Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: L. Zeng, X. Zhang, Y. Pei, L. Zhao, L. Hua, J. Yang and N. Huang, *Digital Discovery*, 2026, DOI: 10.1039/D6DD00056H.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00056H

Developing a Machine-Learning Interatomic Potential for Non-Covalent Interactions in Proteins

Lejia Zeng[1,2], Xintong Zhang[1,2], Yuchan Pei[1,2], Lifeng Zhao[2], Lan Hua[2], Jincai Yang[2], Niu Huang[1,2,]*

[1]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing 102206, China

[2]National Institute of Biological Sciences, 7 Science Park Road, Zhongguancun Life Science Park, Beijing 102206, China

KEYWORDS: Machine Learning Interatomic Potential (MLIP), Non-covalent Interactions (NCI), Active Learning (AL), Quantum Mechanics (QM), Force Field (FF)

* To whom correspondence should be addressed. N. H. (Email) huangniu@nibs.ac.cn (Phone) 86-10-80720645 (Fax) 86-10-80720813

View Article Online
DOI: 10.1039/D6DD00056H

ABSTRACT

Machine learning interatomic potentials (MLIPs) enable efficient modeling of molecular interactions with quantum mechanical (QM) accuracy for complex systems. However, constructing robust and representative training datasets that capture subtle, system-specific interaction motifs remains challenging. Here, a PAirwise Non-covalent Interaction Potential (PANIP) is introduced, an ensemble MLIP model built upon the Neural Equivariant Interatomic Potentials (NequIP) framework and trained on non-covalent interactions (NCIs) between protein-derived fragments. PANIP is trained using an automated multi-fidelity active learning (MFAL) workflow that distills a diverse and information-rich subset from an otherwise prohibitively large pool of fragment dimers extracted from high-resolution protein structures in the Protein Data Bank (PDB). This strategy concentrates high-level QM calculations (ωB97X-D3BJ/def2-TZVPP) on the most informative structures while preserving comprehensive coverage of the data distribution. Applied to dimers constructed from 17 chemically distinct protein fragments (side chains, backbone motif, and water), this workflow yields the PDB Fragment Interaction Dataset (PDB-FRAGID), a condensed yet representative subset comprising only 8.7% of the original 36.3 million-dimer pool while maintaining structural and chemical diversity. Despite this drastic reduction, PANIP retains ωB97X-D3BJ/def2-TZVPP-level accuracy on both equilibrium and non-equilibrium fragment configurations and achieves mean absolute errors below 0.2 kcal/mol on out-of-distribution systems, demonstrating excellent transferability across diverse NCI motifs. Compared to the widely used AIMNet2 potential, PANIP delivers substantially lower errors on

View Article Online
DOI: 10.1039/D6DD00056H

protein-derived fragments and exhibits superior generalization, particularly for charged and strongly interacting dimers. By combining a fragmentation-based energy decomposition scheme with PANIP, protein–ligand binding energies can be estimated at near force-field computational cost while preserving QM-level accuracy for pairwise NCIs, enabling its use as a fragment-based scoring function that rivals specialized docking scoring functions despite being trained solely on fragment dimers. The PANIP models and associated benchmark sets are available at https://github.com/hnlab/PANIP, and the PDB-FRAGID dataset is available at https://github.com/hnlab/PDB-FRAGID.

INTRODUCTION

The accurate modeling of non-covalent interactions (NCIs) within biomolecules at quantum-mechanical (QM) accuracy remains computationally prohibitive for large systems. The integration of QM methods with machine learning (ML) enables the efficient exploration of complex molecular systems[1-3]. Machine learning interatomic potentials (MLIPs) reproduce QM potential energy surfaces (PESs) with high fidelity, often achieving QM-level accuracy at a fraction of the computational cost for small- and medium-sized systems. These developments have opened new avenues for efficiently exploring complex conformational and interaction landscapes in chemistry and biology[4-7].

View Article Online
DOI: 10.1039/D6DD00056H

However, the reliability and transferability of MLIPs depend critically on the quality, diversity, and balance of the underlying training data. Many existing datasets and MLIP frameworks undersample chemical and conformational spaces[8–14], emphasize a limited set of molecular species, or are tuned to specific classes of systems, which restricts their applicability to new interaction motifs and environments[3, 5, 15–18]. Recent efforts have introduced large-scale, chemically diverse datasets and "universal" models that span broad chemical spaces[19, 20]. However, these primarily emphasize universality, and when it comes to particular and subtle interaction motifs crucial for biological systems, important regions may remain underexplored. In particular, NCIs in proteins involve a rich variety of hydrogen bonding, electrostatic, dispersion, cation–π, and sulfur-containing interactions that are strongly context-dependent and sensitive to local geometry. Capturing these motifs at QM accuracy in a manner that generalizes across diverse protein environments remains a major challenge.

Generating new QM datasets tailored to specific systems or interactions is both labor-intensive and computationally expensive, sometimes negating the gains provided by ML acceleration. Moreover, when such system-specific datasets are integrated with existing large-scale datasets, the amount of additional training data should be carefully controlled to avoid excessive training burden while still filling critical gaps in chemical and conformational coverage. A central challenge is therefore to construct optimized, broadly representative training datasets that minimize redundant QM calculations while maximizing coverage of relevant NCIs. Achieving this goal requires intelligent data selection and model training workflows.

View Article Online
DOI: 10.1039/D6DD00056H

The Protein Data Bank (PDB)[21] provides a rich, experimentally grounded source of NCI geometries within biologically functional contexts[22–25]. Unlike synthetic datasets, PDB-derived geometries reflect realistic chemical diversity, spatial complexity, and thermodynamic relevance[22,23]. This makes the PDB an attractive foundation for building MLIPs that target protein NCIs and are designed to generalize across biologically relevant interaction types, including sidechain-sidechain, sidechain-backbone, and protein-water contacts. At the same time, PDB data exhibits inherent biases, for example, over-representation of certain interactions, resolution limitations, and crystallographic artifacts, which must be carefully considered when constructing datasets and training models.

Leveraging the PDB at scale for MLIP training presents two main challenges. First, generating accurate QM reference energies for the hundreds of millions of possible fragment geometries is computationally prohibitive, even with efficient density functional theory (DFT) methods. Second, the distribution of interaction types in the PDB is highly imbalanced: common interactions such as conventional hydrogen bonds, are heavily over-represented, while rarer but functionally important interactions, including certain cation-π, sulfur-aromatic, or ionic contacts, occur relatively infrequently. Naively labeling and training on all available dimers would not only be infeasible, but would also risk biasing model learning. Moreover, processing and integrating such a vast structural and energetic dataset during model training requires substantial computational resources and careful data-management strategies.

To address these challenges, a multi-fidelity active learning (MFAL)[26–28] workflow is employed to construct an efficient, representative dataset and to train an MLIP tailored to protein

View Article Online
DOI: 10.1039/D6DD00056H

NCIs. This hierarchical approach integrates low-cost and high-level QM calculations: the composite density functional method r²SCAN-3c[29] serves as a low-fidelity method for large-scale energetic screening, while a machine learning surrogate identifies a representative and diverse subset for high-level refinement at the ωB97X-D3BJ/def2-TZVPP level[30, 31]. This process yields the PDB Fragment Interaction Dataset (PDB-FRAGID)—a condensed yet representative subset containing only 8.7% of the original pool. Trained on PDB-FRAGID, the resulting ensemble model PANIP achieves accuracy comparable to ωB97X-D3BJ/def2-TZVPP calculations. Built on the Neural Equivariant Interatomic Potentials (NequIP) architecture[32], PANIP preserves spatial symmetries and accurately captures atomic environments. The model was rigorously validated against multiple benchmark sets, including low-energy dimers, optimized dimers, geometries from the Cambridge Structural Database (CSD)[33], and non-equilibrium conformations generated by random sampling. Using this framework, we systematically characterized NCI patterns across the entire PDB, with particular emphasis on underexplored sulfur-driven interactions. Finally, we demonstrate PANIP's potential utility by applying it as a fragment-based[34] scoring function for protein-ligand binding prediction in benchmark systems.

METHODS

Overall Workflow. As illustrated in Figure 1, a systematic workflow was developed to build MLIPs that leverage chemically relevant fragment-fragment interactions derived from high-resolution protein structures in the PDB. Proteins were fragmented into dimeric pairs to generate

View Article Online
DOI: 10.1039/D6DD00056H

an initial dataset, which was first labeled with baseline interaction energies computed at the low-cost r²SCAN-3c level[29]. An MFAL method was designed to iteratively guide the selection of a representative and diverse subset of dimers, ensuring balanced coverage of the full structural distribution. This refined subset, termed the PDB-FRAGID, was recalculated at the higher-accuracy ωB97X-D3BJ/def2-TZVPP level[30, 31]. The curated data was then used to train the PANIP models, resulting in robust MLIPs with QM level accuracy in predicting molecular interaction energies.

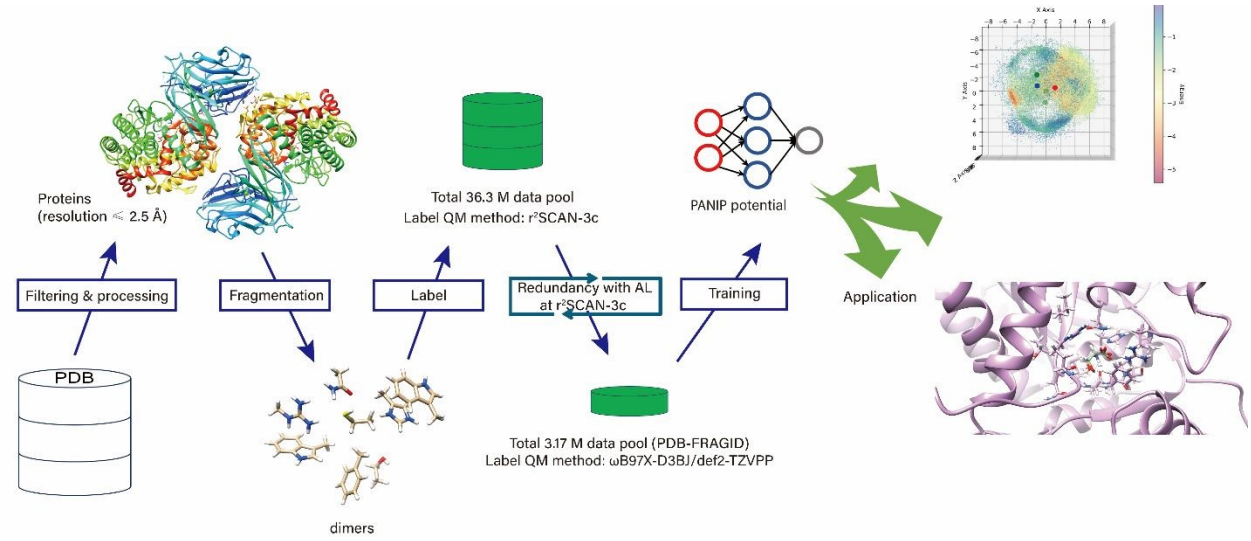

Figure 1. Overall workflow of dataset construction and model training.

Fragmentation Definition and Dimer Extraction. To represent NCIs in proteins, 17 chemically distinct fragment types were defined, encompassing amino acid side chains, backbone segment, and water (Figure 2). Key atoms (e.g., carbons bonded to aromatic rings, thiols, hydroxyls, or ammonium groups) were retained to preserve local electronic environments and minimize

View Article Online
DOI: 10.1039/D6DD00056H

symmetry-related redundancy. These fragments were paired into 153 unique dimer types, forming the basis for NCI analysis.

Initial protein structures were obtained from PDB entries with resolution ≤ 2.5 Å and unique UniProt[35] IDs to reduce redundancy, yielding 29,204 proteins. Missing hydrogens were added using the high-throughput molecular dynamics (HTMD) approach[36], while the original PDB coordinates were preserved without further structural optimization. Protonation states of ionizable residues were assigned according to typical physiological conditions: all three common protonation states were considered for histidine, reflecting its pKa near physiological pH, whereas lysine and arginine side chains were treated as protonated, and aspartic acid side chains as deprotonated.

Interacting fragment pairs were identified using two criteria. First, the shortest heavy-atom distances between 2 and 4 Å (capturing interactions slightly beyond the sum of van der Waals radii for atomic pairs (C=1.70 Å, N=1.55 Å, O=1.52 Å, S=1.80 Å) and excluding covalent bonds (<2 Å)[37]). The 4 Å upper bound was chosen to focus on geometries where NCIs are most significant, while avoiding an overwhelming number of weak, long-range contacts. Second, fragment pairs were required to be separated by at least two residues along the protein sequence to mitigate trivial local contacts arising from backbone connectivity. This yielded a raw set of 36.3 million dimers spanning 153 unique fragment-fragment combinations across the 17 fragment types in protein (see Supporting Information for details).

View Article Online
DOI: 10.1039/D6DD00056H

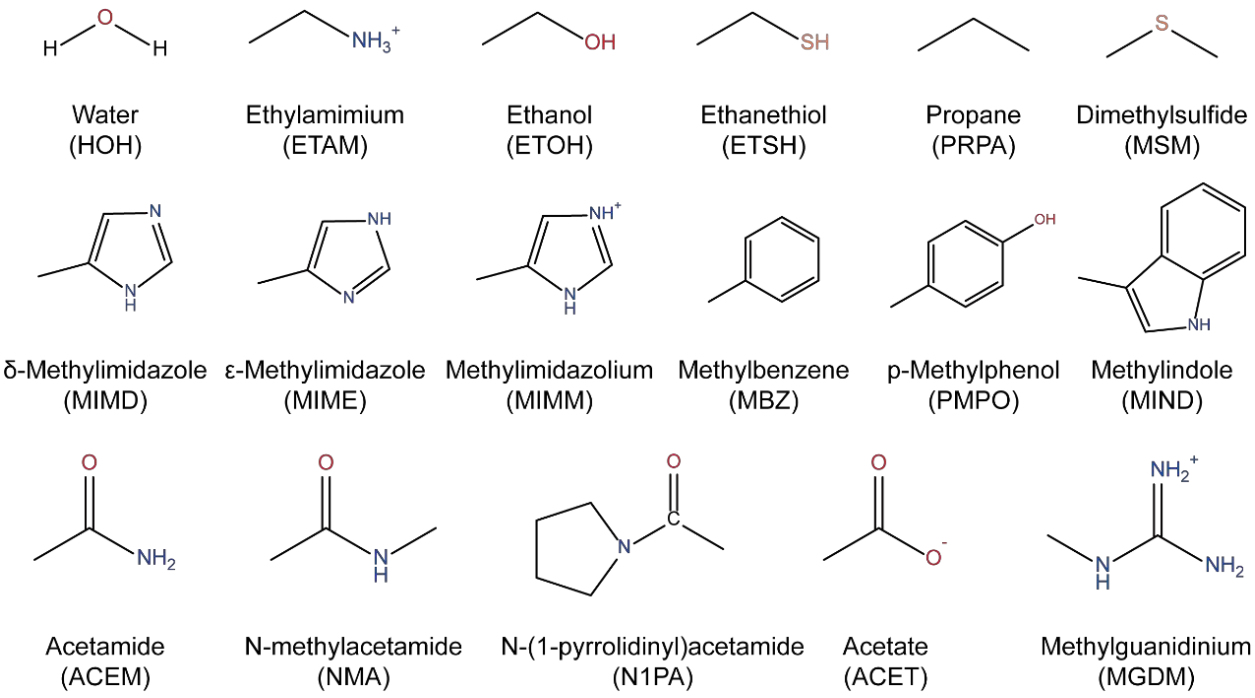

Figure 2. 17 representative chemical groups defined in proteins.

Data Reduction via Multi-Fidelity Active Learning. To balance computational efficiency with accuracy, a multi-fidelity active learning workflow[26 - 28] was designed (Figure S1). MFAL extends active learning by combining low-cost quantum chemical approximations with selected high-fidelity refinements, reducing the need for exhaustive expensive calculations. Specifically, the r²SCAN-3c method[29] served as a low-fidelity method to provide baseline interaction energies for all dimers. r²SCAN-3c combines the r²SCAN meta-GGA functional with atom-pairwise dispersion and basis-set corrections, and was calculated using ORCA 5.0.3[38]. The non-covalent interaction energy (ΔE_{AB}) for each dimer was computed as:

$$\Delta E_{AB} = E(AB) - [E(A) + E(B)] \quad (1)$$

View Article Online
DOI: 10.1039/D6DD00056H

where $E(AB)$, $E(A)$, and $E(B)$ represent the total energy of the dimer and isolated monomers, respectively.

As an initial step, a 2% random bootstrap set was used to train an initial surrogate model based on NequIP[32]. The surrogate model was then iteratively applied to the remaining dimers, and an uncertainty-based acquisition function flagged high-error cases using the normalized criterion $|E_{pred} - E_{ref}|/\sqrt{N} > 0.04$ kcal/mol, where E_{pred} is the predicted energy, E_{ref} is the reference (r²SCAN-3c) energy, and N is the atom count. The threshold of 0.04 kcal/mol per atom follows the established by Smith et al[28], and was validated in preliminary testing as a balanced choice between computational accuracy and cost. In each iteration, a random 2% subset of these flagged structures was added to the training set. This iterative loop continued until fewer than 5% of the unscreened dimers remained as high-error cases, which were then fully included in the final dataset. The final refined subset (PDB-FRAGID) condensed the original pool to a representative core comprising only 8.7% of its size (~3.15 million) while preserving balanced coverage of the full structural distribution.

High-Level QM labeling. All dimers in the PDB-FRAGID dataset were recomputed at the higher-accuracy ωB97X-D3BJ/def2-TZVPP level[30, 31] using ORCA 5.0.3[38]. The ωB97X-D3BJ[30] (a range-separated hybrid meta-GGA density functional with D3(BJ) empirical dispersion correction) coupled with the def2-TZVPP[31] basis set delivers exceptional precision for modeling non-covalent interactions and conformational energies while maintaining a balanced treatment of short- and long-range electronic correlations[30, 31]. This functional was selected for its well-documented accuracy for non-covalent interactions across large-scale benchmarks including

View Article Online
DOI: 10.1039/D6DD00056H

GMTKN55, comparable to the top-performing range-separated hybrids but at lower computational cost[39–41]. The def2-TZVPP basis set, when paired with counterpoise (CP) correction, provides a reliable approximation to the complete basis set (CBS) limit for such systems[39]. The resolution of identity approximation with coulomb, exchange, and correlation contributions (RIJCOSX) approximation[42] was applied to accelerate calculations, with auxiliary basis sets def2/J[43] and def2-TZVPP/C[44] employed for integration and correlation treatments. The ma-def2-TZVPP[45] diffuse function was added only for negatively charged oxygen atoms (e.g., in carboxylate groups like ACET) to improve electron density descriptions. For neutral systems, additional diffuse functions provided no benefit and could even introduce numerical instabilities. The basis set superposition error (BSSE) was corrected using the counterpoise (CP) method[46–48]. It has been shown that in larger basis sets, such as the finite triple-ζ basis employed here (def2-TZVPP), CP-corrected interaction energies yielded greater reliability than their uncorrected values[49]. NCI energies were computed according to Eq.1. These high-fidelity ωB97X-D3BJ/def2-TZVPP interaction energies on PDB-FRAGID serve as the reference data for training PANIP.

Training Protocol of PANIP. PANIP was implemented as an ensemble of NequIP[32], an equivariant graph neural network that explicitly encodes the rotational, translational, and permutational symmetries of atomistic systems. By enforcing E(3)-equivariance in its message-passing layers, NequIP achieves high data efficiency and systematically improved accuracy compared to conventional invariant graph neural networks. This makes it particularly

View Article Online
DOI: 10.1039/D6DD00056H

well-suited for learning NCI energies, where capturing subtle orientation-dependent non-covalent effects is essential. For NequIP training, we modified the default settings as follows: a cutoff radius of 7.5 Å, inclusion of O, N, and S in the chemical symbols list, and a batch size of 512. Early stopping was applied with a patience of 100 epochs on the validation loss. Global energy rescaling was set to *dataset_total_energy_std* (the standard deviation of total energies in the dataset).

To accelerate the construction of the final model, MFAL-selected subsets from each fragment type were first used to train fragment-specific NequIP models. These fragment-wise datasets were then combined to form the unified PDB-FRAGID dataset, which was used to train the final PANIP ensemble (5 models) via five-fold cross-validation. Unless otherwise noted, predictions are reported as ensemble averages. In addition, fragment-specific models generally exhibit slightly higher accuracy on their respective subsets but reduced generalization across fragment types, whereas the unified PANIP model provides robust performance across the full diversity of protein-derived NCIs. Full training protocols, hyperparameters, and model files are available on GitHub (https://github.com/hnlab/PANIP).

Benchmark Datasets. PANIP was evaluated on four benchmark sets derived from both protein and non-protein sources. First, the low-energy representative from the original r²SCAN-3c-labeled pool was selected for each unique dimer type, resulting in 277 candidate structures. These structures were optimized at the r²SCAN-3c level to ensure well-defined equilibrium geometries, forming a dedicated benchmark set. Single-point energy calculations were performed using the ωB97X-D3BJ/def2-TZVPP method on all selected dimer geometries (with and without

View Article Online
DOI: 10.1039/D6DD00056H

optimization). This benchmark set was used to assess model performance in reproducing equilibrium interaction energies. Same-charge dimer pairs were excluded due to frequent wave function convergence failures during optimization.

Second, to evaluate model transferability beyond training data, dimers involving the same fragment types were extracted from CSD[33] using ConQuest[50]. Although the chemical fragment pairs are identical, the geometries originate from small-molecule structures (not protein environments) and thus sample distinct conformational and packing motifs—providing a complementary test of generalization across structural sources. To preserve structural integrity and chemical independence, only dimers featuring single aliphatic bonds (C-C or C-H) were retained. Pairs with the closest interatomic distances between 2 and 4 Å were selected, aligning with NCI criteria established for PDB dimers. Organometallic structures were excluded and hydrogen positions were normalized to standard bond geometries, yielding 33,274 CSD-derived dimers for analysis.

Third, to evaluate performance on non-equilibrium conformations, a biased fragment-pair random sampling method was developed. Fragments were treated as rigid bodies with fixed bond lengths and angles. Initial 3D fragment structures were generated from SMILES strings[51] using the RDKit software package[52]. One fragment was fixed at the origin, and the other was positioned within a 12 Å radius sphere centered on the fixed fragment, with rotational sampling applied. A biased random sampling protocol was implemented using a logarithmic normal distribution for the acceptance probability:

$$x = r_i^{vdw} + r_j^{vdw} - d_{ij} \qquad (2)$$

View Article Online
DOI: 10.1039/D6DD00056H

$$p(x) = \frac{1}{(-x+c)\sigma\sqrt{2\pi}} e^{-\frac{(\ln(-x+c)-\mu)^2}{2\sigma^2}} \quad (3)$$

Where x is the intermolecular overlap (Å), r_i^{vdw} and r_j^{vdw} are the van der Waals radii of atoms i and j, and d_{ij} is their interatomic distance. Parameters were set to σ=0.8, μ=1.0, and c=1.6 —chosen to maximize acceptance probability at an intermolecular overlap of ~0.1 Å (favoring physically realistic conformations) and ensure overlaps remained below 1.6 Å to avoid severe clashes[53, 54] (Figure S2). This coordinate system enabled denser sampling at shorter intermolecular distances and sparser sampling at longer distances. For each dimer type, 100 conformations were sampled, yielding a total of 15,300 conformations. Interaction energies for these dimers were computed using both PANIP and ωB97X-D3BJ/def2-TZVPP for comparison.

Finally, to enable comparison with a well-established ML potential, PANIP was benchmarked against AIMNet2[55]. AIMNet2 is a generalizable model trained on diverse equilibrium and non-equilibrium datasets at the ωB97M-D3/def2-TZVPP level with Becke–Johnson (BJ) damping, offering accuracy comparable to the reference method employed herein (ωB97X-D3BJ/def2-TZVPP)[41]. In addition to the CSD subset, we employed two external benchmarks: (i) a filtered set of 40 systems from GMTKN55[40], with performance reported using WTMAD-2 metrics, and (ii) a filtered dataset of 19,183 structures derived from a recent ChemRxiv benchmark developed for general-purpose ML potentials targeting intermolecular and noncovalent interactions[56]. High-level reference energies for both benchmarks reach or closely approximate CCSD(T)/CBS accuracy. The detailed structural information and raw prediction values for all evaluated systems are provided in our public GitHub repository. Since AIMNet2

View Article Online
DOI: 10.1039/D6DD00056H

predicts total energies, NCI energies were calculated by applying Eq.%1 to the AIMNet2 predicted total energies of the dimer and its corresponding monomers.

Docking and Rescoring. We evaluated the model ensemble's ability to predict the native-like binding pose for three well-established protein-ligand model systems[57, 58]: L99A in complex with indole (PDB 185L)[59], L99A/M102H in complex with phenol (PDB 4I7L)[60], and the M2 isoform of pyruvate kinase (PKM2) in complex with serine (PDB 4B2D)[61]. In addition, we assessed performance across a broader set of complexes—including other L99A variants (e.g., M102Q, M102H, M102E), and additional ligand classes—the results of which are summarized in Table 2. Docking was performed using default parameter settings from an automated platform as described previously[62, 63]. For all docking poses generated by DOCK 3.7[64], binding energies were first calculated using scoring functions based on the united AMBER force field (as in DOCK 3.7).

To compute binding interaction energies with MLIP, both the receptor and ligand were fragmented within a 5 Å distance cutoff around the ligand (Figure 3). Notably, for PKM2-serine, the serine ligand was decomposed into three fragments (ACET, ETAM, and ETOH). The MLIP-estimated binding interaction energy is given by

$$\Delta E_{bind} = \sum_i^{rec} \sum_j^{lig} \Delta E_{pred} + \Delta E_{lnk} - \Delta E_{dup} \qquad (4)$$

where ΔE_{pred} is the interaction energy between protein fragment i and ligand fragment j (computed using MLIP), ΔE_{lnk} accounts for energy contributions from linking atoms

View Article Online
DOI: 10.1039/D6DD00056H

(protein/ligand, green circle in Figure 3), and ΔE_{dup} corrects for double-counted atoms in fragment-fragment interactions (orange circle in Figure 3).

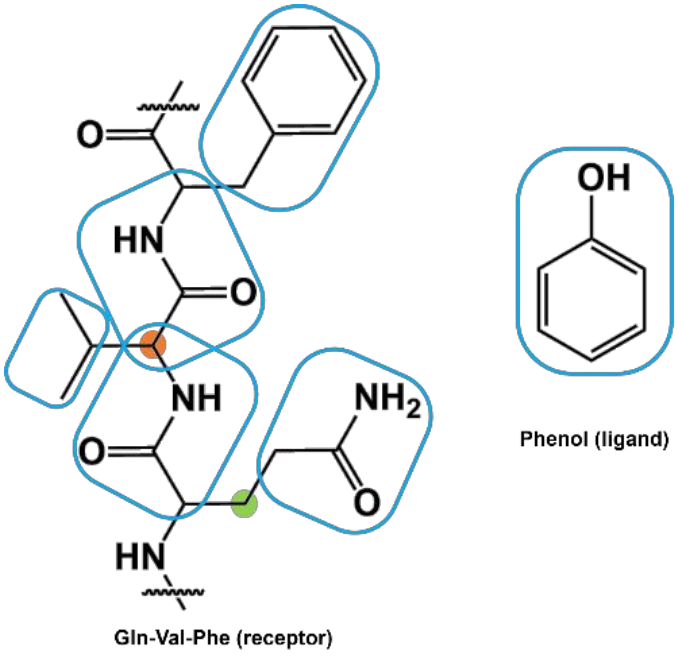

Figure 3. The sketch of protein fragmentation approach. Receptor and ligand are decomposed into fragments (blue boxes), linker atoms (green circle), and duplicated atoms (orange circle)

RESULTS AND DISCUSSION

PDB-FRAGID Dataset Construction. A major challenge in developing MLIPs is constructing training sets that simultaneously provide broad coverage of chemically diverse NCIs and remain computationally tractable for high-level QM labeling. To address this, an iterative data selection strategy using MFAL was used to reduce the initial 36.3 million PDB-derived dimers into a compact, information-rich dataset, PDB-FRAGID. Using r²SCAN-3c as a low-fidelity method, interaction energies were computed for all dimers, and a NequIP-based surrogate model was

View Article Online
DOI: 10.1039/D6DD00056H

iteratively refined to identify high-error dimers for inclusion. This process distilled the dataset to approximately 3.15 million dimers (8.7% of the original pool, Figure 4, Table S1), while preserving coverage across 17 fragment types and 153 dimer combinations and prioritizing chemically challenging systems such as charged and polar fragments (Figure 4). Although r²SCAN-3c offers a computationally efficient baseline and exhibits strong correlation with ωB97X-D3BJ energies (Figure S7), it shows systematic deviations for specific classes of noncovalent interactions. Notable discrepancies are observed for sulfur-containing systems (MAE=0.68 kcal/mol) and ionic contacts (MAE=0.96 kcal/mol), with an overall dimer RMSE near 1 kcal/mol. Therefore, ωB97X-D3BJ/def2-TZVPP calculations were used to generate the final energy labels.

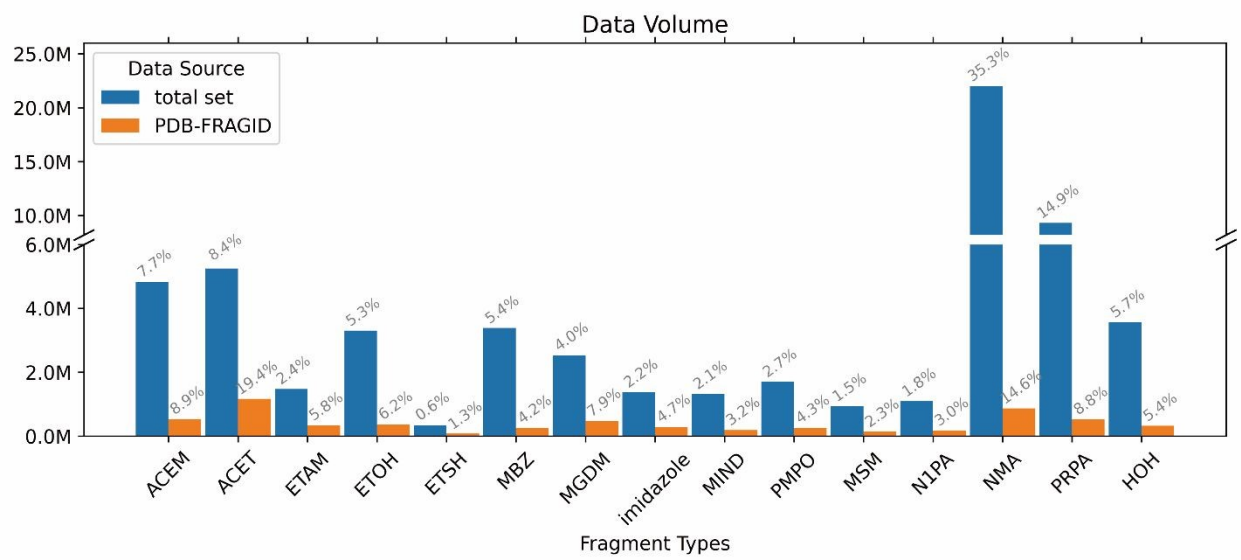

Figure 4. Data Volume for each fragment.

Model Performance on Benchmark Sets. PANIP was evaluated on four benchmark sets to assess accuracy, robustness, and transferability: low-energy dimers, optimized low-energy

View Article Online
DOI: 10.1039/D6DD00056H

dimers, CSD-derived geometries, and non-equilibrium conformations generated by biased random sampling (Figure 5). For the PDB-derived low-energy dimers, PANIP achieved a mean absolute error (MAE) of 0.09 kcal/mol, a root mean squared error (RMSE) of 0.163 kcal/mol, and an R^2=0.999 relative to ωB97X-D3BJ/def2-TZVPP, indicating excellent reproduction of equilibrium NCI energies in the training domain (Figure 5A). After geometry optimization, the MAE and RMSE increased to 0.547 and 1.207 kcal/mol, respectively (Figure S3), but remained within chemically acceptable limits, indicating that the model can capture the energetic landscape near the equilibrium structures (Figure 5B).

On the CSD-derived benchmark set, which probes transferability to small-molecule crystal environments distinct from proteins, PANIP maintained strong performance (MAE 0.171 kcal/mol, RMSE 0.507 kcal/mol, R^2=0.999; Figure 5C), underscoring its generalization beyond PDB geometries. For 15,300 randomly sampled conformations, PANIP achieved an overall MAE of 0.448 kcal/mol, RMSE of 1.372 kcal/mol, and R^2=0.996. When restricted to energetically favorable dimers (8,572 entries with interaction energy < 0 kcal/mol), the MAE and RMSE improved to 0.195 and 0.363 kcal/mol, respectively, with R^2=0.999 (Figure 6a). The repulsive contact subset with interaction energies above 0 kcal/mol (6,728 entries) showed a relatively higher error (RMSE 0.882 kcal/mol; Figure 6b). High-energy outliers (> 100 kcal/mol), representing less than 5% of our active-learning–curated dataset, were systematically underestimated due to severe steric clashes and limited training analogs (e.g., ETSH dimers; Figure 4), but these geometries are rare in realistic biomolecular applications.

View Article Online
DOI: 10.1039/D6DD00056H

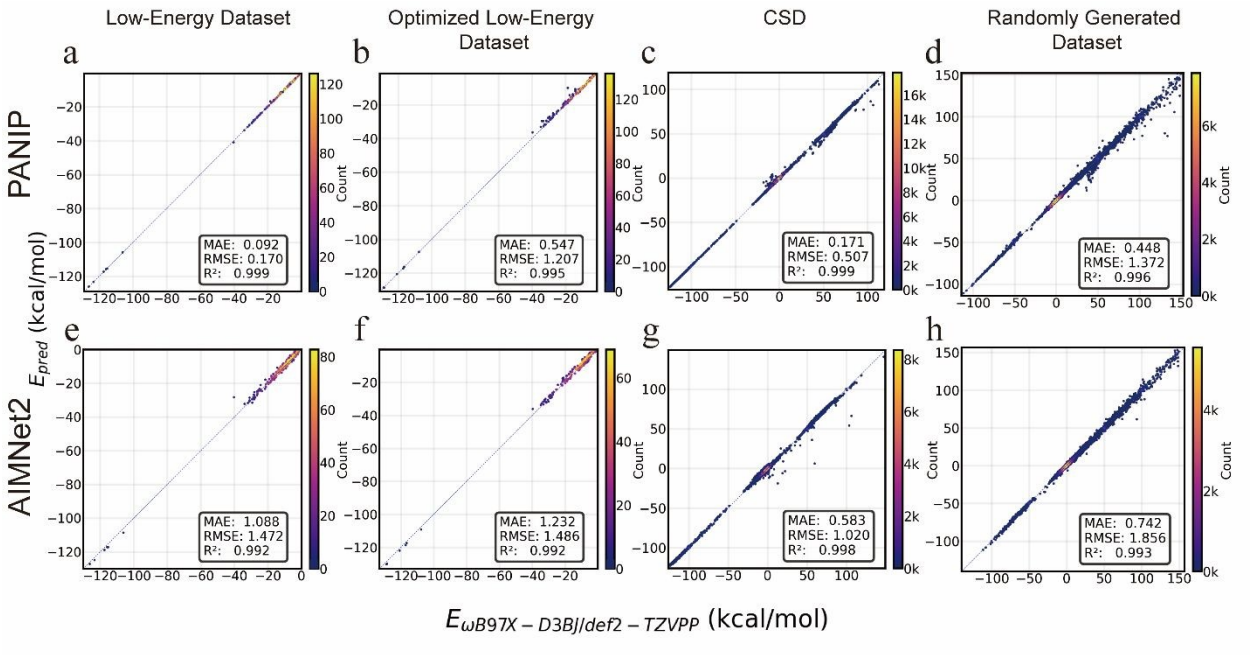

Figure 5. Correlation plots between non-covalent interaction energies calculated at ωB97X-D3BJ/def2-TZVPP level and predicted by PANIP and AIMNet2, E_{pred}, for the low-energy dataset (a, e), optimized low-energy dataset (b, f), CSD subset (c, g), and randomly generated dataset (d, h). The black line indicates perfect prediction (y=x). Color indicates the normalized KDE-based point density.

Computationally, PANIP achieves speedups of over two orders of magnitude relative to direct hybrid DFT calculations. For example, evaluating 15,300 randomly generated structures on a single CPU core requires 463 days and 11 hours for ωB97X-D3BJ/def2-TZVPP, 4 days and 20 hours for r²SCAN-3c, while PANIP completes in only 6 hours and 11 minutes, reflecting near-linear scaling compared to the formal N^4 scaling of hybrid DFT methods (cubic with modern RI/RIJCOSX implementations; Table S2)[30]. Furthermore, we benchmarked PANIP and AIMNet2 across four independent benchmark sets on a single NVIDIA 4090 GPU. In end-to-end

View Article Online
DOI: 10.1039/D6DD00056H

workflows for NCI energy prediction, PANIP reduced wall-clock computation time by approximately 1.3-fold relative to AIMNet2 (Table S3).

Considering that NCIs are diverse and highly sensitive to subtle structural variations, even minor conformational changes can lead to pronounced energy differences[65, 66]. The model's success in predicting unseen conformations underscores the sufficiency of PDB-derived dimers in sampling conformational space, enabling reliable generalization—a critical challenge in developing MLIPs. This emphasizes the importance of representative training sets for capturing diverse conformational and interaction landscapes. Such capabilities are particularly vital for applications in drug discovery and protein engineering, where accurate prediction of NCIs is essential.

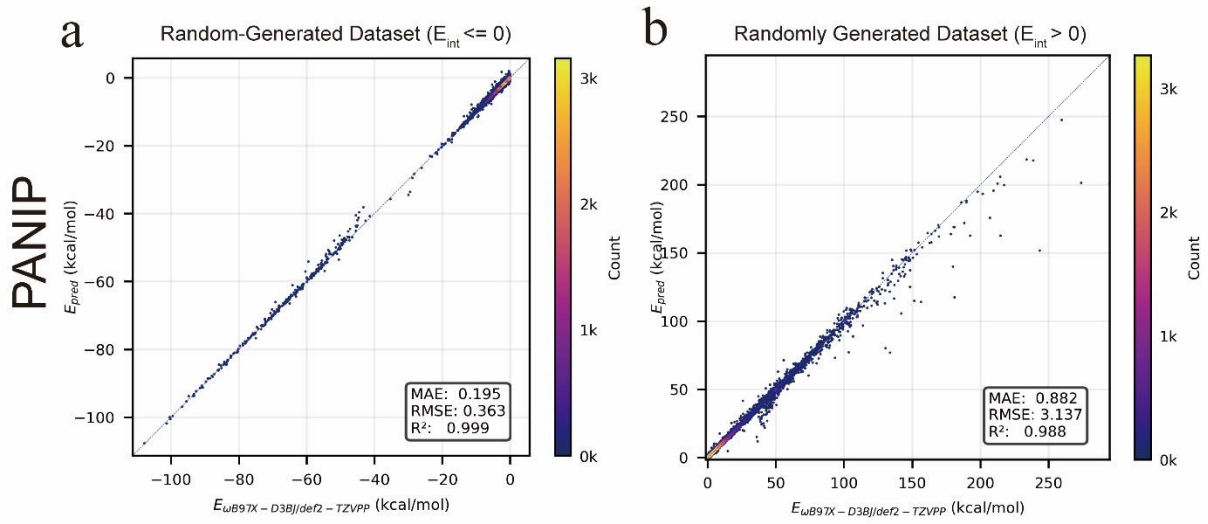

Figure 6. Comparison of noncovalent interaction energies calculated at the ωB97X-D3BJ/def2-TZVPP level and PANIP-predicted values for the randomly generated dataset: (a) samples with interaction energy < 0 kcal/mol, and (b) samples with interaction energies > 0 kcal/mol.

View Article Online
DOI: 10.1039/D6DD00056H

Comparison with AIMNet2. Benchmarking against AIMNet2 on the CSD-derived geometries (Figure 5e, f) demonstrates PANIP's superior performance. On the full CSD subset, AIMNet2 exhibited larger errors (MAE 0.583 kcal/mol, RMSE 1.020 kcal/mol, and R^2=0.998), alongside a tendency to underestimate strongly repulsive dimers. Similar limitations were observed across additional test sets, where AIMNet2 MAEs exceeded 1 kcal/mol for low-energy dimers and optimized dimers, while the randomly generated dataset showed an MAE of 0.742 kcal/mol (Figure 5, Figure S4). Correlation plots for AIMNet2 on the randomly generated dataset, split into attractive and repulsive subsets (Figure S4), further reveal elevated errors relative to PANIP in both interaction categories.

PANIP and AIMNet2 were further evaluated on two independent external datasets featuring high-level reference energies at or near the CCSD(T)/CBS accuracy, where PANIP consistently outperformed AIMNet2. On a filtered subset of GMTKN55[40] curated to match model applicability (40 structures), PANIP achieved a WTMAD-2 of 0.72 kcal/mol, compared to 1.21 kcal/mol for AIMNet2. On a larger intermolecular benchmark containing 19,183 structures sourced from a intermolecular interaction database[56], PANIP yielded an MAE of 0.35 kcal/mol, an RMSE of 0.46 kcal/mol, and R^2=0.99, whereas AIMNet2 returned an MAE of 0.91 kcal/mol, an RMSE of 6.16 kcal/mol, and R^2=–0.20. These outcomes confirm PANIP's robust accuracy and strong linear correlation with reference data, in contrast to the substantial errors and poor correlation observed for AIMNet2. Across both benchmark sets, the filtered structures comprise fragments identical to the fragment defined in Figure 2, as well as a broad range of distinct molecular systems, validating the robustness of PANIP's generalization. Full details of the 14 constituent sub-datasets are summarized in Table S4, and raw prediction values are provided in our public GitHub repository."

View Article Online
DOI: 10.1039/D6DD00056H

These deviations likely stem from the indirect calculation of interaction energies via total energy difference. Unlike QM reference calculations that include counterpoise corrections, MLIPs trained solely on total energies fail to account for such corrections. In contrast, PANIP directly predicts high-level NCI energies, avoids compounded errors from total energy differences, and consistently achieves sub-chemical-accuracy MAEs, including for charged and strongly interacting dimers. These results highlight the importance of high-fidelity NCI training data and a protein-specific design for accurate modeling of biomolecular NCIs.

Exploring NCI Patterns in the PDB. Leveraging PANIP's accuracy and efficiency, the interaction energies were predicted for all 36.3 million PDB-derived dimers, enabling systematic mapping of energy-geometry correlations across the entire dataset (Figure 7, Figure S5). Representative analyses focus on both well-characterized interactions like cation-π interactions, and previously underexplored interactions like dimethyl sulfide–aromatic contacts, to illustrate the model's ability to recover known patterns and reveal new trends. The lowest-energy representative conformations were selected for each interaction pattern and their NCI maps[67] were visualized to highlight regions where weak interactions predominantly occur, a detailed description of the NCI visualization method is provided in the Supporting Information.

For Cation-π Interaction, dimers involving lysine (ETAM) with tyrosine (PMPO) or tryptophan (MIND) were examined, yielding 28,914 unique ETAM-PMPO and 24,098 ETAM-MIND dimers. Spatial distribution of the cation relative to the aromatic ring revealed distinct low-energy geometries with centroid distances varying from 4.5 to 7 Å. Positively charged

View Article Online
DOI: 10.1039/D6DD00056H

ETAM monomers are predominantly localized around the phenolic hydroxyl group of PMPO, forming a ring-shaped low-energy region, as well as another low-energy region above the aromatic π ring (Figure 7a and Figure S5a). Three representative conformations were identified. *Struc_1* with the lowest-energy (-20.41 kcal/mol) features ETAM simultaneously engaging in a cation-π interaction (N atom 2.9 Å from the phenol ring center) and a nonclassical hydrogen bond between the methyl group of ETAM and the hydroxyl group of PMPO. Cation-π interaction dimers accounted for 17.7% of ETAM-PMPO dimers. *Struc_2* (-19.52 kcal/mol) is dominated by a strong NH···O hydrogen bond between ETAM's NH group and PMPO's hydroxyl oxygen, with a prevalence of 19.8%. The conformation involving a CH···π interaction between the methylene groups of ETAM and the aromatic ring of PMPO is less frequent and higher in energy (Figures 4a, *struc_3*).

Due to the enhanced electron density on Trp's six-membered ring, ETAM tends to interact with MIND above this ring (Figure S5b), with centroid distances of 5-6.5 Å. Figure 7b shows the lowest-energy structures of three representative NCI patterns. The lowest-energy structure, *struc_1* (-24.22 kcal/mol), involves a strong cation-π interaction between ETAM's nitrogen atom and the indole ring. Cation-π interaction dimers accounted for 27.6% of ETAM-MIND dimers. Conformation represented by *struc_2* (-15.44 kcal/mol) adopts a "parallel" orientation between ETAM heavy-atom plane and the indole ring, in which hydrogens attached to the three heavy atoms form weak XH···π interactions with the aromatic ring simultaneously. In contrast, *struc_3*

View Article Online
DOI: 10.1039/D6DD00056H

(-10.83 kcal/mol) features a perpendicular arrangement dominated by methylene CH···π interactions accounting for 33.7% of dimers. These findings align with prior studies[68 – 72], supporting the reliability of PANIP for large-scale NCI analysis.

For previously underexplored dimethyl sulfide–aromatic Interactions[73, 74], 78,674 MSM (methionine)–MBZ (phenylalanine) dimers were extracted and analyzed. These revealed five distinct low-energy geometric patterns (Figure%4c, Figure S5c, and Figure S6), spanning interaction distances of 3.5–7%Å with a predominant peak near 5.1%Å and interaction energies ranging from -4.9 to 0.2%kcal/mol (most populated around -1.6 kcal/mol). Mapping the sulfur positions shows that low-energy structures (~ -4 kcal/mol) concentrate in a previously unreported ring-shaped region above the aromatic plane, reflecting stereochemical preferences of sulfur and complementary electrostatic interactions (Figure S5c). Joint analysis of energetic and orientational preferences identifies five low-energy patterns. In the innermost region, C–S bond points toward the ring (Figure S6a), consistent with stereochemical preferences of divalent sulfur approaching the positively polarized aromatic edge (Figure 7c, *struc_5*, -1.52 kcal/mol). Parallel arrangements, in which both the C–S bond and the MSM plane align parallel to the aromatic ring (Figure S6b and S6c), yield the lowest energy (*struc_1*, -4.91 kcal/mol) through complementary electrostatic interactions between sulfur and methyl dipoles. When positioned farther from the aromatic ring, MSM assumes an outward-pointing C–S bond geometry (Figure S6a; *struc_4*, -3.06 kcal/mol), consistent with π electrons interacting with the C–S σ antibonding orbital. A fourth pattern places MSM above the ring with a perpendicular orientation, where one C – S

View Article Online
DOI: 10.1039/D6DD00056H

bond aligns antiparallel to the ring dipole (*struc_2*, -4.43 kcal/mol). Geometries featuring C–H···S contacts are also observed (*struc_3*, -3.21 kcal/mol), reflecting weak hydrogen-bond donation from aromatic C – H groups to sulfur lone pairs. Collectively, these motifs underscore how methionine-like sulfur can engage aromatics through directional stereoelectronic and electrostatic complementarity, providing structural principles relevant to stereospecific recognition in Met-containing binding environments.

View Article Online
DOI: 10.1039/D6DD00056H

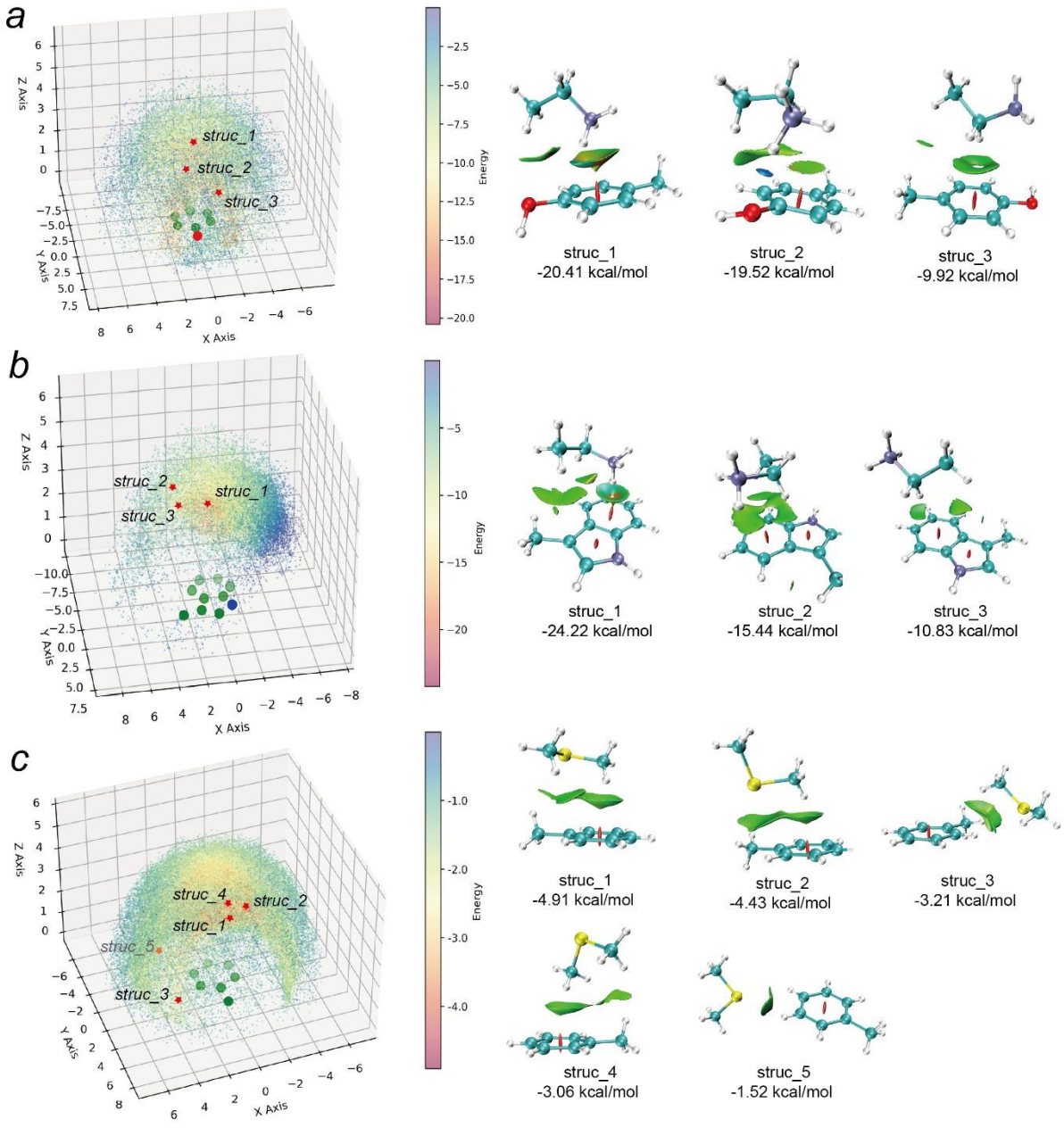

Figure 7. Spatial distribution and representative low-energy structure of ETAM-PMPO (a), ETAM-MIND (b), and MBZ-MSM (c) dimers. In the left column, each scatter point represents the position of the fragment's central atom (N or S) relative to the central fragment. The color of each scatter point corresponds to the energy scale shown in the adjacent color bar. Circles on the xy-plane represent heavy atoms of the central fragment: green for carbon atoms, red for oxygen

View Article Online
DOI: 10.1039/D6DD00056H

atoms, and blue for nitrogen. Red stars mark the positions of the structures shown in the right column. The right column depicts the lowest-energy structures representing characteristic interaction patterns for each dimer. Isosurfaces show NCI maps between the fragments, and the NCI energy of each structure is given below the corresponding image.

Application to Protein-Ligand Complexes. To demonstrate the potential application in biomolecular modeling, PANIP was applied as a fragment-based scoring function for protein-ligand docking on three model systems: (1) indole bound to the apolar L99A mutant (PDB: 185L)[59], (2) phenol bound to the polar L99A/M102H double mutant (PDB: 4I7L)[60], and (3) serine bound to pyruvate kinase M2 (PKM2, PDB: 4B2D)[75] (Figure 8). The interaction energies ΔE_{bind} of all docking poses were calculated using both our PANIP scoring function (Eq.4), AIMNet2 scoring function and the united AMBER force field energy function in DOCK. Pose prediction accuracy was quantified via root mean square deviation (RMSD) of ligand heavy atoms relative to crystallographic poses (Table 1). Native poses extracted from crystal structures were included in the docking pool for validation.

Across these test systems, PANIP achieved substantial improvement in pose ranking compared with a conventional AMBER-based docking score, while delivering overall performance comparable to AIMNet2. For the L99A-indole system, both PANIP and AIMNet2 correctly prioritized the native pose within the top 100 docking poses, while the AMBER-derived DOCK score placed the native pose outside the top 100 and, selected an alternate pose with an RMSD of 0.34 Å as its highest-ranked prediction (Figure 8a-c). For the L99A/M102H-phenol system, PANIP ranked the native pose first, outperforming AIMNet2 (rank 20) and DOCK (No. 76;

View Article Online
DOI: 10.1039/D6DD00056H

Figure 8d-f). For the PKM2-serine system, the serine ligand was decomposed into three fragments (ACET, ETAM, and ETOH) for binding energy estimation. PANIP ranked the native pose third, whereas AIMNet2 and DOCK assigned ranks of 9 and 46. Additionally, the top-scoring pose from PANIP exhibited an RMSD of 0.434 Å, lower than the optimal poses from AIMNet2 (0.54 Å) and DOCK (1.24 Å). The top-scoring PANIP pose forms a hydrogen bond (2.321 Å) between the ligand serine's hydroxyl group and the carbonyl group of ILE469, differing from the native crystal ligand pose in which the serine hydroxyl engages the carbonyl group of ARG43 via a 2.149 Å hydrogen bond (Figure 8g-i). However, the top PANIP-selected pose positioned the serine hydroxyl group near the serine NH3+ fragment, potentially inducing intramolecular repulsion not captured by the pairwise fragment approximation, illustrating a key limitation of the current framework.

Table 1. Binding pose prediction results for three protein-ligand systems: L99A in complex with indole (PDB 185L), L99A/M102H in complex with phenol (PDB 1LI2), and PKM2 in complex with serine (PDB 4B2D). Performance is compared across three scoring functions: PANIP, AIMNet2, and DOCK. The native pose corresponds to the crystal ligand binding pose, while the best prediction denotes the top-scoring pose identified by each scoring function.

Ligand	Protein	PDB ID	Ranking of Native Pose			RMSD of the top-scoring pose (Å)		
			PANIP	AIMNet2	DOCK	PANIP	AIMNet2	DOCK
Indole	L99A	185L	1	1	101	0.19	0.19	0.34
Phenol	L99A/M102H	4I7L	1	21	76	0.18	0.08	0.21

View Article Online
DOI: 10.1039/D6DD00056H

Serine	PKM2	4B2D	3	9	12	0.43	0.54	1.24

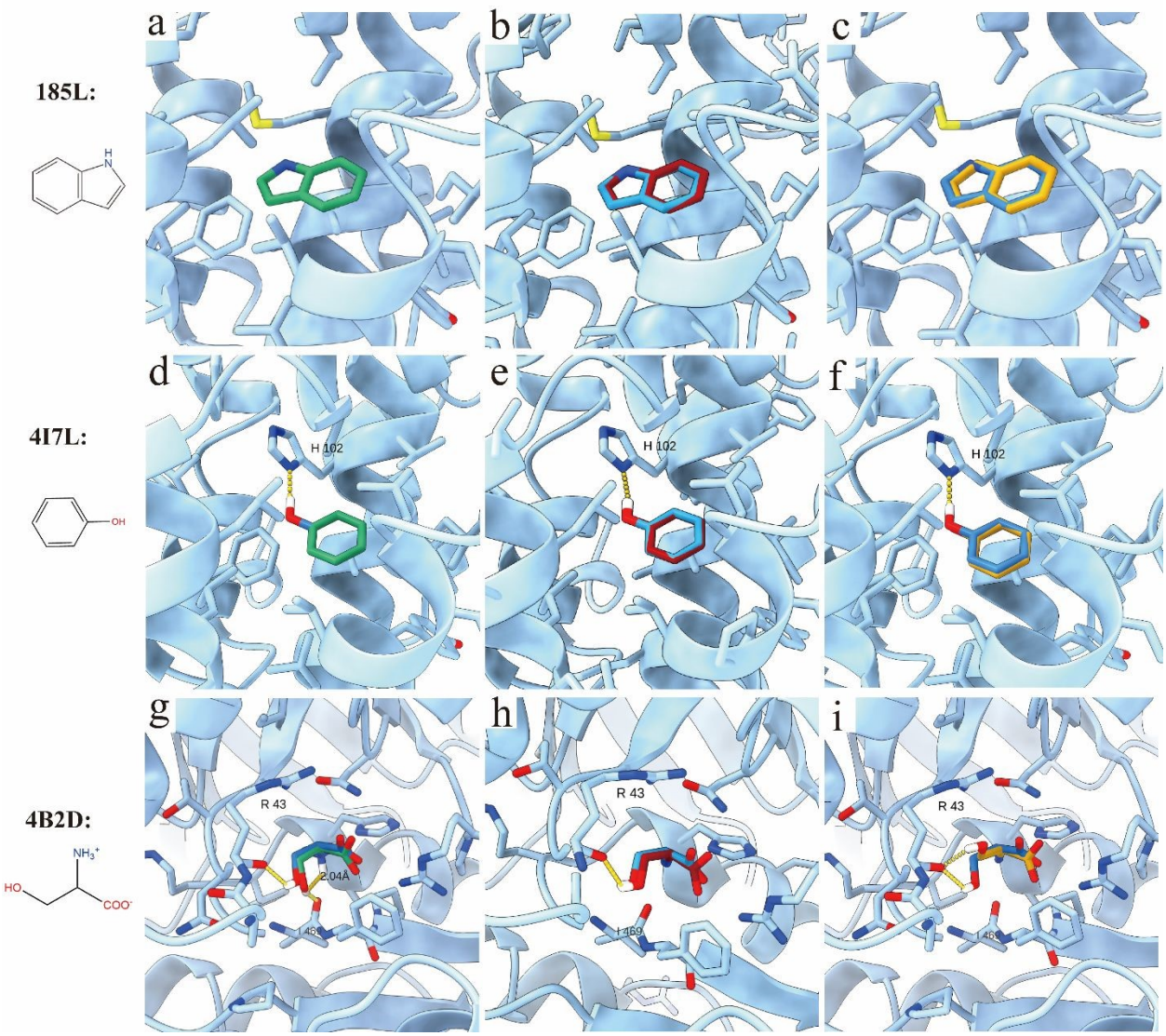

Figure 8. Structures and the top-scoring pose of indole ligand in L99A (top), phenol ligand in L99A/M102H (middle), and serine in PKM2 (bottom) predicted by PANIP (green), AIMNet2 (red), and DOCK (yellow). The native structures are colored in blue for comparison. Structures are visualized in ChimeraX[76].

View Article Online
DOI: 10.1039/D6DD00056H

Additional model systems were evaluated (Table 2), including various alkyl-substituted benzene isomers and phenol binding to L99A and its polar or charged variants (L99A/M102Q, M102H, and M102E), as well as amino acid ligands such as alanine and phenylalanine interacting with PKM2. Across these model systems, PANIP-based scoring function recovered the native pose as the top-ranked solution in 50% of the 22 tested complexes, consistently outperforming the DOCK score in recovering native poses and reducing RMSDs of top-ranked structures. Notably, this performance is achieved without explicit long-range electrostatic or solvation corrections, suggesting that accurate short-range NCI energetics are a dominant factor for pose discrimination in these relatively simple model systems.

Table 2. Binding pose prediction results for 19 protein-ligand systems: ligands in complex with L99A, L99A/M102Q, L99A/102H, L99A/M102E, and PKM2. Performance is compared across three scoring functions: PANIP, AIMNet2, and DOCK. The native pose corresponds to the crystal ligand binding pose, while the best prediction denotes the top-scoring pose identified by each scoring function*.

Ligand	PDB ID	Ranking of Native Pose			RMSD of the best prediction (Å)		
		PANIP	AIMNet2	DOCK	PANIP	AIMNet2	DOCK
L99A							
toluene	4W53	89	93	101	0.43	0.43	0.42
isobutylbenzene	184L	1	1	101	0.36	0.37	0.38
n-Butylbenzene	4W57	1	1	20	1.38	1.42	1.32

View Article Online
DOI: 10.1039/D6DD00056H

benzene	3HH4	1	20	101	0.46	0.46	0.61
ethylbenzene	3HH6	1	63	101	0.44	0.44	0.58
propylbenzene	4W55	18	86	101	0.29	0.43	0.38
sec-butylbenzene	4W56	1	19	5	1.08	1.09	1.11
octylbenzene	4W59	5	13	101	0.19	0.19	0.32
			L99A/M102Q				
benzene	5JWT	19	20	101	0.23	0.23	0.59
ethylbenzene	5JWV	1	22	101	0.29	0.70	0.91
phenol	1LI2	66	61	69	0.27	0.18	0.15
			L99A/M102H				
benzene	4I7J	18	72	101	0.32	0.46	0.50
toluene	4I7K	37	36	101	0.33	0.28	0.64
			L99A/M102E				
benzene	3GUJ	1	1	101	0.23	0.23	0.38
toluene	3GUK	1	1	101	0.43	1.63	0.62
ethylbenzene	3GUL	1	3	95	0.19	0.21	0.35
phenol	3GUO	62	80	56	0.16	0.16	0.22
			PKM2				
alanine	2G50	3	20	30	0.71	0.71	0.71
phenylalanine	4FXJ	1	1	1	0.42	0.45	0.42

* For structures with alternative conformations, the scoring results are reported for the conformation with the highest crystallographic occupancy.

CONCLUSIONS

View Article Online
DOI: 10.1039/D6DD00056H

We present PANIP, a protein-specific MLIP for NCIs, trained on PDB-derived fragment dimers with ωB97X-D3BJ/def2-TZVPP reference energies. A multi-fidelity active learning workflow enabled the construction of PDB-FRAGID, a high-precision, low-redundancy dataset that compresses 36.3 million dimers to ~ 3.15 million representative structures while preserving the diversity of 17 fragment types and 153 dimer combinations. PANIP achieves sub-chemical-accuracy MAEs across multiple benchmark sets, generalizes to CSD-derived and non-equilibrium geometries, and offers near force-field computational cost, making it suitable for large-scale exploration of protein NCI landscapes.

Compared to the widely adopted AIMNet2 potential, PANIP delivers consistently lower errors across diverse benchmark datasets, highlighting the advantages of targeted, high-fidelity NCI training data. Importantly, the benchmark sets contain both fragments matching those defined in this work and numerous distinct structural motifs, thereby validating PANIP's robust generalization. When integrated into a fragment-based energy decomposition scheme, PANIP functions as an effective scoring function for protein–ligand docking. It outperforms conventional AMBER-based scoring in pose ranking across diverse model systems and achieves performance comparable to AIMNet2.

Although the dataset was constructed from only 17 fragment types, these moieties were carefully selected to recapitulate the prevalent chemical environments within proteins, including amino acid side chains, backbone motifs, and water molecule. This compact yet chemically representative fragment set provides sufficient coverage for training PANIP, as evidenced by its reliable transferability to unseen protein–ligand complexes. Although AIMNet2 exhibits inferior accuracy than PANIP on the present benchmark datasets, it remains applicable to a much broader

View Article Online
DOI: 10.1039/D6DD00056H

range of chemical systems. Accordingly, our protein-tailored dataset and PANIP model act as a complementary tool to general-purpose ML potentials, delivering enhanced, targeted accuracy for biomolecular non-covalent interactions.

While the current framework focuses on pairwise interactions, ongoing efforts aim to expand fragment diversity to cover broader chemical spaces, to incorporate long-range electrostatic, multi-body, and solvation effects. For example, the important long-range contribution can be incorporated through analytic or hybrid correction schemes, as exemplified by latent Ewald summation (LES)[77], SO3LR[78], and FENNIX[79]. These methods infer electrostatics and dispersion interactions from short-range features with minimal empirical calibration. Additional strategies include integrating well-established force field correction terms[80, 81], leveraging ML-based correction schemes[82, 83], or employing metamodeling approaches[84]. Together, PDB-FRAGID and PANIP provide a foundation for scalable, QM-accurate modeling of protein NCIs and for refining both data-driven and classical force-field descriptions in biomolecular simulations.

Author contributions

L.Z. and N.H. conceived ideas and wrote the manuscript. L.Z. performed most of the work. X.Z. provided data for Figure 5c and Table 2. Y.P., L.Z., L.H., and J.Y. contributed to exploratory work during the initial stages of the project. All authors reviewed the manuscript.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

View Article Online
DOI: 10.1039/D6DD00056H

The PDB-FRAGID dataset is available at https://github.com/hnlab/PDB-FRAGID and can be downloaded from Zenodo (https://zenodo.org/records/18213106). The trained models of PANIP, benchmark sets and codes used in this paper are available for download from a GitHub repository https://github.com/hnlab/PANIP.

Code availability

Related codes are publicly available at https://github.com/hnlab/PANIP.

Author information

Corresponding Author

*E-mail: huangniu@nibs.ac.cn

ORCID

Niu Huang: 0000-0002-6912-033X

Lejia Zeng: 0009-0004-3410-8404

Funding Sources

Beijing Municipal Science and Technology Commission, Administrative Commission of Zhongguancun Science Park, Grant/Award Number: Z201100005320012

Notes

The authors declare no competing financial interest.

Acknowledgment

View Article Online
DOI: 10.1039/D6DD00056H

This work is supported by Beijing Municipal Science & Technology Commission (Z201100005320012 to Niu Huang) and Tsinghua University.

Abbreviations

AL, active learning; BSSE, basis set superposition; CSD, Cambridge Structural Database; HTMD, high-throughput molecular dynamics; MAE, mean absolute error; MLIPS, machine learning interatomic potentials; ML, machine learning; NCIs, non-covalent interactions; NequIP, Neural Equivariant Interatomic Potentials; PDB, Protein Data Bank; PDB-FRAGID, PDB Fragment Interaction Dataset; PES, potential energy surfaces; PKM2, M2 isoform of pyruvate kinase; RMSD, root mean square deviation; QM, quantum mechanical.

REFERENCES

(1) Han, Y.; Ali, I.; Wang, Z.; Cai, J.; Wu, S.; Tang, J.; Zhang, L.; Ren, J.; Xiao, R.; Lu, Q.; Hang, L.; Luo, H.; Li, J. Machine Learning Accelerates Quantum Mechanics Predictions of Molecular Crystals. Physics Reports 2021, 934, 1–71. https://doi.org/10.1016/j.physrep.2021.08.002.

(2) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. J. Phys. Chem. Lett. 2020, 11 (6), 2336–2347. https://doi.org/10.1021/acs.jpclett.9b03664.

(3) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. Chem. Rev. 2021, 121 (16), 10037–10072. https://doi.org/10.1021/acs.chemrev.0c00868.

View Article Online
DOI: 10.1039/D6DD00056H

(4) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. Nat Rev Chem 2020, 4 (7), 347–358. https://doi.org/10.1038/s41570-020-0189-9.

(5) Behler, J.; Csányi, G. Machine Learning Potentials for Extended Systems: A Perspective. Eur. Phys. J. B 2021, 94 (7), 142. https://doi.org/10.1140/epjb/s10051-021-00156-1.

(6) Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems | Chemical Reviews. https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107 (accessed 2025-01-06).

(7) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. Chem. Sci. 2017, 8 (10), 6924–6935. https://doi.org/10.1039/C7SC02267K.

(8) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; Bergdorf, M.; Klepeis, J. L.; Shaw, D. E. Quantum Chemical Benchmark Databases of Gold-Standard Dimer Interaction Energies. Sci Data 2021, 8 (1), 55. https://doi.org/10.1038/s41597-021-00833-x.

(9) 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 (7), 4192–4202. https://doi.org/10.1021/acs.jctc.0c00121.

View Article Online
DOI: 10.1039/D6DD00056H

(10) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., III. OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features. The Journal of Chemical Physics 2020, 153 (12), 124111. https://doi.org/10.1063/5.0021955.

(11) Ullah, A.; Dral, P. O. Molecular Quantum Chemical Data Sets and Databases for Machine Learning Potentials. ChemRxiv August 21, 2024. https://doi.org/10.26434/chemrxiv-2024-w3ld0.

(12) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. ChemRxiv December 19, 2019. https://doi.org/10.26434/chemrxiv.11365040.v1.

(13) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; De Fabritiis, G.; Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. Sci Data 2023, 10 (1), 11. https://doi.org/10.1038/s41597-022-01882-6.

(14) 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 (17), 1985–1993. https://doi.org/10.1039/B600027D.

(15) A Perspective on Deep Learning for Molecular Modeling and Simulations. https://doi.org/10.1021/acs.jpca.0c04473.

View Article Online
DOI: 10.1039/D6DD00056H

(16) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. Accurate Fourth-Generation Machine Learning Potentials by Electrostatic Embedding. J. Chem. Theory Comput. 2023, 19 (12), 3567–3579. https://doi.org/10.1021/acs.jctc.2c01146.

(17) Kocer, E.; Ko, T. W.; Behler, J. Neural Network Potentials: A Concise Overview of Methods. Annual Review of Physical Chemistry 2022, 73 (Volume 73, 2022), 163–186. https://doi.org/10.1146/annurev-physchem-082720-034254.

(18) Eastman, P.; Pritchard, B. P.; Chodera, J. D.; Markland, T. E. Nutmeg and SPICE: Models and Data for Biomolecular Machine Learning. J. Chem. Theory Comput. 2024, 20 (19), 8583–8593. https://doi.org/10.1021/acs.jctc.4c00794.

(19) Levine, D. S.; Shuaibi, M.; Spotte-Smith, E. W. C.; Taylor, M. G.; Hasyim, M. R.; Michel, K.; Batatia, I.; Csányi, G.; Dzamba, M.; Eastman, P.; Frey, N. C.; Fu, X.; Gharakhanyan, V.; Krishnapriyan, A. S.; Rackers, J. A.; Raja, S.; Rizvi, A.; Rosen, A. S.; Ulissi, Z.; Vargas, S.; Zitnick, C. L.; Blau, S. M.; Wood, B. M. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. arXiv May 13, 2025. https://doi.org/10.48550/arXiv.2505.08762.

(20) Wood, B. M.; Dzamba, M.; Fu, X.; Gao, M.; Shuaibi, M.; Barroso-Luque, L.; Abdelmaqsoud, K.; Gharakhanyan, V.; Kitchin, J. R.; Levine, D. S.; Michel, K.; Sriram, A.; Cohen, T.; Das, A.; Rizvi, A.; Sahoo, S. J.; Ulissi, Z. W.; Zitnick, C. L. UMA: A Family of Universal Models for Atoms. arXiv June 30, 2025. https://doi.org/10.48550/arXiv.2506.23971.

View Article Online
DOI: 10.1039/D6DD00056H

(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Research 2000, 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(22) Ferreira de Freitas, R.; Schapira, M. A Systematic Analysis of Atomic Protein–Ligand Interactions in the PDB. Med. Chem. Commun. 2017, 8 (10), 1970–1981. https://doi.org/10.1039/C7MD00381A.

(23) Burra, P. V.; Zhang, Y.; Godzik, A.; Stec, B. Global Distribution of Conformational States Derived from Redundant Models in the PDB Points to Non-Uniqueness of the Protein Structure. Proceedings of the National Academy of Sciences 2009, 106 (26), 10505–10510. https://doi.org/10.1073/pnas.0812152106.

(24) Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Langer, T.; Wolber, G. The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery. J. Med. Chem. 2008, 51 (22), 7021–7040. https://doi.org/10.1021/jm8005977.

(25) Bank, R. P. D. PDB Statistics: Overall Growth of Released Structures Per Year. https://www.rcsb.org/stats/growth/growth-released-structures (accessed 2025-01-21).

(26) Hernandez-Garcia, A.; Saxena, N.; Jain, M.; Liu, C.-H.; Bengio, Y. Multi-Fidelity Active Learning with GFlowNets. arXiv September 1, 2024. https://doi.org/10.48550/arXiv.2306.11715.

View Article Online
DOI: 10.1039/D6DD00056H

(27) Peherstorfer, B.; Willcox, K.; Gunzburger, M. Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization. SIAM Rev. 2018, 60 (3), 550–591. https://doi.org/10.1137/16M1082469.

(28) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. The Journal of Chemical Physics 2018, 148 (24), 241733. https://doi.org/10.1063/1.5023802.

(29) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A “Swiss Army Knife” Composite Electronic-Structure Method. J. Chem. Phys. 2021, 154 (6), 064103. https://doi.org/10.1063/5.0040021.

(30) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. Phys. Chem. Chem. Phys. 2008, 10 (44), 6615–6620. https://doi.org/10.1039/B810189B.

(31) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. Journal of Computational Chemistry 2011, 32 (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(32) 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 (1), 2453. https://doi.org/10.1038/s41467-022-29939-5.

View Article Online
DOI: 10.1039/D6DD00056H

(33) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. Acta Cryst B 2016, 72 (2), 171–179. https://doi.org/10.1107/S2052520616003954.

(34) Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M. The Multiple Roles of Computational Chemistry in Fragment-Based Drug Design. J Comput Aided Mol Des 2009, 23 (8), 459–473. https://doi.org/10.1007/s10822-009-9284-1.

(35) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Research 2019, 47 (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.

(36) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. J. Chem. Theory Comput. 2016, 12 (4), 1845–1852. https://doi.org/10.1021/acs.jctc.6b00049.

(37) Smith, J. G.; McGraw-Hill Companies. Organic Chemistry; McGraw-Hill Education, 2016.

(38) Neese, F. The ORCA Program System. WIREs Computational Molecular Science 2012, 2 (1), 73–78. https://doi.org/10.1002/wcms.81.

(39) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. Molecular Physics 2017, 115 (19), 2315–2372. https://doi.org/10.1080/00268976.2017.1333644.

View Article Online
DOI: 10.1039/D6DD00056H

(40) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. Phys. Chem. Chem. Phys. 2017, 19 (48), 32184–32215. https://doi.org/10.1039/C7CP04913G.

(41) Najibi, A.; Goerigk, L. The Nonlocal Kernel in van Der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. J. Chem. Theory Comput. 2018, 14 (11), 5725–5738. https://doi.org/10.1021/acs.jctc.8b00842.

(42) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, Approximate and Parallel Hartree–Fock and Hybrid DFT Calculations. A 'Chain-of-Spheres' Algorithm for the Hartree–Fock Exchange. Chemical Physics 2009, 356 (1), 98–109. https://doi.org/10.1016/j.chemphys.2008.10.036.

(43) Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. Phys. Chem. Chem. Phys. 2006, 8 (9), 1057–1065. https://doi.org/10.1039/B515623H.

(44) Hill, J. G.; Platts, J. A. Auxiliary Basis Sets for Density-Fitted MP2 Calculations: Correlation-Consistent Basis Sets for the 4d Elements. J. Chem. Theory Comput. 2009, 5 (3), 500–505. https://doi.org/10.1021/ct8005584.

(45) Zheng, J.; Xu, X.; Truhlar, D. G. Minimally Augmented Karlsruhe Basis Sets. Theor Chem Acc 2011, 128 (3), 295–305. https://doi.org/10.1007/s00214-010-0846-z.

View Article Online
DOI: 10.1039/D6DD00056H

(46) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. Molecular Physics 1970, 19 (4), 553–566. https://doi.org/10.1080/00268977000101561.

(47) Simon, S.; Duran, M.; Dannenberg, J. J. How Does Basis Set Superposition Error Change the Potential Surfaces for Hydrogen-bonded Dimers? The Journal of Chemical Physics 1996, 105 (24), 11024–11031. https://doi.org/10.1063/1.472902.

(48) van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. State of the Art in Counterpoise Theory. Chem. Rev. 1994, 94 (7), 1873–1885. https://doi.org/10.1021/cr00031a007.

(49) 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 (1), 49–57. https://doi.org/10.1021/ct400149j.

(50) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. Acta Cryst B 2002, 58 (3), 389–397. https://doi.org/10.1107/S0108768102003324.

(51) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. 1988, 28 (1), 31–36. https://doi.org/10.1021/ci00057a005.

View Article Online
DOI: 10.1039/D6DD00056H

(52) RDKit. http://www.rdkit.org/ (accessed 2023-05-23).

(53) Limpert, E.; Stahel, W. A.; Abbt, M. Log-Normal Distributions across the Sciences: Keys and Clues: On the Charms of Statistics, and How Mechanical Models Resembling Gambling Machines Offer a Link to a Handy Way to Characterize Log-Normal Distributions, Which Can Provide Deeper Insight into Variability and Probability—Normal or Log-Normal: That Is the Question. BioScience 2001, 51 (5), 341–352. https://doi.org/10.1641/0006-3568(2001)051%255B0341:LNDATS%255D2.0.CO;2.

(54) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a Visualization System for Exploratory Research and Analysis. J Comput Chem 2004, 25 (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(55) Anstine, D. M.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs. Chem. Sci. 2025, 16 (23), 10228–10244. https://doi.org/10.1039/D4SC08572H.

(56) Nayal, K. S.; Cho, I.; Isayev, O. Benchmarking Universal Machine-Learned Interatomic Potentials for Intermolecular and Noncovalent Interactions. ChemRxiv 2026 (0218). https://doi.org/10.26434/chemrxiv.15000203/v1.

(57) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A Model Binding Site for Testing Scoring Functions in Molecular Docking. J Mol Biol 2002, 322 (2), 339–355. https://doi.org/10.1016/s0022-2836(02)00777-5.

View Article Online
DOI: 10.1039/D6DD00056H

(58) Li, Y.; Bao, M.; Yang, C.; Chen, J.; Zhou, S.; Sun, R.; Wu, C.; Li, X.; Bao, J. Computer-Aided Identification of a Novel Pyruvate Kinase M2 Activator Compound. Cell Proliferation 2018, 51 (6), e12509. https://doi.org/10.1111/cpr.12509.

(59) Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. Biochemistry 1995, 34 (27), 8576–8588. https://doi.org/10.1021/bi00027a007.

(60) Merski, M.; Shoichet, B. K. The Impact of Introducing a Histidine into an Apolar Cavity Site on Docking and Ligand Recognition. J Med Chem 2013, 56 (7), 2874–2884. https://doi.org/10.1021/jm301823g.

(61) Chaneton, B.; Hillmann, P.; Zheng, L.; Martin, A. C. L.; Maddocks, O. D. K.; Chokkathukalam, A.; Coyle, J. E.; Jankevics, A.; Holding, F. P.; Vousden, K. H.; Frezza, C.; O'Reilly, M.; Gottlieb, E. Serine Is a Natural Ligand and Allosteric Activator of Pyruvate Kinase M2. Nature 2012, 491 (7424), 458–462. https://doi.org/10.1038/nature11540.

(62) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. J. Med. Chem. 2006, 49 (23), 6789–6801. https://doi.org/10.1021/jm0608356.

(63) Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated Docking Screens: A Feasibility Study. J. Med. Chem. 2009, 52 (18), 5712–5720. https://doi.org/10.1021/jm9006966.

View Article Online
DOI: 10.1039/D6DD00056H

(64) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. Journal of Computational Chemistry 1992, 13 (4), 505–524. https://doi.org/10.1002/jcc.540130412.

(65) Priya Gnanasekar, S.; Arunan, E. Molecular Beam and Spectroscopic Techniques: Towards Fundamental Understanding of Intermolecular Interactions/Bonds. 2017. https://doi.org/10.1039/BK9781782621737-00259.

(66) Vioglio, P. C.; Chierotti, M. R.; Gobetto, R. Solid-State NMR Techniques for the Study of Intermolecular Interactions.

(67) Lu, T.; Chen, Q. Visualization Analysis of Weak Interactions in Chemical Systems. In Comprehensive Computational Chemistry (First Edition); Yáñez, M., Boyd, R. J., Eds.; Elsevier: Oxford, 2024; pp 240–264. https://doi.org/10.1016/B978-0-12-821978-2.00076-3.

(68) Dougherty, D. A. Cation-Pi Interactions Involving Aromatic Amino Acids. J Nutr 2007, 137 (6 Suppl 1), 1504S-1508S; discussion 1516S-1517S. https://doi.org/10.1093/jn/137.6.1504S.

(69) Infield, D. T.; Rasouli, A.; Galles, G. D.; Chipot, C.; Tajkhorshid, E.; Ahern, C. A. Cation-π Interactions and Their Functional Roles in Membrane Proteins. Journal of Molecular Biology 2021, 433 (17), 167035. https://doi.org/10.1016/j.jmb.2021.167035.

View Article Online
DOI: 10.1039/D6DD00056H

(70) Gallivan, J. P.; Dougherty, D. A. Cation-π Interactions in Structural Biology. Proceedings of the National Academy of Sciences 1999, 96 (17), 9459–9464. https://doi.org/10.1073/pnas.96.17.9459.

(71) Wu, R.; McMahon, T. B. Investigation of Cation-π Interactions in Biological Systems. J. Am. Chem. Soc. 2008, 130 (38), 12554–12555. https://doi.org/10.1021/ja802117s.

(72) Ma, J. C.; Dougherty, D. A. The Cation−π Interaction. Chem. Rev. 1997, 97 (5), 1303–1324. https://doi.org/10.1021/cr9603744.

(73) Pal, D.; Chakrabarti, P. Non-Hydrogen Bond Interactions Involving the Methionine Sulfur Atom. Journal of Biomolecular Structure and Dynamics 2001, 19 (1), 115–128. https://doi.org/10.1080/07391102.2001.10506725.

(74) Allen, F. H.; Bird, C. M.; Rowland, R. S.; Raithby, P. R. Resonance-Induced Hydrogen Bonding at Sulfur Acceptors in R1R2C=S and R1CS2− Systems. Acta Crystallographica Section B 1997, 53 (4), 680–695. https://doi.org/10.1107/S0108768197002656.

(75) Serine is a natural ligand and allosteric activator of pyruvate kinase M2 | Nature. https://www.nature.com/articles/nature11540 (accessed 2025-01-15).

(76) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. Protein Science 2021, 30 (1), 70–82. https://doi.org/10.1002/pro.3943.

View Article Online
DOI: 10.1039/D6DD00056H

(77) Kim, D.; Wang, X.; Vargas, S.; Zhong, P.; King, D. S.; Inizan, T. J.; Cheng, B. A Universal Augmentation Framework for Long-Range Electrostatics in Machine Learning Interatomic Potentials. J. Chem. Theory Comput. 2025, 21 (24), 12709–12724. https://doi.org/10.1021/acs.jctc.5c01400.

(78) Kabylda, A.; Frank, J. T.; Suárez-Dou, S.; Khabibrakhmanov, A.; Medrano Sandonas, L.; Unke, O. T.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields. J. Am. Chem. Soc. 2025, 147 (37), 33723–33734. https://doi.org/10.1021/jacs.5c09558.

(79) Plé, T.; Lagardère, L.; Piquemal, J.-P. Force-Field-Enhanced Neural Network Interactions: From Local Equivariant Embedding to Atom-in-Molecule Properties and Long-Range Effects. Chem. Sci. 2023, 14 (44), 12554–12569. https://doi.org/10.1039/D3SC02581K.

(80) Páll, S.; Zhmurov, A.; Bauer, P.; Abraham, M.; Lundborg, M.; Gray, A.; Hess, B.; Lindahl, E. Heterogeneous Parallelization and Acceleration of Molecular Dynamics Simulations in GROMACS. The Journal of Chemical Physics 2020, 153 (13), 134110. https://doi.org/10.1063/5.0018516.

(81) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for

View Article Online
DOI: 10.1039/D6DD00056H

Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. Computer Physics Communications 2022, 271, 108171. https://doi.org/10.1016/j.cpc.2021.108171.

(82) Grumet, M.; von Scarpatetti, C.; Bučko, T.; Egger, D. A. Delta Machine Learning for Predicting Dielectric Properties and Raman Spectra. J. Phys. Chem. C 2024, 128 (15), 6464–6470. https://doi.org/10.1021/acs.jpcc.4c00886.

(83) Nováček, M.; Řezáč, J. PM6-ML: The Synergy of Semiempirical Quantum Chemistry and Machine Learning Transformed into a Practical Computational Method. December 6, 2024. https://doi.org/10.26434/chemrxiv-2024-3nwwv-v3.

(84) Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning | npj Computational Materials. https://www.nature.com/articles/s41524-024-01339-x (accessed 2025-03-13).

View Article Online
DOI: 10.1039/D6DD00056H

Data availability

The PDB-FRAGID dataset is available at https://github.com/hnlab/PDB-FRAGID and can be downloaded from Zenodo (https://zenodo.org/records/18213106). The trained models of PANIP, benchmark sets and codes used in this paper are available for download from a GitHub repository https://github.com/hnlab/PANIP.