ represents the distance from the origin of the feature space to each data point. Then, the amplitude space is clustered into $m \times m$ bins, where $m$ is the number of sections that the data set is partitioned along feature score and CCSD $t_2$ axes. The initial choice of $m$ was 100. Therefore, the amplitude space was divided into 10,000 bins. Each amplitude is assigned to a bin according to its feature score $S_i$ and the value of the CCSD $t_2$ amplitude. Next, a common cutoff will be set to all the bins, such that the defined percentage of amplitudes will be selected when it is applied. If a bin has less or equal number of amplitudes than the cutoff, all the amplitudes are used for training. If a bin has amplitudes more than the cutoff, then the number of amplitudes that are equal to the cutoff will be randomly selected for training. Overall, ten models were trained with 5, 10, 20, 30, 40, 50, 60, 70, 80, and 90% of the total number of amplitudes. Figure S4 in the Supporting Information Section S-V shows a heatmap of the amplitude grouping to bins.

**4.2. Clustering to Bins.** The CB approach is similar to SB, but instead of assigning a score, $k$-means clustering[50] is used for assigning amplitudes to bins. $k$-means clustering is an unsupervised learning algorithm whose goal is to assign the data points to $k$ number of clusters by minimizing the difference between the data point and the centroid (mean) of the cluster that has been assigned to. The algorithm starts by assigning centroids to the clusters. The number of clusters $k$ is treated as a model hyperparameter that should be defined (vide infra). After the initial assessment of CB where a value of 100 clusters was selected for $k$, a hyperparameter optimization was performed to determine the best number of $k$ clusters. For that purpose, the $k$-means cluster algorithm was executed 10 times for each clustering with different centroid seeds, and the best clustering was selected each time. To assess the performance, ten models were trained that contained 5, 10, 20, 30, 40, 50, 60, 70, 80, and 90% of the total amplitudes using the CB scheme.

We used the 30 scaled features multiplied by the weight for each feature and the CCSD $t_2$ amplitude value for clustering into $k$ bins. Then, amplitudes were selected for training from each bin using a similar approach to the SB method by setting a common cutoff value. Both SB and CB approaches attempt to select an amplitude subspace that uniformly represents the full amplitude space, which can reduce the amplitudes with values closer to zero but can also avoid model overfitting.

After the optimum percent of amplitudes was determined using the above-mentioned procedure, the effect of the number of bins was assessed by creating models with different numbers


https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

of bins and calculating their accuracy. For the SB approach, six models were trained by selecting 20% of the total amplitudes from 50 × 50, 100 × 100, 150 × 150, 200 × 200, 250 × 250, and 300 × 300 bins. For the CB approach, six models were trained with 5% amplitudes chosen from 50, 100, 150, 200, 250, and 300 bins (see Supporting Information, Section S-VIII).

**4.3. Large Amplitudes.** The LA approach is different from the SB and CB approaches since it uses only data points with MP2 amplitude magnitude greater than a defined cutoff value. All amplitudes below this value are set to zero, while the rest are predicted by DDCC. To further validate this notion, we calculated the correlation coefficient between the MP2 amplitudes and the CCSD $t_2$ amplitudes (STO-3G basis set, LMOs). A correlation coefficient of 0.9330 was obtained when all amplitudes were taken into account (Figure 3), but when amplitudes with magnitudes greater than 0.01 were considered, the correlation coefficient increased to 0.9943. For the LA approach, 11 models were trained with data points corresponding to MP2 amplitudes within the $10^{-2}$−$10^{-7}$ range.

**4.4. EC Scheme.** The EC scheme shares similarities with the LA scheme since both approaches use a cutoff value for selecting training data points. There are two main differences between LA and EC. First, EC uses the MP2 correlation energy $e^{ab}_{ij(\mathrm{MP2})}$ as a cutoff criterion, instead of the amplitude value as in LA. For a two-electron excitation $ij \rightarrow ab$, the MP2 correlation energy $e^{ab}_{ij(\mathrm{MP2})}$ is given as

$$e^{ab}_{ij(\mathrm{MP2})} = \langle ij||ab\rangle t^{ab}_{ij(\mathrm{MP2})} \tag{8}$$

Optimum performance was found when a cutoff of $1 \times 10^{-7}$ $E_h$ (14.3% of the total amplitudes) was used for the STO-3G basis set, $1 \times 10^{-7.5}$ $E_h$ (5.76% of the total amplitudes) for models trained with data from the cc-pVDZ basis set, and $1 \times 10^{-8}$ $E_h$ (4.47% of the total amplitudes) for models with data from aug-cc-pVDZ. All amplitudes with MP2 correlation energy greater than the cutoff value were selected for training. The second difference is related to the selection of the amplitudes that are "predicted" by the DDCC scheme. Specifically, after careful inspection of the opposite-spin (OS) MP2 correlation energy captured by an amplitude subspace (see Supporting Information, Section S-VI), we set a cutoff threshold based on the OS MP2 correlation energy percentage ($p_{\mathrm{cutoff}}$). All $ij \rightarrow ab$ correlation energies $e^{ab}_{ij(\mathrm{MP2})}$ of test molecules are arranged in ascending order, and CCSD $t_2$ predictions are made for the amplitudes with the largest magnitude for $e^{ab}_{ij(\mathrm{MP2})}$ that capture a $p_{\mathrm{cutoff}}$ percent of the total MP2 OS correlation energy. In Section 5, we discuss the effect of different $p_{\mathrm{cutoff}}$ values on the model accuracy.

**4.5. Probabilistic Selection.** The PS approach uses a weighting mechanism where each amplitude from the initial amplitude space is assigned a weight $W_n$ defined as

$$W_n = t^{ab}_{ij(\mathrm{MP2})} \times (e_i + e_j) \tag{9}$$

where $e_k$ is the energy of occupied orbital $k$.

Once each amplitude has been assigned a weight, the probability $p_n$ is computed as

$$p_n = \frac{W_n}{\sum_n W_n} \tag{10}$$

In eq 9, $p_n$ is the probability assigned to the $n$th amplitude. Once the weights are converted to probabilities, a certain percentage of the initial amplitudes are selected in accordance with the assigned probabilities. For the PS approach, the MP2 amplitude was used as an initial weight, as it is the feature most highly correlated to the CCSD amplitude, and the percentage of amplitudes selected from the initial amplitude space was treated as a hyperparameter to be optimized to reduce the error of the model.

## 5. RESULTS AND DISCUSSION

**5.1. Comparison of CMO and LMO Models.** Table 1 shows the MAE for models trained with CMOs and LMOs.

**Table 1. Average MAE (in m$E_h$) of DDCC Models Trained with Canonical or Localized Orbitals and with Data from the Same Molecule (Individual Model) or with Data from All Five Molecules (Mixed Model)[a]**

| | canonical orbitals | | localized orbitals | |
|---|---|---|---|---|
| | individual model | mixed model | individual model | mixed model |
| diethyl ether | 11.90 (2.26) | 12.41 (2.83) | 0.67 (0.09) | 0.77 (0.15) |
| divinyl ether | 11.86 (2.25) | 11.80 (3.72) | 1.08 (0.17) | 1.98 (0.44) |
| vinyl alcohol | 4.22 (1.04) | 7.17 (2.04) | 0.34 (0.13) | 0.46 (0.47) |
| methyl vinyl ether | 8.36 (1.89) | 10.34 (2.50) | 0.76 (0.44) | 1.29 (0.47) |
| propenol | 7.77 (1.27) | 9.34 (1.73) | 0.60 (0.05) | 0.69 (0.13) |
| **average** | **8.82** | **10.21** | **0.69** | **1.04** |

[a]Standard deviation for each MAE calculation is given in parentheses.

These DDCC models are grouped as "individual models" (i.e., models that are trained with data from a specific molecule and tested on conformers of the same molecule) and as "mixed models", where heterogeneous data from all five molecules are used for training. These MAEs show that there is a clear gain in both accuracy and transferability from the use of localized orbitals. All individual and mixed models showed an increase in accuracy by an order of magnitude when LMOs were introduced in DDCC. For example, the average MAE of the individual and mixed models with CMOs is 8.82 and 10.21 m$E_h$, respectively, while the corresponding deviations from the LMO models are only 0.69 and 1.04 m$E_h$, respectively. LMOs also demonstrate increased transferability since the difference between mixed and individual models is 0.35 m$E_h$, while the same difference with CMOs is 1.39 m$E_h$. We also note that the deviations from LMO individual and mixed models are comparable, apart from divinyl ether (difference of 0.90 m$E_h$).

**5.2. Evaluation of the Amplitude Selection Schemes.** The bar plot of Figure 4 shows the distribution of all amplitudes (black bars) as well as 20% of the amplitudes selected from the SB, CB, LA, PS, and EC schemes (yellow bars). These amplitudes were calculated with the STO-3G basis set and using LMOs. In addition to the above-mentioned schemes, a distribution for 20% of randomly selected data points is also shown in Figure 4e for comparison with the data-driven selection schemes. The SB, CB, LA, PS, and EC schemes select almost all the amplitudes with considerably larger magnitudes as shown in Figure 4(a), (b), (c), (d), and (f) respectively, whereas the random model omits a significant number of amplitudes with larger magnitudes.

A comparison of MAE for the calculation of energy for test conformers with different amplitude selection schemes with



https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

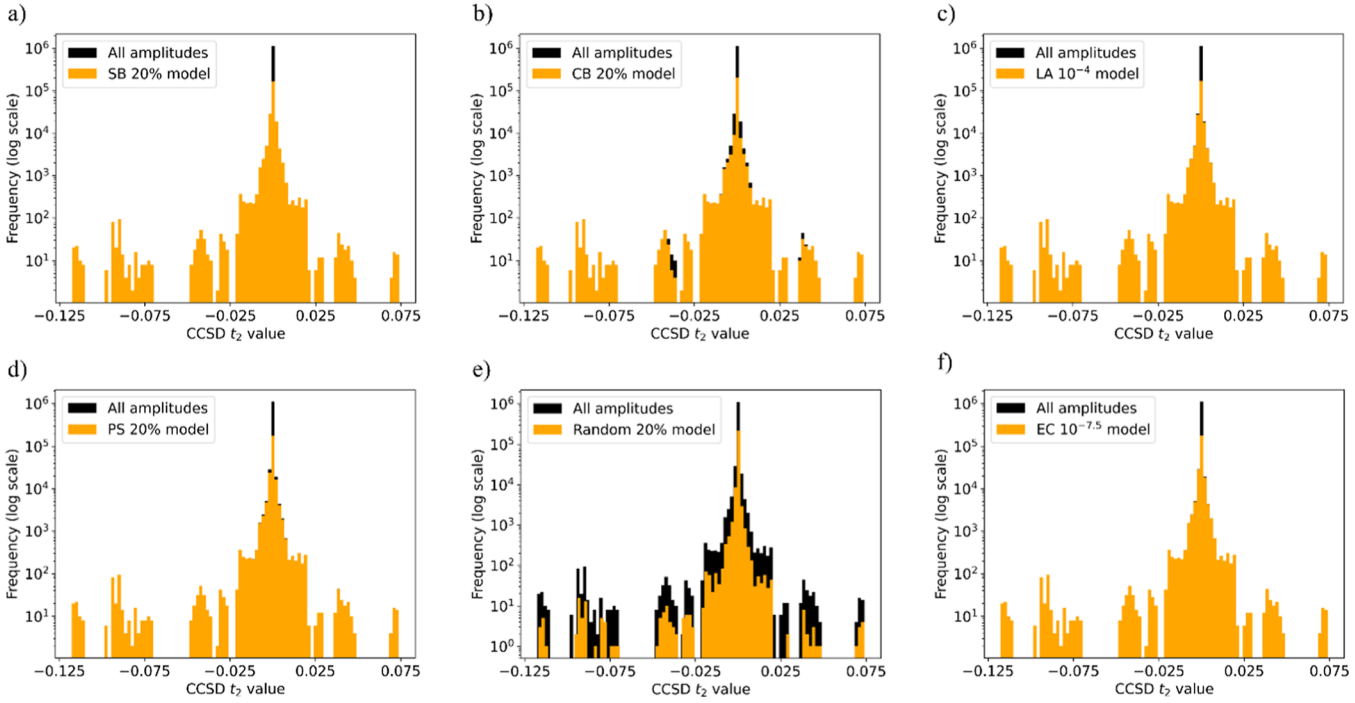



**Figure 4.** Distribution of selected CCSD $t_2$ amplitudes (yellow) compared with the distribution of the total amplitude space for (a) SB, (b) CB, (c) LA, (d) PS, (e) random, and (f) EC schemes.

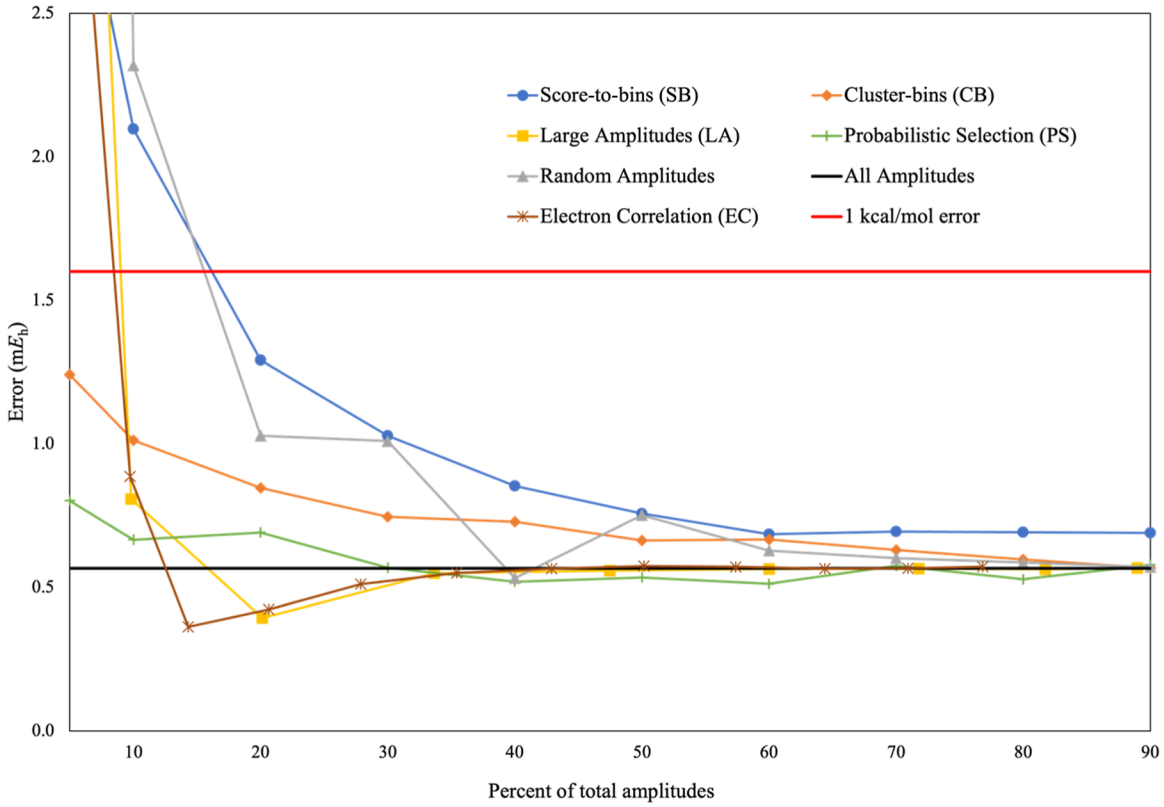



**Figure 5.** MAE vs percentage of amplitudes selected for SB (blue), CB (orange), LA (yellow), EC (brown), and PS (green) schemes. MAE for the random selection approach was also depicted in gray. CCSD amplitudes for training and testing were calculated using the STO-3G basis set and LMO formalism.

LMOs is shown in Figure 5. It is evident that the CB and PS models provide an accuracy below 1 kcal/mol with only 5% of the total amplitudes. Models with the LA approach provide an accuracy below 1 kcal/mol error with 9.81% of the total amplitudes. Both SB and randomly selected amplitude models have an accuracy below 1 kcal/mol with 20% of the total amplitudes. Overall, with LMOs, all schemes achieve less than 1 kcal/mol accuracy with a significantly lower number of



https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

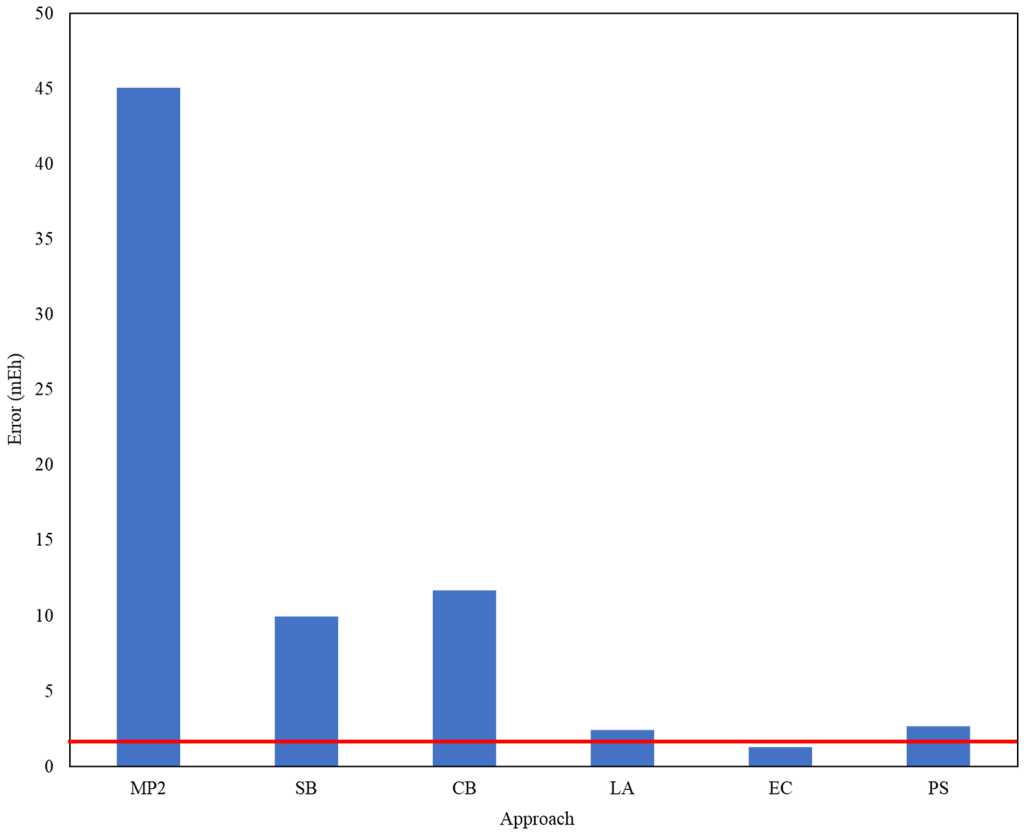



**Figure 6.** Bar plot with the MAEs for models trained with SB, CB, LA, EC, and PS schemes calculated with the cc-pVDZ basis set. The difference between MP2 and CCSD correlation energies is shown as MP2 (left bar). The 1 kcal/mol threshold is indicated by a horizontal red line.

amplitudes. Interestingly, the CB, LA, and random-amplitude models converge to the accuracy of the all-amplitude model as the number of training data points systematically approaches the full amplitude space (100%). SB models approach a plateau at around 60% of the total amplitudes. We would also like to note that the PS model with LMOs has the best performance across all different amplitude percentages, from 5 to 90% (vide infra). On the contrary, all CMO models with amplitude selection failed to achieve below 1 kcal/mol accuracy (see Supporting Information, Section S-VII). An important observation is the behavior of the LA models, which achieve significantly less error (6.95 m$E_h$) than the all-amplitude model (10.19 m$E_h$) at 18% of the total amplitudes, but the error gradually increases and approaches asymptotically the all-amplitude model when the number of amplitudes increases.

Next, we performed an analysis to investigate the effect of the number of bins on the accuracy of SB and CB models with LMO formalism. Results of this analysis are listed in the Supporting Information, Section S-VIII. The best number of bins for SB models was 300 × 300 bins and that for CB models was 50 bins. These results will be used for the calculations with the cc-pVDZ basis set.

The behavior of the LA models shows an interesting pattern. Errors from LA models are large at low data percentages (less than 10% of the total data points used for training). This large error can be attributed to neglecting contributions from a large number of amplitudes for the correlation energy calculation. The accuracy is increased when more amplitudes are added (15%). Upon incremental addition above 20%, the errors start to increase and asymptotically reach the error of the all-amplitude model. Thus, with 15% of the total number of amplitudes, we avoid overfitting of the DDCCSD/LA model. This is visually shown in Figure 3, where the $R^2$ value between all MP2 and CCSD amplitudes is 0.9330, which increases to 0.9943 when only the large amplitudes are considered.

The EC scheme shows behavior similar to that of the LA scheme. At low percentages, the error is significantly high due to a lack of an adequate number of amplitudes to account for the accurate estimation of the total correlation energy. The EC scheme approaches the lowest error of 0.363 m$E_h$ with only 14.3% (when the cutoff is $1 \times 10^{-7}$ m$E_h$) of the total amplitudes. Upon further addition of amplitudes, the MAE increases asymptotically due to overfitting and eventually approaches the all-amplitude model error when 42.9% (cutoff is $1 \times 10^{-9}$ m$E_h$) of the total amplitudes are used.

Models trained with PS schemes at low percentages (between 5 and 20%) predict the CCSD energy with greater accuracy than the other schemes considered in this study. Further addition of amplitudes above 30% does not deviate from the accuracy of the all-amplitude model, and the highest accuracy for the PS scheme is at 60% (0.523 m$E_h$). The PS scheme has the lowest range of variation of MAE, thus, it is the most consistent scheme out of all the amplitude selection schemes we have used for STO-3G.

**5.3. Basis Set Effects.** Next, we explored basis set effects by keeping the same hyperparameters (e.g., number of bins and amplitude percentage) as they were obtained from the smaller STO-3G basis set. For example, the percentages for the CB and SB models were 5% and 20%, respectively, since those were the lowest percentage values that provided MAEs below 1 kcal/mol. The number of bins used in the CB and SB models was 50 and 300 × 300, respectively. For the LA model, a cutoff


https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

Table 2. MAE (in m$E_h$) Calculated for LA, EC, and PS Schemes with Different Cutoff Values (cc-pVDZ Basis Set)[a]

| LA scheme | | | EC scheme | | PS scheme | |
|---|---|---|---|---|---|---|
| cutoff | MAE | percent amplitudes | $p_{cutoff}$ | MAE | cutoff (percent) | MAE |
| $1 \times 10^{-2}$ | 493.563 (99.836) | 0.003% | 99.1% | 1.406 (0.503) | 2% | 5.253 (4.675) |
| $1 \times 10^{-3}$ | 112.333 (36.765) | 0.330% | 99.15% | 1.288 (0.767) | 4% | 3.027 (2.423) |
| $1 \times 10^{-4}$ | 2.447 (1.531) | 4.654% | 99.2% | 1.370 (0.975) | 6% | 2.674 (1.765) |
| $1 \times 10^{-5}$ | 8.837 (1.648) | 21.801% | 99.25% | 1.565 (1.164) | 8% | 6.850 (5.231) |
| $1 \times 10^{-6}$ | 8.754 (1.338) | 49.024% | 99.3% | 1.953 (1.205) | 10% | 10.501 (6.019) |

[a]Cutoff values for the LA scheme are set for the magnitude of MP2 amplitudes, whereas the cutoffs for the EC scheme are set for the correlation energy of each data point. For the PS scheme, cutoffs are the percentages of data points selected from the total amplitude space. Percentages of amplitudes selected for training EC scheme models were fixed at 5.76% of the total number of amplitudes. The standard deviation for each MAE calculation is given in parentheses.

value of $1 \times 10^{-4}$ was applied since the model with $1 \times 10^{-4}$ cutoff provides the lowest error for the STO-3G basis set.

MAEs for models trained with SB, CB, LA, EC, and PS schemes calculated with the cc-pVDZ basis set are shown in Figure 6, together with the MAE between MP2 and CCSD correlation energies (45.04 m$E_h$, see bar with label "MP2"). The model trained with 5% of the total amplitude space using the CB approach has an MAE of 11.68 m$E_h$, while the SB model with 20% of amplitudes has an MAE of 10.06 m$E_h$. Additional models that utilize larger data percentages of the total amplitudes were trained (CB with 20%, 30% with SB), but the accuracy remained close to or above 10 m$E_h$ (10.55 and 9.941 m$E_h$ for CB and SB, respectively). Since the addition of larger amounts of training data (larger percentages) increases the computational effort without decreasing the model deviations, we are not considering the CB and SB approaches in the following sections.

As mentioned above, the first model for the LA approach with the cc-pVDZ basis set was trained using the cutoff value for the best LA model for the STO-3G basis set ($1 \times 10^{-4}$). CCSD energies were predicted with an accuracy of 2.447 m$E_h$ using only 4.65% of the amplitudes with this model. Next, models were trained and tested with the LA approach by using different cutoff values to identify the best value for the cc-pVDZ basis set (see Table 2, left column). As the number of amplitudes is decreased by increasing the cutoff, the MAE increases. A smaller cutoff increases the number of amplitudes for training and testing but the MAE also increases (~8.8 m$E_h$). Therefore, from this analysis, we found that the accuracy of the models converges at the $1 \times 10^{-4}$ cutoff. Similar behavior for MAE was observed for STO-3G models (Figure 5).

The EC model that was tested with a $p_{cutoff}$ of 99.15% of the total MP2 correlation was able to provide the lowest errors (see the middle column of Table 2). This model was able to predict the CCSD energy with an error of 1.288 m$E_h$.

For the PS approach, the model trained with 60% of the total amplitudes resulted in a large deviation (117.160 m$E_h$). To further evaluate the performance at lower percentages of amplitudes, we trained models with the PS approach from 1 to 10% (see Table 2, right column), and the lowest MAE (2.674 m$E_h$) was achieved with only 6% of the total number of amplitudes. Therefore, the PS offers a useful roadmap for avoiding model overfitting.

We selected the best scheme from the cc-pVDZ basis set calculations, which is the EC scheme, for the aug-cc-pVDZ basis set calculations. The results for the aug-cc-pVDZ basis set with different cutoff values are summarized in Table S7 of the Supporting Information, Section S-IX. The average difference between MP2 and CCSD correlation energy for the aug-cc-pVDZ basis set is 42.532 m$E_h$. The EC model was trained using 4.47% of the total number of amplitudes (amplitudes with $|e^{ab}_{ij(MP2)}|$ greater than $1 \times 10^{-8} E_h$). The best cutoff for testing the EC model was 98.8%, and the MAE for this cutoff is 1.946 m$E_h$.

## 6. CONCLUSIONS

The DDCC model offers an alternative approach to the acceleration of CC theory. As the system and basis set sizes increase, the number of training data that are needed increases and eventually becomes the bottleneck for the application of DDCC to larger molecular systems. The main objective of this study was the $t_2$ amplitude space truncation, an important task that can reduce the required quantum chemical data and training time, while also increase model accuracy. For this purpose, we introduced five amplitude selection schemes that can be grouped into two categories. The first category includes "data-driven" schemes such as the SB, the CB, and the PS. The second category contains two schemes that are based on chemical arguments and in particular on the two-electron excitation parameter $t_2$ (LA) or the two-electron excitation energy $\epsilon^{ab}_{ij}$ (EC). Results obtained from the STO-3G basis set allowed us to explore the behavior and accuracy of these schemes as well as to refine model hyperparameters. We found that models that utilized 15% of the total amplitude set together with the LA, EC, and PS schemes were able to provide similar or better accuracy when compared with the model trained with all available amplitudes. These three schemes were also able to provide reasonable accuracy when used together with the data obtained from the cc-pVDZ basis set, but further assessment with the larger aug-cc-pVDZ basis set revealed that only the EC scheme was able to provide satisfactory results. Finally, our initial hypothesis of reduction of overfitting by amplitude space truncation was proven, since errors start to increase after reaching a minimum for all basis sets when LA and PS approaches were used. We are currently exploring physics-based approaches that can be coupled together with the amplitude selection schemes discussed here that can further increase the applicability of DDCC schemes. Another future direction is the exploration of electron correlation effects far from the equilibrium geometries, to develop models with improved transferability for strongly correlated systems.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

Code is available at https://github.com/Varunasp93/DDCC.git.



https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947

## Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.3c06600.

Information about the features of the current model, hyperparameter optimization, and performance of the ML models (PDF)

Geometries used for training and testing the models (TXT)

## AUTHOR INFORMATION


### Corresponding Author

**Konstantinos D. Vogiatzis** − *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996-1600, United States*; orcid.org/0000-0002-7439-3850; Email: kvogiatz@utk.edu

### Authors

**P. D. Varuna S. Pathirage** − *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996-1600, United States*

**Justin T. Phillips** − *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996-1600, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.3c06600


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant no. 2143354 (CAREER: CAS-Climate). We also acknowledge the Infrastructure for Scientific Applications and Advanced Computing (ISAAC) of the University of Tennessee for computational resources.

## REFERENCES


(1) Tew, D. P.; Klopper, W.; Helgaker, T. Electron correlation: The many-body problem at the heart of chemistry. *J. Comput. Chem.* **2007**, *28* (8), 1307−1320.

(2) Townsend, J.; Kirkland, J. K.; Vogiatzis, K. D. Post-Hartree-Fock methods: configuration interaction, many-body perturbation theory, coupled-cluster theory. In *Mathematical Physics in Theoretical Chemistry*; House, J. E., Ed.; Elsevier, 2019, pp 63−117.

(3) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157* (6), 479−483.

(4) Hattig, C.; Klopper, W.; Kohn, A.; Tew, D. P. Explicitly correlated electrons in molecules. *Chem. Rev.* **2012**, *112* (1), 4−74.

(5) Kong, L.; Bischoff, F. A.; Valeev, E. F. Explicitly correlated R12/F12 methods for electronic structure. *Chem. Rev.* **2012**, *112* (1), 75−107.

(6) Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S. Linear-scaling methods in quantum chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds., 2007; Vol. *23*, p 1.

(7) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138* (3), 034106.

(8) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144* (2), 024109.

(9) 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* (13), 134101.

(10) Ma, Q.; Werner, H.-J. Explicitly correlated local coupled-cluster methods using pair natural orbitals. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8* (6), No. e1371.

(11) Vogiatzis, K. D.; Barnes, E. C.; Klopper, W. Interference-corrected explicitly-correlated second-order perturbation theory. *Chem. Phys. Lett.* **2011**, *503* (1−3), 157−161.

(12) Stoll, H. The correlation energy of crystalline silicon. *Chem. Phys. Lett.* **1992**, *191* (6), 548−552.

(13) Vogiatzis, K. D.; Klopper, W.; Friedrich, J. Non-covalent Interactions of $CO_2$ with Functional Groups of Metal-Organic Frameworks from a CCSD (T) Scheme Applicable to Large Systems. *J. Chem. Theory Comput.* **2015**, *11* (4), 1574−1584.

(14) Datta, D.; Gordon, M. S. A massively parallel implementation of the CCSD (T) method using the resolution-of-the-identity approximation and a hybrid distributed/shared memory parallelization model. *J. Chem. Theory Comput.* **2021**, *17* (8), 4799−4822.

(15) Jones, G. M.; Story, B.; Maroulas, V.; Vogiatzis, K. D. *Molecular Representations for Machine Learning*; American Chemical Society, 2023.

(16) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **2021**, *121* (16), 9759−9815.

(17) 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* (1), 5024.

(18) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8* (4), 3192−3203.

(19) 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* (24), 241733.

(20) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10* (1), 2903.

(21) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401.

(22) Qiao, Z.; Christensen, A. S.; Welborn, M.; Manby, F. R.; Anandkumar, A.; Miller, T. F. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (31), No. e2205221119.

(23) Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G. A.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R.; et al. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **2021**, *155* (20), 204103.

(24) Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336−2347.

(25) Manzhos, S. Machine learning for the solution of the Schrödinger equation. *Mach. Learn.: Sci. Technol.* **2020**, *1* (1), 013002.

(26) 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* (23), 4861−4871.

(27) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J. R.; Miller, T. F. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021**, *154* (6), 064108.

(28) Han, R.; Rodriguez-Mayorga, M.; Luber, S. A machine learning approach for MP2 correlation energies and its application to organic compounds. *J. Chem. Theory Comput.* **2021**, *17* (2), 777−790.

(29) Margraf, J. T.; Reuter, K. Making the coupled cluster correlation energy machine-learnable. *J. Phys. Chem. A* **2018**, *122* (30), 6343−6348.

(30) Ikabata, Y.; Fujisawa, R.; Seino, J.; Yoshikawa, T.; Nakai, H. Machine-learned electron correlation model based on frozen core approximation. *J. Chem. Phys.* **2020**, *153* (18), 184108.
(31) Townsend, J.; Vogiatzis, K. D. Data-driven acceleration of the coupled-cluster singles and doubles iterative solver. *J. Phys. Chem. Lett.* **2019**, *10* (14), 4129−4135.
(32) Jones, G. M.; Pathirage, P. D. V. S.; Vogiatzis, K. D. Data-driven acceleration of coupled-cluster and perturbation theory methods. In *Quantum Chemistry in the Age of Machine Learning*; Dral, P. O., Ed.; Elsevier, 2023, pp 509−529.
(33) 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* (7), 3504−3511.
(34) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38* (6), 3098−3100.
(35) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter* **1988**, *37* (2), 785−789.
(36) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54* (2), 724−728.
(37) Hehre, W. J.; Ditchfield, R.; 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* (5), 2257−2261.
(38) 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* (18), 184108.
(39) Townsend, J.; Vogiatzis, K. D. Transferable MP2-based machine learning for accurate coupled-cluster energies. *J. Chem. Theory Comput.* **2020**, *16* (12), 7453−7461.
(40) Hehre, W. J.; Stewart, R. F.; Pople, J. A. Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals. *J. Chem. Phys.* **1969**, *51* (6), 2657−2664.
(41) 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* (5), 2769−2773.
(42) Dunning, T. H.; Hay, P. J. Gaussian basis sets for molecular calculations. In *Methods of Electronic Structure Theory*; Schaefer, H. F., Ed.; Springer, 1977, pp 1−27.
(43) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96* (9), 6796−6806.
(44) Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*; IEEE, 1995; pp 278−282.
(45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(46) Agarawal, V.; Roy, S.; Shrawankar, K. K.; Ghogale, M.; Bharathi, S.; Yadav, A.; Maitra, R. A hybrid coupled cluster-machine learning algorithm: Development of various regression models and benchmark applications. *J. Chem. Phys.* **2022**, *156* (1), 014109.
(47) Agarawal, V.; Patra, C.; Maitra, R. An approximate coupled cluster theory via nonlinear dynamics and synergetics: The adiabatic decoupling conditions. *J. Chem. Phys.* **2021**, *155* (12), 124115.
(48) Patra, C.; Agarawal, V.; Halder, D.; Chakraborty, A.; Mondal, D.; Halder, S.; Maitra, R. A synergistic approach towards optimization of coupled cluster amplitudes by exploiting dynamical hierarchy. *ChemPhysChem* **2023**, *24* (4), No. e202200633.
(49) Agarawal, V.; Roy, S.; Chakraborty, A.; Maitra, R. Accelerating coupled cluster calculations with nonlinear dynamics and supervised machine learning. *J. Chem. Phys.* **2021**, *154* (4), 044110.
(50) Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28* (2), 129−137.


https://doi.org/10.1021/acs.jpca.3c06600
J. Phys. Chem. A 2024, 128, 1938−1947