
Δ -ML Ensembles for Selecting Quantum Chemistry Methods to Compute Intermolecular Interactions

Austin M. Wallace*

School of Chemistry and Biochemistry
Georgia Institute of Technology
Atlanta, GA 30332-0400
awallace43@gatech.edu

C. David. Sherrill†

School of Chemistry and Biochemistry
Georgia Institute of Technology
Atlanta, GA 30332-0400
sherrill@gatech.edu

Giri P. Krishnan

Center for Artificial Intelligence in Science and Engineering
Georgia Institute of Technology
Atlanta, GA 30308
giri@gatech.edu

Abstract

Ab initio quantum chemical methods for accurately computing interactions between molecules have a wide range of applications but are often computationally expensive. Hence, selecting an appropriate method based on accuracy and computational cost remains a significant challenge due to varying performance of methods. In this work, we propose a framework based on an ensemble of Δ -ML models trained on features extracted from a pre-trained atom-pairwise neural network to predict the error of each method relative to all other methods including the “gold standard” coupled cluster with single, double, and perturbative triple excitations at the estimated complete basis set limit [CCSD(T)/CBS]. Our proposed approach provides error estimates across various levels of theories and identifies the computationally efficient approach for a given error range utilizing only a subset of the dataset. Further, this approach allows comparison between various theories. We demonstrate the effectiveness of our approach using an extended BioFragment dataset, which includes the interaction energies for common biomolecular fragments and small organic dimers. Our results show that the proposed framework achieves very small mean-absolute-errors below 0.1 kcal/mol regardless of the given method. Furthermore, by analyzing all-to-all Δ -ML models for present levels of theory, we identify method groupings that align with theoretical hypotheses, providing evidence that Δ -ML models can easily learn corrections from any level of theory to any other level of theory.

1 Introduction

Accurate quantum mechanical (QM) computations of intermolecular interactions are valuable to identify the most probable crystal structure for organic molecules,[1, 2] understanding protein-ligand interactions involved in binding,[3, 4], modeling nucleotide stacking,[5, 6] and developing intermolecular force-fields.[7–9] Although many methods exist to compute interaction energies, the trade-off of accuracy and computational cost drives the choice of specific pairings of methods and

*Center for Computational Molecular Science and Technology and School of Chemistry and Biochemistry

†Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry, and School of Computational Science and Engineering

basis sets for quantum mechanical calculations. Any specific method/basis set pair is called the level of theory. CCSD(T)/CBS[10] is considered the gold standard level of theory for interaction energies,[11]; however, it scales as $\mathcal{O}(N^7)$, making it very expensive.

Within QM, the interaction energy quantifies how attractive or repulsive two molecules are to each other. More formally, the interaction energy can be defined in a supermolecular approach through

$$\Delta E_{\text{int}} = E_{IJ} - E_I - E_J, \quad (1)$$

where IJ represents the energy of a dimer while I and J represent the energies of the isolated monomers. The types of non-covalent interactions that impact the interaction energy are electrostatics, van der Waals forces, hydrogen bonds, exchange-repulsion—akin to steric energies—and polarization.

QM interaction energies are quite sensitive to electron correlation, basis set size, and counterpoise corrections (CP).[12] Consequently, predicting interaction energies from lower levels of theory, such as Hartree-Fock (HF), can lead to significant errors, while sometimes inexpensive methods relying on error cancellation like SAPT0/jun-cc-pVDZ can yield reasonably accurate results in certain chemical systems, while failing at others, like $\pi - \pi$ aromatic systems.[13] For small systems, high-accuracy methods like CCSD(T)/CBS can be computed; however, the scaling of $\mathcal{O}(N^7)$ makes these methods intractable for most practical applications. Hence, high-throughput screening approaches largely rely on the most inexpensive QM methods like HF, MP2, or DFT even at the cost of accuracy. With hundreds of levels of theory available, selecting an appropriate one for any particular set of chemical systems becomes a significant challenge, especially for novice users. In this work, we demonstrate the effectiveness of Δ -ML neural network models which leverage pre-trained models for QM interaction energies and predict the difference between lower accuracy method and higher accuracy method, providing a significant computational gain without major loss in accuracy. The Δ -ML neural network models can be trained on a small subset of the data and provide strong generalization, enabling potential use in large-scale screening of molecules.

1.1 Key Contributions

- Our framework identifies appropriate levels of theory for a given system through a combination of compute time estimators and Δ -ML error predictions.
- Hierarchical clustering of the Δ -ML Ensemble demonstrates these models capture theoretical relationships between methods, providing evidence for the effectiveness of applying Δ -ML models to identify computationally efficient levels of theory for chemical system(s).

Related Works: Machine-learned Δ -correction models have emerged as a potential approach to predict the result of accurate methods from less expensive methods using neural networks or machine learning.[14–18] Such Δ -ML methods allow capturing expensive electron correlation effects[14] and basis set effects. Oftentimes only a very small percentage of the dataset is needed to be computed at the higher level of theory.[14, 19] In such methods, the objective is to predict the difference (or Δ) between the target high-level of theory interaction energy (E_{high}) and a low-level of theory (E_{low}) using machine learning methods. This task assumes that there are computationally inexpensive functions that can capture more expensive functions, such as, high-level electron correlation in terms of molecular features relating to the geometry and pre-training on other properties.

Interaction energies present unique challenges in which approximate levels of theory can yield overbinding or underbinding due to combinations of incomplete correlation effects, basis set truncation errors, and types of interactions based on the chemical system.[12, 20, 13] As a result, naive models trained to predict total energies do not necessarily yield accurate predictions for interaction energies. The present Δ -ML models address this issue by focusing directly on the discrepancies in E_{int} , exploiting the smoother error landscape of the delta compared to the total energy.

The present work targets developing Δ -ML deep neural network models to predict interaction energies of one level of theory from another. Generally, Δ -ML models are targeting a single level of theory to a reference level of theory; however, the present work expands this to 80 levels of theory to acquire additional insight into how levels of theory compare for interaction energies. Typically, one would want to predict the expected error from a lower-level of theory to a higher-level of theory, but one could also ask how well can one map from any level of theory to another. The models do not require interaction energies as inputs to compute the error; hence, an additional application of these models is to estimate how inaccurate a level of theory would be if computed prior to any quantum calculations.

2 Methods

Dataset: The present work leverages data accumulated through various different works[21, 22, 22–26, 24, 27] to investigate how 80 different levels of theory perform at predicting intermolecular interaction energies on small organic molecules. More specifics on the subsets are available in Table S1. The dataset contains 3816 dimers with reference data at approximately “silver standard” interaction energies [DW-CCSD(T**)-F12/aug-cc-pVDZ]. However, to acquire the gold standard energies, the present work computed a subset of 3324 dimers with CCSD(T)/CBS/CP for higher quality reference energies. Methods are paired with specific Dunning’s augmented, correlation consistent double, triple, or quadruple- ζ basis sets[28, 29]—cc-pVDZ, aug-cc-pVDZ, aug-cc-pVTZ, and aug-cc-pVQZ. From herein, the present work will refer to this dataset as BFDB-Ext, containing 250K quantum interaction energy computations made easily accessible through this work. Due to the dataset on small organic dimers up to 38 atoms consisting of H, C, N, O, and S, the developed models are not guaranteed to generalize to significantly larger molecular systems like biomolecules.

Δ -corrected Models Ensembles from Pre-trained Models: To provide a reliable recommendation of an appropriate level of theory for computing intermolecular interaction energies, it is necessary to estimate the errors associated with each method relative to established reference values based on experimental measurements or computational benchmarks. Using BFDB-Ext, models can be trained to estimate the error for a given dimer using a particular level of theory where $E_{\text{IE},\text{ref}}$ is CCSD(T)/CBS/CP. For each level of theory, a separate Δ -model is trained to predict the error through

$$\Delta E_{\text{pred}} \approx E_{\text{IE},x} - E_{\text{IE},\text{ref}}, \quad (2)$$

where $E_{\text{IE},x}$ is the interaction energy at the specified level of theory.

We employ a pre-trained model originally developed for predicting dimer interaction energies on a substantially larger and more diverse dataset. This allows for our framework to be applicable to smaller datasets which may have limited chemical diversity. A recent atomic-pairwise neural network (AP-Net2) model is a 2.6M parameter pre-trained model that employs message-passing networks to predict monomer properties and subsequently SAPT0/jun-cc-pVDZ interaction energies.[30] AP-Net2 was trained the Splinter dataset[31] of over 1.6 million datapoints from over 9000 unique dimers, primarily targeting describing protein-ligand interactions. Since BFDB-Ext molecules resemble those in the Splinter dataset, AP-Net2 embeddings are well suited for Δ -corrected model for BFDB-Ext dataset.

Hyperparameter search identified a five-layer network (details in Supplement) as sufficient to achieve errors below 0.1 kcal/mol. Models were trained for 100 epochs on a 40/60 train/test split using mean squared error (MSE) between levels of theory as the loss, with inputs taken from the penultimate embeddings of AP-Net2. To train all-to-all Δ -ML models requires approximately 450 walltime hours with 8 cores on a Xeon 6226 CPU. In future works, the total number of levels of theory for larger datasets would be limited based on some methods having similar error distributions and allowing the approach to generalize to more data.

Compute Time Estimators: Alongside error estimation, we fit a polynomial to compute times using water clusters and small organics from BFDB-Ext. This task is necessary for downstream applications of the error estimating model by restricting recommended levels of theory to those that are computable by the end user. Otherwise, the error estimator would always recommend using CCSD(T)/CBS/CP energies, although in reality this is not desirable nor realistically computable for many chemical systems.

3 Results & Discussion

Model Performance: Selected Δ AP-Net2 models are shown in Figure 2a demonstrating performance predicting electron correlation corrections from a base level of theory to the reference, which are estimated CCSD(T)/CBS/CP energies in this case. Particularly different classes of methods—HF, MP2, SAPT, B3LYP, and B2PLYP—are included in the primary table (full list included in Table S2). Even HF/aug-cc-pVDZ/CP can be corrected from an MAE of 2.89 kcal mol⁻¹ to 0.08 kcal mol⁻¹, albeit still having a max error of 4.09 kcal mol⁻¹. Meanwhile, other levels of theory that have better baseline errors can also be corrected to roughly the same accuracy, but smaller max errors. For example, MP2/aug-cc-pVQZ/CP has a baseline MAE of 0.21 kcal mol⁻¹ and a max unsigned

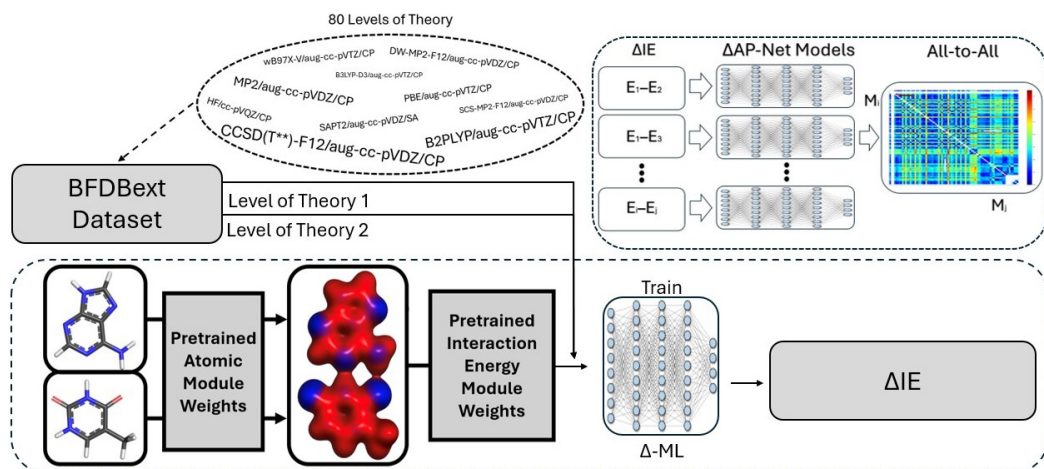


Figure 1: Overview of methodology of using the BFDText to train 80x80 Δ AP-Net2 models for predicting from any level of theory in the dataset to another level of theory.

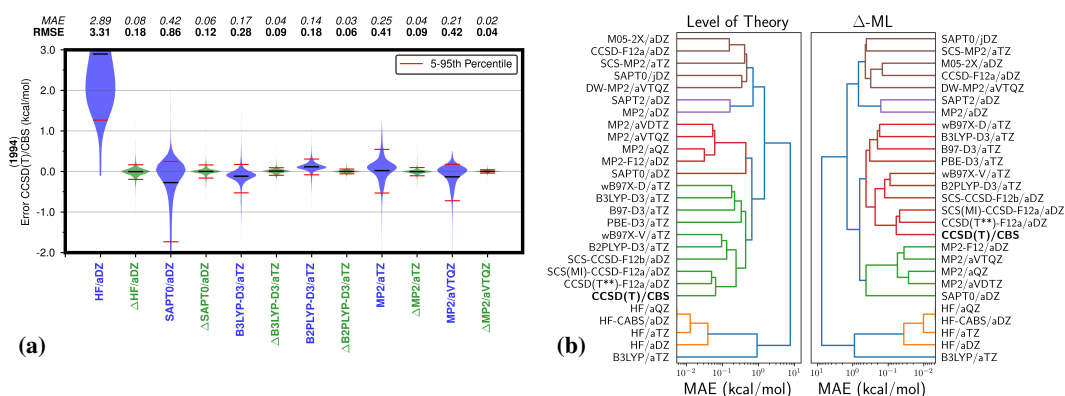


Figure 2: (a) BFDText dataset test error distributions for select levels of theory with respect to an estimated CCSD(T)/CBS/CP reference. The black horizontal line represents the mean error and the red horizontal lines represent the 5th and 95th percentiles. The uncorrected level of theory IE errors are in blue, while the Δ AP-Net2 plus level of theory IE errors are in green. (b) Dendrogram of select methods Δ AP-Net2 model predicted error estimations ordered by MAE. Note the clusters of methods are nearly identical as the all-to-all M_1 to M_2 dendrogram in the SI, meaning that the models are accurately predicting any M_1 to M_2 . All levels of theory here are using CP.

error of $3.56 \text{ kcal mol}^{-1}$, but after applying a Δ AP-Net2 model, the MAE is reduced to 0.02 and max error to $0.73 \text{ kcal mol}^{-1}$. The models accurately predict errors (below $<0.1 \text{ Kcal}$) on the test set, effectively learning the mapping from one level of theory to the reference. Here we tested the generalization only within the same chemical spaces and further work could extend this framework to evaluate generalization to disparate chemical spaces where the mapping might be more complex.

Level of Theory Hierarchies: Clustering of the MAE from all-to-all predictions, we evaluated how well the Δ -ML ensemble captures the relationships between different levels of theory compared to theoretical expectations. As shown in Figure 2b, the dendrograms from both the Δ -ML and theoretical expectation show strong alignment. This shows that the Δ -ML models capture relationships between levels of theory, further validating the approach (see SI for details).

Time Estimation: While predicting the exact compute time for a given level of theory would require detailed knowledge of the hardware and software implementation, a rough estimate can be acquired by fitting polynomials to accurately predict the log of the compute times. The practical goal of this task is to filter out levels of theory that are beyond the user's computational budget. To this end, polynomial expressions detailed in the Appendix are of the available singlepoint energy computations on water clusters and small organic molecules from the BFDText dataset. The resulting fitting

RMSEs are shown in Table S3. While the fits are not perfect, they reasonably filter out levels of theory that are too expensive for given systems.

4 Conclusion

The present work has demonstrated that Δ -ML models can be trained to predict the error of a given level of theory from any other level of theory. Particularly, the models are able to use one of the cheapest levels of theory, HF/aug-cc-pVDZ/CP, to predict CCSD(T)/CBS/CP reference value with a surprisingly small MAE of 0.08 kcal mol⁻¹. Even more interesting is that these models are able to predict between any two levels of theory with similar accuracy even when the methods themselves quite differently like DFT to wavefunction methods on these systems. Furthermore, when combining the ensemble of Δ -ML models with the compute time estimators, users can rely on data instead of chemical intuition to select an appropriate level of theory for their desired accuracy, computational cost, and chemical system(s). To enhance generalization, this framework can be applied to datasets with more chemical diversity and likely fewer levels of theory. A next step of this work is to unify the usage of error and time estimators to enable large-scale screening applications critical for material or drug discovery.

References

- [1] Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *WIREs Comput. Mol. Sci.* **2016**, 7, e70057, None.
- [2] Borca, C. H.; Glick, Z. L.; Metcalf, D. P.; Burns, L. A.; Sherrill, C. D. Benchmark Coupled-Cluster Lattice Energy of Crystalline Benzene and Assessment of Multi-Level Approximations in the Many-Body Expansion. *J. Chem. Phys.* **2023**, 158, 234102.
- [3] Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with Aromatic Rings in Chemical and Biological Recognition. *ChemInform* **2003**, 34, e70057, None.
- [4] Parrish, R. M.; Sitkoff, D. F.; Cheney, D. L.; Sherrill, C. D. The Surprising Importance of Peptide Bond Contacts in Drug-Protein Interactions. *Chem. - Eur. J.* **2017**, 23, 7887–7890.
- [5] Hill, G.; Forde, G.; Hill, N.; Lester, W. A.; Andrzej Sokalski, W.; Leszczynski, J. Interaction energies in stacked DNA bases? How important are electrostatics? *Chem. Phys. Lett.* **2003**, 381, 729–732, None.
- [6] Parker, T. M.; Hohenstein, E. G.; Parrish, R. M.; Hud, N. V.; Sherrill, C. D. Quantum-Mechanical Analysis of the Energetic Contributions to π Stacking in Nucleic Acids Versus Rise, Twist, and Slide. *J. Am. Chem. Soc.* **2013**, 135, 1306–1316.
- [7] McDaniel, J. G.; Schmidt, J. R. Physically-Motivated Force Fields From Symmetry-Adapted Perturbation Theory. *J. Phys. Chem. A* **2013**, 117, 2053–2066.
- [8] Vleet, M. J. V.; Misquitta, A. J.; Schmidt, J. R. New Angles On Standard Force Fields: Toward a General Approach for Treating Atomic-Level Anisotropy. *J. Chem. Theory Comput.* **2018**, 14, 739–758.
- [9] Schriber, J. B.; Nascimento, D. R.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. CLIFF: A Component-Based, Machine-Learned, Intermolecular Force Field. *J. Chem. Phys.* **2021**, 154, 184110.
- [10] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A 5th-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **1989**, 157, 479–483.
- [11] Řezáč, J.; Hobza, P. Describing Noncovalent Interactions Beyond the Common Approximations: How Accurate Is the Gold Standard, CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, 9, 2151–2155.

- [12] Burns, L. A.; Marshall, M. S.; Sherrill, C. D. Comparing Counterpoise-Corrected, Uncorrected, and Averaged Binding Energies for Benchmarking Noncovalent Interactions. *J. Chem. Theory Comput.* **2014**, *10*, 49–57.
- [13] Schriber, J. B.; Wallace, A. M.; Cheney, D. L.; Sherrill, C. D. Levels of symmetry-adapted perturbation theory (SAPT). II. Convergence of interaction energy components. *J. Chem. Phys.* **2025**, *163*, 084114, None.
- [14] 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.
- [15] Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ -machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory. *J. Chem. Phys.* **2021**, *154*, 051102.
- [16] Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **2019**, *150*, 131103.
- [17] Vinod, V.; Zaspel, P. Benchmarking data efficiency in Δ -ML and multifidelity models for quantum chemistry. *J. Chem. Phys.* **2025**, *163*, 024134.
- [18] Huang, Y.; Hou, Y.-F.; Dral, P. O. Active delta-learning for fast construction of interatomic potentials and stable molecular dynamics simulations. *Machine Learning: Science and Technology* **2025**, *6*, 035004.
- [19] Song, K.; Li, J. The neural network based Δ -machine learning approach efficiently brings the DFT potential energy surface to the CCSD(T) quality: a case for the OH + CH₃OH reaction. *Phys. Chem. Chem. Phys.* **2023**, *25*, 11192–11204.
- [20] Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. Levels of Symmetry Adapted Perturbation Theory (SAPT). I. Efficiency and Performance for Interaction Energies. *J. Chem. Phys.* **2014**, *140*, 094106.
- [21] Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 88–96.
- [22] Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis Set Convergence of the Coupled-cluster Correction, $\delta_{MP2}^{CCSD(T)}$: 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.
- [23] Burns, L. A.; Vázquez-Mayagoitia, Á.; Sumpter, B. G.; Sherrill, C. D. Density-Functional Approaches to Noncovalent Interactions: A Comparison of Dispersion Corrections (DFT-D), Exchange-Hole Dipole Moment (XDM) Theory, and Specialized Functionals. *J. Chem. Phys.* **2011**, *134*, 084107.
- [24] Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7*, 2197–2203.
- [25] Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- [26] Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

- [27] Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDdb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017**, *147*, 161727.
- [28] 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.
- [29] Woon, D. E.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994**, *100*, 2975–2988, None.
- [30] Glick, Z. L.; Metcalf, D. P.; Glick, C. S.; Spronk, S. A.; Koutsoukas, A.; Cheney, D. L.; Sherrill, C. D. A Physics-aware Neural Network for Protein-ligand Interactions with Quantum Chemical Accuracy. *Chem. Sci.* **2024**, *15*, 13313–13324.
- [31] Spronk, S. A.; Glick, Z. L.; Metcalf, D. P.; Sherrill, C. D.; Cheney, D. L. A Quantum Chemical Interaction Energy Dataset for Accurately Modeling Protein-Ligand Interactions. *Sci. Data* **2023**, *10*, 619.

A Technical Appendices and Supplementary Material

A.1 Model Details

The Δ -ML models used within this work are based on the atom-pairwise message passing neural networks developed in previous work.[30] These consist of an atomic module that learns to predict atomic charges, dipoles and quadruples through message-passing neural networks. This module uses 3 message passes, 8 Bessel functions, and a cutoff distance of 5.0 Å. The update and readout functions are dense feed-forward neural networks with 3 three hidden layers with 256, 128, and 64 neurons. The last layer has a linear operation to reach the last hidden layer of size 8 or 1 for update and readout, respectively. The intermolecular atomic-pairwise module that has been adapted for the Δ -ML models use the same defaults as AP-Net2, except for predicting a single energy instead of 4 and dropping the multipolar electrostatics. The Δ -ML update and readout layers use the same hidden layer sizes as the atomic module.

Table 1: Datasets used in training Δ -ML models. For each dataset, we provide the total number of dimers (Size), the number of heavy atoms in the largest dimer (Largest), relevant references, and a brief description.

Database	Size	Largest	Ref.	Description
<i>Curves & Surfaces</i>				
HBC6	118	6	[21, 22]	dissoc. curves of doubly hydrogen-bonded (HB) complexes
NBC10ext	183	12	[22–24]	dissoc. curves of dispersion-bound (DD) complexes
<i>Small Dimers</i>				
S22	22	-	[25, 26, 24]	
<i>Extracted from Biological Systems</i>				
SSI	3372	20	[27]	peptide sidechain-sidechain complexes
BB1	100	20	[27]	peptide sidechain-sidechain complexes
<i>Total</i>	3816	20		

Method	Basis Set	Mode
B2PLYP-D3	aug-cc-pVTZ	CP
DW-CCSD(T**)-F12	aug-cc-pVDZ	CP
CCSD(T**)-F12a	aug-cc-pVDZ	CP
MP2	aug-cc-pVTQZ	CP

CCSD-F12a	aug-cc-pVDZ	CP
HF-CABS	aug-cc-pVDZ	CP
SCS(MI)-MP2	cc-pVQZ	CP
DW-MP2	cc-pVQZ	CP
SCS(N)-MP2	cc-pVQZ	CP
SCS-MP2	cc-pVQZ	CP
HF	cc-pVQZ	CP
MP2	cc-pVQZ	CP
SCS(MI)-MP2	aug-cc-pVTZ	CP
DW-MP2	aug-cc-pVTZ	CP
SCS(N)-MP2	aug-cc-pVTZ	CP
SCS-MP2	aug-cc-pVTZ	CP
HF	aug-cc-pVTZ	CP
MP2	aug-cc-pVTZ	CP
SCS(MI)-CCSD-F12a	aug-cc-pVDZ	CP
SCS(MI)-CCSD-F12b	aug-cc-pVDZ	CP
DW-MP2	aug-cc-pVDZ	CP
SCS-CCSD-F12b	aug-cc-pVDZ	CP
MP2-F12	aug-cc-pVDZ	CP
CCSD-F12b	aug-cc-pVDZ	CP
SCS(N)-MP2	aug-cc-pVDZ	CP
CCSD(T**)-F12b	aug-cc-pVDZ	CP
SCS-MP2-F12	aug-cc-pVDZ	CP
SCS-MP2	aug-cc-pVDZ	CP
DW-MP2-F12	aug-cc-pVDZ	CP
SCS-CCSD-F12a	aug-cc-pVDZ	CP
HF	aug-cc-pVDZ	CP
MP2	aug-cc-pVDZ	CP
SCS(N)-MP2-F12	aug-cc-pVDZ	CP
SCS(MI)-MP2	aug-cc-pVDTZ	CP
DW-MP2	aug-cc-pVDTZ	CP
SCS(N)-MP2	aug-cc-pVDTZ	CP
SCS-MP2	aug-cc-pVDTZ	CP
MP2	aug-cc-pVDTZ	CP
SCS(MI)-MP2	aug-cc-pVQZ	CP
DW-MP2	aug-cc-pVQZ	CP
SCS(N)-MP2	aug-cc-pVQZ	CP
SCS-MP2	aug-cc-pVQZ	CP
HF	aug-cc-pVQZ	CP
MP2	aug-cc-pVQZ	CP
SCS(MI)-MP2	aug-cc-pVTQZ	CP
DW-MP2	aug-cc-pVTQZ	CP
SCS(N)-MP2	aug-cc-pVTQZ	CP
SCS-MP2	aug-cc-pVTQZ	CP
SAPT0	aug-cc-pVDZ	SA
SAPT0	jun-cc-pVDZ	SA
sSAPT0	aug-cc-pVDZ	SA
sSAPT0	jun-cc-pVDZ	SA
SCS-SAPT0	jun-cc-pVDZ	SA
SAPT2	aug-cc-pVDZ	SA
SAPT2+	aug-cc-pVDZ	SA
B3LYP	aug-cc-pVTZ	unCP
B3LYP-D2	aug-cc-pVTZ	unCP
B3LYP-D3	aug-cc-pVTZ	unCP
B2PLYP	aug-cc-pVTZ	unCP
B2PLYP-D2	aug-cc-pVTZ	unCP

B2PLYP-D3	aug-cc-pVTZ	unCP
B97	aug-cc-pVTZ	unCP
wB97X-D	aug-cc-pVTZ	unCP
M05-2X	aug-cc-pVDZ	unCP
PBE	aug-cc-pVTZ	unCP
PBE-D2	aug-cc-pVTZ	unCP
PBE-D3	aug-cc-pVTZ	unCP
B97-D2	aug-cc-pVTZ	unCP
B97-D3	aug-cc-pVTZ	unCP
B2PLYP	aug-cc-pVTZ	CP
B3LYP	aug-cc-pVTZ	CP
B3LYP-D3	aug-cc-pVTZ	CP
B97-D3	aug-cc-pVTZ	CP
M05-2X	aug-cc-pVDZ	CP
PBE	aug-cc-pVTZ	CP
PBE-D3	aug-cc-pVTZ	CP
wB97X-D	aug-cc-pVTZ	CP
wB97X-V	aug-cc-pVTZ	CP
wB97X-V	aug-cc-pVTZ	unCP
CCSD(T)	CBS	CP

Table 2: List of all levels of theory, basis sets, and modes used in the

Level of Theory	Train RMSE [log(s)]	Test RMSE [log(s)]
MP2	0.1542	0.1855
HF	0.1048	0.1175
B2PLYP-D3	0.1518	0.1444
B3LYP-D3	0.1966	0.1875
PBE-D3	0.2005	0.1817
M05-2X	0.2148	0.2021
wB97X-V	0.2025	0.1851
wB97X-D	0.1812	0.1531
FNO-CCSD	0.1811	0.1687
FNO-CCSD(T)	0.2404	0.1916

Table 3: Summary of polynomial fitting errors for different levels of theory

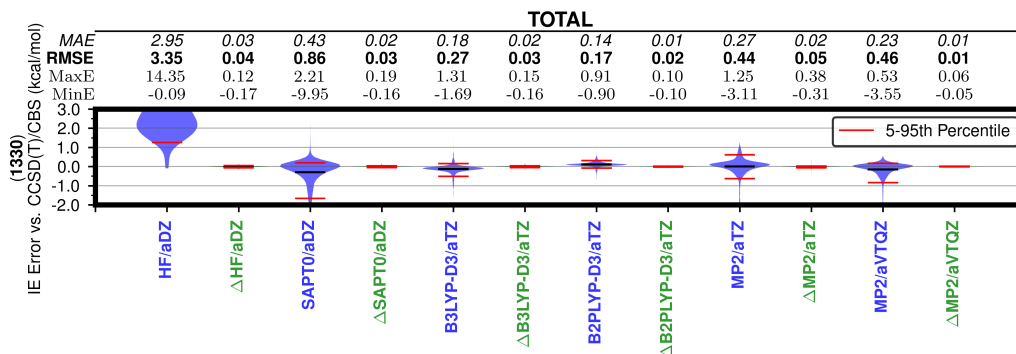


Figure 3: BFDExt dataset train error distributions for select levels of theory with respect to an estimated CCSD(T)/CBS/CP reference. The black horizontal line represents the mean error and the red horizontal lines represent the 5th and 95th percentiles. The uncorrected level of theory IE errors are in blue, while the δ AP-Net2 plus level of theory IE errors are in green.

