Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

On improving experimental binding affinity predictions with synthetic data

Kevin Ryczko

kevin.ryczko@sandboxaq.com

SandboxAQ https://orcid.org/0000-0001-6933-3856

Phyo Phyo Zin
SandboxAQ

Jordan Crivelli-Decker
SandboxAQ

Punit Jha
SandboxAQ

Harish Ramadas
SandboxAQ

Lorenzo Boninsegna
SandboxAQ

Justin Overhulse
SandboxAQ

Benjamin Shields
SandboxAQ

Pablo Lemos
SandboxAQ

Sasaank Bandi
Maarten van Damme
Amogh Sood
Lee Huntington
Mary Pitman
Martin Ganahl
Andrea Bortolato
Ly Le

Article

Keywords:

Posted Date: June 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9861984/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Some of the authors are shareholders in SandboxAQ.

On improving experimental binding affinity predictions with synthetic data

Kevin Ryczko[1*], Phyo Phyo Zin[1], Jordan Crivelli-Decker[1], Ly Le[1], Punit K. Jha[1], Harish Ramadas[1], Lorenzo Boninsegna[1], Justin Overhulse[1], Benjamin J. Shields[1], Pablo Lemos[1], Sasaank Bandi[1,2†], Maarten van Damme[1], Amogh Sood[1], Lee Huntington[1], Mary Pitman[1], Martin Ganahl[1], Andrea Bortolato[1*]

[1]SandboxAQ, Palo Alto, CA, United States.
[2]Microsoft Research, Cambridge, MA, United States.

*Corresponding author(s). E-mail(s): kevin.ryczko@sandboxaq.com; andrea.bortolato@sandboxaq.com;
†All work was done at SandboxAQ

Abstract

The success of deep learning binding affinity prediction models depends critically on expanding experimental data with reliable synthetic data. We extend the Structurally Augmented IC50 Repository (SAIR) with ≈80K absolute free energy perturbation (AFEP) calculations and present two distinct data splits, SAIR-FEP and SAIR-OOD (out-of-distribution), to simulate realistic drug discovery scenarios. We compare sequence-based proteochemometric (PCM) models and state-of-the-art, structure-based deep learning models and demonstrate that PCM models can be enhanced by physics-based descriptors. While structure-based deep learning methods capture finer geometric detail, their performance is highly sensitive to the input structure. By filtering for high-confidence, co-folded complexes, we show that the performance improves predictably, whereas training on all complexes blindly limits performance gains. Finally, using the SAIR-OOD split, we demonstrate that simultaneous training on synthetic and experimental data improves performance on publicly available, experimental benchmarks. For the OpenBind benchmark, our models yield the lowest root mean squared errors and a 1.85x improvement of Spearman ranking correlation when compared to the original AEVPLIG model training on experimental data alone. These results provide a clear strategy for using synthetic data to advance experimental binding affinity predictions.

1 Introduction

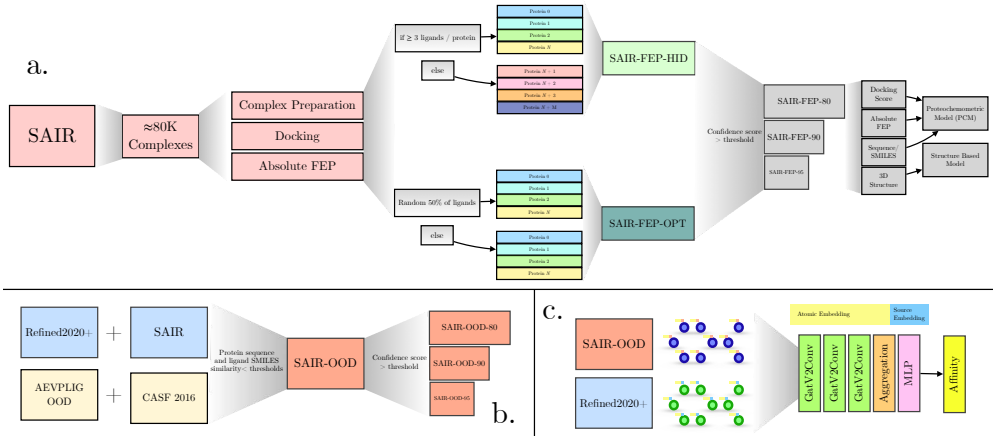


Fig. 1 Overview of the work presented. a) The construction of the SAIR-FEP dataset, which includes computed docking and absolute free energies for ≈80K complexes. Hit identification (SAIR-FEP-ID) and lead optimization (SAIR-FEP-OPT) splits are done to evaluate models on different drug discovery tasks. The data is then further filtered using the co-folding confidence score and then used for training downstream models. b) The construction of the SAIR-OOD dataset, which requires filtering out complexes that have a high protein or ligand similarity with test sets evaluated (AEVPLIG-OOD and CASF 2016). This dataset is then further filtered using the co-folding confidence score. c) Depiction of the AEVPLIG model with the additional source embedding, allowing for training on synthetic and experimental complexes simultaneously.

Predicting the binding affinity between small molecules and their protein targets is a fundamental task in drug discovery, guiding both hit identification and lead optimization. While experimental techniques provide accurate affinity measurements, they remain resource-intensive and therefore cannot evaluate the vast chemical and proteomic spaces relevant to modern therapeutic design [1, 2]. To address these challenges, physics-based and machine-learning methods have become essential for prioritizing candidate compounds and accelerating drug discovery.

Historically, structure-based approaches such as molecular docking, scoring functions, and physics-informed simulations have served as the primary computational tools for estimating binding strength [3–5]. Although widely used, these methods often struggle with limited accuracy, complex parametrization, imperfect energy approximations, and challenges in handling protein flexibility [6]. More recently, deep learning (DL) models have gained momentum by learning nonlinear relationships directly from biochemical data. Models leveraging 3D protein–ligand structures, graph neural networks (GNNs), or molecular descriptors have demonstrated improved predictive performance [7–10], yet they remain heavily dependent on curated training data and carefully engineered molecular representations in order to be effective.

Structure-based deep learning models leverage 3D information from protein–ligand complexes to learn interaction patterns that correlate with binding strength. Early

successes include voxelized 3D convolutional neural networks such as K_{DEEP} [7, 11, 12] and atomic convolutional networks [9], which demonstrated that CNNs can capture spatially local interactions in binding pockets. GNN architectures have also achieved strong performance by encoding complexes as interaction graphs; recent examples include AEVPLIG [13] and GEMS [14], both of which operate on a ligand graph, and [15], which fuses protein and ligand graphs together. Hybrid physics–ML methods further integrate physics-based features with learned representations to improve generalization [16].

Despite these advances, structure-based models rely heavily on large quantities of high-quality structural data and accurate binding affinity data. While open-access resources exist (such as the RCSB Protein Data Bank [17], and PDBBind [18]), these sources are often imperfect and contain significant biases that may undermine models trained on them. For example, several analyses have demonstrated that deep networks may exploit biases in datasets such as PDBbind [18], often memorizing instead of learning true intermolecular interactions [19]. Recent advances in computational structure-prediction approaches have now made it possible to generate large amounts of accurate structural data to augment existing datasets or even to predict the bound conformation of a protein-ligand complex [20]. This, in theory, should enable the generation of additional high-quality structural data for training, thereby improving the performance of structure-based models. However, recent work [21] shows that naively training on this data does not yield more performant models, and that traditional deep learning scaling laws are not observed. Thus, it remains unclear which approach is most effective for integrating synthetic data into the training sets of DL affinity models.

To overcome limitations in structural data, one can turn to SMILES and sequence-driven approaches, such as proteochemometric (PCM) models. Models such as ChemBoost [22] and WideDTA [23] treat protein sequences and SMILES strings as textual inputs, learning "chemical language" embeddings to predict affinity. More recent models integrate graph-based representations with sequence features, such as GNNSeq [24], which combines graph neural networks with sequence-derived descriptors to achieve competitive performance without requiring protein–ligand complex structures. However, binding affinity depends critically on structural information, so structure-based models are expected to better capture the underlying physics, leading to better generalization.

Overall, the literature illustrates a diverse landscape of affinity prediction methods, each with different assumptions about data availability and molecular representation. Structure-based DL offers fine-grained geometric information, while language-based approaches provide flexibility and broad applicability. However, systematic comparisons under controlled conditions remain limited in the literature thus far. In addition, regarding structure-based approaches, it remains unclear what the best practices are for training a model on synthetic structures. This motivates our benchmarking study, which evaluates language-model–based predictors against DL and classical structure-based approaches using consistent datasets, metrics, and evaluation protocols.

In this work, we train PCM and structure-based models (AEVPLIG) on the Structurally Augmented IC50 Repository (SAIR) dataset [25], evaluating the impact of including augmented structural data during training and comparing their performance in predicting experimentally derived binding affinity data. We first compare them directly on a subset of SAIR, called SAIR-FEP, where we've computed $\approx$ 80K AFEP calculations along with various docking scores. We investigate the performance of PCM models using multiple permutations of physics-informed features to identify the most effective feature-stacking strategies for binding affinity prediction. Next, we study the performance of structure-based models on publicly available benchmarks when trained on both experimental and synthetic data simultaneously. To do so, we create another subset of SAIR called SAIR-OOD, in which all complexes that overlap with the experimental complexes were removed to avoid data leakage (fig. 1). We aim to understand model performance across various splits and to probe the underlying data to identify best practices for machine learning applied to synthetic data for binding affinity prediction.

2 Methods

Name	N_{proteins}	N_{ligands}	$N_{\text{complexes}}$
SAIR	5149	734359	1043300
SAIR-FEP-HID$_{\text{train}}$	1107	48461	55083
SAIR-FEP-HID$_{\text{test}}$	505	22053	23519
SAIR-FEP-HID-80$_{\text{train}}$	818	29765	32918
SAIR-FEP-HID-85$_{\text{train}}$	609	20019	21775
SAIR-FEP-HID-90$_{\text{train}}$	319	8571	9195
SAIR-FEP-HID-90$_{\text{test}}$	152	3895	4015
SAIR-FEP-HID-95$_{\text{train}}$	89	1435	1480
SAIR-FEP-HID-95$_{\text{test}}$	43	849	863
SAIR-FEP-OPT$_{\text{train}}$	1223	33762	36880
SAIR-FEP-OPT$_{\text{test}}$	1223	34358	37506
SAIR-OOD	5089	723555	1031912
SAIR-OOD-80	3219	399054	536118
SAIR-OOD-85	2297	261812	339582
SAIR-OOD-90	1222	102144	120614
SAIR-OOD-95	285	15940	17076
Refined2020+	11130	14400	18022
Refined2020++	4822	7294	8240
AEVPLIG-OOD	189	277	291
CASF 2016	156	193	193

Table 1 Details of training and test sets used in this report. The values 80/85/90/95 indicate the confidence value (of the co-folding model) threshold applied. For example, SAIR-OOD-90 indicates that a confidence score threshold of 0.9 was applied.

2.1 Curation process of SAIR-FEP

2.1.1 The SAIR Dataset

The SAIR dataset [25] consists of predicted co-folded structures generated from Boltz-1x [20] of 5,244,285 structures across 1,048,857 protein-ligand complexes sourced from ChEMBL [26, 27] and BindingDB [28, 29]. This structural data is also paired with experimentally derived binding data to train binding affinity models.

2.1.2 Filtering of SAIR

In this work, we filter the SAIR dataset to ensure the highest quality structures are used in our analyses. For each complex, 5 conformations were generated with Boltz-1x, with multiple entries possible for certain complexes due to assay variability in IC50 measurements. We selected only the top-ranked structure for each protein-ligand complex with an aggregated confidence score (estimating reliability via pLDDT and ipTM) greater than 0.5, yielding 1,754,697 complexes. These were further filtered to retain only biochemical assays, proteins with at least 10 measurements, and those with sufficient pIC50 variability ($\geq$2.5 log units). The filtered data had 1,641 unique proteins and 100,617 entries. Next, we focus on curating 2 train/test splits that emulate realistic

drug discovery scenarios. SAIR-FEP-HID (Hit Identification): This split evaluates zero-shot model generalization across unseen protein targets. We partitioned the data protein-wise: high-count proteins ($\geq$3 ligands) were randomly allocated to 70% for training and 30% for testing. To maximize the challenge of target generalization, all remaining low-count proteins (<3 ligands) were exclusively assigned to the test set, ensuring the model is evaluated on rare targets. SAIR-FEP-OPT (Lead Optimization): This split evaluates model performance when proteins are shared but ligands are split randomly between the training and test sets. This emulates a lead optimization campaign, where the target and some hit molecules are known, and the model must identify better molecules against established hits. Dataset statistics for these splits are shown in table 1.

2.1.3 Complex Preparation

Co-folded protein-ligand structures were initially prepared by separating the ligand and protein into their apo forms. Next, ligands were processed with RDKit [30] to ensure correct protonation states, tautomer assignment, and to resolve stereochemistry. The binding poses generated by co-folding were reserved for Gnina scoring [12] and AFEP calculations [31]. Proteins were prepared to ensure proper protonation of key residues, repair missing atoms or side chains, and perform short energy minimizations to resolve steric clashes or unrealistic geometries.

2.1.4 Docking Rescoring Using CNN and Vinardo Scoring Functions

Prior to absolute binding free energy calculations, all complexes were rescored using Gnina [12] with both a convolutional neural network (CNN) based scoring function and the Vinardo empirical scoring function [32]. This rescoring step served as an intermediate triage layer to benchmark machine-learning-based and classical docking scores on the same physics-enriched structural ensemble and to enable direct comparison with subsequent AFEP-derived binding free energies. The CNN evaluates protein–ligand complexes represented as three-dimensional voxel grids to predict binding affinity and pose quality directly from learned spatial interaction patterns, while Vinardo is a physics-inspired empirical scoring function derived from AutoDock Vina [33] that estimates binding affinity as a weighted sum of intermolecular interaction terms optimized for robustness and transferability.

2.1.5 Free energy calculations

For each protein-ligand complex, free energy calculations were performed using the AQFEP protocol described in [31]. AQFEP uses an absolute free energy perturbation calculation based on the double-decoupling alchemical protocol. The double-decoupling approach is considered the "gold standard" for absolute free-energy calculations by ensuring thermodynamic consistency, accurate sampling of the free-energy landscape, and broad applicability across a variety of systems. Using an absolute free energy calculation, which directly estimates the binding free energy of the given ligand-protein pair, the procedure requires much less human guidance

than the more typical relative free energy calculations. Unlike relative FEP, which requires a congeneric ligand series and predefined alchemical mappings between similar compounds, AQFEP is an absolute FEP method that evaluates each ligand independently, making it directly applicable to heterogeneous libraries spanning diverse chemotypes and binding modes. This independence enables more flexible benchmarking across targets and chemistries, and facilitates systematic comparison with docking and machine–learning–based scoring methods in large-scale virtual screening workflows. Convergence was assessed independently for both the bound and unbound legs using predefined criteria, and only complexes satisfying all convergence requirements were retained for downstream analysis. Binding free energies are reported as ΔG (kcal/mol), with associated uncertainties reflecting the statistical error of the free energy estimates. In total, roughly 157,000 GPU hours (NVIDIA Tesla T4) were used for the FEP calculations.

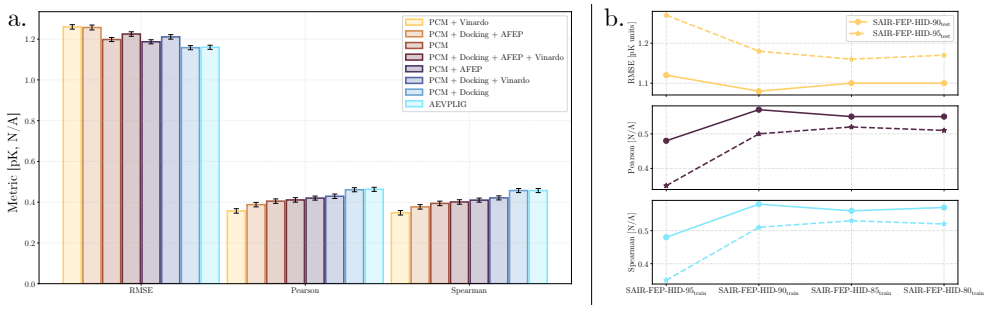


Fig. 2 a) Metrics of PCM with the addition of docking and absolute free energy perturbation values as input features, as well as AEVPLIG evaluated on the SAIR-FEP-HID test set. Physics-based data enhances PCM predictions. b) Metrics tracked across the SAIR-FEP-HID-90/95$_{\text{test}}$ sets for different training sets using AEVPLIG. Performance increases with the addition of complexes with high confidence (> 0.9), but plateaus with those of lower confidence.

2.2 Curation of SAIR-OOD

As mentioned in previous literature [19], benchmarks on structure-based affinity models have revealed biases in the underlying training data, highlighting data leakage and the need for proper held-out test sets to rigorously evaluate these models. To address this, we evaluated the impact of training on two datasets: Refined2020+ [13] (which is a specialized split of PDBBind2020 [18]) and SAIR [25]. To better understand our model's generalization performance and the impact of synthetic data, we evaluated on two commonly used test sets in the literature: AEVPLIG-OOD [13], and CASF 2016 [34]. However, initial analyses revealed high structural similarity between our training sets and the two test sets. As a result, we curated a second dataset, SAIR-OOD, that addresses this limitation. The curation approach is described below.

To minimize data leakage, we removed entries from the training set that had high structural or sequence similarity to the evaluation sets, as done in previous work [13].

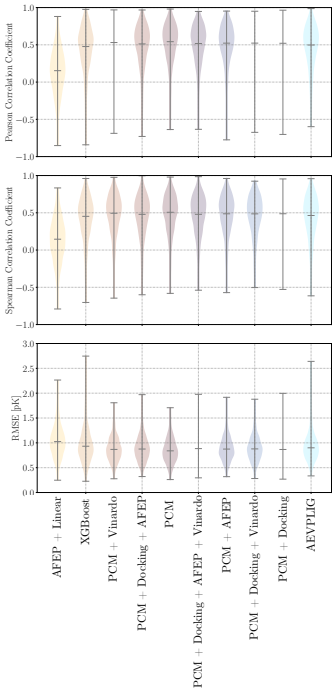


Fig. 3 Comparison of Pearson/Spearman correlation coefficient and root mean squared error (RMSE) distributions across model architectures and feature integration strategies for the SAIR-FEP-OPT split for proteins with ≥ 10 ligands. Distributions represent the performance across individual protein targets. AFEP-Linear and XGBoost represent "local" models trained independently for each protein, whereas AEVPLIG and PCM-based models utilize "global" training across the entire dataset. While AEVPLIG achieves the tightest distribution overall, the PCM baseline outperforms the variants augmented with additional physics-based scoring features.

To calculate protein similarity, we used MMSeqs2 [35], and to calculate ligand similarity, we used Tanimoto similarity between entries' Morgan fingerprints, as included in RDKit version 2025.09.03 [30]. If the protein or the ligand had a similarity value > 0.5, the complex was removed from the training set. With SAIR, 5 complexes were predicted per complex. However, since structure-based models require a single structure, we chose to use the complex with the lowest Vina score [4]; this left 1,043,300 unique structures. Filtering these complexes based on sequence and SMILES similarity left 1,031,912 structures. In addition, we further filtered this set based on the confidence scores output by Boltz-1, which are included in the SAIR dataset. We considered confidence values of 0.8, 0.85, 0.9, and 0.95, yielding 4 unique training datasets: SAIR-OOD-80/85/90/95. The number of entries, along with the unique number of proteins and ligands, is shown in table 1.

In addition to the filtering of SAIR, we also filtered the entries of the Refined2020+ [13] to eliminate overlap with the CASF 2016 [34] test set. We refer to this training data as Refined2020++.

2.3 Training of Proteochemometric Models

We developed a Proteochemometric (PCM) modelling framework to predict the binding affinity (potency) of small molecules against multiple protein targets. This approach utilizes a supervised regression architecture that simultaneously learns from ligand chemical space and protein sequence space.

2.3.1 Feature Representation

Molecules were represented as Simplified Molecular Input Line Entry System (SMILES) strings and featurized using Morgan fingerprints (2048 bits, count-based). Protein targets were represented by their primary amino acid sequences. To capture high-dimensional biological context, we employed pre-computed embeddings from the ESM-2 (Evolutionary Scale Modeling) transformer-based language model [36]. These embeddings provide a dense numerical representation of protein sequences, capturing evolutionary and structural information. Physics-based features: To evaluate the contribution of physics-based descriptors, we systematically trained models with docking scores, including minimized affinities and deep-learning-based scoring (CNNscore, CNNaffinity) from gnina [12] as well as free energy of binding and solvation descriptors (from AQFEP).

2.3.2 Model Architecture and Optimization

PCM models were implemented with Pytorch [37] using an ensemble approach that included Gaussian Mixture Network (GMN) [38], Mean Variance Estimation (MVE), and Random Forest (RF). Hyperparameter selection was conducted via an automated Bayesian optimization framework using cross-validation. The search space encompassed critical parameters, including feature bit counts and network layer dimensions, while the preprocessor block remained frozen to maintain feature consistency. The optimization process was initialized with an $n = 1$ random sampling phase, followed by 9 iterations of Bayesian search directed by a surrogate probability model. The objective function was set to maximize the Spearman correlation coefficient. Upon completion, the configuration yielding the best validation performance was selected, and the final models were refitted on the full training dataset to maximize data utility and stability.

2.4 Training of structure-based models

The structure-based model investigated in this work is a GNN called AEVPLIG [13] which operates on molecular graphs. Briefly, the nodes and edges in the graph represent the heavy atoms of the ligand. The heavy atoms of the protein pocket are used to construct node-level embeddings but are not explicitly included in the graph passed to the GNN. An element of the embedding for heavy atom i from a particular ligand is given by

$$g_i^t = \sum_{j \in \text{Protein atoms of type } t} e^{-\eta(r_{ij}-r_s)^2} f_{\text{cutoff}}(r_{ij}) \tag{1}$$

where

$$f_{\text{cutoff}}(r_{ij}) = \begin{cases} \frac{1}{2}(\cos(\pi r_{ij}/r_{\text{cutoff}}) + 1) & \text{if } r_{ij} \leq r_{\text{cutoff}} \\ 0 & \text{if } r_{ij} > r_{\text{cutoff}}, \end{cases} \tag{2}$$

r_{ij} is the radial distance between ligand atom i and protein pocket atom j. One defines a set of values for η and r_s yielding a vector embedding for the set of protein atoms of type t. All of the different protein atoms are then concatenated to form the full input vector, $\mathbf{g_i}$. When training on synthetic and experimental data simultaneously, there is an additional embedding vector $\mathbf{s}$ that is concatenated with the atomic input vectors $\mathbf{g_i}$. This embedding represents the source of the data and the structure it is derived from, and can take 2 forms: synthetic or experimental.

We train the models for 200 epochs using a batch size of 128, a hidden dimension of 256, 3 attention heads, a learning rate of 0.0001229, and the leaky ReLU activation functions throughout. The ensemble model consisted of 5 models, all of which were trained using different seeds of the same data. We refer the curious reader to the original manuscript for more details [13]. When training on synthetic and experimental data simultaneously, an additional embedding vector is concatenated with the atomic input vectors. This embedding represents the source of the data and the structure it is derived from, and can take 2 forms: synthetic or experimental.

3 Results

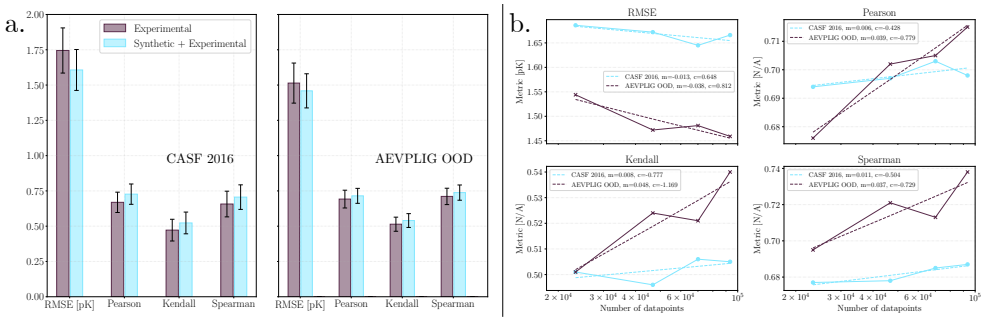


Fig. 4 a) Test metrics comparing AEVPLIG models trained on experimental data alone versus models trained on synthetic and experimental data simultaneously. Consistent performance increases are seen when training on synthetic data. b) Scaling curves for various metrics versus the number of high-quality complexes (confidence score > 0.9) being trained on. Traditional scaling laws hold, with model performance consistently improving as dataset size increases.

3.1 Model performance on SAIR-FEP

3.1.1 Including docking and FEP results improves predictions from proteochemometric models

To establish baseline performance for feature-based models, we first evaluated the PCM model using the SAIR-FEP-HID train/test split. As mentioned in section 2, an array of physics-based scoring methods was included in this split to evaluate if including these metrics improves model performance. These values were appended to the PCM model's existing features. We compare this model to a state-of-the-art structure-based model, AEVPLIG, without additional physics-based features included during training. As shown in fig. 2a, we observe improvements across all metrics when including physics-based scores as features. 3 of 4 models that included docking scores (outputs of gnina docking, including those from the convolutional neural network (CNN)) all improved over the baseline model. However, because some gnina metrics depend on the CNN model's output, data leakage may occur, potentially leading to results that do not generalize. In some cases, combining different scoring functions (e.g., PCM + Docking + AFEP) yields worse metrics than using them independently (e.g., PCM + Docking or PCM + AFEP). This indicates that including multiple scoring functions could introduce contradictions, inevitably leading to poor model performance. AEVPLIG outperformed most PCM models and was on par with the best-performing PCM model, including docking scores, indicating that the non-linear mappings learned directly from structural data are sufficient to produce a high-quality model. Interestingly, including physics-based docking and AFEP features in the node embeddings of AEVPLIG models did not improve performance on the test set.

3.1.2 Quality of co-folded structures impacts downstream model performance

Our initial analysis revealed that AEVPLIG was the best-performing model on the SAIR-FEP-HID split, indicating that augmented structural data can improve downstream model performance. We now consider the effects of the quality of the underlying structural data used during model training. To start, we filter SAIR-FEP-HID with confidence scores of 0.8, 0.85, 0.9, and 0.95, each yielding its own train/test split. Confidence scores were obtained from the predicted local distance difference test (pLDDT) head of the Boltz-1x model. This was introduced for alphafold models and has been shown to correlate well with binding poses [39, 40]. We then monitor the performance of AEVPLIG models on the 0.9 and 0.95 confidence-score test splits (called SAIR-FEP-HID-90_{test} and SAIR-FEP-HID-95_{test}, respectively) as we add more data with lower confidence scores. As shown in fig. 2b, we observe saturation of all metrics as we include more data, which is contrary to scaling law behaviour that is typically observed in deep learning models. An interesting observation is that for SAIR-FEP-HID-90_{test} the best metrics observed were for the training set SAIR-FEP-HID-85_{train}, and for SAIR-FEP-HID-95_{test} the best metrics observed were with the training SAIR-FEP-HID-90_{train}. This suggests that a careful balance between dataset size and the quality of co-folded structures used during training affects model performance.

3.1.3 Model performance in a lead optimization scenario

We now consider the SAIR-FEP-OPT data split, in which proteins are shared across the training and testing sets, but ligands are split 50/50. As discussed in section 2, this aims to test model performance in cases where the target and hit molecules are known, and to find new molecules that surpass current hits in binding affinity. Here, we consider independent models trained for each protein and a model trained across all proteins. For the per-protein models, we use linear regression with ΔG from the AFEP calculations as the only input, serving as the physics-based baseline. In addition, we train XGBoost [41] models on Morgan fingerprints with 1024 bits and radius of 2, computed with RDKit [30]. For all XGBoost models, we use a max tree depth of 4, 100 models, and a learning rate of 0.1. For the all-protein model, we use PCM variants as described in section 3.1.1 and AEVPLIG. In all cases, we compute the Pearson and Spearman correlation coefficients and the root mean squared error (RMSE) for each protein with 10 or more ligands, and aggregate the coefficients across proteins. As shown in fig. 3, the physics-based linear regression model—relying solely on AFEP-derived ΔG values without data-driven calibration—naturally yields the lowest average correlation of ≈ 0.15 with a broad spread across targets. This behaviour is consistent with its sensitivity to small co-folding errors and the omission of protein reorganization and ligand entropic contributions in the fast AQFEP protocol, underscoring its value as a conservative baseline and the importance of thorough retrospective validation before prospective use. The PCM model, without additional physics-based features, was the top-performing architecture among those evaluated. Interestingly, the inclusion of physics-based metrics (Docking and AFEP) into the PCM variants resulted in a performance degradation compared to the base model.

Although the mean correlations are ≈ 0.5 for the models, there are cases in which negative correlations are seen. In future work, we plan to further investigate these cases to understand if this is a model or data issue. The AEVPLIG model barely surpasses the local XGBoost model in this case, suggesting that more complex structure-based models are not always the best choice.

3.2 Model performance on SAIR-OOD

3.2.1 Training on synthetic data improves the performance of models on experimental benchmarks

Our previous analyses demonstrated that utilizing high-quality, structurally augmented data improves model performance on affinity prediction tasks. However, an open question in the literature is whether combining augmented data and experimentally derived structural data can improve model generalization performance. To address this, we allow the model to train on both experimental and synthetic data simultaneously and evaluate the model on two commonly used model generalization benchmarks used in the field. To accomplish this, we modified the AEVPLIG model architecture by introducing an additional embedding layer and adding an additional index to the data such that the correct embedding is selected during training and inference. This embedding was then concatenated to the node embeddings before applying the graph attention layers. A depiction of this is shown in fig. 1c. The PCM model requires no modification to train across synthetic and experimental data simultaneously. Although it performed well during training on Refined2020++ and testing on CASF 2016, we found that adding more data did not improve model performance. An initial baseline AEVPLIG model was trained on experimental data alone and compared to a model trained on both experimental and co-folded structures. All models were evaluated on CASF 2016 and AEVPLIG-OOD (see section 2). As shown in fig. 4a, models trained on experimental data alone yield RMSEs from 1.5-1.7 pK units, and correlation statistics ranging from 0.51-0.71. Importantly, after adding synthetic data (SAIR-OOD-85/90) to the training set alongside experimental (Refined2020+/Refined2020++) data, we observe performance increases on both test sets. For CASF 2016, we observe an 8% decrease in the RMSE, and a 9-11% increase in the correlation metrics compared to the baseline. For AEVPLIG-OOD, we find a 4% decrease in the RMSE, and a 3-5% increase in the correlation metrics. Although the correlation increases may seem modest, they can have significant performance increases in downstream drug discovery tasks. For example, consider the increase in hits identified during a virtual screening campaign. With a total library size of 1,000,000 and 1000 true hits, if the improvement from the Pearson correlation of 0.669 to 0.727, as shown in Table 3 for CASF 2016, the number of hits goes from 27 to 38 when filtering out the top 0.01% (1000 molecules). This is a 41% increase in the number of hits, which is a substantial gain even with a modest improvement in correlation. This result suggests that including high-quality synthetic structural data improves the performance of structure-based affinity models on challenging out-of-distribution test sets.

3.2.2 Neural Scaling with high-quality structural data in the low data regime

To understand a model's scaling performance, one typically assesses its test-set performance as the number of training points increases. For structure-based deep learning models trained on synthetic protein-ligand complexes, this and previous work suggest that naively increasing the dataset size does not lead to performance improvements [21]. However, given that our other analyses revealed that the quality of the structure data is a key variable in model performance, we sought to replicate this pattern in our data and evaluate whether adjusting the quality of structures could recover this finding. To evaluate this, we train models with different dataset sizes by changing the confidence score filter (see table 1). We then monitor the performance of AEVPLIG on CASF 2016 and AEVPLIG-OOD. Compared with the experimental baseline, models trained with more data also show performance gains on the test sets. However, similar to what is observed in fig. 2b, the metrics do not continue to improve as the number of complexes increases.

Since adding low-confidence structures to the data does not yield typical model scaling behaviour, we evaluated whether improving the quality of structures in the training set could help recover it. To do so, we randomly sample 20%, 40%, 60%, and 80% of SAIR-OOD-90, and train on the selected samples along with the Refined2020+ and Refined2020++ data. Next, we evaluate the performance of the AEVPLIG model on CASF 2016 and AEVPLIG-OOD as a function of dataset size. Applying this quality filter allows us to roughly recover typical model scaling behaviour as seen in fig. 4b. When comparing the differences in the curves between CASF 2016 and AEVPLIG-OOD, the curves for AEVPLIG-OOD have slopes with magnitudes 3-6 times larger than those of CASF 2016. This indicates that the data aggregation strategy used to construct SAIR [25] may exhibit stronger overlap with the complexes in AEVPLIG-OOD than with those in CASF 2016. From these scaling curves, one can estimate how many complexes are needed to observe a particular metric value. As an example, to achieve a Pearson correlation coefficient of 0.9 for AEVPLIG-OOD would require $\approx 32 \times 10^6$ complexes. For CASF 2016, this many complexes would yield a Pearson correlation of 0.723, indicating, once again, that the data aggregation strategy used in SAIR is currently inadequate for this test split. Taken together, these results suggest that by including only high-quality co-folded structures during training, we can recover traditional model scaling behaviour. However, further experimentation with larger datasets is needed to confirm these findings. This paves the way for large-scale computational data augmentation to build affinity prediction models that perform effectively on extreme generalization tasks.

3.2.3 Performance on the EV-A71 2A Protease target from OpenBind

To further understand the generalizability and applicability of the models trained in this work, we ran inference on the recent OpenBind affinity prediction benchmark [42] using the AEVPLIG model trained on SAIR-OOD and Refined2020+. The OpenBind EV-A71 2A protease benchmark is a dense, single-target structure–affinity dataset

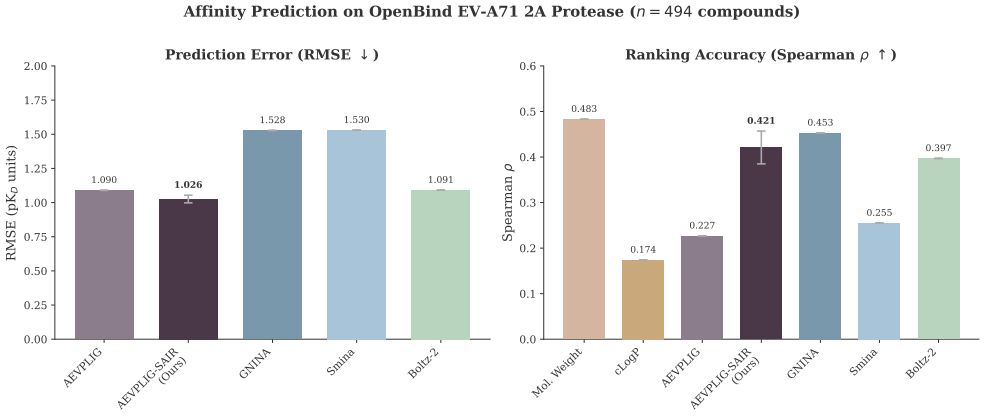


Fig. 5 Root mean squared errors (RMSEs) and Spearman rank correlation for various methods on the OpenBind EV-A71 2A Protease benchmark. The AEVPLIG model trained on SAIR and Refined2020+ achieves the lowest RMSE and highest Spearman rank correlation among AI based methods. The error bars are the standard error computed via bootstrapping.

released by the OpenBind Consortium. The dataset contains 925 crystallographic protein–ligand binding events across 699 compounds, of which 494 have experimental affinity measurements derived from filtered dose–response data. Ground-truth structures were determined by X-ray crystallography at the Diamond Light Source.

Following the official evaluation protocol, predictions were made for all 649 structures with paired crystallographic coordinates and affinity labels. Compound-level metrics were computed by grouping structures by canonical SMILES and averaging predictions for compounds represented by multiple crystal structures. Performance was assessed using RMSE and Spearman rank correlation coefficient, computed over 494 unique compounds. We evaluated the model twice, once with the source embedding for synthetic data and once with the source embedding for experimental data. As shown in fig. 5, we find that our models yield the lowest RMSE when compared to the physics-based and AI-based methods included in the current benchmark. Interestingly, this result was found when using the synthetic data embedding. In addition, we find that, among the AI-based methods, that our AEVPLIG model (with the experimental source embedding) had the highest Spearman ranking coefficient. We see a 1.85x improvement when compared to the original AEVPLIG model training on experimental data alone. Although the error bars show that the new AEVPLIG model is on-par with gnina, it still does not outperform molecular weight, indicating that there is still future work to be done.

4 Conclusion

In this work, we investigate the role of incorporating synthetic data in improving deep-learning-based experimental binding affinity predictions for protein-ligand complexes. We first extend the SAIR dataset by adding additional physics-based data. We also create two additional data splits to aid future researchers in benchmarking deep-learning affinity models. Next, we demonstrate that augmenting experimentally derived structural data with synthetic data, generated by a co-folding model, can improve downstream model performance on challenging out-of-distribution prediction tasks.

We initially create a core dataset, SAIR-FEP, in which we perform docking and AFEP calculations for $\approx 80,000$ complexes and split the data into train/test sets to simulate realistic drug discovery scenarios. Next, we curate a train/test split, called SAIR-FEP-HID, which is intended to be a true out-of-domain split in which the target is not included in the training set. With this set, we train PCM models and show that including additional physics-based features (i.e., docking scores and AFEP values) improves test-set performance. We also compared the PCM models to a DL approach, AEVPLIG, which demonstrates competitive performance alongside the highest-performing physics-augmented PCM variant. However, introducing more co-folded complexes with lower confidence scores into AEVPLIG does not improve test-set metrics, contrary to the typical scaling laws observed in deep learning models.

The second train/test split of SAIR-FEP, called SAIR-FEP-OPT, is intended to simulate a lead-optimization scenario. Here, the training and testing sets have the same targets but different ligands. With this split, we compare the correlation of models trained independently for each protein to that of a model trained across all proteins. Among the global models, the PCM base model achieved the highest mean Pearson correlation. It notably outperformed variants that incorporated additional physics-based features such as docking and AFEP values.

We also curate SAIR-OOD, an additional split in which we remove SAIR entries with similar complexes from the CASF 2016 and AEVPLIG-OOD test sets to prevent data leakage. We then train AEVPLIG models on synthetic and experimental data simultaneously and show that including synthetic data improves test set metrics on the experimental data. For the OpenBind benchmark, our models yield the lowest root mean squared errors and a 1.85x improvement of Spearman ranking correlation when compared to the original AEVPLIG model training on experimental data alone. We also investigate model performance as more, lower-confidence co-folded structures are added to the training set and again observe metric saturation. To test if typical scaling laws are observed, we randomly select entries from the high-confidence complexes in increasing amounts and monitor the test metrics. Here, we find that adding more complexes improves performance, indicating typical scaling-law behaviour of DL models for binding affinity prediction.

This work demonstrates that including synthetic data in various ways generally improves the predictive performance of models for experimental binding affinities. However, we show that one cannot blindly include synthetic data in their training. To achieve performance improvements, careful curation of the co-folded structures is

required. Taken together, these results show that by including only high-quality co-folded structures during training, we can recover traditional model scaling behaviour. This paves the way for large-scale computational data augmentation efforts to build affinity prediction models that perform effectively on extreme generalization tasks common in drug discovery.

5 Data Availability

The docking and absolute free energy values for both SAIR and PDBBind can be found in a comma separated value format in the Github repository found here: https://github.com/sandbox-quantum/on-improving-experimental-binding-affinity-predictions-with-synthetic-data. The AEVPLIG-OOD train/test split is found here: https://github.com/isakvals/OOD-Test. The structures from SAIR can be downloaded from Hugging Face: https://huggingface.co/datasets/SandboxAQ/SAIR and the structures from PDBBind (including CASF 2016) can be downloaded from here: https://www.pdbbind-plus.org.cn/.

6 Code Availability

To train the AEVPLIG models, we used the following Github repository: https://github.com/weitse-hsu/AEV-PLIG-refined. A patch can be applied for our model by using the code found here: https://github.com/sandbox-quantum/on-improving-experimental-binding-affinity-predictions-with-synthetic-data. Scripts to train proteochemetric and XGBoost models can also be found in this repo.

7 Author Contributions

K. R., P. P. Z. and A. B. prepared data, trained models, wrote and reviewed the manuscript. J. C-D. advised, wrote, and reviewed the manuscript, L. L and P. K. J. contributed the docking and AFEP calculations, wrote and reviewed the manuscript. H. R., L. B., J. O., B. J. S., P. L., S. B. and M. G. advised and reviewed the manuscript. M. v. D. contributed to data preparation, advised, and reviewed the manuscript. A. S., L. H., and M. P. developed the AFEP method, advised, and reviewed the manuscript.

References

[1] Cooper, A.: Biophysical Chemistry vol. 24. Royal Society of Chemistry, ??? (2011)

[2] Freire, E.: A thermodynamic approach to the affinity optimization of drug candidates. Chemical Biology & Drug Design **74**(5), 468–472 (2009)

[3] Morris, G.M., *et al.*: Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry **19**(14), 1639–1662 (1998)

[4] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry **31**(2), 455–461 (2010)

[5] Mobley, D.L., Gilson, M.K.: Predicting binding free energies: frontiers and benchmarks. Annual review of biophysics **46**(1), 531–558 (2017)

[6] Cournia, Z., Allen, B., Sherman, W.: Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. Journal of Chemical Information and Modeling **57**(12), 2911–2937 (2017)

[7] Jiménez, J., *et al.*: Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. Journal of Chemical Information and Modeling **58**(2), 287–296 (2018)

[8] Feinberg, E.N., *et al.*: Potentialnet for molecular property prediction. ACS Central Science **4**(11), 1520–1530 (2018)

[9] Gomes, J., et al.: Atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint arXiv:1703.10603 (2017)

[10] Satorras, V.G., Hoogeboom, E., Welling, M.: E(n) equivariant graph neural networks (2021)

[11] Zheng, L., Fan, J., Mu, Y.: Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS omega **4**(14), 15956–15965 (2019)

[12] McNutt, A.T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., Koes, D.R.: Gnina 1.0: molecular docking with deep learning. Journal of cheminformatics **13**(1), 43 (2021)

[13] Valsson, Í., Warren, M.T., Deane, C.M., Magarkar, A., Morris, G.M., Biggin, P.C.: Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. Communications Chemistry **8**(1), 41 (2025)

[14] Graber, D., Stockinger, P., Meyer, F., Mishra, S., Horn, C., Buller, R.: Resolving data bias improves generalization in binding affinity prediction. Nature Machine Intelligence, 1–13 (2025)

[15] Zhou, G., Yuan, S., Xu, Q., Li, H., Chen, H., Shen, J.: A multi-geometric graph fusion network for protein–ligand affinity prediction. Physical Chemistry Chemical Physics **27**(45), 24629–24640 (2025)

[16] Kaneriya, A., Samudrala, M., Ganesh, H., Moran, J., Dandibhotla, S., Dakshanamurthy, S.: Structurenet: Physics-informed hybridized deep learning framework for protein–ligand binding affinity prediction. Bioengineering **12**(5), 505 (2025)

[17] Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al.: The rcsb protein data bank: integrative view of protein, gene and 3d structural information. Nucleic acids research, 1000 (2016)

[18] Wang, R., Fang, X., Lu, Y., Yang, C.-Y., Wang, S.: The pdbbind database:

methodologies and updates. Journal of medicinal chemistry **48**(12), 4111–4119 (2005)

[19] Libouban, P.-Y., Aci-Sèche, S., Gómez-Tamayo, J.C., Tresadern, G., Bonnet, P.: The impact of data on structure-based binding affinity predictions using deep neural networks. International Journal of Molecular Sciences **24**(22), 16120 (2023)

[20] Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., Barzilay, R.: Boltz-1 democratizing biomolecular interaction modeling. bioRxiv (2025) https://doi.org/10.1101/2024.11.19.624167 https://www.biorxiv.org/content/early/2025/05/06/2024.11.19.624167.full.pdf

[21] Hsu, W.-T., Grevtsev, S., Herz, A.M., Douglas, T., Magarkar, A., Biggin, P.C.: Can ai-predicted complexes teach machine learning to compute drug binding affinity? Journal of Chemical Information and Modeling (2025)

[22] Özçelik, R., Öztürk, H., Özgür, A., Ozkirimli, E.: Chemboost: A chemical language based approach for protein–ligand binding affinity prediction. Molecular Informatics **40**(5), 2000212 (2021)

[23] Öztürk, H., Ozkirimli, E., Özgür, A.: Widedta: prediction of drug-target binding affinity. arXiv preprint arXiv:1902.04166 (2019)

[24] Dandibhotla, S., Samudrala, M., Kaneriya, A., Dakshanamurthy, S.: Gnnseq: A sequence-based graph neural network for predicting protein–ligand binding affinity. Pharmaceuticals **18**(3), 329 (2025)

[25] Lemos, P., Beckwith, Z., Bandi, S., Van Damme, M., Crivelli-Decker, J., Shields, B.J., Merth, T., Jha, P.K., De Mitri, N., Callahan, T.J., et al.: Sair: Enabling deep learning for protein-ligand lnteractions with a synthetic structural dataset. bioRxiv, 2025–06 (2025)

[26] Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., *et al.*: Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research **40**(D1), 1100–1107 (2012)

[27] Zdrazil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S., Veij, M., Ioannidis, H., Lopez, D.M., Mosquera, J.F., *et al.*: The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic acids research **52**(D1), 1180–1192 (2024)

[28] Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K.: Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research **35**(suppl_1), 198–201 (2007)

[29] Liu, T., Hwang, L., Burley, S.K., Nitsche, C.I., Southan, C., Walters, W.P., Gilson, M.K.: Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. Nucleic acids research **53**(D1), 1633–1644 (2025)

[30] Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., Gedeck, P., Jones, G., Kawashima, E., Nealschneider, D., et al.: rdkit/rdkit: 2025_03_1 (q1 2025) release. Zenodo (2025)

[31] Crivelli-Decker, J.E., Beckwith, Z., Tom, G., Le, L., Khuttan, S., Salomon-Ferrer, R., Beall, J., Gómez-Bombarelli, R., Bortolato, A.: Machine learning guided AQFEP: A fast and efficient absolute free energy perturbation solution for virtual screening. Journal of Chemical Theory and Computation **20**(16), 7188–7198 (2024) https://doi.org/10.1021/acs.jctc.4c00399

[32] Quiroga, R., Villarreal, M.A.: Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. PloS one **11**(5), 0155183 (2016)

[33] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry **31**(2), 455–461 (2010)

[34] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., Wang, R.: Comparative assessment of scoring functions: the casf-2016 update. Journal of chemical information and modeling **59**(2), 895–913 (2018)

[35] Steinegger, M., Söding, J.: Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)

[36] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al.: Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv (2022)

[37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[38] Bishop, C.M.: Mixture density networks (1994)

[39] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., *et al.*: Accurate structure prediction of biomolecular interactions with alphafold 3. Nature **630**(8016), 493–500 (2024)

[40] Omidi, A., Møller, M.H., Malhis, N., Bui, J.M., Gsponer, J.: Alphafold-multimer accurately captures interactions and dynamics of intrinsically disordered protein regions. Proceedings of the National Academy of Sciences **121**(44), 2406407121 (2024)

[41] Chen, T.: Xgboost: A scalable tree boosting system. Cornell University (2016)

[42] Delft, F., AlQuraishi, M., Brennan, P., Chodera, J., Čondić-Jurkić, K., Deane, C., DiMaio, F., Griffen, E., Murcko, M., Shotton, E., Stuart, D., Velankar, S., Walsh, M.: OpenBind's First Release: A Structure–affinity Dataset for Structure-based AI. https://doi.org/10.5281/zenodo.20026661 . https://openbind.uk/news/blog-openbinds-first-release-a-structure-affinity-dataset-for-structure-based-ai/ Accessed 2026-05-25

Description	Training set	Testing set	RMSE ↓ [pK Units]	Spearman ↑	Pearson ↑
PCM + Vinardo	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.260 ± 0.011	0.348 ± 0.011	0.357 ± 0.011
PCM + Docking + AFEP	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.257 ± 0.012	0.377 ± 0.011	0.388 ± 0.011
PCM	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.198 ± 0.010	0.394 ± 0.011	0.405 ± 0.011
PCM + Docking + AFEP + Vinardo	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.225 ± 0.011	0.401 ± 0.011	0.411 ± 0.011
PCM + AFEP	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.187 ± 0.010	0.410 ± 0.010	0.420 ± 0.010
PCM + Docking + Vinardo	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.211 ± 0.011	0.421 ± 0.011	0.429 ± 0.010
PCM + Docking	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.160 ± 0.010	0.457 ± 0.010	0.461 ± 0.010
AEVPLIG	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.160 ± 0.010	0.457 ± 0.010	0.463 ± 0.010
AEVPLIG	SAIR-FEP-HID-80$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.172 ± 0.029	0.519 ± 0.028	0.525 ± 0.027
AEVPLIG	SAIR-FEP-HID-85$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.157 ± 0.027	0.524 ± 0.028	0.533 ± 0.026
AEVPLIG	SAIR-FEP-HID-90$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.181 ± 0.030	0.503 ± 0.029	0.508 ± 0.029
AEVPLIG	SAIR-FEP-HID-95$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.273 ± 0.030	0.346 ± 0.033	0.351 ± 0.031
AEVPLIG	SAIR-FEP-HID-80$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.103 ± 0.063	0.557 ± 0.064	0.566 ± 0.057
AEVPLIG	SAIR-FEP-HID-85$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.096 ± 0.059	0.550 ± 0.069	0.563 ± 0.061
AEVPLIG	SAIR-FEP-HID-90$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.085 ± 0.062	0.570 ± 0.067	0.582 ± 0.061
AEVPLIG	SAIR-FEP-HID-95$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.125 ± 0.067	0.483 ± 0.074	0.484 ± 0.068

Table 2 Metrics tracked for PCM and AEVPLIG models on various train/test splits of the SAIR-FEP-HID split. 80/85/90/95 indicates the confidence score filter applied to the co-folded data.

Test Dataset	Train Dataset	Model	RMSE ↓ [pK Units]	Pearson ↑	Kendall ↑	Spearman ↑
CASF v2016	Refined2020++	AEVPLIG	1.745 ± 0.160	0.669 ± 0.072	0.472 ± 0.077	0.657 ± 0.091
CASF v2016	Refined2020++	PCM	1.690 ± 0.160	0.690 ± 0.072	-	0.680 ± 0.080
AEVPLIG OOD	Refined2020+	AEVPLIG	1.514 ± 0.142	0.692 ± 0.063	0.514 ± 0.050	0.711 ± 0.058
AEVPLIG OOD	Refined2020+	PCM	1.800 ± 0.110	0.490 ± 0.068	-	0.520 ± 0.080
CASF v2016	SAIR-OOD-95 + Refined2020+	AEVPLIG	1.713 ± 0.149	0.684 ± 0.070	0.482 ± 0.068	0.669 ± 0.080
AEVPLIG OOD	SAIR-OOD-95 + Refined2020+	AEVPLIG	1.505 ± 0.129	0.690 ± 0.058	0.501 ± 0.053	0.697 ± 0.059
CASF v2016	SAIR-OOD-90 + Refined2020++	AEVPLIG	1.642 ± 0.154	0.708 ± 0.076	0.512 ± 0.074	0.694 ± 0.037
CASF v2016	SAIR-OOD-90 + Refined2020++	PCM	1.690 ± 0.140	0.690± 0.060	-	0.700 ± 0.060
AEVPLIG OOD	SAIR-OOD-90 + Refined2020+	AEVPLIG	1.471 ± 0.123	0.709 ± 0.054	0.513 ± 0.052	0.711 ± 0.056
AEVPLIG OOD	SAIR-OOD-90 + Refined2020+	PCM	1.800 ± 0.110	0.510 ± 0.080	-	0.520 ± 0.100
CASF v2016	SAIR-OOD-85 + Refined2020++	AEVPLIG	**1.607 ± 0.146**	**0.727 ± 0.072**	**0.523 ± 0.077**	**0.706 ± 0.087**
AEVPLIG OOD	SAIR-OOD-85 + Refined2020+	AEVPLIG	1.515 ± 0.131	0.683 ± 0.055	0.506 ± 0.051	0.703 ± 0.058
CASF v2016	SAIR-OOD-80 + Refined2020++	AEVPLIG	1.722 ± 0.173	0.647 ± 0.079	0.446 ± 0.074	0.618 ± 0.092
AEVPLIG OOD	SAIR-OOD-80 + Refined2020+	AEVPLIG	1.495 ± 0.149	0.698 ± 0.061	0.522 ± 0.053	0.718 ± 0.061
CASF v2016	SAIR-OOD + Refined2020++	PCM	1.840 ± 0.180	0.630 ± 0.070	-	0.640 ± 0.080
AEVPLIG OOD	SAIR-OOD + Refined2020+	AEVPLIG	1.476 ± 0.126	0.712 ± 0.059	0.527 ± 0.056	0.719 ± 0.063
AEVPLIG OOD	SAIR-OOD + Refined2020+	PCM	1.780 ± 0.120	0.580 ± 0.080	-	0.520 ± 0.100
CASF v2016	20% SAIR-OOD-90 + Refined2020++	AEVPLIG	1.686 ± 0.158	0.694 ± 0.075	0.501 ± 0.074	0.677 ± 0.088
AEVPLIG OOD	20% SAIR-OOD-90 + Refined2020+	AEVPLIG	1.544 ± 0.136	0.676 ± 0.056	0.501 ± 0.051	0.695 ± 0.058
CASF v2016	40% SAIR-OOD-90 + Refined2020++	AEVPLIG	1.672 ± 0.158	0.697 ± 0.074	0.496 ± 0.074	0.678 ± 0.087
AEVPLIG OOD	40% SAIR-OOD-90 + Refined2020+	AEVPLIG	1.472 ± 0.128	0.702 ± 0.059	0.524± 0.054	0.721± 0.061
CASF v2016	60% SAIR-OOD-90 + Refined2020++	AEVPLIG	1.645 ± 0.176	0.703 ± 0.079	0.506 ± 0.076	0.685 ± 0.091
AEVPLIG OOD	60% SAIR-OOD-90 + Refined2020+	AEVPLIG	1.481 ± 0.124	0.705± 0.057	0.521 ± 0.055	0.713 ± 0.062
CASF v2016	80% SAIR-OOD-90 + Refined2020++	AEVPLIG	1.666 ± 0.157	0.698 ± 0.072	0.505 ± 0.073	0.687 ± 0.085
AEVPLIG OOD	80% SAIR-OOD-90 + Refined2020+	AEVPLIG	**1.459± 0.121**	**0.715± 0.053**	**0.540 ± 0.049**	**0.738 ± 0.054**

Table 3 Metrics for AEVPLIG models trained on various splits of the SAIR-OOD data split. 80/85/90/95 indicates the confidence score filter applied to the co-folded data.

A Results in Tabular Format

B How modest improvements in correlation lead to significant improvement for enrichment factors

To begin, we assume there are a certain number of actives, N_{actives}, from a library of molecules with size N_{total}. The true and predicted affinities, y and $\hat{y}$, are connected via a standard bi-variate normal distribution function, $\Phi_2(y, \hat{y}; \rho)$, where ρ is the correlation between y and $\hat{y}$. To find the number of hits, we can integrate this function, i.e.

$$N_{\text{hits}} = \int_{k}^{\infty} \int_{\hat{k}}^{\infty} \Phi(y, \hat{y}; \rho) dy d\hat{y} \tag{3}$$

where $k, \hat{k}$, change depending on the threshold taken. Typical choices include 0.01%, 0.1%, 1%, 5% which correspond to values of $k, \hat{k} = 3.719, 3.090, 2.326, 1.645$. These are the number of standard deviations above the mean that a true/predicted structure must exceed to be included in the cutoff. In section B, we compute the number of hits

Table 4 Affinity prediction on the OpenBind EV-A71 2A protease benchmark ($n = 494$ compounds).

Method	RMSE (pK_D)	Spearman ρ
Molecular Weight	—	0.483
cLogP	—	0.174
AEV-PLIG	1.090	0.227
AEVPLIG-SAIR (Ours, idx=0=synthetic)	**1.026**	0.394
AEVPLIG-SAIR (Ours, idx=1=experimental)	1.171	**0.421**
GNINA	1.528	0.453
Smina	1.530	0.255
AQAffinity	1.632	0.117
Boltz-2	1.091	0.397

Cutoff	N_{hits}, $\rho = 0.669$	N_{hits}, $\rho = 0.727$	% Increase
0.01%	27	38	41%
0.1%	136	182	34%
1%	460	566	23%
5%	786	870	11%

Table 5 Enrichment performance gains when correlation improves by $\approx 10\%$, which is what is observed in this manuscript. The total number of compounds $N_{\text{total}} = 1,000,000$ and the total number of true actives, $N_{\text{actives}} = 1000$.

found from a library of $N_{\text{total}} = 1,000,000$ and $N_{\text{actives}} = 1000$. In a realistic setting, we would likely downselect to 1000 molecules (0.01% cutoff), and in this scenario we see a 41% improvement in the estimated number of hits found (27→38).