$\triangle$-ML Ensembles for Selecting Quantum Chemistry Methods to Compute Intermolecular Interactions

Austin M. Wallace[*]
Department of Chemistry
Georgia Institute of Technology
Atlanta, GA 30318
awallace43@gatech.edu

C. David. Sherrill[†]
Department of Chemistry
Georgia Institute of Technology
Atlanta, GA 30318
sherrill@gatech.edu

Giri P. Krishnan[‡]
ARTISAN
Georgia Institute of Technology
Atlanta, GA 30318
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

[*]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
[‡]Georgia Tech Center for Artificial Intelligence in Science and Engineering

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML4PS.

trade-off of accuracy and computational cost drives the choice of specific pairings of methods and 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

$$\Delta E_{\text{int}} = E_{IJ} - E_I - E_J, \tag{1}$$

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 $E_{\mathrm{IE,ref}}$ is CCSD(T)/CBS/CP. For each level of theory, a separate Δ-model is trained to predict the error through

$$\Delta E_{\mathrm{pred}} \approx E_{\mathrm{IE,x}} - E_{\mathrm{IE,ref}}, \tag{2}$$

where $E_{\mathrm{IE,x}}$ 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 ΔAP-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

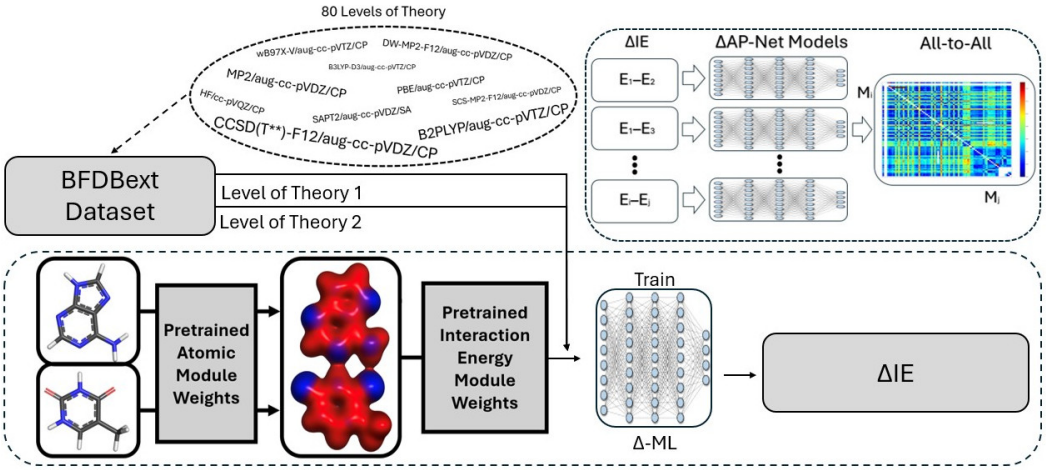

Figure 1: Overview of methodology of using the BFDBext to train 80x80 ΔAP-Net2 models for predicting from any level of theory in the dataset to another level of theory.

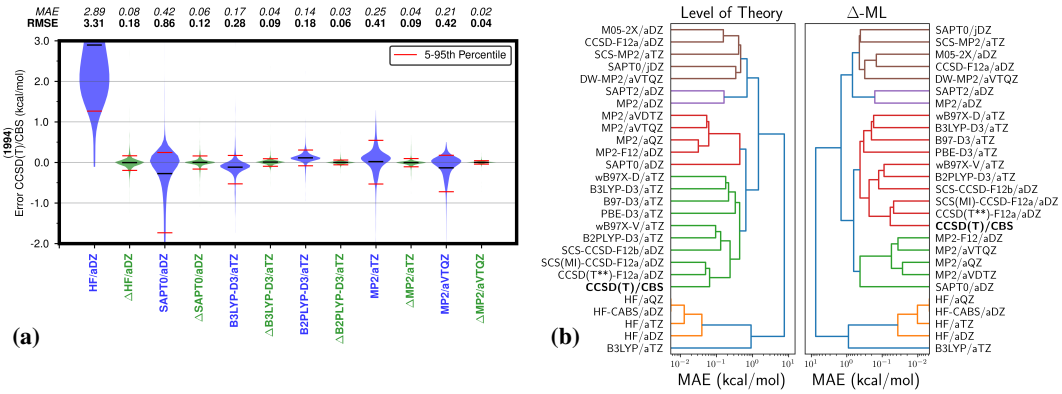

Figure 2: (a) BFDBExt dataset test error distributions for select levels of theory with respect to an estimated CCSD(T)/CBS/CP reference. The black horizontal line represents the mean error and the red horizontal lines represent the 5th and 95th percentiles. The uncorrected level of theory IE errors are in blue, while the ΔAP-Net2 plus level of theory IE errors are in green. (b) Dendogram of select methods ΔAP-Net2 model predicted error estimations ordered by MAE. Note the clusters of methods are nearly identical as the all-to-all M1 to M2 dendogram in the SI, meaning that the models are accurately predicting any M1 to M2. All levels of theory here are using CP.

error of 3.56 kcal mol^{-1}, but after applying a ΔAP-Net2 model, the MAE is reduced to 0.02 and max error to 0.73 kcal mol^{-1}. The models accurately predict errors (below <0.1 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 BFDBext 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 (BFDb): 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
Curves & Surfaces				
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
Small Dimers				
S22	22	-	[25, 26, 24]	
Extracted from Biological Systems				
SSI	3372	20	[27]	peptide sidechain-sidechain complexes
BBI	100	20	[27]	peptide sidechain-sidechain complexes
Total	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

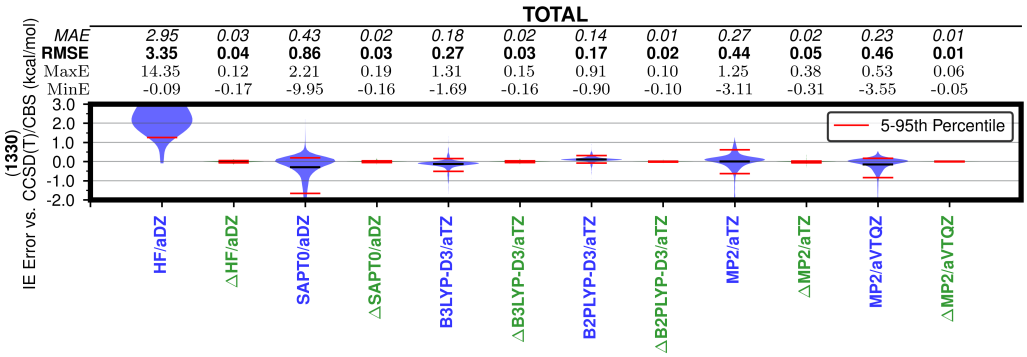

Figure 3: BFDBExt dataset train error distributions for select levels of theory with respect to an estimated CCSD(T)/CBS/CP reference. The black horizontal line represents the mean error and the red horizontal lines represent the 5th and 95th percentiles. The uncorrected level of theory IE errors are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

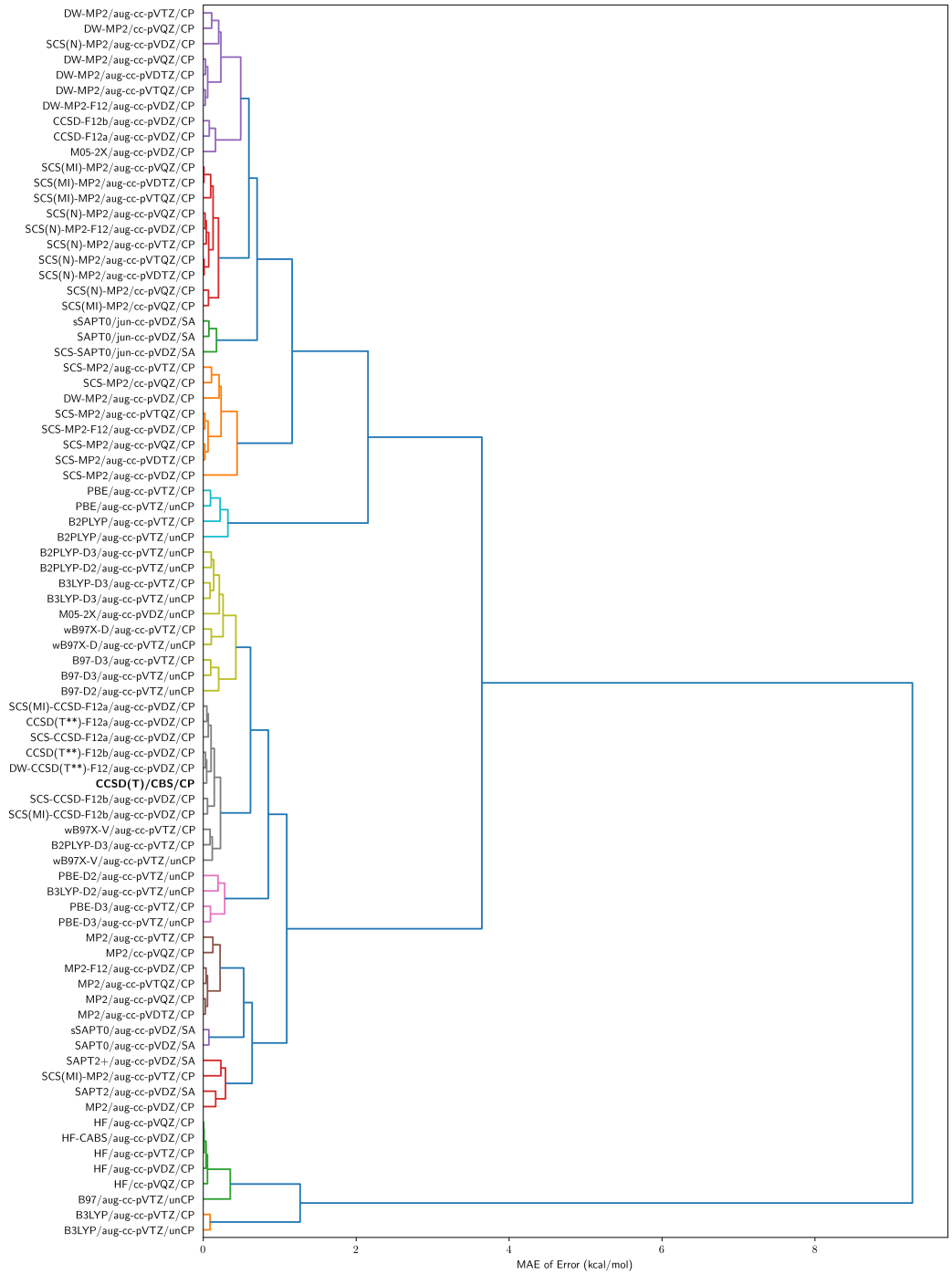

Figure 4: Dendogram of all-to-all δAP-Net2 model predicted error estimations ordered by MAE. Note the clusters of methods are nearly identical as the all-to-all M1 to M2 dendogram in the SI, meaning that the models are accurately predicting any M1 to M2.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduces provide the broad outline of the $Delta$-ML framework and its salient abilities. The method, results sections and figures support the claims made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: At several sections in the manuscript, we mention that our proposed approach is evaluated in one dataset and its generalization across other datasets are known. Given the page limit, we did not include limitation section, instead these are mentioned in other sections.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: All the code for the publication is in https://anonymous.4open.science/r/QCMLForge-1381/ and data available https://anonymous.4open.science/r/BFDBext-Dataset-C168. The datasets and pre-trained model used in this work is publicly available. The details of the neural network required to obtain the results are provided in the manuscript.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the code for the publication is in https://anonymous.4open.science/r/QCMLForge-1381/. The datasets and pre-trained model used in this work is publicly available. The BFDBExt data is available at https://anonymous.4open.science/r/BFDBext-Dataset-C168

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All these details were provided in the method section and the supplements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We used violin plots to capture the distribution wherever applicable

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The work provides sufficient estimates on compute time required to generate the presented figures and key results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The present study complies with the NeurIPS Code of Ethics for it produces a framework for creating Δ-ML models on theoretically generated data through proper licensing.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This work is a proof of concept for a larger results which has broader impacts. In this work, we only demonstrate the feasibility of a particular approach and would need additional supporting work to make claims on broader impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Datasets and models are either previously published or based on typical neural networks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The associated repository (https://anonymous.4open.science/r/QCMLForge-1381/) has an MIT license and the data is available publically at https://anonymous.4open.science/r/BFDBext-Dataset-C168.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new asset is released

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: crowdsourcing nor research with human subjects was not involved in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used for any of the core design, methodology or interpreting the results.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.