ON IMPROVING EXPERIMENTAL BINDING AFFINITY PREDICTIONS WITH SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

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 physics-based computations and present two distinct data splits, SAIR-FEP and SAIR-OOD. With SAIR-FEP, we perform ≈80K absolute free energy perturbation calculations (AFEP) and curate two train/test splits to simulate realistic drug discovery scenarios. The free energy of binding and other physics-based computations are then used as either input features. We compare the performance of proteochemometric and state-of-the-art structure-based deep learning models and show that including physics-based features improves predictions, and that the quality of the structure plays a key role in their performance. For SAIR-OOD, we remove SAIR entries that overlap with complexes in public-facing benchmarks and demonstrate that simultaneous training on synthetic and experimental data improves performance on public-facing, experimental benchmarks.

1 INTRODUCTION

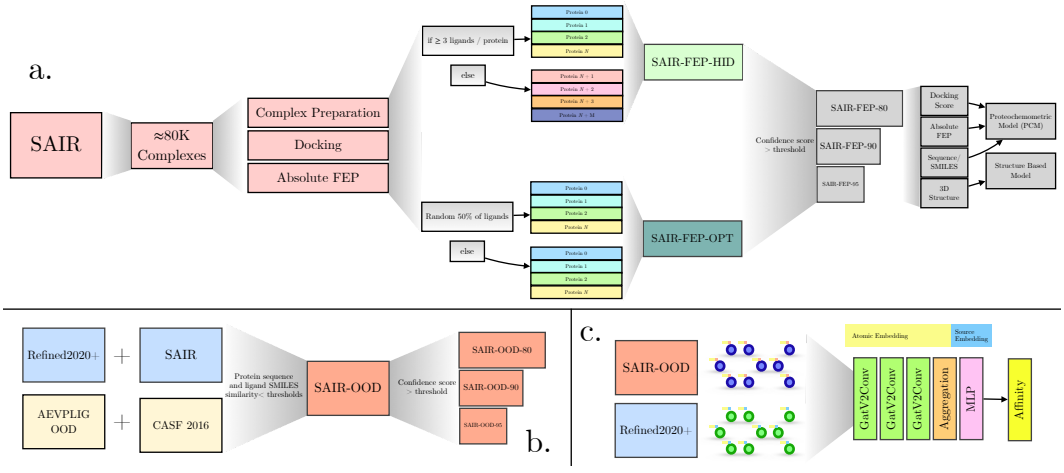

Figure 1: Overview of the work presented. a) The construction of the SAIR-FEP dataset, which includes computed docking and absolute free energies for ≈80K complexes. Hit identification (SAIR-FEP-ID) and lead optimization (SAIR-FEP-OPT) splits are done to evaluate models on different drug discovery tasks. 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 experimen-

tal techniques provide accurate affinity measurements, they remain resource-intensive and therefore cannot evaluate the vast chemical and proteomic spaces relevant to modern therapeutic design (Cooper, 2011; Freire, 2009). To address these challenges, physics-based and machine-learning methods have become essential for prioritizing candidate compounds and accelerating early-stage 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 (Morris et al., 1998; Trott & Olson, 2010a; Mobley & Gilson, 2017).

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 (Jiménez et al., 2018; Feinberg et al., 2018; Gomes et al., 2017; Satorras et al., 2021), 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. GNN architectures have achieved strong performance by encoding complexes as interaction graphs; recent examples include AEVPLIG Valsson et al. (2025) and GEMS Graber et al. (2025), both of which operate on a ligand graph, and Zhou et al. (2025), which fuses protein and ligand graphs together. Hybrid physics–ML methods further integrate physics-based features with learned representations to improve generalization Kaneriya et al. (2025).

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 Rose et al. (2016), and PDBBind Wang et al. (2005)), these sources are often imperfect and contain significant biases that may undermine models trained on them. 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 Wohlwend et al. (2025). This, in theory, should allow for the generation of additional high-quality structural data to train on, improving the performance of structure-based models. Another avenue to overcome limitations in structural data is to turn to SMILES and sequence-driven approaches, such as proteochemometric (PCM) models. Models such as ChemBoost Özçelik et al. (2021) and WideDTA Öztürk et al. (2019) 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 Dandibhotla et al. (2025), which combines graph neural networks with sequence-derived descriptors to achieve competitive performance without requiring protein–ligand complex structures.

In this work, we train proteochemometric and structure-based models on the Structurally Augmented IC50 Repository (SAIR) dataset (Lemos et al., 2025), evaluating the impact of including augmented structural data during training and comparing their performance in predicting experimentally derived binding affinity data. We first compare a feature-based model to a structure-based DL model 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 the PCM models with the addition of physics-based data during training. 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. 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

2.1 OVERVIEW OF DATASETS

As shown in fig. 1, we curate three new splits of SAIR. The splits SAIR-FEP-HID and SAIR-FEP-OPT include additional docking and AFEP data and are meant to emulate hit identification and lead optimization scenarios. The third split, called SAIR-OOD, removes all similar proteins and ligands from SAIR, when considering the test sets, to ensure no data leakage and to evaluate models in out-of-domain scenarios. For more information about the curation of these datasets we refer the reader to section A.

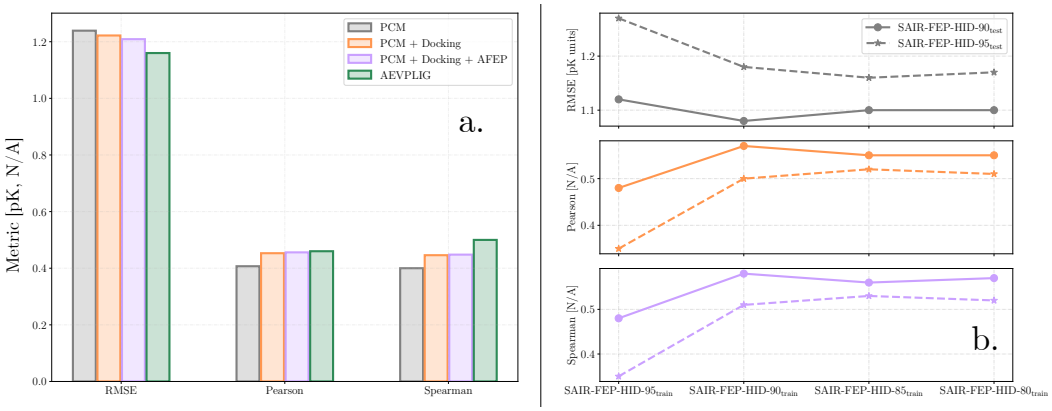

Figure 2: a) Metrics of PCM with the addition of docking and absolute free energy perturbation values as input features, as well as AEVPLIG. Physics-based data enhances PCM predictions. b) Metrics tracked across the SAIR-FEP-90/95$_{\text{test}}$ sets for different training sets using AEVPLIG. Performance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

2.2 Overview of models

In this work, we compare a proteochemometric (PCM) modeling framework to a structure-based, deep learning framework on the various splits outlined above. Briefly, the PCM model is a Gaussian mixture network Bishop (1994) that operates on Morgan fingerprints (for the ligand) and ESM-2 Lin et al. (2022) embeddings (for the protein) which are concatenated. Additional docking and AFEP quantities are also concatenated with these embeddings. The deep learning model used in this report is AEVPLIG Valsson et al. (2025), which is an attention-based GNN where the initial node embeddings are radial atomic environment vectors from TorchANI Gao et al. (2020). For more information, we refer the reader to section A.

3 Results

3.1 Model performance on SAIR-FEP

3.1.1 Including docking and FEP results improves predictions from proteochemometric models

In order to establish baseline performance of feature-based models, we first evaluated the performance of the PCM model using the SAIR-FEP-HID train/test split. 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 in feature-based models. When considering the RMSE, adding in docking and AFEP yields only a 2.5% reduction. However, when we consider the Pearson and Spearman correlation metrics, we see improvements of 12% and 10% over the baseline PCM model. This analysis also revealed that the AEVPLIG model outperforms the best PCM model on all test sets and all metrics, indicating that high-dimensional structure-based features are a rich basis in which correlations can be learned. Interestingly, including physics-based features for docking and AFEP into node embeddings of AEVPLIG models did not result in improvements on the test set.

3.1.2 Quality of co-folded structures impacts 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 Boltz-1x model (see Methods). We then monitor the performance of AEVPLIG models on the 0.9 and 0.95 confidence-

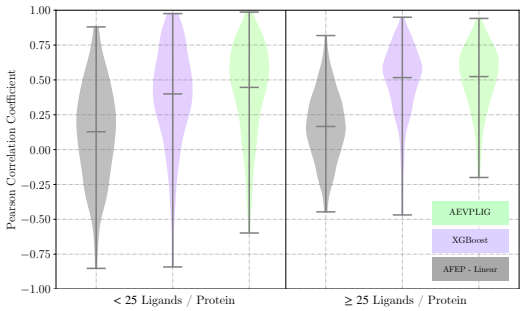

Figure 3: A comparison of distributions of Pearson correlation coefficients for different model architectures and strategies for the SAIR-FEP-OPT split. Linear regression applied to ΔG values (AFEP - Linear) and XGBoost models were trained for each protein, whereas AEVPLIG was trained across all proteins. AEVPLIG achieves the highest average Pearson correlation coefficient in both ligand count cases, but the XGBoost models achieve similar performance when the number of ligands per protein is ≥ 25.

score test splits (called SAIR-FEP-HID-90$_{\text{test}}$ and SAIR-FEP-HID-95$_{\text{test}}$, 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$_{\text{test}}$ the best metrics observed were for the training set SAIR-FEP-HID-85$_{\text{train}}$ and for SAIR-FEP-HID-95$_{\text{test}}$ the best metrics observed were with the training SAIR-FEP-HID-90$_{\text{train}}$. This suggests that a careful balance between dataset size and the quality of co-folded structures included during training has an impact on 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. 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 Chen (2016) models on Morgan fingerprints. For the all-protein model, we use AEVPLIG. In all cases, we compute the Pearson correlation coefficient for each protein and aggregate the coefficients across proteins for those with < 25 or ≥ 25 ligands. We show these distributions in fig. 3. Looking at this figure, AEVPLIG shows the highest average correlation in both cases, proteins with < 25 ligands and proteins with ≥ 25 ligands. Interestingly, the average correlation of XGBoost models with ≥ 25 ligands is comparable to AEVPLIG, though with a wider distribution. These results indicate that, in the low-data regime (i.e., $<$25 ligands/protein), a model trained on many protein-ligand systems can improve correlation with new ligands given a known protein. With more ligands, a single-protein modeling strategy using XGBoost yields results comparable to the many-protein approach.

3.2 Model performance on SAIR-OOD

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

We now 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 as shown in fig. 1c. 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, with the addition of synthetic data into the training set (SAIR-OOD-85/90) 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.

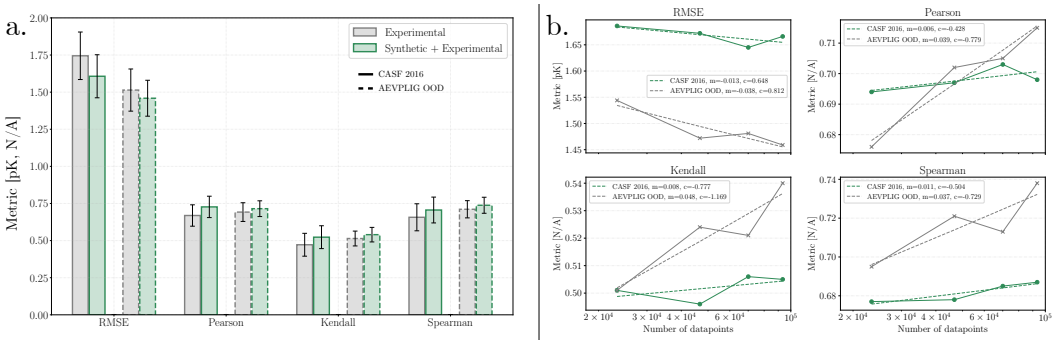

Figure 4: a) Test metrics comparing AEVPLIG models trained on experimental data alone versus models trained on synthetic and experimental data simultaneously. Consistent performance increases are seen when training on synthetic data. b) Scaling curves for various metrics versus the number of high-quality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

For AEVPLIG-OOD, we find a 4% decrease in the RMSE, and a 3-5% increase in the correlation metrics. 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 Scaling laws are recovered with high-quality structural data

Since the addition of data that contains low-confidence structures does not yield typical model scaling behavior, we then evaluated whether improving structure quality in the training set could help to recover traditional model scaling behavior. 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 behavior 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 CASF 2016. This indicates that the data aggregation strategy taken to construct SAIR Lemos et al. (2025) may have a stronger overlap with the complexes in AEVPLIG-OOD than with complexes from CASF 2016. Taken together, these results show that by including only high-quality co-folded structures during training, we can recover traditional model scaling behavior. This paves the way for large-scale computational data augmentation efforts to build affinity prediction models that can perform effectively in extreme generalization tasks.

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. First, we extend the SAIR dataset by adding additional physics-based data and create two additional data splits to aid future researchers in benchmarking deep-learning affinity models. Second, 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. 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 into 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 behavior. 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.

A Appendix

A.1 Results in Tabular Format

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	14537	18022
Refined2020++	4822	7346	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.

Description	Training set	Testing set	RMSE ↓ [pK Units]	Spearman ↑	Pearson ↑
PCM	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.239	0.400	0.407
PCM + Docking	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.222	0.446	0.453
PCM + Docking + AFEP	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.209	0.448	0.456
AEVPLIG	SAIR-FEP-HID$_{\text{train}}$	SAIR-FEP-HID$_{\text{test}}$	1.16	0.47	0.46
AEVPLIG	SAIR-FEP-HID-80$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.17	0.52	0.51
AEVPLIG	SAIR-FEP-HID-85$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.16	0.53	0.52
AEVPLIG	SAIR-FEP-HID-90$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.18	0.51	0.50
AEVPLIG	SAIR-FEP-HID-95$_{\text{train}}$	SAIR-FEP-HID-90$_{\text{test}}$	1.27	0.35	0.35
AEVPLIG	SAIR-FEP-HID-80$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.10	0.57	0.55
AEVPLIG	SAIR-FEP-HID-85$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.10	0.56	0.55
AEVPLIG	SAIR-FEP-HID-90$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.08	0.58	0.57
AEVPLIG	SAIR-FEP-HID-95$_{\text{train}}$	SAIR-FEP-HID-95$_{\text{test}}$	1.12	0.48	0.48

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	RMSE ↓ [pK Units]	Pearson ↑	Kendall ↑	Spearman ↑
CASF v2016	Refined2020++	1.745 ± 0.160	0.669 ± 0.072	0.472 ± 0.077	0.657 ± 0.091
AEVPLIG OOD	Refined2020+	1.514 ± 0.142	0.692 ± 0.063	0.514 ± 0.050	0.711 ± 0.058
CASF v2016	SAIR-OOD-95 + Refined2020+	1.713 ± 0.149	0.684 ± 0.070	0.482 ± 0.068	0.669 ± 0.080
AEVPLIG OOD	SAIR-OOD-95 + Refined2020+	1.505 ± 0.129	0.690 ± 0.058	0.501 ± 0.053	0.697 ± 0.059
CASF v2016	SAIR-OOD-90 + Refined2020++	1.642 ± 0.154	0.708 ± 0.076	0.512 ± 0.074	0.694 ± 0.037
AEVPLIG OOD	SAIR-OOD-90 + Refined2020+	1.471 ± 0.123	0.709 ± 0.054	0.513 ± 0.052	0.711 ± 0.056
CASF v2016	SAIR-OOD-85 + Refined2020++	**1.607 ± 0.146**	**0.727 ± 0.072**	**0.523 ± 0.077**	**0.706 ± 0.087**
AEVPLIG OOD	SAIR-OOD-85 + Refined2020+	1.515 ± 0.131	0.683 ± 0.055	0.506 ± 0.051	0.703 ± 0.058
CASF v2016	SAIR-OOD-80 + Refined2020++	1.722 ± 0.173	0.647 ± 0.079	0.446 ± 0.074	0.618 ± 0.092
AEVPLIG OOD	SAIR-OOD-80 + Refined2020+	1.495 ± 0.149	0.698 ± 0.061	0.522 ± 0.053	0.718 ± 0.061
CASF v2016	20% SAIR-OOD-90 + Refined2020++	1.686 ± 0.158	0.694 ± 0.075	0.501 ± 0.074	0.677 ± 0.088
AEVPLIG OOD	20% SAIR-OOD-90 + Refined2020+	1.544 ± 0.136	0.676 ± 0.056	0.501 ± 0.051	0.695 ± 0.058
CASF v2016	40% SAIR-OOD-90 + Refined2020++	1.672 ± 0.158	0.697 ± 0.074	0.496 ± 0.074	0.678 ± 0.087
AEVPLIG OOD	40% SAIR-OOD-90 + Refined2020+	1.472 ± 0.128	0.702 ± 0.059	0.524± 0.054	0.721± 0.061
CASF v2016	60% SAIR-OOD-90 + Refined2020++	1.645 ± 0.176	0.703 ± 0.079	0.506 ± 0.076	0.685 ± 0.091
AEVPLIG OOD	60% SAIR-OOD-90 + Refined2020+	1.481 ± 0.124	0.705± 0.057	0.521 ± 0.055	0.713 ± 0.062
CASF v2016	80% SAIR-OOD-90 + Refined2020++	1.666 ± 0.157	0.698 ± 0.072	0.505 ± 0.073	0.687 ± 0.085
AEVPLIG OOD	80% SAIR-OOD-90 + Refined2020+	**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.2 Additional details for SAIR splits

A.2.1 Filtering of SAIR

In this work, we filter the SAIR dataset Lemos et al. (2025) 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 by retaining only biochemical assays, proteins with at least 10 measurements, and those with sufficient pIC50 variability ($\geq$2.5 log pIC50 range).

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. In the first train/test split, called SAIR-FEP-HID, we aim to evaluate model generalization across proteins, as seen in hit identification. To create the split, we select high-count proteins ($\geq$3 ligands), randomly allocating 70% to training and 30% to testing. The remaining low-count proteins ($<$3 ligands) are also included in the test set.

In the second train/test split, called SAIR-FEP-OPT, we aim to evaluate model performance when the proteins are shared between training and test sets; however, the ligands are split randomly between the training and test sets. This split emulates a lead optimization campaign, where the target and some hit molecules are known, and the goal is to identify other molecules with improved binding affinity relative to the known hits. Dataset statistics for these splits are shown in table 1. For details about complex preparation, docking, and AFEP calculations, we refer the reader to section A.

A.2.2 Docking Rescoring Using CNN and Vinardo Scoring Functions

Prior to absolute binding free energy calculations, all complexes were rescored using Gnina McNutt et al. (2021) with both a convolutional neural network (CNN) based scoring function and the Vinardo empirical scoring function Quiroga & Villarreal (2016). 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 Trott & Olson (2010b) that estimates binding affinity as a weighted sum of intermolecular interaction terms optimized for robustness and transferability.

While both CNN-based and Vinardo scoring functions provide computationally efficient estimates of binding favorability, they rely on static representations of protein-ligand interactions and limited conformational sampling. Neither approach explicitly accounts for solvent reorganization, protein flexibility, or entropic contributions to binding.

A.2.3 Free energy calculations

For each protein-ligand complex, free energy calculations were performed using the AQFEP protocol described in Crivelli-Decker et al. (2024). 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, avoids error accumulation from poorly defined perturbation pathways, 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.

A.2.4 CURATION OF SAIR-OOD

As mentioned in previous literature Libouban et al. (2023), 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+ Valsson et al. (2025) (which is a specialized split of PDBBind2020 Wang et al. (2005)) and SAIR Lemos et al. (2025). 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 Valsson et al. (2025), and CASF 2016 Su et al. (2018). 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 between our training and evaluation sets, we remove entries from the datasets where high similarity between the training and evaluation sets was identified for the protein or ligand, as done in previous work Valsson et al. (2025). To calculate protein similarity, we used MMSeqs2 Steinegger & Söding (2017), and to calculate ligand similarity, we used Tanimoto similarity between entries' Morgan fingerprints, as included in RDKit version 2025.09.03 Landrum et al. (2025). If the protein or the ligand had a similarity value that was > 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 Trott & Olson (2010a); 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+ Valsson et al. (2025) to eliminate overlap with the CASF 2016 Su et al. (2018) test set. We refer to this training data as Refined2020++.

A.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 Landrum et al. (2025) to ensure correct protonation states, tautomer assignment, and to resolve stereochemistry. The binding poses generated by co-folding were reserved for Gnina scoring McNutt et al. (2021) and AFEP calculations Crivelli-Decker et al. (2024). 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.

A.4 ADDITIONAL DETAILS ABOUT MODELS

A.4.1 TRAINING OF PROTEOCHEMOMETRIC MODELS

We developed a Proteochemometric (PCM) modeling 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.

A.4.2 FEATURE REPRESENTATION

Ligand Encoding: Molecules were represented as Simplified Molecular Input Line Entry System (SMILES) strings and featurized using Morgan fingerprints (2048 bits, count-based).

Protein Encoding: 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 Lin et al. (2022). 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 McNutt et al. (2021) as well as free energy of binding and solvation descriptors (from AQFEP).

A.4.3 Model Architecture and Optimization

PCM models were implemented with Pytorch Paszke et al. (2019) using a Gaussian Mixture Network (GMN) Bishop (1994). The selection of final model hyperparameters was conducted through an automated Bayesian optimization framework. For the model architecture, a search space was defined for critical parameters, including feature bit counts and network layer dimensions. 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 R^2 coefficient on an 80/20 molecule split. Upon completion of the optimization cycles, the hyperparameter set associated with the best validation performance was selected. The final production models were subsequently refitted using these optimized parameters on the full training dataset to maximize data utility and model stability.

A.4.4 Training of structure-based models

The structure-based model investigated in this work is a GNN called AEVPLIG Valsson et al. (2025) 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, $\mathbf{g_i}$. We refer the curious reader to the original manuscript for more details Valsson et al. (2025). 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.

References

Christopher M. Bishop. Mixture density networks. 1994. URL https://api.semanticscholar.org/CorpusID:118227751.

Tianqi Chen. Xgboost: A scalable tree boosting system. *Cornell University*, 2016.

Alan Cooper. *Biophysical chemistry*. Number 24. Royal Society of Chemistry, 2011.

Jordan E. Crivelli-Decker, Zane Beckwith, Gary Tom, Ly Le, Sheenam Khuttan, Romelia Salomon-Ferrer, Jackson Beall, Rafael Gómez-Bombarelli, and Andrea Bortolato. 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, aug 2024. doi: 10.1021/acs.jctc.4c00399. URL https://doi.org/10.1021/acs.jctc.4c00399.

Somanath Dandibhotla, Madhav Samudrala, Arjun Kaneriya, and Sivanesan Dakshanamurthy. Gnnseq: A sequence-based graph neural network for predicting protein–ligand binding affinity. *Pharmaceuticals*, 18(3):329, 2025.

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

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

Xiang Gao, Farhad Ramezanghorbani, Olexandr Isayev, Justin S Smith, and Adrian E Roitberg. Torchani: a free and open source pytorch-based deep learning implementation of the ani neural network potentials. *Journal of chemical information and modeling*, 60(7):3408–3415, 2020.

Joseph Gomes et al. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603*, 2017.

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, pp. 1–13, 2025.

José Jiménez 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.

Arjun Kaneriya, Madhav Samudrala, Harrish Ganesh, James Moran, Somanath Dandibhotla, and Sivanesan Dakshanamurthy. Structurenet: Physics-informed hybridized deep learning framework for protein–ligand binding affinity prediction. *Bioengineering*, 12(5):505, 2025.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, Peter Gedeck, Gareth Jones, Eisuke Kawashima, Dan Nealschneider, et al. rdkit/rdkit: 2025_03_1 (q1 2025) release. *Zenodo*, 2025.

Pablo Lemos, Zane Beckwith, Sasaank Bandi, Maarten Van Damme, Jordan Crivelli-Decker, Benjamin J Shields, Thomas Merth, Punit K Jha, Nicola De Mitri, Tiffany J Callahan, et al. Sair: Enabling deep learning for protein-ligand lnteractions with a synthetic structural dataset. *bioRxiv*, pp. 2025–06, 2025.

Pierre-Yves Libouban, Samia Aci-Sèche, Jose Carlos Gómez-Tamayo, Gary Tresadern, and Pascal Bonnet. The impact of data on structure-based binding affinity predictions using deep neural networks. *International Journal of Molecular Sciences*, 24(22):16120, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

David L Mobley and Michael K Gilson. Predicting binding free energies: frontiers and benchmarks. *Annual review of biophysics*, 46(1):531–558, 2017.

Garrett M. Morris 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.

Rıza Özçelik, Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Chemboost: A chemical language based approach for protein–ligand binding affinity prediction. *Molecular Informatics*, 40 (5):2000212, 2021.

Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

Peter W Rose, Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R Bradley, Cole H Christie, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, et al. The rcsb protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic acids research*, pp. gkw1000, 2016.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. 2021. URL https://api.semanticscholar.org/CorpusID:231979049.

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

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

Oleg Trott and Arthur J. Olson. 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, 2010a.

Oleg Trott and Arthur J Olson. 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, 2010b.

Ísak Valsson, Matthew T Warren, Charlotte M Deane, Aniket Magarkar, Garrett M Morris, and Philip C Biggin. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, 8(1):41, 2025.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2025. doi: 10.1101/2024.11.19.624167. URL https://www.biorxiv.org/content/early/2025/05/06/2024.11.19.624167.

Guoqiang Zhou, Shili Yuan, Qianya Xu, Haoran Li, Huaming Chen, and Jun Shen. A multi-geometric graph fusion network for protein–ligand affinity prediction. *Physical Chemistry Chemical Physics*, 27(45):24629–24640, 2025.