JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

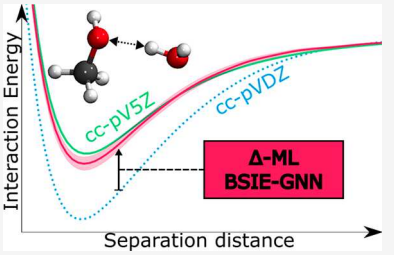
Single-Point Extrapolation to the Complete Basis Set Limit through Deep Learning

Soren Holm, Pablo A. Unzueta, Keiran Thompson, and Todd J. Martínez*

Cite This: https://doi.org/10.1021/acs.jctc.2c01298 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) offers an attractive method for making predictions about molecular systems while circumventing the need to run expensive electronic structure calculations. Once trained on ab initio data, the promise of ML is to deliver accurate predictions of molecular properties that were previously computationally infeasible. In this work, we develop and train a graph neural network model to correct the basis set incompleteness error (BSIE) between a small and large basis set at the RHF and B3LYP levels of theory. Our results show that, when compared to fitting to the total potential, an ML model fitted to correct the BSIE is better at generalizing to systems not seen during training. We test this ability by training on single molecules while evaluating on molecular complexes. We also show that ensemble models yield better behaved potentials in situations where the training data is insufficient. However, even when only fitting to the BSIE, acceptable performance is only achieved when the training data sufficiently resemble the systems one wants to make predictions on. The test error of the final model trained to predict the difference between the cc-pVDZ and cc-pV5Z potential is 0.184 kcal/mol for the B3LYP density functional, and the ensemble model accurately reproduces the large basis set interaction energy curves on the S66x8 dataset.



INTRODUCTION

Today, we can readily compute properties of molecular systems with up to hundreds of atoms using ab initio methods such as Hartree−Fock (HF) or Kohn−Sham Density Functional Theory (DFT). These methods have proven very successful in computational chemistry, and DFT is, in many cases, the preferred method due to its pragmatic accuracy/cost tradeoff. The computational cost of most ab initio methods scales poorly with the number of basis functions, and one is therefore often forced to compromise on the quality of the basis set to stay within a certain cost limit for a given system. Some strategies to circumvent this steep computational scaling while still using large basis sets include incremental methods (e.g., the focal point method)[1−3] or the many-body expansion scheme (i.e., partitioning the target system into more manageable molecular fragments).[4−6]

In practice, any finite basis set is incomplete, and the use of a finite basis set introduces a basis set incompleteness error (BSIE).[7] This BSIE is especially vexing for the atom-centered Gaussian basis sets in widespread use (and to which we confine our attention in the following), while it tends to be more well-behaved for plane-wave basis sets. For sufficiently large basis sets, such as the quintuple zeta cc-pV5Z basis set, the BSIE is well-behaved and its effect may be considered negligible.[8,9] Such large basis sets are usually prohibitively expensive, and most work uses smaller or even minimal basis sets. Unfortunately, these smaller basis sets can introduce significant error.

Computational chemists have long sought to correct for the BSIE of small basis set calculations, and many schemes such as Feller-type basis set extrapolation[10] (eq 1) have been proposed to extrapolate the energies from a series of small basis sets to the complete basis set (CBS) limit.[7,11−17] These extrapolation methods work by considering an assumed functional form expressing how the electronic energy depends on the basis set cardinal number; for example,

$$E_X = E_{\mathrm{CBS}} + \alpha e^{-\beta X} \tag{1}$$

where X is the basis set cardinal number, E_X and E_{CBS} are the electronic energies using a basis with cardinal number X and a complete basis set, respectively. Using a series of calculations with progressively larger basis sets, the assumed functional form enables extrapolation to the CBS limit.

The major drawback of extrapolation schemes is that they require calculations with large basis sets, typically of triple-ζ quality or greater.[18,19] Hence, in situations where one is limited to a smaller basis set and the need to correct the BSIE is the greatest, extrapolation is not an available option for correcting the BSIE.

For basis set corrections to become routinely available, the correction must cost less to compute than the small basis set calculation. Such a method could be of great utility, since it

Special Issue: Machine Learning for Molecular Simulation

Received: December 22, 2022

would allow computational chemists to study larger systems without having to worry about small basis set effects. An important early step in this direction was the geometric counterpoise (gCP) method.[20] The gCP method assumes an explicit functional form for the BSIE, depending only on element types, atomic coordinates, and a set of parameters determined by fitting. Machine learning (ML) is finding widespread use in computational chemistry, because there is a general need for computing molecular properties much less expensively than ab initio methods would otherwise allow. In the Δ-ML[21−25] approach, the goal is to use Machine Learning (ML) techniques to fit the difference between an inexpensive low-level and a more expensive high-level computation. In this paper, we explore a Δ-ML approach to correct the BSIE for HF or DFT.

This strategy imposes an important question: Why try to fit the BSIE rather than the complete potential itself as defined with a large basis set? Since there is already an extensive amount of work published that shows modern ML techniques fitting total potentials to a high accuracy, fitting the difference between two potentials might seem like more work than necessary. However, it is possible that the error of any computational method has a simple structure that is easier to learn than the total potential itself. It is even possible that if the error is divided into its constituent pieces, such as the BSIE and the correlation error, then the individual constituents themselves are even easier to learn. Indeed, the success of gCP suggests that this is true for the present case of BSIE. It has also been shown that Δ-ML schemes can learn these components with either less data or simpler ML techniques,[23,26] while also yielding corrections that are more robust than when trying to learn the total potential.

ML methods for potential energy surface (PES) fitting is a field in rapid development. Since 2007 with the introduction of the Behler−Parinello[27] network, the most successful approaches utilize artificial neural networks (NNs). In the Behler−Parinello approach, one constructs a descriptor that captures the local environment of each atom in a way that respects the translational and rotational symmetries by using atom-centered symmetry functions.[28−32] The descriptors are then used as the input to a feed-forward NN, and the final prediction is taken as either the sum or the mean of the atomic contributions. While there is current debate on the best descriptors for quantum chemistry,[33−38] an alternative approach is to use the naturally occurring graph structure of molecules to construct feature descriptors. Much of the recent development in the field has focused on constructing graph neural network[39,40] (GNN) models that respect these symmetries.

GNNs for learning molecular properties became widespread when the "message-passing" framework was introduced in 2017.[41] In this framework, "messages" of information are exchanged between adjacent atoms in the graph structured input representation, with the goal of constructing atomic descriptors meaningful to the final prediction task. Many different GNN models have been proposed that differ in how the messages are constructed, how the messages are aggregated, and how the aggregated messages are used to update the atomic representations. Two of the more successful models for PES fitting are PhysNet[42] and SchNet,[43] which use only nuclear charges and interatomic distances as input. More-advanced GNNs that utilize angular information also exist such as DimeNet,[44] and recently equivariant GNNs such as E(n)-GNN[45−48] have proven useful for predicting vectorial quantities, such as forces.

In this work, we demonstrate the use of GNNs to provide a geometric correction for the BSIE. We start by building a simplified GNN model based on SchNet. We then train this model on an in-house dataset with DFT and HF energies calculated in a sequence of basis sets (STO-3G, cc-pVDZ, cc-pVTZ, cc-pVQZ, and cc-pV5Z). Fitting our model directly to electronic energy in the cc-pV5Z basis set leads to large errors. However, we find that using Δ-ML significantly reduces errors and standard deviations. After the model is shown to sufficiently outperform traditional extrapolation schemes, we move on to constructing an ensemble model to correct BSIE for S66x8 and large hydrocarbon chains.

METHODS

Our model applies a GNN to correct the BSIE based on molecular geometry. Predicting the BSIE correction of molecular systems is related to potential energy surface (PES) fitting, and the GNN model structure inherits directly from the SchNet and PhysNet models. The model takes atomic numbers ($\mathbf{Z}$) and coordinates ($\mathbf{R}$) as input, and outputs the sum of two contributions. The first contribution, $E_{\text{AtomConst}}$, adds up a constant atomic contribution specific to each element. This part is intended to capture the linear component of the BSIE between basis sets, which will allow the GNN to fit exclusively to the nonlinear geometry-dependent part of the BSIE. The constant atomic contributions are found through linear regression over the training dataset, and these weights are frozen while the GNN fits to the residual BSIE. The model is then trained to predict the BSIE between a smaller basis set (e.g., cc-pVDZ) and the target basis set cc-pV5Z:

$$\tilde{E}_{\text{cc-pV5Z}}(\mathbf{Z}, \mathbf{R}) = E_{\text{cc-pVDZ}}(\mathbf{Z}, \mathbf{R}) + E_{\text{ML}}(\mathbf{Z}, \mathbf{R}) \tag{2}$$

$$E_{\text{ML}}(\mathbf{Z}, \mathbf{R}) = E_{\text{AtomConst}}(\mathbf{Z}) + E_{\text{GNN}}(\mathbf{Z}, \mathbf{R}) \tag{3}$$

The nonlinear part of the BSIE, E_{GNN}, is computed by the GNN. This part of the model uses the atom-type information along with geometric information and operates on a graphic representation $G = (A,E)$ of the system with a predefined connectivity. The atoms of the system constitute the nodes of the graph, and the atomic number is the initial node-associated information $A = [Z_1,Z_2,\cdots,Z_N]$. In the graphic representation, there is an edge between all pairs of atoms that are separated by a distance, r_{ij}, less than the cutoff distance R_c. For the final model, the cutoff distance is chosen to be 6 Å. Associated with each edge, there is an edge embedding, which encodes the distance E separating the atoms ($E = [e_{ij}|r_{ij} < R_c]$.

The embeddings e_{ij} are constructed using $N_r = 32$ equally spaced radial basis functions (RBFs) centered between 0 and the cutoff distance R_c (eq 5). The functional form of the RBFs is a product between a Gaussian and a cutoff function (eq 6), as in the Behler−Parinello[27] network.

$$e_{ij} = [U_1(r_{ij}) \cdots U_{N_r}(r_{ij})] \tag{4}$$

$$U_k(r_{ij}) = f_c(r_{ij}) \times e^{-\eta(r_{ij}-R_k)^2} \tag{5}$$

$$f_c(r_{ij}) = 0.5 \times \left(\cos\left(\frac{\pi r_{ij}}{R_c}\right) + 1\right) \qquad r_{ij} \in [0, R_c] \tag{6}$$

The R_c and N_r hyperparameters were tested across a range of values, and no improvement to the validation error was

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

observed when making them larger than the chosen values. The width parameter η of the RBFs was set to ${N_r}^2/{R_c}^2$ for the chosen number of basis functions and cutoff radius (η = 28.44 Å^{-2} in our model). The centers and widths of the RBFs are not optimized during model training.

To predict the geometry-dependent BSIE contribution, the GNN model relies on higher-order atom embeddings that include geometric information. We used an embedding layer (Figure 1A) to map atomic numbers ($\mathbf{Z}$) to initial zeroth-order

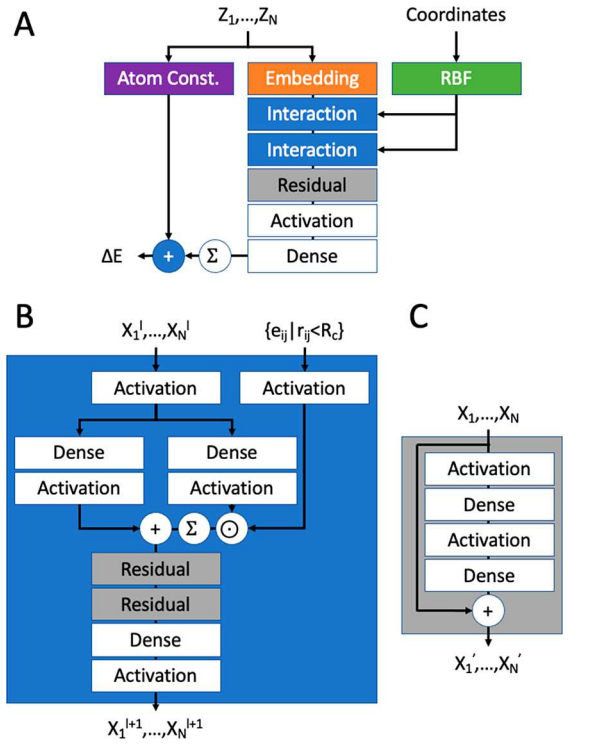


Figure 1. Overview of the GNN model structure applied in this paper. (A) Overall structure of the model; the GNN part of the model consists of two passes through interaction blocks. (B) Structure of the interaction block; the interaction block applied closely resembles the interaction block from the PhysNet model, with the biggest difference being the absence of the gated skip connection. (C) Structure of the residual block as it is used throughout the model.

atom embeddings ($\mathbf{X^0}$), following SchNet and PhysNet. The interaction blocks that follow are responsible for incorporating geometric information into the higher-order atom embeddings ($\mathbf{X^i}$, $i > 0$). All atom embeddings have the same dimension, and for this model, the dimensionality was chosen to be N_d = 32.

The design of the interaction block closely follows PhysNet. In Figure 1B, the algorithm can be understood by following an atom center with the index i. After an initial preactivation, the path of the atom embeddings splits into two. The atom we are following, x_i, follows the left-side path. Its immediate neighbor atoms, $\{x_j | j \in N(i)\}$, follow the right-side path.

The edge embeddings enter the right-side path of the interaction block to construct the neighborhood representation. The edge embedding is used to form an attention mask through a learnable linear transformation A. The attention mask is then applied to the corresponding intermediate embeddings of adjacent atoms through the Hadamard product $\tilde{x}_j \circ Ae_{ij}$. Next, the dimensionality of the neighborhood information (the right-side path) is reduced through element-wise summation of the intermediate embeddings over the atoms. Finally, through element-wise addition, the intermediate center atom embedding, $\tilde{x}_i$ (left-side path), combines with the neighborhood information.

In PhysNet and our model, many of the skip connections are made using residual blocks[49] as depicted in Figure 1C. When training deep neural networks, residual blocks have been shown to improve the training performance by allowing the gradient to flow unimpeded throughout the model. If the only path of gradient flow to early parameters in the model is through all of the later dense layers, then the model can be difficult to train, due to the vanishing/exploding gradient problem.[50]

Our model is significantly smaller than the original PhysNet, both in terms of the depth of the NN and the number of parameters (19 617 trainable parameters, compared to more than 10^6 trainable parameters in PhysNet). The width of the layers is reduced by a factor of 2 to N_d = 32. Inside the interaction block, there are only two residual blocks instead of three. While the original PhysNet model has two additional residual layers after each interaction block, our model has no such residual blocks. The original PhysNet model is five interaction blocks deep, but our model only has two interaction blocks. The interaction block itself also differs slightly from the original structure. Our interaction block does not have gated skip connections, since these were not found to help our comparatively shallow artificial NN. The Swish[51] activation function was chosen over the shifted-softplus activation used in the original PhysNet model.

All the intermediate dense layers preserve the dimensionality of the atom embeddings. This is, in part, due to the use of skip connections throughout the model. Therefore, the shape of the network is completely defined by the dimensionality of the atom, N_d, and edge embeddings, N_r. The parameters of all matrices, W, are initialized using the Glorot Normal[50] scheme. The bias vectors b are initialized to zero. The matrices A and W_{out} and the vectors b_{out} are initialized to zero. The zeroth-order atom embeddings are initialized with a uniform distribution in the range $[-3^{1/2} \cdots 3^{1/2}]$. All initializations are performed as proposed for the original PhysNet model. The model is implemented in TensorFlow using the Keras interface.[52] We validated our hyperparameter choices for the model by examining the validation errors. More detail is given in the Results section that quantifies our findings.

The model is trained by minimizing the mean squared error (MSE) loss function, L:

$$L = \frac{1}{n}\sum_{i}^{n}(\mathrm{BSIE}_i - y_i)^2 \tag{7}$$

We use the Adam optimizer,[53] with the learning parameters set to those proposed in the original paper, to minimize the loss function. The loss function is evaluated over minibatches, each of which has a size of 64 data points. During training, the model accuracy is monitored on validation data, and training is stopped once the validation accuracy has not improved for 200 epochs. The parameters yielding the best validation error are used as the final model parameters. The model was only evaluated on the test dataset after the final parameters were found.

All data points used for training or evaluation are computed using the density-fitted restricted Hartree−Fock code in Psi4.[54] For each geometry, the energy is computed with the basis sets

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

STO-3G, cc-pVDZ, cc-pVTZ, cc-pVQZ, and, cc-pV5Z.[55] We use the def2-svp-jkfit auxiliary basis for the STO-3G primary basis set and the cc-pVXZ-JKFIT auxiliary basis set for the cc-pVXZ primary basis sets (X = DTQ5).[56] We use cc-pV5Z as the target basis set, due to computational affordability and literature precedents.[57] However, one should modify their target basis set for the desired application. The convergence threshold value is set to 10^{-8} for both the energy change and the RMS of the orbital gradient criteria.

For the purpose of training our model, we constructed an in-house (GDB-BSIE) dataset[58] that consisted of all unique molecules with up to seven heavy atoms from the GDB-11[59,60] and GDB-13[61] databases. The dataset contains more than 13 000 molecules, which are further sampled to yield more than 140 000 data points. For each molecule, we include in the dataset the B3LYP/cc-pVDZ optimized structure as well as 10 additional conformations sampled using molecular dynamics (MD). To sample the conformational space, NVT MD simulations at 1000 K were run with B3LYP/cc-pVDZ, sampling geometries every 100 fs intervals for an additional 10 geometries per molecule. Optimization of molecular structures and MD was performed using TeraChem.[62−64] Initial structures (to seed B3LYP/cc-pVDZ geometry optimization) were obtained from SMILES representations of the molecules by using RDKit[65] and the Merck Molecular Force Field.[66] Some of the high-temperature MD trajectories resulted in bond cleavage or overly stretched bonds. We discarded these geometries (less than 1%) by filtering out structures containing bond lengths more than 1.3 times greater than the equilibrium bond length. All conformations of each respective molecule were found to be within 90 kcal/mol of the structure labeled as the minimum energy geometry (see Figure S9 in the Supporting Information).

The elements represented in the GDB-BSIE dataset are H, C, N, O, F, S, and Cl, and the model trained of this dataset is therefore limited to organic molecules containing only subsets of these elements. If one were to apply this method to molecules containing any other elements, new training data including that element would have to be generated. Furthermore, it is possible that heavier elements have structural motifs accessible to them that require special consideration to correct the BSIE. Such elements would also demand that the BSIE−GNN model can recognize the special cases, and it is possible that the model would need further refinement. Model structures that incorporate higher-order information, such as angles, already exist and might be applicable in this situation. In the Results section, we show how hypervalent sulfur provides a special case, but this case is recognized and treated correctly by the BSIE-GNN, even when relying only on interatomic distances.

The GDB BSIE data set was split by molecular identity into training, validation, and test partitions. Validation and testing were performed on molecules not seen during training. The split ratios were 70% training set, 20% validation set, and 10% test set (see Table 1).

For further testing of the GNN model's performance, we constructed two additional datasets. The first dataset is composed of hydrocarbons (only C and H) up to 15 carbon atoms and intends to test the model's ability to generalize to molecules larger than the ones seen in the GDB development dataset. Specifically, the dataset contains all the straight chain hydrocarbons within the size constraint, and the polycyclic aromatic hydrocarbon (PAH) molecules benzene, naphthalene, phenanthrene, phenalene and anthracene. All these geometries are also optimized at the B3LYP/cc-pVDZ level of theory. The second test dataset is an extension of the S66x8 dataset to include more geometries for each complex. In total, this dataset[67] includes 100 geometries (S66x100) for each complex. We emphasize geometries that are close to the equilibrium distance (r_e) by equidistant spacing of the intermolecular distance in increments of 0.025 from $0.7r_e$ to $3.0r_e$, with additional data points further away from the equilibrium distance ($3.25r_e$, $3.5r_e$, $4.0r_e$, $5.0r_e$, $10.0r_e$, $50.0r_e$, and $100.0r_e$).

Table 1. Number of Molecules and Data Points of the GDB-BSIE Dataset in the Training, Validation, and Test Partitions

	Training	Validation	Test	Total
Number of molecules	9227	2631	1321	13 179
Number of data points	100 537	28 685	14 426	143 648

RESULTS

We begin our discussion by benchmarking Feller-type three-point extrapolation schemes and the GNN models using the GDB-BSIE dataset. From this benchmark test, we note that models based on Δ-ML outperform extrapolation schemes while simultaneously eliminating systematic biases present in Feller-type extrapolation. Using the B3LYP Δ-ML model based on double-ζ basis sets, we further validate our model against the intermolecular dataset, S66x100, and a large hydrocarbon dataset. The Δ-ML models yield predictions within 0.2 kcal/mol of cc-pV5Z results, offering a promising alternative for basis set extrapolation.

COMPARISON OF FELLER CBS EXTRAPOLATION TO GNN MODELS

First, we performed the Feller-type three-point extrapolation as a control to compare each GNN model's performance. The mean absolute error (MAE) is reported for each partition of GDB-BSIE dataset so that a comparison can be made to the GNN model's test performance. A summary of the model performance after training on the GDB dataset is provided in Table 2.

As expected, the Feller-type extrapolation exhibits a large MAE (more than 16 kcal/mol) for both RHF and B3LYP when extrapolating from single/double/triple-ζ to quintuple-ζ ([S,D,T]ζ→5ζ), and the error is too large for this method to be useful. At first glance, the [D,T,Q]ζ→5ζ extrapolation for B3LYP falls short of chemical accuracy with a MAE of 1.31 kcal/mol. This echoes earlier difficulties with conventional extrapolation methods for density functional theory.[68] Figure 2C shows that a large part of this error comes from a relatively small set of outliers, i.e., a systematic bias. We have examined the molecules with large error in detail and these are all hypervalent compounds with sulfur. One could improve the Feller-type extrapolation with a correction term proportional to the number of sulfur atoms in the molecule. However, we did not pursue this further since the Feller-type extrapolation is, in any case, too computationally demanding, especially compared to the GNN correction method that we develop here. If the [D,T,Q]ζ→5ζ extrapolation is performed for RHF, then the method is within chemical accuracy with a MAE of 0.187 kcal/mol with a STD of the absolute errors of 0.288 kcal/mol. This level of accuracy will be the target of our GNN

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 2. Mean Absolute Errors for Predictions of the NN Model and Feller-Type Extrapolation on the GDB-BSIE Partitions (Standard Deviation of Absolute Errors Given in Parentheses)[a]

	Mean absolute error (kcal/mol)		
	TRAINING	VALIDATION	TEST
Feller RHF [S,D,T]$\zeta\rightarrow5\zeta$	16.9 (3.24)	16.8 (3.21)	16.9 (3.14)
Feller RHF [D,T,Q]$\zeta\rightarrow5\zeta$	0.203 (0.367)	0.192 (0.343)	0.187 (0.288)
GNN Model RHF 5ζ	1.87 (1.53)	2.67 (2.37)	2.66 (2.43)
GNN Model RHF Δ(Sζ,5ζ)	0.713 (0.590)	0.963 (0.881)	0.977 (0.916)
GNN Model RHF Δ(Dζ,5ζ)	0.119 (0.0988)	0.151 (0.135)	**0.151** (0.134)
Feller B3LYP [S,D,T]$\zeta\rightarrow5\zeta$	21.0 (3.71)	21.0 (3.83)	20.9 (3.80)
Feller B3LYP [D,T,Q]$\zeta\rightarrow5\zeta$	1.33 (0.666)	1.33 (0.636)	1.31 (0.547)
GNN Model B3LYP 5ζ	1.78 (1.56)	2.37 (2.06)	2.38 (2.15)
GNN Model B3LYP Δ(Sζ,5ζ)	0.670 (0.556)	0.897 (0.818)	0.928 (0.844)
GNN Model B3LYP Δ(Dζ,5ζ)	0.149 (0.123)	0.182 (0.161)	**0.184** (0.163)

[a]The model is separately trained to fit either the BSIE error between cc-pVDZ and cc-pV5Z (Δ(Dζ,5ζ)), the BSIE between STO-3G and cc-pV5Z (Δ(Sζ,5ζ)), or the total potential of cc-pV5Z (5ζ). The Feller type three-point extrapolation targets the cc-pV5Z energy and extrapolates from either cc-pV[D,T,Q]Z ([D,T,Q]$\zeta\rightarrow5\zeta$) or from sto-3G, cc-pVDZ, and cc-pVTZ ([S,D,T]$\zeta\rightarrow5\zeta$). The best performing models are indicated by boldface font.

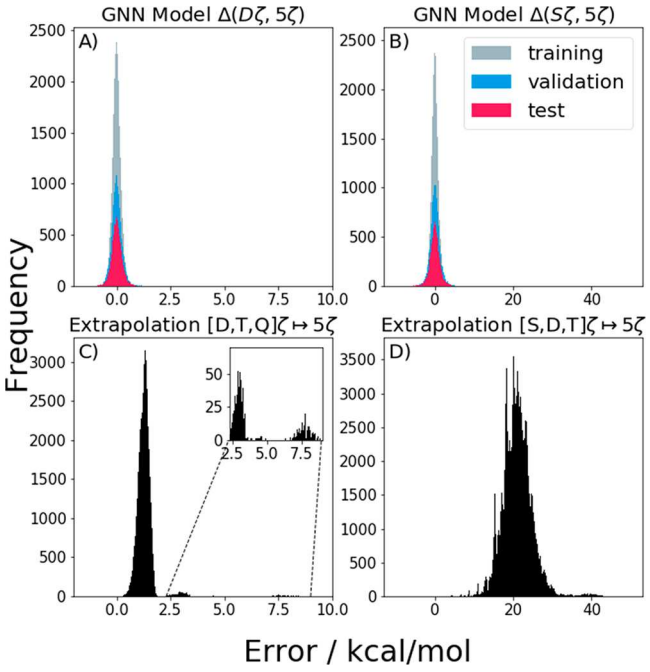


Figure 2. Histogram of the prediction errors of the GNN model (panels (A) and (B)) and the Feller-type extrapolation (panels (C) and (D)) for the B3LYP potential. The first column of plots compares the model trained on the Δ(Dζ,5ζ) incompleteness error and the [D,T,Q]$\zeta\rightarrow5\zeta$ Feller-type extrapolation. The second column compares the NN model trained on the Δ(Sζ,5ζ) incompleteness error and the [S,D,T]$\zeta\rightarrow5\zeta$ Feller-type extrapolation.

correction method, and we would hope the method could achieve this accuracy from a cc-pVDZ or smaller basis set calculation.

Next, we present the performance of the GNN model trained to reproduce the target cc-pV5Z basis set potential energy (i.e., not using Δ-ML approach). We are not able to achieve chemical accuracy training the model directly to either the RHF or B3LYP cc-pV5Z potential energy, with testing errors of 2.66 and 3.35 kcal/mol, respectively (Table 2). Even though these results appear to improve on the Feller-type [S,D,T]$\zeta\rightarrow5\zeta$ extrapolation, they do not achieve chemical accuracy.

The accuracy improves as the Δ-ML model is applied by fitting to the difference between a smaller basis set and the cc-pV5Z one. Training the same model architecture to predict the difference between the single-ζ and quintuple-ζ energies, i.e., the Δ(Sζ,5ζ) incompleteness error, achieves a test MAE of 0.977 and 0.928 kcal/mol for the RHF and B3LYP potentials, respectively. Figure 2B shows the distribution of errors for the GNN-based Δ(Sζ,5ζ) extrapolations and Figure 2D shows the distribution of errors for the Feller-type [S,D,T]$\zeta\rightarrow5\zeta$ extrapolation. Comparing these two distributions, one can see that the GNN model achieves a much narrower error distribution with a STD of the absolute error of 0.844 kcal/mol. Even though the desired level of accuracy (that of the [D,T,Q]$\zeta\rightarrow5\zeta$ extrapolation) is not achieved, this GNN model might be useful in some contexts since it only requires a minimal basis set ab initio calculation.

When we instead train the model to predict the difference between the double-ζ and quintuple-ζ energies, i.e., the Δ(Dζ,5ζ) incompleteness error, MAEs of 0.151 and 0.184 kcal/mol are achieved for RHF and B3LYP, respectively. We tried to further decrease the errors in the RHF Δ(Dζ,5ζ) model by modifying the hyperparameters in our GNN model through N_r, R_c, the atomic attribute dimensionality, and number of interaction blocks. Any modification of N_r or R_c from the original values yielded errors larger than 0.151 kcal/mol. However, we did notice modest improvements by altering the atomic attribute dimensionality and number of interaction blocks. For example, by changing the atomic attribute dimensionality from 32 to 64, the model achieved a MAE of 0.146 kcal/mol. This marginal gain comes at the cost of quadrupling the overall size of the model and was kept fixed at 32 for training efficiency. In addition, increasing the number of interaction blocks from two to three also improves the model to a MAE of 0.146 kcal/mol but increases the number of trainable parameters by 50%. Thus, the number of interaction blocks was also fixed to two. As a final test, we trained the original SchNet model on our GDB-BSIE dataset for RHF Δ(Dζ,5ζ), yielding a MAE of 0.159 kcal/mol. While containing significantly more trainable parameters than our model, it was not able to find a set of parameters that performed dramatically better, which indicates that a MAE of 0.14−0.16 kcal/mol is most likely optimal, based on our dataset.

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

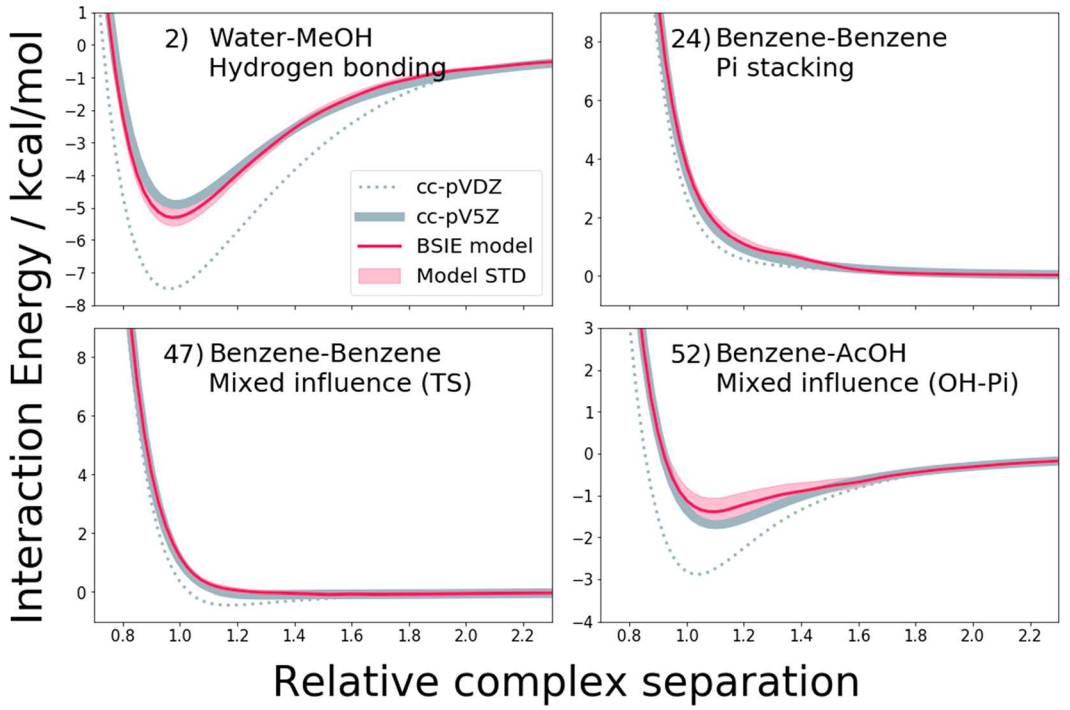


Figure 3. Interaction energy curves of four representative test complexes of the S66x100 dataset. The model correction is obtained from an ensemble of 30 models trained to fit the B3LYP/Δ(Dζ,5ζ) potential of the GDB-BSIE + S66 dataset. The close spaced data points are obtained from the S66x100 extension to the S66x8 dataset.

In the Δ(Dζ,5ζ) setting, the level of accuracy exceeds that of [D,T,Q]ζ→5ζ extrapolation, and achieves chemical accuracy at a much less-expensive computational cost, compared to the extrapolation method. The increased error for the B3LYP prediction compared to RHF is somewhat surprising, since DFT results converge faster than HF or correlated wave function methods.[19] One would assume that the DFT starting point at cc-pVDZ is closer to the converged value. Apparently, this is not so clear and could be related to the difficulties observed in applying conventional extrapolation methods to DFT. It has been suggested that this may be due to basis set dependence in functionals with empirical parameters (i.e., when the functional parameters have been optimized for molecules in a specific basis set).[68,69] Comparing Figures 2A and 2C, we see that the Δ(Dζ,5ζ) GNN model centers the error distribution on zero and eliminates the outliers observed for the [D,T,Q]ζ→5ζ extrapolation. As mentioned above, the outliers in the [D,T,Q]ζ→5ζ extrapolation are molecules containing hypervalent atoms (which are exclusively sulfur atoms in the GDB-BSIE dataset). We finally note that corrections to post-HF methods such as MP2 is outside the current scope of this paper (i.e., one should not expect the trends we observe in solely RHF or DFT to transfer neatly to MP2). We are currently examining the potential for this GNN strategy to correct BSIE in post-HF methods.

S66x8 INTERACTION ENERGIES

The Basis Set Superposition Error (BSSE) is a consequence of basis set incompleteness that manifests as too strong noncovalent interactions and/or too short equilibrium distances. The superposition error arises due to an imbalance in the expressive power of a basis set when comparing across different scales of separation for interacting units. At close range, the basis functions on neighboring noncovalently interacting units are available to atoms in the interacting region, but these functions are not available at longer separations. In the CBS limit, there is no imbalance between the two regimes as the basis is complete, regardless of the complex separation distance.

There exist methods for approximately dealing with the BSSE without performing calculations in large basis sets, and the most famous one is the Boys and Bernardi Counterpoise Correction[70] (CP). The CP method approximately identifies the imbalanced energy contribution by calculating the stabilization of the separate molecular components in the presence of the basis functions of neighboring components. This method is therefore only applicable in situations where the system can be easily separated into molecular components, and it would not be usable for larger systems with intramolecular BSSE. Other methods have been developed, such as Atomic Counterpoise[71] and the gCP[20] method, and those methods are more appropriate for use when there are no clear separable subunits of the overall system. Similarly, our GNN model does not require identification of separable subunits. However, application to molecular complexes traditionally used to study BSSE represents a good test of the model.

We test our ML approach in its ability to correct for the BSSE by applying the GNN model to the S66x8 dataset[72] of molecular complexes. The GNN model did not generalize well to molecular complexes when only trained on the GDB-BSIE dataset (see Figures S4, S5, and S6 in the Supporting Information), but similar deficiencies were observed for SchNet (trained on the QM9 dataset) and the ANI-1 model (trained on the ANI dataset), as shown in Figures S7 and S8 in the Supporting Information. Figure S7 shows the default single ANI-1 model performance for several molecules in the S66x100 dataset, along with the ensemble models provided

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

by TorchANI.[73] Figure S8 shows the same comparisons using the default SchNet model.[74] Both these models were used without any additional training or hyperparameter tuning. The ML potentials generated from these models are qualitatively incorrect, as might be expected since they were not explicitly trained on intermolecular complexes.

We included data points of 50 complexes from our S66x100 dataset in the training/validation dataset, and the remaining 16 complexes are used for testing. The S66x100 data points were repeated to achieve a total weight of 25% within the total training dataset that also contains the GDB-BSIE data. This way, we are upsampling the minority dataset (S66x100) such that we apply more weight to these structures during the training procedure.

While the GNN HF results from Table 2 are encouraging, we carry out our further tests with the B3LYP density functional as a more illustrative use case. Figure 3 shows the B3LYP/Δ(Dζ,5ζ) model corrected interaction energy curves for four representative complexes of the test partition. The correction shown is the mean of an ensemble of models with 30 members, and the ensemble effectively cancels high-frequency components of the potential produced by the individual models (Figures S1 and S2 in the Supporting Information). The overall correction to the interaction energies is generally successful, although one can see a small bump in the potential for benzene−benzene π-stacking (complex 24, at approximately $1.4r_e$). The corrected potential of pyridine-ethyne (complex 65 in Figure S1) displays the largest qualitative deviation from the target potential of all the complexes in the test partition, but the model correction still improves upon the cc-pVDZ potential. Notably, the same approach was applied to fit the Δ(Sζ,5ζ) and 5ζ models. However, the quality of the corrections to the S66 interaction energies degrades significantly, and the use of an ensemble model is not enough to eliminate the deficiencies (see Figures S2 and S3 in the Supporting Information).

In Figure 4, we compare the magnitude of errors in interaction energies (IEs) compared to the B3LYP/cc-pV5Z potential. Figure 4A shows a single model trained on the Δ(Dζ,5ζ) potential of the GDB-BSIE dataset, and Figure 4B shows the errors of the 30-member ensemble model trained on GDB-BSIE and S66x100 data. Although the errors in Figures 4A and 4B look comparable, the quality of corrections obtained from the ensemble model that has seen S66x100 data is significantly better, in terms of the smoothness of the potential (see Figure S4 and S1). The magnitude of the errors of the ensemble model are comparable to those of the CP corrections shown in Figure 4C, but the GNN model is much more broadly applicable since the system does not have to be separable into molecular components. Finally, Figure 4D shows the gCP correction method which produces IE errors significantly larger in magnitude than of the ensemble model.

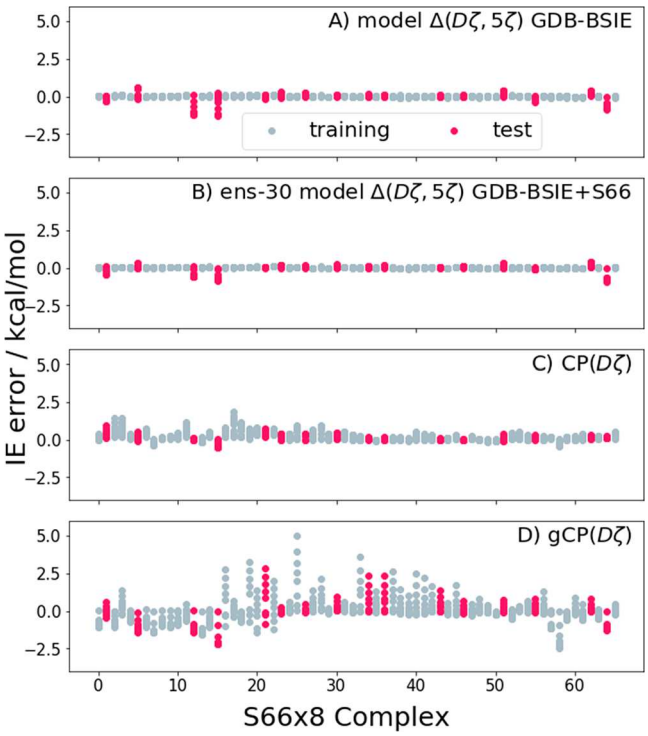


Figure 4. Errors in IEs compared the B3LYP/cc-pV5Z potential. (A) Single model trained on the B3LYP/Δ(Dζ,5ζ) potential of the GDB-BSIE dataset. (B) Ensemble of 30 models trained on the B3LYP/Δ(Dζ,5ζ) potential of the GDB-BSIE + S66 dataset. (C) Counterpoise correction applied to B3LYP/cc-pVDZ basis set calculations. (D) gCP correction applied to B3LYP/cc-pVDZ calculations.

EXAMINING LARGER MOLECULES

The ultimate goal of using NNs for PES fitting is to apply the trained model in a regime where direct electronic structure computations are too expensive. We further test our models on larger systems to gauge the generalizability to regimes previously not seen. The computational cost of a cc-pV5Z calculation limits the energy calculations to rather small molecules, and the GDB-BSIE dataset only has molecules with up to seven heavy atoms. We do include some larger systems with up to 16 heavy atoms from the S66x100 dataset, but this dataset does not add much chemical diversity to the training data. We constructed a test data set of linear hydrocarbons and polycyclic aromatic hydrocarbons (PAHs) of increasing sizes up to 15 carbon atoms, where all molecules containing eight or more carbons are not found in the training data. Figure 5 shows how the GNN model trained on the GDB-BSIE + S66x100 dataset performs on the hydrocarbon test dataset.

In Figure 5, we see that the ensemble model produces marginally larger uncertainties for molecules containing more heavy atoms than the train/validation data, but the accuracy of the ensemble mean prediction is almost unaffected by molecular size. The large uncertainty in the correction for methane is explained by the fact that methane is not seen during training of the models. Outside of the region where the models have seen a lot of data (3−7 heavy atoms), any individual model displays some bias. For the larger linear hydrocarbons, the bias is well-behaved (linear as a function of molecule size), and good corrections can be achieved using ensemble models, which allow the Δ(Dζ,5ζ) ensemble to make chemically accurate predictions for molecules of larger sizes that are not represented in training. Furthermore, the prediction uncertainties for PAHs are much larger than the linear hydrocarbons, but again the ensemble mean is well-predicted and well within chemical accuracy.

CONCLUSION

In conclusion, we have demonstrated the use of GNNs to correct BSIE. We started by developing our own dataset, GDB-BSIE, which consisted of molecules from GDB11 and GDB13 calculated at RHF and B3LYP with basis sets ranging from STO-3G to cc-pV5Z. We show that the Feller-type basis set

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

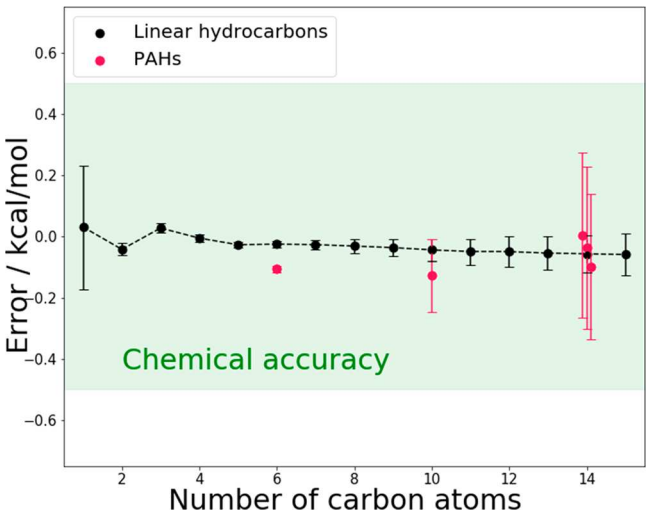


Figure 5. Performance of ensemble GNN models on predicting Δ(Dζ,5ζ) B3LYP potential for larger hydrocarbon molecules. The models are trained on GDB-BSIE + S66x100. The ensemble model is an equally weighted average over all 30 trained models. The error bars are obtained by computing the standard deviation over a thousand unique ensembles of size 10 sampled from the 30 trained models without replacement. The threshold for chemical accuracy (±0.5 kcal/mol) is shown with the green shaded region.

extrapolation works well using [D,T,Q] ζ→5ζ extrapolation for RHF and B3LYP. However, we find that, for our dataset, Feller-type extrapolations contain a systematic shift and large errors for molecules containing hypervalent atoms. We present a Δ-ML scheme, correcting the incompleteness error between the cc-pVDZ and cc-pV5Z basis sets, which yields chemical accuracy on par with or better than the Feller extrapolations for both RHF and B3LYP with small standard deviation in the errors. Furthermore, we show this method of basis set corrections yields more well-behaved potentials than a Δ-ML approach starting from the STO-3G basis set, or the prediction of the total cc-pV5Z potential.

We thoroughly tested our GNN models on two key metrics: intermolecular energies and size extensivity. For the intermolecular benchmarks, we show that individual models yield qualitatively incorrect behavior across different intermolecular separations in the S66x8 dataset when trained solely on GDB-BSIE. Augmenting the dataset with some S66x100 data yields individual models with the correct behavior, but these are not as smooth as expected, with respect to separation distance.

We remedy the nonsmoothness by constructing ensemble models which display the correct behavior in the PES and provide uncertainties for the final predicted potentials. Given the limited data in S66x100, we applied an upsampling procedure to give more weight to the intermolecular complexes during training. Best practices for upsampling minority classes in machine learning is still an open question and improvements here might remove the residual small corrugation seen in some of the intermolecular PESs.[75,76]

We also tested our model in regions of chemical space with larger molecules that are not represented in our training and validation data. We show results with chemical accuracy on long linear hydrocarbons, and polycylic aromatic hydrocarbons up to 15 heavy atoms, albeit with somewhat larger uncertainties as the molecules get larger.

Overall, we have shown that one can achieve cc-pV5Z accuracy for RHF or B3LYP at the computational cost of cc-pVDZ in tandem with a GNN to correct the basis set incompleteness error. The Δ-ML approach shows evidence of more-reliable predictions across chemical space when the training data are very limited, as they are in this case. In the future, we anticipate that one could also extract the learned features from a GNN model which contribute to accurate predictions and try to relate the weights back to a physics-based model. This could allow for the development of a constrained physics-informed model, which would likely yield more well-behaved potentials and more reliable extrapolation from limited sets of training data than current GNNs are able to achieve.

■ ASSOCIATED CONTENT

Data Availability Statement

The following information is available on publicly hosted sites (GitHub and Zenodo): code for the GNN model is available at https://github.com/mtzgroup/BSIE-GNN. Datasets are available on Zenodo at 10.5281/zenodo.7402847 (S66x100 with geometries and HF/B3LYP energies in STO-3G, cc-pVDZ, cc-pVTZ, cc-pVQZ, and cc-pV5Z basis sets) and 10.5281/zenodo.7402871 (GDB-BSIE dataset with geometries and HF/B3LYP energies in STO-3G, cc-pVDZ, cc-pVTZ, cc-pVQZ, and cc-pV5Z basis sets).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c01298.

Potential energy curves from GNN models; energy distribution (relative to local minimum) of sampled structures in GDB-BSIE dataset (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Todd J. Martínez − *Department of Chemistry and The PULSE Institute, Stanford University, Stanford, California 94305, United States; SLAC National Accelerator Laboratory, Menlo Park, California 94024, United States;* orcid.org/0000-0002-4798-8947; Email: todd.martinez@stanford.edu, toddjmartinez@gmail.com

Authors

Soren Holm − *Department of Chemistry and The PULSE Institute, Stanford University, Stanford, California 94305, United States; SLAC National Accelerator Laboratory, Menlo Park, California 94024, United States*

Pablo A. Unzueta − *Department of Chemistry and The PULSE Institute, Stanford University, Stanford, California 94305, United States; SLAC National Accelerator Laboratory, Menlo Park, California 94024, United States;* orcid.org/0000-0002-0371-4805

Keiran Thompson − *Department of Chemistry and The PULSE Institute, Stanford University, Stanford, California 94305, United States; SLAC National Accelerator Laboratory, Menlo Park, California 94024, United States;* orcid.org/0000-0003-3531-5857

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c01298

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (N00014-18-1-2659 and N00014-21-1-2151). P.A.U. acknowledges support by the National Science Foundation MPS-Ascend Postdoctoral Research Fellowship, under Grant No. 2213324.

■ REFERENCES

(1) Csaszar, A. G.; Allen, W. D.; Schaefer, H. F., III In pursuit of the ab initio limit for conformational energy prototypes. *J. Chem. Phys.* **1998**, *108*, 9751−9764.
(2) East, A. L.; Allen, W. D. The heat of formation of NCO. *J. Chem. Phys.* **1993**, *99*, 4638−4650.
(3) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, δ MP2 CCSD (T): Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. *J. Chem. Phys.* **2011**, *135*, 194102.
(4) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: a route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112*, 632−672.
(5) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336−1343.
(6) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, 064113.
(7) Schwenke, D. W. On one-electron basis set extrapolation of atomic and molecular correlation energies. *Mol. Phys.* **2012**, *110*, 2557−2567.
(8) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(9) Wilson, A. K.; van Mourik, T.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon. *J. Mol. Struct. (THEOCHEM)* **1996**, *388*, 339−349.
(10) Feller, D. Application of systematic sequences of wave functions to the water dimer. *J. Chem. Phys.* **1992**, *96*, 6104−6114.
(11) Varandas, A. J. C. Straightening the Hierarchical Staircase for Basis Set Extrapolations: A Low-Cost Approach to High-Accuracy Computational Chemistry. *Annu. Rev. Phys. Chem.* **2018**, *69*, 177−203.
(12) Feller, D. Benchmarks of improved complete basis set extrapolation schemes designed for standard CCSD(T) atomization energies. *J. Chem. Phys.* **2013**, *138*, 074103−074103.
(13) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N_2, and H_2O. *Chem. Phys. Lett.* **1998**, *286*, 243−252.
(14) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639−9646.
(15) Kutzelnigg, W. Theory of the expansion of wave functions in a gaussian basis. *Int. J. Quantum Chem.* **1994**, *51*, 447−463.
(16) Varandas, A. J. C. Basis-set extrapolation of the correlation energy. *J. Chem. Phys.* **2000**, *113*, 8880−8887.
(17) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. Estimates of the ab initio limit for $\pi-\pi$ interactions: The benzene dimer. *J. Am. Chem. Soc.* **2002**, *124*, 10887−10893.
(18) Jensen, F. *Introduction to Computational Chemistry*, 3rd Edition; John Wiley & Sons, 2017.
(19) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc. A* **2014**, *372*, 20120476.
(20) Kruse, H.; Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **2012**, *136*, 154101−154101.
(21) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(22) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. *J. Chem. Theory Comput.* **2021**, *17*, 826−840.
(23) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12*, 4536−4546.
(24) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.
(25) Zhang, P.; Shen, L.; Yang, W. Solvation Free Energy Calculations with Quantum Mechanics/Molecular Mechanics and Machine Learning Models. *J. Phys. Chem. B* **2019**, *123*, 901−908.
(26) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity statistical machine learning for molecular crystal structure prediction. *J. Phys. Chem. A* **2020**, *124*, 8065−8078.
(27) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401−146401.
(28) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(29) 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*, 3192−3203.
(30) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.
(31) 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*, 2903.
(32) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495−4501.
(33) Vassilev-Galindo, V.; Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *J. Chem. Phys.* **2021**, *154*, 094119.
(34) Pozdnyakov, S. N.; Willatt, M. J.; Bartok, A. P.; Ortner, C.; Csanyi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125*, 166001.
(35) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 1−36.
(36) Ramakrishnan, R.; von Lilienfeld, O. A., Machine learning, quantum chemistry, and chemical space. In *Reviews in Computational Chemistry*, Vol. *432*; Parrill, A. L., Lipkowitz, K. B., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017, pp 225−256.
(37) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(38) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.
(39) Kipf, T. N.; Welling, M., Semi-supervised classification with graph convolutional networks. 2016, *arXiv Preprint* arXiv:1609.02907.

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(40) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R. Relational inductive biases, deep learning, and graph networks. 2018, *arXiv preprint* arXiv:1806.01261.
(41) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. In *34th International Conference on Machine Learning*, Sydney, Australia, 2017; pp 1263−1272.
(42) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(43) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722−241722.
(44) Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. 2020, *arXiv preprint* arXiv:2003.03123.
(45) Satorras, V. G.; Hoogeboom, E.; Welling, M. *E*(*n*) Equivariant Graph Neural Networks. In *International Conference on Machine Learning*, PMLR, 2021; pp 9323−9332.
(46) Batatia, I.; Batzner, S.; Kovács, D. P.; Musaelian, A.; Simm, G. N.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. The Design Space of E (3)-Equivariant Atom-Centered Interatomic Potentials. 2022, *arXiv preprint* arXiv:2205.06643.
(47) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.
(48) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics. 2022, *arXiv preprint* arXiv: 2204.05249.
(49) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9908* LNCS; pp 630−645 (2016).
(50) Glorot, X.; Bengio, Y., Understanding the difficulty of training deep feedforward neural networks. *JMLR Workshop and Conference Proceedings*, Vol. *9*; 2010, pp 249−256.
(51) Ramachandran, P.; Zoph, B.; Le, Q. V., Searching for Activation Functions. In *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, 2017.
(52) *mtzgroup/BSIE-GNN*. Available via the Internet at: https://github.com/mtzgroup/BSIE-GNN.
(53) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*, 2014, *1412*, 6980.
(54) Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D. Psi4: an open-source ab initio electronic structure program. *WIRes: Comp. Mol. Sci.* **2012**, *2*, 556−565.
(55) Woon, D. E.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358−1371.
(56) Weigend, F. A fully direct RI-HF algorithm: Implementation, optimized auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285−4291.
(57) Witte, J.; Neaton, J. B.; Head-Gordon, M. Push it to the limit: Characterizing the convergence of common sequences of basis sets for intermolecular interactions as described by density functional theory. *J. Chem. Phys.* **2016**, *144*, 194306.
(58) Holm, S.; Unzueta, P.; Martinez, T. J. *GDB-BSIE Dataset*, DOI: 10.5281/zenodo.7402871.
(59) Fink, T.; Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.
(60) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Ang. Chem. Int. Ed.* **2005**, *44*, 1504−1508.
(61) 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*, 8732−8733.
(62) Ufimtsev, I. S.; Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **2008**, *4*, 222−231.
(63) Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *J. Chem. Theory Comput.* **2009**, *5*, 1004−1015.
(64) Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619−2628.
(65) RDKit: Open-source cheminformatics. https://www.rdkit.org (accessed December 13).
(66) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(67) Holm, S.; Unzueta, P.; Martinez, T. J. *S66x100 Dataset*, DOI: 10.5281/zenodo.7402847.
(68) Kraus, P. Basis set extrapolations for density functional theory. *J. Chem. Theor. Comp.* **2020**, *16*, 5712−5722.
(69) Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn-Sham basis set limit. *J. Chem. Phys.* **2002**, *116*, 7372.
(70) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553−566.
(71) Jensen, F. An atomic counterpoise method for estimating inter- and intramolecular basis set superposition errors. *J. Chem. Theory Comput.* **2010**, *6*, 100−106.
(72) Goerigk, L.; Kruse, H.; Grimme, S. Benchmarking density functional methods against the S66 and S66x8 datasets for non-covalent interactions. *ChemPhysChem* **2011**, *12*, 3421−3433.
(73) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408−3415.
(74) Schütt, K.; Kessel, P.; Gastegger, M.; Nicoli, K.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.
(75) Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Art. Intell.* **2016**, *5*, 221−232.
(76) Johnson, J. M.; Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, DOI: 10.1186/s40537-019-0192-5

https://doi.org/10.1021/acs.jctc.2c01298
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX