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

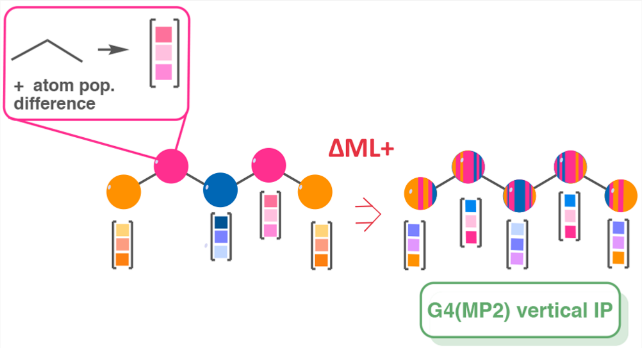
Quantitative Prediction of Vertical Ionization Potentials from DFT via a Graph-Network-Based Delta Machine Learning Model Incorporating Electronic Descriptors

Published as part of The Journal of Physical Chemistry virtual special issue "MQM 2022: The 10th Triennial Conference on Molecular Quantum Mechanics".

Sarah Maier,* Eric M. Collins, and Krishnan Raghavachari*

Cite This: *J. Phys. Chem. A* 2023, 127, 3472−3483 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: While accurate wave function theories like CCSD(T) are capable of modeling molecular chemical processes, the associated steep computational scaling renders them intractable for treating large systems or extensive databases. In contrast, density functional theory (DFT) is much more computationally feasible yet often fails to quantitatively describe electronic changes in chemical processes. Herein, we report an efficient delta machine learning (ΔML) model that builds on the Connectivity-Based Hierarchy (CBH) scheme—an error correction approach based on systematic molecular fragmentation protocols—and achieves coupled cluster accuracy on vertical ionization potentials by correcting for deficiencies in DFT. The present study integrates concepts from molecular fragmentation, systematic error cancellation, and machine learning. First, we show that by using an electron population difference map, ionization sites within a molecule may be readily identified, and CBH correction schemes for ionization processes may be automated. As a central feature of our work, we employ a graph-based QM/ML model, which embeds atom-centered features describing CBH fragments into a computational graph to further increase accuracy for the prediction of vertical ionization potentials. In addition, we show that the incorporation of electronic descriptors from DFT, namely electron population difference features, improves model performance well beyond chemical accuracy (1 kcal/mol) to approach benchmark accuracy. While the raw DFT results are strongly dependent on the underlying functional used, for our best models, the performance is robust and much less dependent on the functional.

1. INTRODUCTION

In recent decades, the field of theoretical quantum chemistry has made tremendous strides toward the development of methods that can be applied on fairly large molecules to achieve reasonable accuracy without incurring prohibitive computational costs.[1] Today, hybrid and range separated DFT methods are ubiquitous in computational studies of large molecules, and fast semiempirical methods are becoming increasingly accurate.[2,3] Nonetheless, computational results on par with those obtained from highly accurate, correlated methods like coupled cluster theory are still largely unattainable if one wishes to tackle extensive databases or consider systems with more than a handful of heavy atoms.[4−7]

Before the debut of high-speed computing, calculating highly accurate thermodynamic properties of molecular systems was largely impractical. In 1970, John Pople introduced the isodesmic bond separation (IBS) scheme (example in Figure 1), improving the accuracy of calculated thermodynamic properties using simple theoretical models, e.g., Hartree−Fock theory with a moderate basis set.[8] In an IBS reaction, bonds between heavy atoms are extracted as molecular fragments containing two heavy atoms, and all formal bond types are preserved. For these types of reactions, errors specific to local molecular units are well-balanced in reactants and products, and highly accurate reaction energies may be achieved even

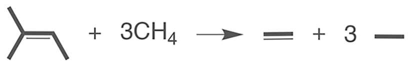


Figure 1. Isodesmic bond separation reaction scheme for 2-methylbut-2-ene.

Received: December 16, 2022
Revised: March 21, 2023
Published: April 4, 2023

https://doi.org/10.1021/acs.jpca.2c08821

with lower levels of theory. As such, the heat of an IBS reaction is a measure of the departure from additivity of bond energies. The advent of IBS schemes revealed that appropriately balanced reaction energies could be exploited for error cancellation, improving accuracy in computational thermochemistry.

Many researchers have furthered the foundational work by Pople, developing improved methods such as the hybridization-based homodesmotic scheme by George and coworkers.[9] Various derivative methods founded on error cancellation and bond type matching have evolved from the IBS scheme, including the hyperhomodesmotic, semihomodesmotic, quasihomodesmotic, and homomolecular homodesmotic schemes to name a few.[10,11] In an effort to lend order to this myriad of methods, the generalized and systematic Connectivity-Based Hierarchy (CBH) scheme was developed in our group.[12] CBH defines an intuitive protocol by which corrections to low-level theoretical methods may be constructed from IBS-like reaction schemes based solely on connectivity and bond-types. Central to CBH is a hierarchy of reaction schemes, whose sequential levels incorporate larger portions of the molecular environment, thus providing greater error cancellation leading to higher accuracy.

The CBH protocol is well-defined, and a thorough explanation of the method and its applications can be found in a string of papers by Raghavachari and co-workers.[13−16] While advancing through the ranks of the hierarchy, fragment size increases systematically, with CBH-0 units consisting of a single heavy atom, CBH-1 units consisting of two heavy atoms, and CBH-2 units consisting of one heavy atom along with all heavy atoms in its immediate bonding environment. Larger fragments capture larger portions of the molecular environment and are expected to provide better error cancellation between reactants and products. An example of the CBH-2 fragmentation scheme is shown in Figure 2.

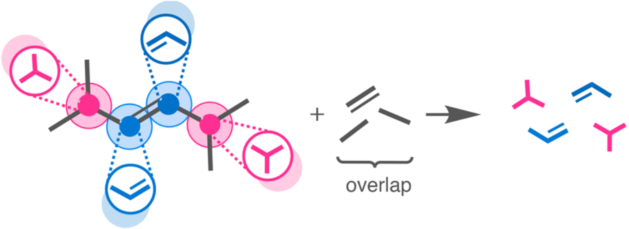


Figure 2. Atom-centered CBH-2 fragmentation scheme for 2,5-dimethylhex-3-ene.

Despite its simplicity, CBH achieves high accuracy. Chemically meaningful correction schemes are attained using the basic building blocks of chemical structure. The method has been applied successfully to predict a range of thermochemical problems with coupled cluster accuracy, including heats of formation of neutral and charged organic molecules, bond dissociation energies, pK_as, and redox potentials.[13−17]

Complementary techniques distinct from the usual tools of a theoretical chemist have also shown promise in the ongoing battle between accuracy and efficiency. In particular, several studies have achieved high accuracy in chemical prediction using schemes based on machine learning (ML).[18−24] ML techniques offer reasonable accuracy along with fast computational speeds.[25,26] Various ML models have been developed to predict a range of chemical properties; however, many are trained on DFT data and thus cannot compete with the high-level accuracy of correlated QM methods. Methods which attain such high-level accuracy are named "chemically accurate" (typically defined to be ±1 kcal/mol) and are needed for a strong match to experiment.

Alongside IBS-like reaction schemes, hybrid QM/ML delta machine learning (ΔML) techniques have emerged to correct for deficiencies in low-level electronic structure calculations.[27−34] DFT is capable of capturing a high fraction of the true molecular energy. However, there is a portion of the energy which can only be captured using the most sophisticated correlated methods.[3,35,36] Unfortunately, determining this portion of the energy can be prohibitive for larger molecules. ΔML has shown promise in its ability to accurately and efficiently determine differences between such highly accurate methods and DFT. Under the ΔML framework, ML is used to generate a correction term, Δ, for a low-level calculation:

$$\Delta^{ML} \approx E_{high\text{-}level} - E_{low\text{-}level} \tag{1}$$

While developed in completely different contexts, both CBH and ΔML view low-level methods such as DFT as a suitable foundation to be exploited for more accurate chemical prediction. While CBH is capable of correcting many of the deficiencies of low-level methods, it may fail when CBH fragments cannot fully capture the molecular environment. While CBH corrects errors specific to *local* molecular units, ML has, in principle, no such restriction. In a 2021 study, we drew a direct connection between ΔML and CBH with our FragGraph model, which encodes CBH fragments as descriptors in a graph neural network regime.[37] We demonstrated that molecular units created during CBH fragmentation function as effective molecular descriptors in the prediction of atomization energies when coupled to ΔML strategies and graph neural networks.

In the present study, we apply our methods more broadly and demonstrate their performance for an electronic property. We adopt a graph ΔML method for the prediction of vertical ionization potentials. In a 2020 study, CBH was used to calculate redox potentials of 46 C-, O-, N-, Cl-, F-, and S-containing molecules with an accuracy within ∼0.09 V of G4.[14] In the current work, additional improvements to the CBH-Redox method are made via two distinct routes. First, CBH-Redox requires electron loss to be localized on a particular fragment. This fragmentation protocol requires a chemist's intuition to determine the most likely site of oxidation, and so is ill-suited for direct automation. Included in the current work is an automated method to identify oxidation sites using an electron population difference map constructed from atomic electron populations of neutral and ionized species. Second, we merge our expertise in QM calculations with new advancements in ML to develop a ΔML method which uses a graph model along with CBH-like features to predict vertical ionization potentials (IPs) with high-level accuracy. As a final point, we take full advantage of ΔML by adding features taken from low-level electronic structure calculations to our models. By incorporating QM-based features, we observe a significant improvement in performance.

2. METHODS

2.1. CBH-Redox.

In the CBH framework, a molecule is broken down into smaller units according to the fragmentation

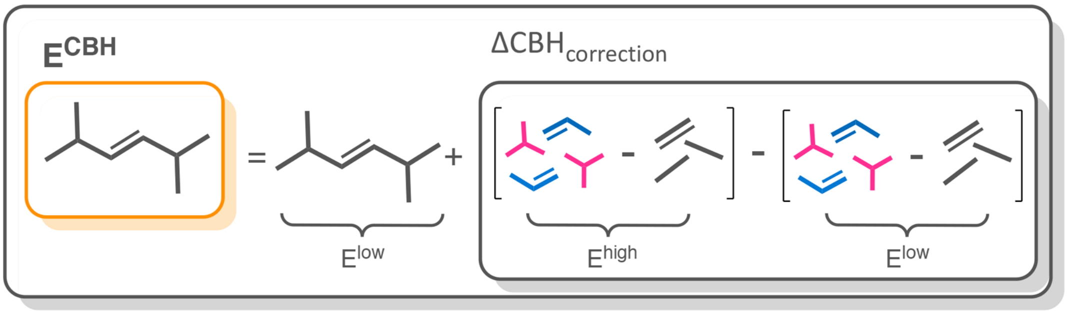


Figure 3. Illustration of the calculation of $\Delta CBH_{correction}$ and E^{CBH} for 2,5-dimethylhex-3-ene.

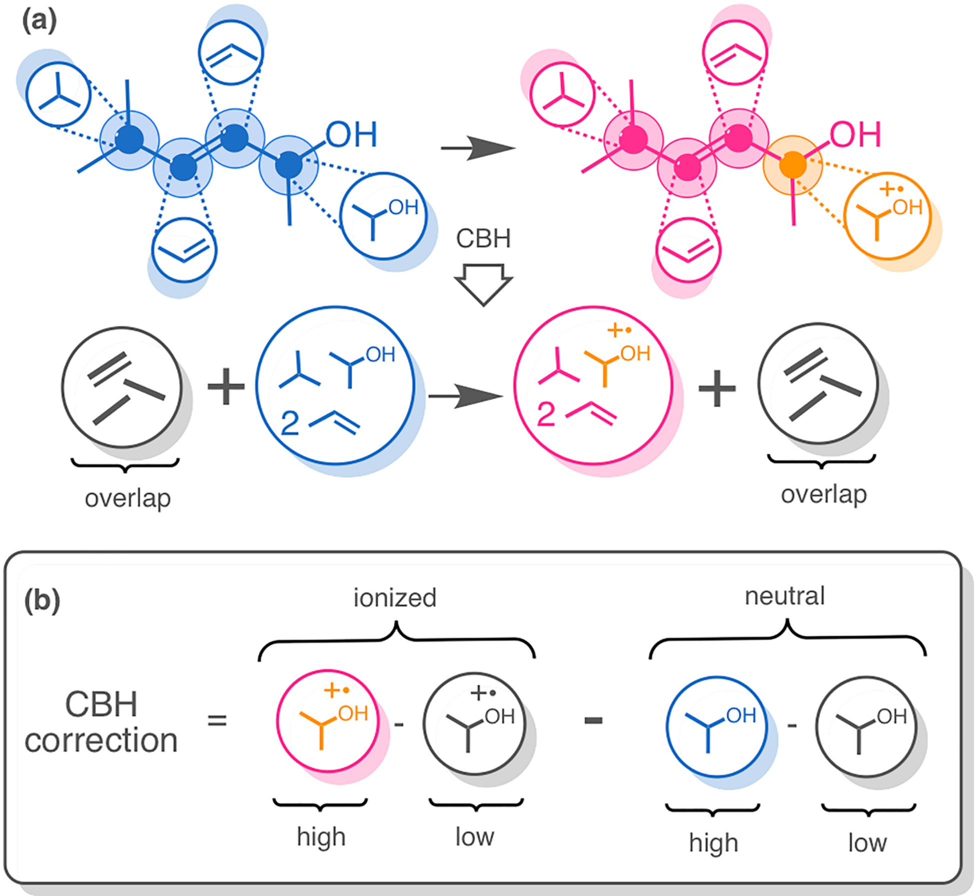


Figure 4. (a) Ionization process for 2,5-dimethylhexane modeled with CBH-2 before elimination of common fragments and (b) the net CBH-2 correction to the ionization energy after elimination of common fragments.

scheme prescribed by a particular rung in the hierarchy, and resultant fragments are used to construct a correction to the total low-level energy. The CBH correction and approximate high-level energy are calculated as

$$E_{high}(full) - E_{low}(full) \approx \sum_{i} E_{high}(i) - \sum_{i} E_{low}(i) = \Delta CBH_{correction} \quad (2)$$

$$E_{High}(full) \approx E_{Low}(full) + \Delta CBH_{correction} = E^{CBH} \quad (3)$$

where $E_{high}(full)$ is the energy of the full molecule calculated at the high-level of theory, $E_{low}(full)$ is the energy of the full molecule calculated at the low-level of theory, $E_{high}(i)$ is the energy of the ith fragment calculated at the high-level of theory, $E_{low}(i)$ is the energy of the ith fragment calculated at the low-level of theory, and $\Delta CBH_{correction}$ is the total CBH correction to the full low-level energy. Hydrogens are added as needed to preserve the original hybridization. CBH is an energy correction which considers the local molecular environment of each fragment, removing a fragment's low-level energy and replacing it with its high-level energy. Figure 3 shows how $\Delta CBH_{correction}$ and E^{CBH} are calculated for the molecule 2,5-dimethylhex-3-ene. CBH is a chemically intuitive approach to obtain highly accurate thermochemical calculations, using structure-based information to derive local corrections to the electronic environment of a molecule.

As one advances through the hierarchy, i.e., CBH-0 → CBH-1 → CBH-2, etc., the protocol systematically generates fragments of increasing size, with CBH-0 units consisting of a single heavy atom, CBH-1 units consisting of two heavy atoms, and CBH-2 units consisting of one heavy atom along with all heavy atoms in its immediate bonding environment. CBH improves the accuracy of QM calculation by exploiting the systematic cancellation of error that occurs in reactions that balance local chemical environments.

The CBH protocol for calculating accurate energies can be extended to calculating accurate reaction energies. For example, the CBH-2 protocol for an ionization process is shown in Figure 4. For the molecular processes considered in this work, the initial and final structures differ only by a single electron. This similarity in structure leads to a cancellation of fragments between reactants and products, and only a small number of high-level calculations of fragment molecules are needed, providing a considerable computational advantage. The CBH correction for ionization after elimination of common fragments is illustrated in Figure 4b. Our previous implementation of CBH-Redox required a chemist's expertise in order to determine the most appropriate fragment to undergo oxidation.[14] In the current work, we present an automated method for identifying sites of oxidation, expanding the applicability of our protocol. Specifically, the most likely site for electron loss is determined by taking the difference of atomic populations for the neutral and ionized species. In this way, a population difference map is created whereby electron loss can be localized to atoms (Figure 5). The atom which experiences the greatest loss in electron population serves as the center of the CBH-2 fragment which undergoes ionization. If atoms have identical populations, as in Figure 5, the atom listed earliest in the xyz file is taken as the site of ionization. Natural population analysis (NPA), which has been shown to be less sensitive to basis set choice, is used for calculating atomic populations.[38,39]

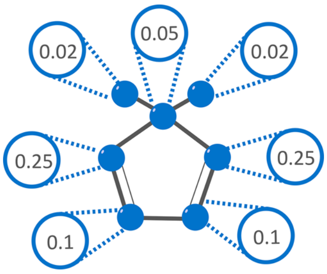


Figure 5. 5,5-dimethylcyclopenta-1,3-diene (with carbon atoms in blue). The numbers in circles show the loss in atomic population (NPA) on the heavy atoms due to ionization (B3LYP-D3BJ).

2.2. Data Set and Electronic Structure Methods.

Molecules from the QM7b data set were chosen as the focus for the present study.[40,41] The QM7b database contains 7,211 molecules with a maximum size of 7 C, N, O, S, and Cl atoms. The size of the molecules is large enough that the CBH-2 approach lends a considerable computational advantage. Using these reference molecules, we compare the performance of DFT, CBH-corrected DFT (CBH-Redox), and ΔML using graph neural networks.

All calculations were performed using the Gaussian 16 suite of programs.[42] G4(MP2), which is known to reproduce experimental thermochemical data within ∼1 kcal/mol was chosen as the reference theory for the calculation of vertical IPs.[43] The B3LYP-D3BJ functional was chosen as the low-level method to test overall performance of CBH-Redox and ΔML.[44−48] In previous studies, adding dispersion slightly improved the performance of CBH.[49] In order to determine the effect of basis set size, we ran low-level calculations with the 6-31G(d) basis set as well as with the larger 6-31G(2df,p) basis set. An unrestricted Kohn−Sham wave function was used for all radical species. Additional calculations were also run at the ωB97X-D/6-31G(2df,p) level of theory to test functional dependence.[50]

Optimizations of the full molecules and the fragments were performed at the B3LYP/6-31G(2df,p) level of theory to match G4(MP2), with frequencies of the full molecules being scaled by 0.9854. This is the same level of theory used to determine the geometries and thermochemical corrections in the G4 family of methods, leading to balanced energy comparisons between DFT and G4(MP2).[51,52] All structures were verified to be local minima. Following optimization, single-point calculations of full molecules and fragments were performed at the low and high levels of theory for both ionized and neutral species. In the present work, we consider only vertical ionization potentials, as opposed to adiabatic ionization potentials which are more relevant for redox potential calculations. As such, only a single optimization was needed for each molecule.

2.3. ΔML: CBH and Electronic State Difference Descriptors.

In a 2021 study, we demonstrated the advantages of our ΔML FragGraph model for predicting atomization energies.[37] In this work, we were able to achieve chemical accuracy, with an out-of-sample mean absolute error (MAE) well below 1 kJ/mol compared to target G4(MP2) calculated energies for molecules in the relatively large QM9 data set of ∼130,000 systems. In the FragGraph model, heavy atoms are represented as nodes in a computational graph, and covalent bonds are represented as edges, although fully connected (FC) graphs are also possible. Local descriptions of each atom's chemical environment are then numerically encoded to form node-wise descriptors. In the above-mentioned study, vector representations of CBH-2 fragments were obtained by passing fragments to the pretrained mol2vec model and were then embedded in nodes in the computational graph.[53] These CBH descriptors run parallel to various other circular fingerprints, i.e., Morgan fingerprints and extended connectivity fingerprints (ECFP).[54−56]

Since there is less data for training in the present study, we elected to simplify our ML model, being cautious of overfitting. Therefore, we simplify our input as well as the architecture itself. First, we describe initial CBH-2 fragments via the traditional atomic features which contribute to the ECFP fingerprint, i.e., atomic number, number of attached atoms, number of attached heavy atoms, number of attached hydrogens, if the atom is in a ring, and if the atom is aromatic, encoded as one-hot vectors. Additional connectivity information encoded in CBH fragments is presumed to be captured via the convolutional steps of the graph network. DFT calculated bond distances were used as edge features. We also decreased the complexity of the ML model, reducing the number of parameters by an order of magnitude.

CBH-2 fragment features based on structure alone proved sufficient for our previous study, in part because atomization energies are highly dependent on structure. Ionization processes, however, have an additional layer of complexity since they involve a change in charge state and multiplicity. We assert that *ML models trained to produce reaction energies*

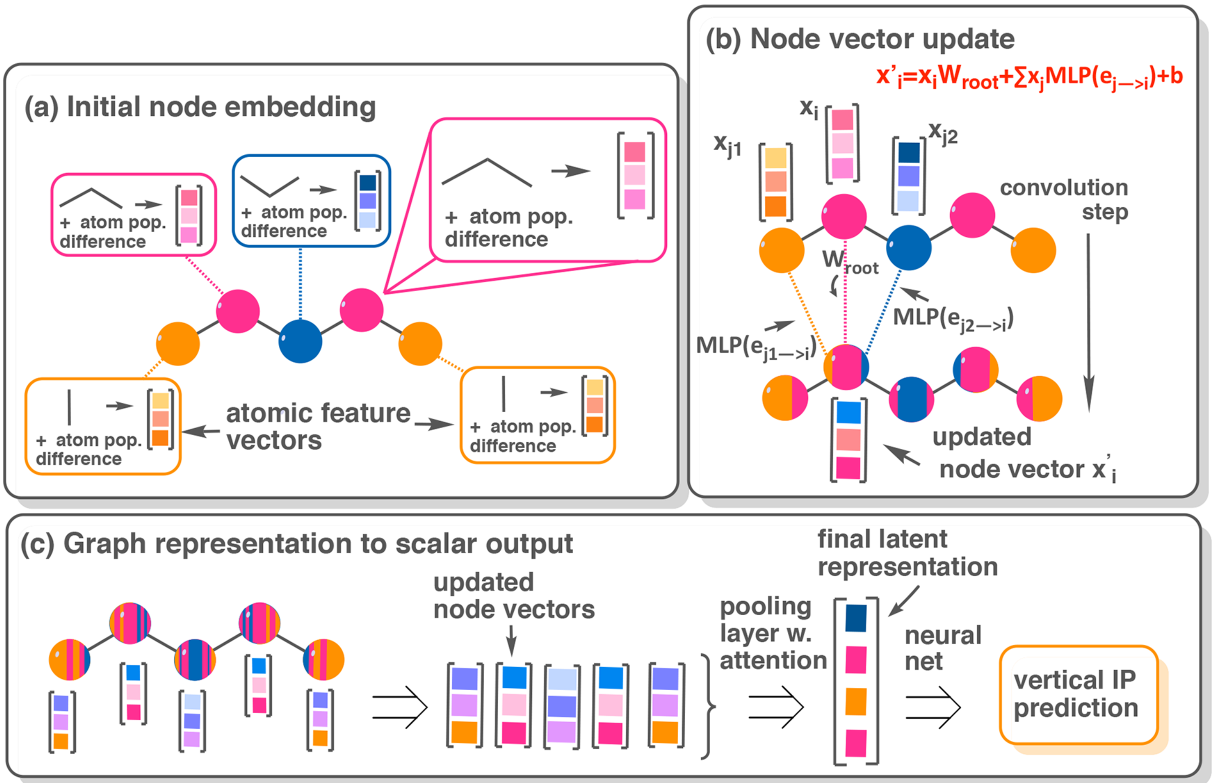


Figure 6. Steps taken in ΔML protocol. (a) Atomic feature vectors containing structural and electronic information embedded in graph network. (b) Illustration of node updates via a graph convolutional step. (c) Illustration of final model layers which take a graph representation with updated node vectors and output a scalar (IP). Color indicates feature vector class, i.e., which CBH-2 fragment and population difference contributes to it. Color striations illustrate information that has been passed between neighbors in convolutional steps.

involving a change in charge/multiplicity may be improved by augmenting its descriptors with electronic information. Our ΔML model requires low-level energy calculations; thus atom- and bond-centered electronic features may be conveniently calculated for little additional computational cost.

Central to any ionization process is the loss of an electron. From the CBH-Redox protocol, we see that an oxidation site may be identified by calculating atomic populations and constructing a charge difference map. Extending our CBH-Redox protocol for ML, we use atomic population differences as systematic attributes to boost learning via ΔML. More specifically, for learning ionization potentials (IP), we enhance our CBH-2 based features with information describing the change in atomic electronic populations (Figure 6a). Electron loss is thus smeared over all atoms in a molecule, rather than on a single fragment. A similar approach is found in the DFT-LOC method for calculating IPs.[57]

2.4. ΔML: Graph Model Architecture and Training Procedure.

The Python Spektral library was used to build the graph network employed for the present study.[58] Spektral is a Python library for graph deep learning, based on the Keras API and TensorFlow 2.[59−61] The model consists of three edge-conditioned convolutional (ECC) layers from the paper by Simonovsky and Komodakis as implemented in Spektral, followed by a global pooling layer and a single dense neural network.[62,63] Batch normalization was used between each ECC layer. Each ECC layer makes vector updates (Figure 6b) for a given node i according to eq 4,

$$x_i' = x_i\mathbf{W}_{root} + \sum_{j\in N(i)} x_j\mathbf{MLP}(e_{j\to i}) + b \tag{4}$$

where x_i contains attributes for node i, $\mathbf{W}_{root}$ is a weight matrix, x_j contains attributes for node j, which is contained in the neighborhood of node i, **MLP** is a multilayer perceptron that outputs an edge-specific weight as a function of edge attributes for the edge connecting node i and j, and b is the bias term. Only covalent bonds were considered edges. Each ECC layer was implemented with a channel dimension of 64, along with the ReLU activation function. The output of the convolutional steps was passed to a GlobalAttentionPool layer from the paper by Li et al. as implemented in Spektral whose output channel dimension was 32.[63] Finally, the output of the pooling layer was passed through a single dense neural network in order to obtain the final IP prediction (Figure 6c). A batch size of 64 was used.

ML models were trained to reproduce the difference between G4(MP2) and DFT calculated IPs, where the IP is calculated as the difference between neutral and cationic energies. The data was split 70:10:20 for training, validation, and testing, respectively. The mean absolute error was chosen as the loss function for model training. The Adamax optimizer in Keras, a variant of Adam, was used in training, and the initial learning rate was set to 0.001.[64] During optimization, if the validation error did not improve by the user-provided threshold (0.0001 kcal/mol) within 25 epochs, then the model weights were reset to those yielding the lowest recorded error on the validation set, and the learning rate was decreased by half. Optimization stopped when the learning rate decreased

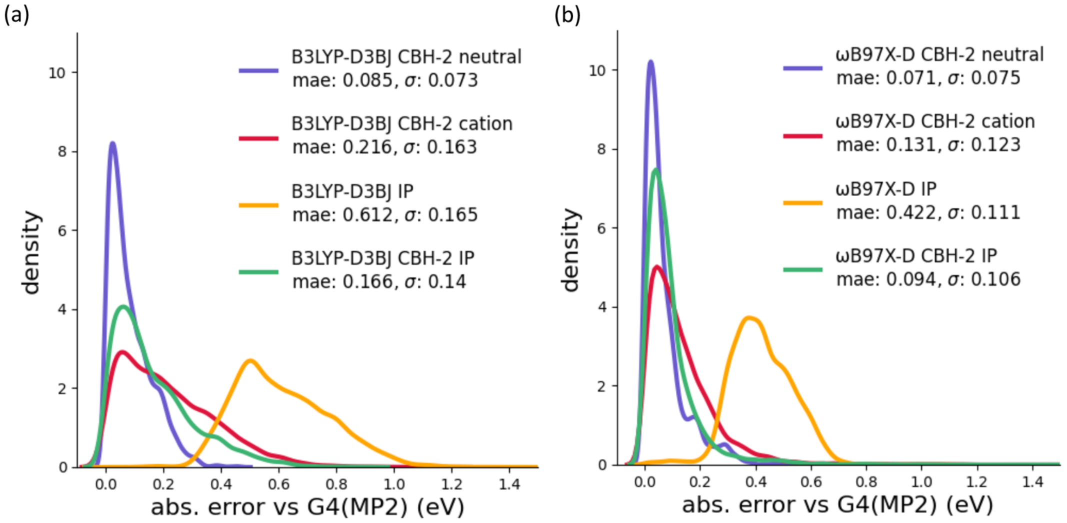


Figure 7. Kernel density estimation plots showing distributions of absolute errors against G4(MP2) reference for CBH-2 corrected neutral energies (purple), CBH-2 corrected cationic energies (red), uncorrected IPs (yellow), and CBH-2 corrected IPs (green) for (a) B3LYP-D3BJ/6-31G(2df,p) and (b) ωB97X-D/6-31G(2df,p) for 7,174 molecules

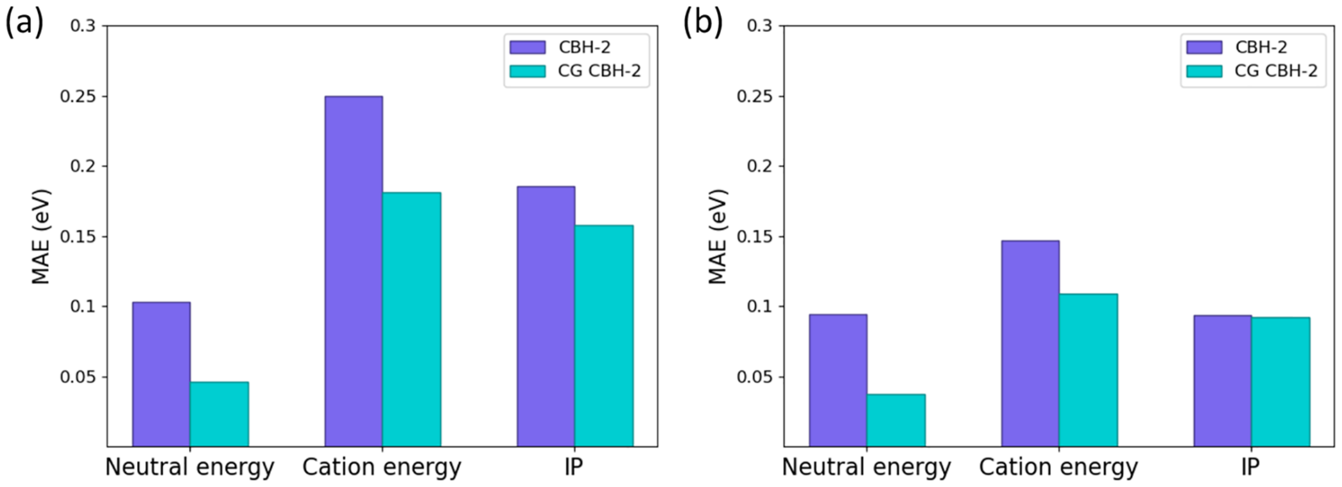


Figure 8. Comparison of MAEs for CBH-2 (purple) and CG-CBH-2 (cyan) for (a) B3LYP-D3BJ/6-31G(2df,p) and (b) ωB97X-D/6-31G(2df,p) for 2,482 molecules.

to 10^{-7}, resulting in ~500−600 training epochs. In order to test the model sensitivity with respect to training data, an ensemble of 10 models was trained, whose members were each trained on a random distribution of 70% of the total data set. Results for individual runs, as well as averages are given below as well as in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. CBH-2 Energies and IPs. Fragments formed from CBH-2, the second rung of CBH, typically provide sufficiently accurate corrections to DFT energies while lowering computational cost. We examined the performance of CBH-2 in correcting QM7b neutral and cation DFT energies, as well as in correcting vertical ionization potentials. 37 species, being too small to perform CBH-2, were omitted during analysis, leaving a total of 7,174 pairs of neutral and cationic molecules, with CBH-2 IPs calculated as the difference between the CBH-2 neutral and cation energies.

Figure 7a illustrates the distribution of absolute errors (eV) for CBH-2 corrected B3LYP-D3BJ/6-31G(2df,p) energies compared to G4(MP2). Errors for the QM7b data set are higher than previously reported benchmarks of CBH-2, which typically range between ~0.04 and 0.09 eV. In general, CBH-2 performs better on neutral species, with a MAE of 0.085 eV. The CBH-2 error for cations is approximately double that of the neutral species. This effect is likely due to a mismatch of charge distribution between fragments and the full molecule. A larger CBH-2 error indicates a departure from strict bond additivity, which is expected to be greater for molecules having a net charge. Nonetheless, CBH-2 decreases the overall MAE of B3LYP IPs versus G4(MP2) by more than 0.4 eV (0.166 vs 0.612 eV).

In previous studies, CBH was shown to have a slight dependence on low level functional choice. Figure 7b shows kernel density estimation plots for absolute errors of CBH-2 corrected ωB97X-D/6-31G(2df,p) energies compared to G4(MP2). Here again CBH-2 neutral energies have less error than CBH-2 cation energies, though poor performance

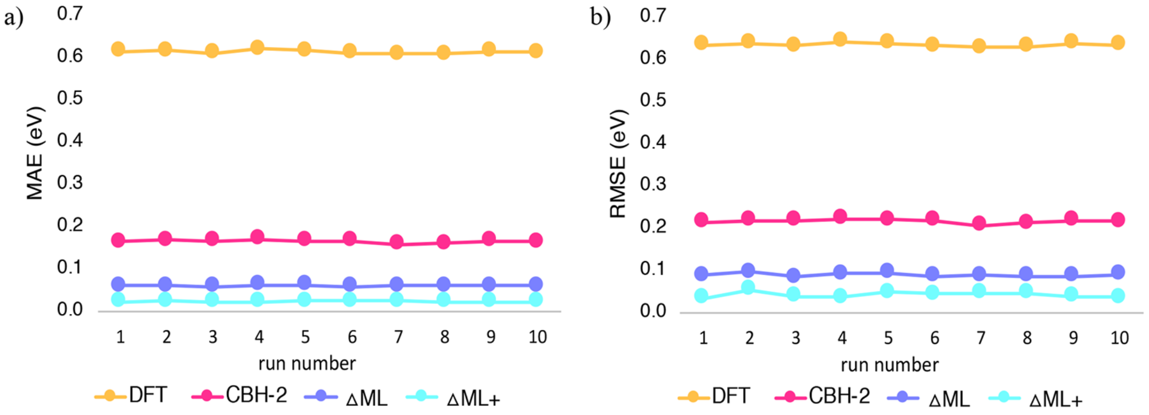


Figure 9. (a) MAE and (b) RMSE plots comparing the performance of uncorrected B3LYP-D3BJ/6-31G(2df,p) (yellow), CBH-2 corrected (pink), ΔML corrected (purple), and ΔML+ corrected (cyan) values for 10 random distributions of 20% of the QM7b data set. For the ML models, this 20% represents the test set.

for cations is far less severe for ωB97X-D. Improvement in cation energies leads to an increase in accuracy for CBH-2 IPs, with a MAE of 0.09 eV. This error falls much closer to the previous CBH-2 benchmark on redox reactions. Even in the absence of CBH corrections, ωB97X-D IPs have a greatly reduced MAE as compared to B3LYP-D3BJ, which likely contributes to improved CBH-2 performance.

We also evaluated CBH's performance when the smaller 6-31G(d) basis set was used in the low level. Overall, results remained similar between basis sets, with MAEs of 0.173 and 0.095 eV for B3LYP-D3BJ/6-31G(d) and ωB97X-D/6-31G(d), respectively, compared to 0.166 and 0.094 eV for B3LYP-D3BJ/6-31G(2df,p) and ωB97X-D/6-31G(2df,p), respectively. This result is useful, indicating that large improvements to accuracy can be made even when lower levels of theory are chosen.

3.2. Coarse-Grained CBH.

One major advantage of CBH is the generality of its fragmentation scheme. However, schemes which cut strong or delocalized bonds may result in fragments unrepresentative of the local bonding environment in the parent molecule. For example, fragments formed by cutting the NO bond of a nitro group may give potentially erroneous energy corrections. Approximately 2,500 molecules of the QM7b data set contain at least one such problematic functional group, namely nitro groups, sulfoxides, nitriles, and alkynes. To mitigate the errors associated with these groups, we adopted a coarse-grained version of CBH (CG-CBH) in which S═O, N═O, C≡N, and C≡C bonds are all kept intact.

The effect of coarse-graining is shown in Figure 8 for CBH-2 corrected B3LYP-D3BJ/6-31G(2df,p) and ωB97X-D/6-31G(2df,p) energies. We see a reduction in error for neutral species, decreasing by ~0.06 eV for both functionals. Using CG-CBH-2, errors in cation energies are reduced by ~0.07 eV and ~0.05 eV for B3LYP-D3BJ and ωB97X-D, respectively. CG-CBH-2 cation energies still display a higher overall error, stemming from a greater mismatch in charge distribution between full molecules and their fragments. Compared to absolute energies, CG-CBH-2 IPs show a smaller overall improvement, with errors decreasing by ~0.03 eV for B3LYP-D3BJ and ωB97X-D errors largely unchanged. It is hypothesized that CBH-2 IPs, being reaction energies rather than total molecular energies, already benefit from a large degree of error cancellation, resulting in a diminished impact from coarse-graining.

3.3. ΔML.

CBH-2 corrects for systematic errors inherent to local chemical units, bringing uncorrected DFT IP errors down substantially. Wishing to push accuracy thresholds further, we introduce a progression of methods aimed at systematically improving DFT IP prediction, incorporating established QM principles and more recent ML techniques. From previous results, we see that CBH offers a systematic and intuitive take on DFT correction schemes, illustrating that much of DFT's failing lies in the treatment of local bonding environments. With this in mind and as a final step in the progression toward chemical accuracy, we next tested the performance of graph neural networks in producing corrections to DFT. By encoding node and edge features, graph networks naturally leverage connectivity information, much like CBH. However, our ML protocol includes an electronic descriptor coming from a population analysis on the full molecule and knows about bond distances rather than bond orders. Thus, our ML-protocol does not have the limitations resulting from the multiple resonance structures of CBH fragments.

Figure 9 compares MAEs and root mean squared errors (RMSE) for uncorrected DFT (B3LYP-D3BJ/6-31G(2df,p)) as well as for several different correction schemes over a series of test runs, each considering 10 random distributions of 20% of the QM7b data set (~1,500 molecules). This 20% represents the test set in the case of ML models. Three correction schemes are shown, namely CBH-2, ΔML, which features only structure-based node descriptors, and ΔML+, which features structure-based node descriptors along with atomic population differences as calculated by DFT on the full molecule. For ΔML(+) models, the remaining portion of the data set was used in training and validation. Thus, the figure illustrates the sensitivity of each scheme to which molecules are included in the test set as well as the sensitivity to which molecules are used in training/validation.

Each subsequent scheme edges closer toward chemical accuracy. Along the increasingly sophisticated series of methods from uncorrected DFT → CBH-2 → ΔML → ΔML+, the average MAEs shrink from 0.611 eV → 0.164 eV → 0.062 eV → 0.024 eV. The RMSE likewise decreases successively. Models which use ML display steady MAEs and RMSEs across runs. Importantly, the incorporation of electronic population difference features reduces the ΔML model MAE by nearly two-thirds, and the RMSE is reduced by more than half. By encoding ionization processes at the atomic (node) level via NPA population differences, model performance improves. Additionally, unlike in CBH, ionization is not localized to one portion of the molecule, giving a more realistic picture of electron loss.

The advantage of using DFT atomic features in the model is highlighted in Figure 10, which shows the distribution of signed errors for uncorrected, ΔML corrected, and ΔML+ corrected DFT IPs for molecules in the test set of run 1. After ΔML corrections are applied, errors are centered around zero and exhibit a narrowed distribution. The spread of errors is decreased even further for ΔML+, illustrating the value of DFT-based descriptors. DFT-based atomic descriptors grant a tremendous advantage during model training as well. Training and validation errors for the ΔML+ model fall significantly below that of ΔML. Moreover, the variance between training and validation MAE is lesser for the ΔML+ model, indicating a more general model. Training and validation curves are provided in Figure S1.

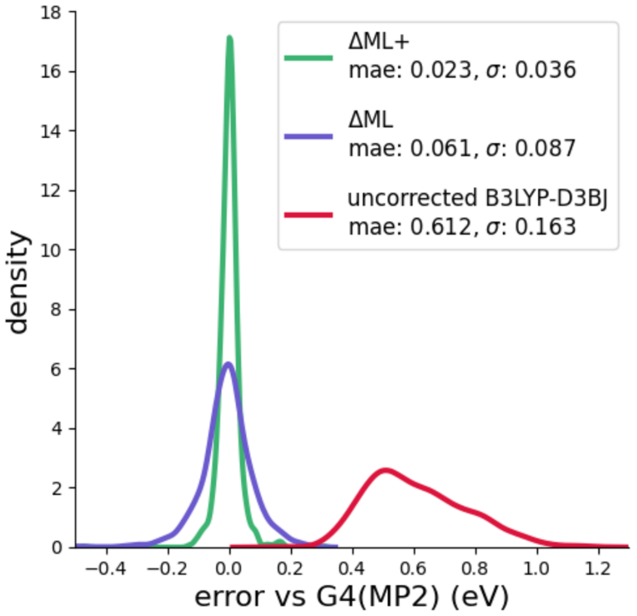


Figure 10. Kernel density estimation plots showing distributions of errors for uncorrected DFT (red), ΔML corrected (purple), and ΔML+ corrected (green) IPs for ~1500 molecules from run 1 test set.

The following analyses provide further insight into the ΔML + model and its predictions. Figure 11a displays the overall correlation between G4(MP2) and ΔML+ corrected DFT IPs for the test set molecules in run 1 (~1,500 molecules). Uncorrected DFT values are shown in gray. ML removes outliers and decreases the overall spread, with ΔML+ corrected IPs having an R^2 value of essentially unity. Most values lie along the diagonal, with the color gradient indicating density of points. Figure 11b displays the correlation in the predicted deltas themselves. It is worth noting that model performance remains consistent across the full range of corrections (~1 eV).

Finally, the functional dependence of ΔML+ was tested. Figure 12 compares the MAEs and RMSEs for uncorrected B3LYP-D3BJ and ωB97X-D IPs as well as for ΔML+ corrected IPs. ΔML+ (B3LYP-D3BJ) and ΔML+ (ωB97X-D) were trained on B3LYP-D3BJ and ωB97X-D calculated IP data, respectively. Error metrics deviate substantially between uncorrected DFT functionals (average MAE across 10 runs of 0.61 eV for B3LYP-D3BJ and 0.42 eV for ωB97X-D). However, the ΔML+ corrected counterparts are well matched (0.024 eV for B3LYP-D3BJ and 0.026 eV for ωB97X-D), which is notable because the raw errors are very different. Moreover, deviations in MAE are no greater than 0.004 eV between the two functionals considering all 10 runs. *Thus, ΔML shows promise in reducing the variation between results calculated with different DFT methods.* Additionally, ML models are not especially sensitive to the choice of basis set, and results supporting this conclusion are shown in Tables S1 and S2. This again is a valuable result, indicating that low-level methods may be used to generate training data without substantial loss in accuracy.

Deep learning models are highly complex and are often considered "black box" methods. As such, techniques which

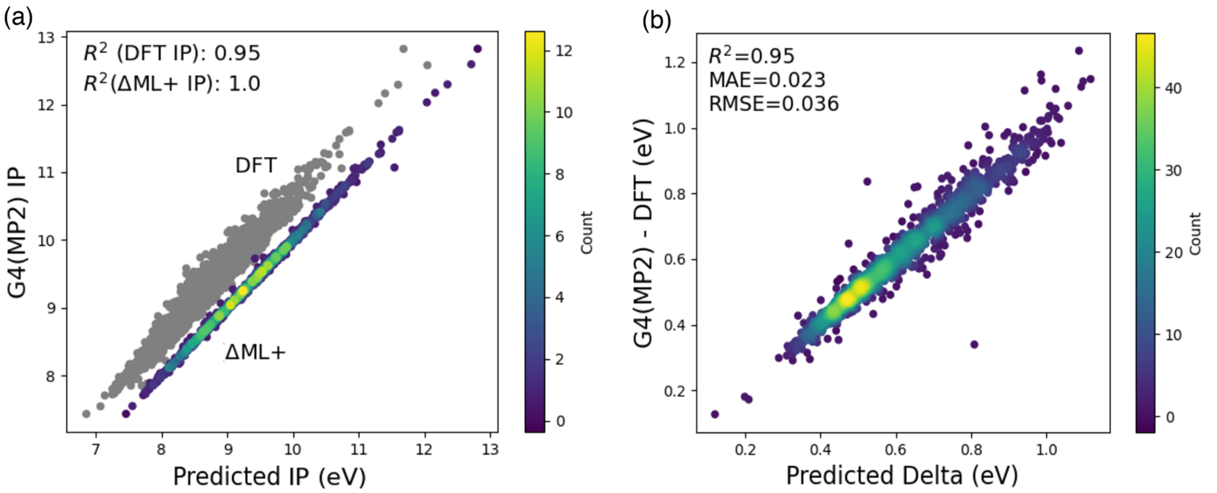


Figure 11. (a) ΔML+ predicted B3LYP-D3BJ/6-31G(2df,p) IPs plotted against G4(MP2) reference IPs (in color). Uncorrected DFT values are plotted in gray for comparison. (b) ΔML+ predicted deltas for DFT (B3LYP-D3BJ/6-31G(2df,p)) IPs versus target correction. ~1500 molecules from the run 1 test set were used in each plot.

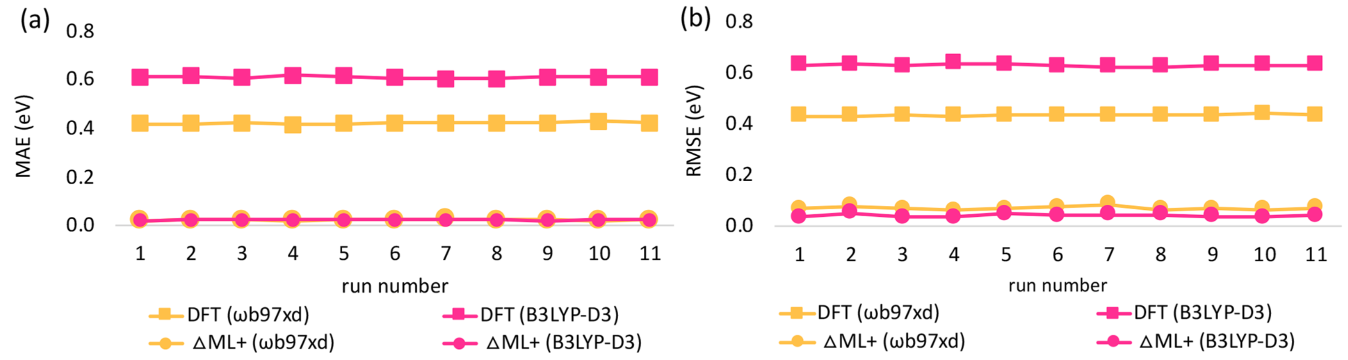


Figure 12. (a) MAE and (b) RMSE plots comparing uncorrected B3LYP-D3BJ/6-31G(2df,p) (pink squares) and ωB97X-D/6-31G(2df,p) (yellow squares) IPs as well as ΔML+ corrected (circles) IPs. The 6-31G(2df,p) basis set was used in all cases shown above.

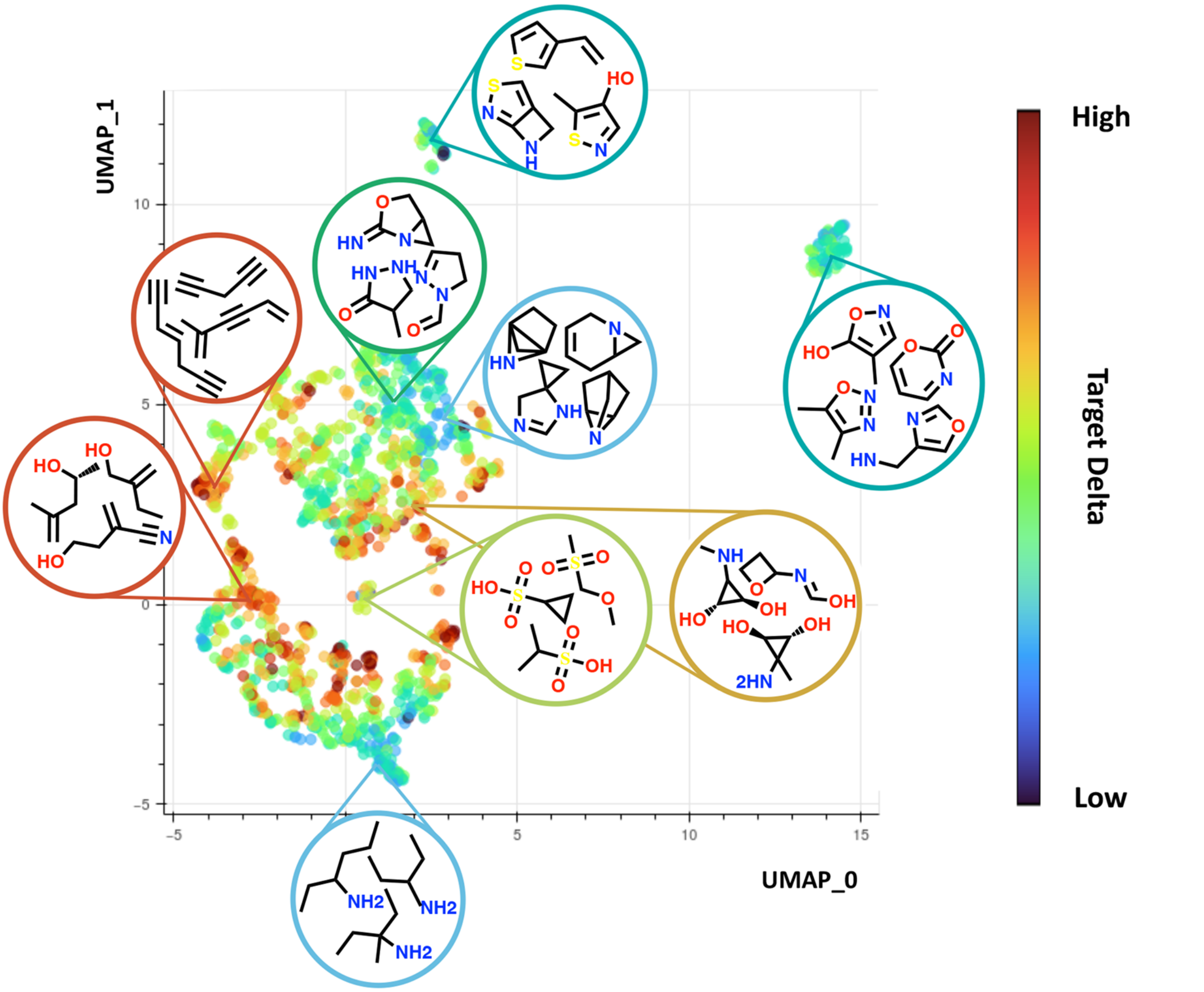


Figure 13. UMAP projection to two dimensions of latent space of last internal layer of ΔML+ network trained with B3LYP-D3/6-31G(2df,p) as the low-level theory. Color indicates size of target delta, with red symbolizing a higher delta than blue.

assist in model interpretability are highly valuable. 2D visualization of the internal network is one way to decipher predictions made by deep learning models. Each intermediate layer of the network, between the input and the output layer, encodes a latent (or hidden) representation of the input. This latent representation expresses learned relationships between data points. However, this high dimensional representation must be condensed before visualization is possible. Thus, we carried out Uniform Manifold Approximation and Projection (UMAP) dimension reduction analysis (n_neighbors = 15, min_dist = 0.2) on the last internal layer of the ΔML+ model in order to visualize the relationships learned by the model.[65]

UMAP is a dimension reduction technique that seeks to preserve a data set's overall structure and can be used to show how a model organizes data prior to output prediction.

Figure 13 shows the UMAP projection of the model's last internal layer to two dimensions for molecules contained in the test set, with color indicating the size of the predicted delta. The latent space organizes the test set according to size of the predicted delta as well as by trends in chemical structure, showing that the network has learned intelligible relationships, and that its internal representation is general enough to capture the structure−activity relationships of unseen molecules. Proximate groupings of data points display similar chemical makeup and have similar predicted corrections. In general, cyclic compounds are found on the upper half of the plot, while acyclic compounds are found on the lower half. Rather distinct clusters are evident for hydrocarbon chains, non-sulfur-containing aromatic heterocycles, sulfur-containing heterocycles, and sulfones. Interestingly, the model manages to separate aromatic heterocycles from nonaromatic heterocycles, with each group making up a totally distinct cluster, despite having similar predicted deltas. Overall, model visualization provides a birds-eye view of the relationship between morphing chemical structure and DFT accuracy.

4. CONCLUSIONS

By integrating concepts from the quantum chemist's toolbox, i.e., molecular fragmentation, systematic error cancellation, and machine learning, chemical accuracy is achievable. Though its performance is often unsatisfactory, DFT error is systematic, providing an excellent springboard to achieve high-level results via systematic correction schemes. A large portion of DFT error can be traced to the incorrect treatment of the local chemical environment, and correction schemes which leverage this fact can be extremely useful to the quantum chemist.

In this study we have shown that by using an electron population difference map, ionization sites within a molecule may be readily identified, and CBH correction schemes for ionization processes may be automated. In addition, we show that the incorporation of electronic descriptors from DFT, namely electron population difference features, improves model performance beyond chemical accuracy (1 kcal/mol) to approach benchmark accuracy. While the raw DFT results are strongly dependent on the underlying functional used, the performance of our best ΔML models is robust and much less dependent on the functional. The sensitivity of the results on the basis set used also appears to be substantially reduced. Finally, as an exciting new tool, ML is capable of capturing the relationship between DFT accuracy and chemical structure.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.2c08821.

> MAEs and RMSEs for all 10 optimization runs calculated with each method, training/validation plots for the ML methods, distribution of error plots, (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Sarah Maier − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States*;
Email: sarmaier@iu.edu

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States*;
orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Author

Eric M. Collins − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States*;
orcid.org/0000-0002-9113-1705

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpca.2c08821

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from the National Science Foundation Grant CHE-2102583 at Indiana University. The Big Red 3 supercomputing facility at Indiana University was used for most of the calculations in this study.

■ REFERENCES

(1) Grimme, S.; Schreiner, P. R. Computational chemistry: The fate of current methods and future challenges. *Angew. Chem., Int. Ed.* **2018**, *57*, 4170−4176.

(2) Christensen, A. S.; Kubar, T.; Cui, Q.; Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **2016**, *116*, 5301−5337.

(3) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.

(4) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. W3 theory: Robust computational thermochemistry in the kj/mol accuracy range. *J. Chem. Phys.* **2004**, *120*, 4129−4141.

(5) Christiansen, O. Coupled cluster theory with emphasis on selected new developments. *Theor. Chem. Acc.* **2006**, *116*, 106−123.

(6) Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *WIRES Comput. Mol. Sci.* **2016**, *6*, 292−310.

(7) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Headgordon, M. A 5th-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.

(8) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular orbital theory of electronic structure of organic compounds. 5. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796−4801.

(9) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. *Theoret. Chim. Acta.* **1975**, *38*, 121−129.

(10) Pieniazek, S. N.; Clemente, F. R.; Houk, K. N. Sources of error in DFT computations of C-C bond formation thermochemistries: Pi ->sigma transformations and error cancellation by DFT methods. *Angew. Chem., Int. Ed.* **2008**, *47*, 7746−7749.

(11) Wheeler, S. E.; Houk, K. N.; Schleyer, P. V. R.; Allen, W. D. A hierarchy of homodesmotic reactions for thermochemistry. *J. Am. Chem. Soc.* **2009**, *131*, 2547−2560.

(12) Ramabhadran, R. O.; Raghavachari, K. Theoretical thermochemistry for organic molecules: Development of the generalized connectivity-based hierarchy. *J. Chem. Theory. Comput.* **2011**, *7*, 2094−103.

(13) Debnath, S.; Sengupta, A.; Raghavachari, K. Eliminating systematic errors in DFT via connectivity-based hierarchy: Accurate bond dissociation energies of biodiesel methyl esters. *J. Phys. Chem. A* **2019**, *123*, 3543−3550.

(14) Maier, S.; Thapa, B.; Raghavachari, K. G4 accuracy at DFT cost: Unlocking accurate redox potentials for organic molecules using systematic error cancellation. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4439−4452.

(15) Sengupta, A.; Raghavachari, K. Prediction of accurate thermochemistry of medium and large sized radicals using connectivity-based hierarchy (CBH). *J. Chem. Theory Comput.* **2014**, *10*, 4342−4350.

(16) Thapa, B.; Raghavachari, K. Accurate pKa evaluations for complex bio-organic molecules in aqueous media. *J. Chem. Theory. Comput.* **2019**, *15*, 6025−6035.

(17) Collins, E. M.; Sengupta, A.; AbuSalim, D. I.; Raghavachari, K. Accurate thermochemistry for organic cations via error cancellation using connectivity-based hierarchy. *J. Phys. Chem. A* **2018**, *122*, 1807−1812.

(18) Artrith, N.; Butler, K. T.; Coudert, F. X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry comment. *Nat. Chem.* **2021**, *13*, 505−508.

(19) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(20) Mater, A. C.; Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.

(21) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(22) Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R. Schnet - a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(23) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5*, aav6490.

(24) Zubatyuk, R.; Smith, J. S.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **2021**, *12*, 4870.

(25) Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(26) von Lilienfeld, O. A.; Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nat. Commun.* **2020**, *11*, 4895.

(27) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Muller, K. R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.

(28) Gupta, A. K.; Raghavachari, K. Three-dimensional convolutional neural networks utilizing molecular topological features for accurate atomization energy predictions. *J. Chem. Theory Comput.* **2022**, *18*, 2132−2143.

(29) King, D. S.; Truhlar, D. G.; Gagliardi, L. Machine-learned energy functionals for multiconfigurational wave functions. *J. Phys. Chem. Lett.* **2021**, *12*, 7761−7767.

(30) Qiao, Z. R.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, 124111.

(31) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The delta-machine learning approach. *J. Chem. Theory. Comput.* **2015**, *11*, 2087−96.

(32) Ruth, M.; Gerbig, D.; Schreiner, P. R. Machine learning of coupled cluster (t)-energy corrections via delta (delta)-learning. *J. Chem. Theory Comput.* **2022**, *18*, 4846−4855.

(33) Sun, G.; Sautet, P. Toward fast and reliable potential energy surfaces for metallic pt clusters by hierarchical delta neural networks. *J. Chem. Theory Comput.* **2019**, *15*, 5614−5627.

(34) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546−1559.

(35) Friesner, R. A. Ab initio quantum chemistry: Methodology and applications. *P. Natl. Acad. Sci. USA* **2005**, *102*, 6648−6653.

(36) Mata, R. A.; Suhm, M. A. Benchmarking quantum chemical methods: Are we heading in the right direction? *Angew. Chem., Int. Ed.* **2017**, *56*, 11011−11018.

(37) Collins, E. M.; Raghavachari, K. A fragmentation-based graph embedding framework for qm/ml. *J. Phys. Chem. A* **2021**, *125*, 6872−6880.

(38) Philips, J. J.; Hudspeth, M. A.; Browne, P. M.; Peralta, J. E. Basis set dependence of atomic spin populations. *Chem. Phys. Lett.* **2010**, *495*, 146−150.

(39) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural-population analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.

(40) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−3.

(41) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Muller, K.-R.; Anatole von Lilienfeld, O Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

(42) Frisch, M. J.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 16*, revision a.-03; Gaussian, Inc.: Wallingford, CT, 2016.

(43) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **2007**, *127*, 124105.

(44) Becke, A. D. Density-functional thermochemistry. 1. The effect of the exchange-only gradient correction. *J. Chem. Phys.* **1992**, *96*, 2155−2160.

(45) Becke, A. D. Density-functional thermochemistry. 5. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.* **1997**, *107*, 8554−8560.

(46) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements h-pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(47) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.

(48) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B* **1988**, *37*, 785−789.

(49) Sengupta, A.; Raghavachari, K. Solving the density functional conundrum: Elimination of systematic errors to derive accurate reaction enthalpies of complex organic reactions. *Org. Lett.* **2017**, *19*, 2576−2579.

(50) Chai, J. D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.

(51) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gn theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 810−825.

(52) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(53) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(54) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *10*, 283−299.

(55) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(56) *Pipeline pilot*, version 7.5; Accelrys, Inc.: San Diego, CA, 2000.

(57) Li, G. Q.; Rudshteyn, B.; Shee, J.; Weber, J. L.; Coskun, D.; Bochevarov, A. D.; Friesner, R. A. Accurate quantum chemical calculation of ionization potentials: Validation of the DFT-loc approach via a large data set obtained from experiments and benchmark quantum chemical calculations. *J. Chem. Theory Comput.* **2020**, *16*, 2109−2123.

(58) Grattarola, D.; Alippi, C. Graph neural networks in tensorflow and keras with spektral. *Ieee Comput. Intell M* **2021**, *16*, 99−106.

(59) Abadi, M.; Barham, P.; Chen, J. M.; Chen, Z. F.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. *Proceedings of Osdi'16:12th Usenix Symposium on Operating Systems Design and Implementation* **2016**, 265−283.

(60) Chollet, F.; Keras, O. https://github.com/fchollet/keras (accessed 04/01/20).

(61) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. Tensorflow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/ (accessed 04/01/20).

(62) Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. *Proc. IEEE CVPR* **2017**, 29−38.

(63) Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated graph sequence neural networks. *CoRR* **2015**, *abs/1511.05493*

(64) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. In *Proc. ICLR* **2015**.

(65) McInnes, L. H. J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426 (accessed 2022-10-01).

■ NOTE ADDED AFTER ASAP PUBLICATION

This article originally published with an incorrect version of Figure 4 and the wrong Figure 5 file. The correct files published April 6, 2023.