Generalization Beyond Benchmarks: Evaluating Learnable Protein-Ligand Scoring Functions on Unseen Targets

Anonymous Author(s)
Affiliation
Address
email

Abstract

As machine learning becomes increasingly central to molecular design, it is vital to ensure the reliability of learnable protein–ligand scoring functions on novel protein targets. While many scoring functions perform well on standard benchmarks, their ability to generalize beyond training data remains a significant challenge. In this work, we evaluate the generalization capability of state-of-the-art scoring functions on dataset splits that simulate evaluation on targets with a limited number of known structures and experimental affinity measurements. Our analysis reveals that the commonly used benchmarks do not reflect the true challenge of generalizing to novel targets. We also investigate whether large-scale self-supervised pretraining can bridge this generalization gap and we provide preliminary evidence of its potential. Furthermore, we probe the efficacy of simple methods that leverage limited test-target data to improve scoring function performance. Our findings underscore the need for more rigorous evaluation protocols and offer practical guidance for designing scoring functions with predictive power extending to novel protein targets.

1 Introduction

Structure-based drug discovery seeks to identify molecules that bind with high affinity and selectivity to a target protein based on structural information of the protein and ligand. In this setting, machine learning (ML) has been applied in two major directions. First, ML models are increasingly used to generate or predict energetically favorable protein-ligand conformations (binding poses) [1–4], complementing or replacing traditional docking algorithms. Second, ML supports the evaluation of these poses through learnable scoring functions, which estimate binding affinities and guide the ranking of candidate compounds [5–8]. Scoring functions play a central role in molecular docking, by evaluating predicted binding poses, and in virtual screening, by prioritizing compounds from large chemical libraries [9]. Beyond structure-based approaches, ML is also applied in structure-free settings to screen compounds without requiring explicit protein–ligand complexes [10]. In recent years, ML models have achieved impressive performance on standard benchmarks [11, 12], driving growing interest in their integration into real-world drug discovery pipelines.

However, the ability of ML-based scoring functions to generalize to novel protein targets remains an important and increasingly relevant challenge, as many therapeutically significant proteins are poorly represented in current datasets. PLINDER [13], a database that clusters protein–ligand complexes according to their structural similarity, provides a useful view of this problem. In the PDBbind database [14, 15], for example, some medically important targets, including the APOE4 variant [16] and the KEAP1 Kelch domain [17], have only a few available complexes. APOE4, a major genetic risk factor for Alzheimer's disease [16], appears as a small cluster of only two complexes with

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

nearly identical ligands. KEAP1, an important target in cancer research [18], is represented by just 18 highly similar pockets in PDBbind. Beyond naturally occurring proteins, advances in protein design, increasingly accelerated by machine learning [19], are beginning to produce proteins with novel binding sites [20]. Designing such sites de novo offers advantages over repurposing natural proteins: it can enable functions not found in nature and allows integration of engineering principles such as tunability, controllability, and modularity directly into the scaffold. In synthetic biology, for example, de novo proteins are being developed to create new metabolic or signaling pathways, where small-molecule modulators could provide external control over these functions [21]. Scoring models must therefore be able to evaluate potential small-molecule interactions in scenarios that lie far outside their training data distribution. As the prevalence of such cases grows, understanding model performance on structurally novel, low-data targets becomes essential.

Despite encouraging benchmark results, the ability of current models to generalize to unseen targets or unfamiliar chemical scaffolds remains uncertain [22, 23]. Several studies have revealed data leakage and inflated performance estimates [7, 24–26], primarily driven by biases in training datasets where protein families such as kinases and proteases are overrepresented and many complexes share high structural and ligand similarity. Widely used benchmarks, including CASF-2016 [27], DUD-E [28], and DEKOIS2.0 [29], often contain target–ligand combinations resembling those in the training data (see Figure 1), which leads to overly optimistic performance estimates and obscures the true difficulty of generalizing to novel proteins. Many of those failures stem from dataset and evaluation biases rather than inherent model limitations. Overcoming these biases is essential for reliable deployment of ML-based scoring functions in drug discovery.

Contributions. The contributions of this work are threefold. First, we present a systematic evaluation of two top-performing machine learning–based scoring functions: GEMS [7] and GenScore [8]. To rigorously assess their ability to generalize, we construct a strict series of dataset splits that limit pocket similarity between training and test sets, providing a framework for out-of-distribution (OOD) evaluation of scoring functions. In contrast to previous leakage analyses, which have mainly focused on docking and pose-generation tasks [13, 30, 31], our work targets the scoring problem directly and evaluates two models that have not been tested under such stringent conditions. Second, to explore potential solutions, we further investigate whether large-scale self-supervised pre-training with ATOMICA [32] embeddings can improve generalization by capturing atomic-scale interaction features across molecular modalities. These representations show preliminary promise in bridging the performance gap on novel targets. Finally, we investigate the scenario where a small number of experimental affinity measurements for the target ligand is available. We explore how to incorporate such measurements in validation or fine-tuning setups. Overall, our analysis highlights critical limitations in current evaluation practices and introduces new directions for developing more robust and generalizable scoring functions for real-world drug discovery applications.

2 Related work

Computational methods for protein-ligand scoring. Structure-based drug discovery (SBDD) uses three-dimensional structural information of biomolecules—typically proteins—to identify or develop new drugs. A central task in SBDD is identifying ligands that bind to a target protein with high affinity and specificity. A critical component of this process is scoring, which refers to estimating the strength of a protein–ligand interaction given a binding pose. Scoring enables two key applications: docking, where the goal is to identify the most likely binding pose of a ligand, and virtual screening, where compounds are ranked by predicted binding affinity to prioritize the most promising candidates.

Traditional physics-based scoring functions vary significantly in their computational cost and accuracy. Molecular mechanics (MM) methods (e.g., [33–35]) are relatively efficient and capture basic physical interactions but they offer limited accuracy for binding affinity prediction. More sophisticated approaches exist, among which quantum mechanical (QM) methods (e.g., [36, 37]) improve estimates by explicitly accounting for electronic effects. Free energy perturbation (FEP) [38], in contrast, is not merely a scoring function but a simulation-based method that can provide rigorous estimates of ligand binding free energies through extensive conformational sampling. However, both QM and FEP approaches are far too computationally expensive to be applied in large-scale virtual screening.

Machine learning for protein–ligand scoring. Recent machine learning (ML) models aim to offer a combination of speed and accuracy by learning scoring functions directly from structural data. Modern architectures such as graph neural networks [39–42], graph transformers [8, 43], and equivariant models that respect three-dimensional geometric symmetries [42, 44] can model complex interaction patterns directly from raw atomic coordinates. Some models also explore pose-free scoring [10], i.e., predicting binding affinity directly from the protein sequence or structure and ligand representation (e.g., SMILES) without requiring a bound pose; however, these approaches fall outside the scope of this work.

Most ML-based models are trained on PDBbind crystal structures [14, 15], commonly using either experimental binding affinity labels or predicting protein–ligand interatomic distance distributions with a mixture density network (MDN) [8], whose output probabilities are used to compute the final score. Another notable approach used denoising score matching for unsupervised binding energy prediction [45]. While ML scoring models have achieved strong benchmark performance, their ability to generalize to unseen targets remains an open question.

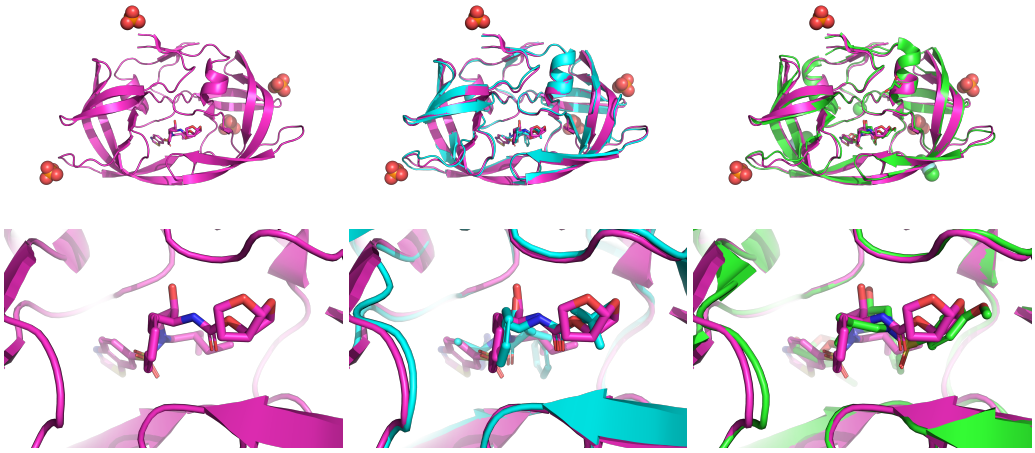


Figure 1: The full proteins (**top**) and their pockets (**bottom**) for the CASF test complex 3O9I (**left**) and the PDBbind CleanSplit training complexes 1BDQ (**middle**) and 4YE3 (**right**) aligned to 3O9I. Proteins are shown as cartoons, ligands as sticks, and remaining heteroatoms as spheres. While CleanSplit removes the most stringent similarities, a certain degree of similarity remains in the dataset. With our PLINDER-based data splitting we go one step further and move entire clusters into a separate test set, aiming to evaluate performance on clusters that are entirely absent from training.

Information leakage in protein–ligand datasets. Many structure-based protein–ligand benchmarks suffer from ligand leakage, where structurally similar or identical binders appear in both training and test sets with similar affinity [46]. Another issue is that binders and non-binders are drawn from different sources (e.g., ChEMBL vs. ZINC in DEKOIS [29, 47]), enabling models to exploit dataset-specific biases [48, 49]. These artifacts affect benchmarks such as DUD-E [28] and DEKOIS2.0 [29]. While more recent efforts like LIT-PCBA [50] and BayesBind [47] attempt to mitigate such artifacts, they still face duplication and leakage issues [51].

Leakage on the protein level is an equally important concern: high sequence or structural similarity between training and test proteins can produce overly optimistic performance estimates. Several approaches have been proposed to address this issue, evolving over time toward stricter and more targeted splits: PoseBusters [30] found that performance of many docking models degrades under stricter sequence-level splits; DockGen [31] reduced redundancy using ECOD [52] domain splits; and PLINDER [13] provided clustering based on different metrics (pocket, ligand, or interaction similarity), allowing tailored splitting according to the task, and demonstrated that DiffDock [4] underperformed in these de-leaked settings. PDBbind CleanSplit, introduced alongside the simple GEMS model [7], was designed to remove the most stringent similarities between PDBbind and the CASF benchmark. On this split, GEMS outperformed larger, more complex models, suggesting that some of the apparent progress on CASF may have resulted from the greater capacity of the complex models to memorize and overfit.

Our study concentrates on protein-side generalization to novel targets, emphasizing evaluation scenarios where the receptor itself is absent from training.

Large-scale self-supervised pretraining. Self-supervised pretraining on large unlabeled protein and molecular datasets has proven highly effective for downstream biochemical tasks. Protein language models such as ESM2 [53] and ANKH [54] capture biophysical properties relevant for protein function, while chemical language models such as ChemBERTA [55] learn representations of small molecules. We adopt the GEMS scoring function [7], which enhances graph neural networks with protein and chemical language model embeddings to improve binding affinity prediction.

Beyond sequence and chemical models, large-scale structural pretraining has also recently emerged; ATOMICA [32] is a geometric deep learning model trained on over two million interaction complexes using self-supervised denoising and masking. ATOMICA learns atomic-scale representations across proteins, small molecules, ions, lipids, and nucleic acids, yielding a compositional latent space that encodes shared physicochemical principles. In this work, we hypothesize and test whether the breadth of ATOMICA's training enables better generalization in protein–ligand scoring tasks.

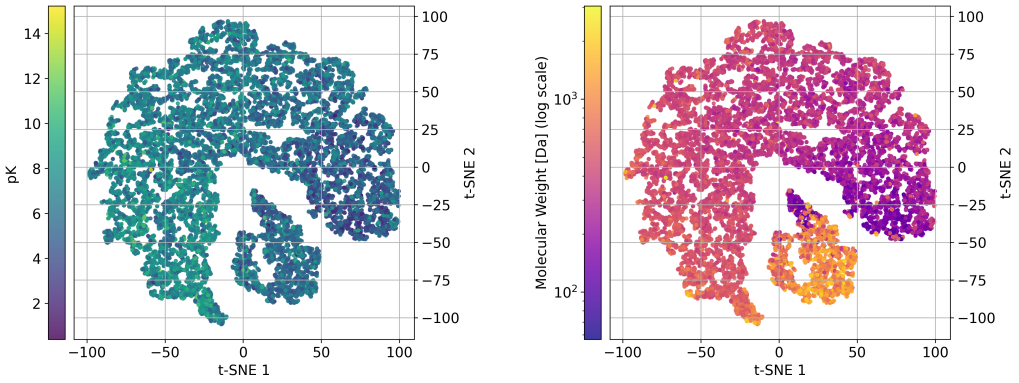


Figure 2: t-SNE projections of ATOMICA embeddings of ligand–pocket interactions from PDBbind, colored by experimental affinity (**left**) and molecular weight (**right**). The large crescent-shaped cluster shows clear gradients of affinity and molecular weight, while the smaller, well-separated cluster contains most of the largest ligands, indicating that ATOMICA space captures both properties.

3 Experimental setup

This section outlines the datasets, splitting strategies, training procedures, and evaluation metrics used in our study. We also describe methods for leveraging limited test-distribution data and for visualizing ATOMICA embeddings in two dimensions, providing intuition into the structure of the learned representation space.

Datasets. We use PDBbind (v.2020) [14, 15] as the primary dataset for training, given its wide adoption and availability of experimental binding affinity labels (approximately 19,000 complexes). For benchmarking, we employ the well-established CASF-2016 dataset [27], comprising 285 high-quality complexes. Additionally, we use PLINDER [13] for clustering and splitting PDBbind complexes. Finally, we make use of ATOMICA [32] embeddings (`ATOMICA_pretrained`), obtained via self-supervised pretraining with masking and denoising tasks on the Cambridge Structural Database (CSD) [56], which comprises roughly 1.75 million structures, and on Q-BioLiP [57, 58], with approximately 300,000 entries. While Q-BioLiP contains complexes overlapping with PDBbind, ATOMICA relies only on self-supervised tasks of coordinate denoising and masked block identity prediction, using information available at inference. We therefore do not consider this a source of data leakage.

Train–test splitting. We first evaluate models using the original PDBbind CleanSplit benchmark, testing on CASF targets as a baseline (Table 1). This establishes a reference point for performance

when near-duplicate complexes are removed from the training dataset based on a joint assessment of ligand, pose, and pocket similarity.

To construct stricter OOD evaluations, we further prepare our own train–test splits based on PLINDER's pocket-level clustering using lDDT [13] (cluster type `pocket_lddt__50__community`). Seven CASF targets were selected to represent diverse protein families (PDB IDs: 1NVQ, 1SQA, 2P15, 2VW5, 3DD0, 3F3E, 3O9I). For each target, all complexes from its cluster are assigned to a separate test set, while the remaining data is used for training and validation. This approach ensures that the test sets contain pockets that are structurally distinct from those seen during training, providing a more stringent evaluation of OOD generalization (see Figure 1). We also apply CleanSplit filtering to the training and validation sets of our OOD splits, enabling evaluation on CASF without obvious leakage and enabling direct comparison between CASF and OOD performance (see Figure 3). Throughout this work, we refer to clusters by the PDB ID of one of the complexes used to retrieve them (e.g., the "1NVQ cluster"). Visual inspection suggests that clusters usually contain highly similar pockets, often preserving the protein fold. Nonetheless, the clustering procedure can, in principle, detect similar pockets arising in different folds.

Training. To create the training datasets, the preprocessed graph representations of PDBbind published alongside GEMS and GenScore were filtered based on the complex IDs. Models were then trained in a 5-fold cross-validation (CV) setting with label-based stratification. To enable direct comparisons, the original GEMS and GenScore implementations were trained on the same 5-fold CV splits, using their respective published training scripts and setups. To generate predictions on the test datasets, the five models obtained from cross-validation were combined into an ensemble via prediction averaging.

Evaluation metrics. We focused on evaluating scoring power, quantified by the Pearson correlation between predicted and true affinities. The affinity labels are based on experimental measurements of K_i, K_d, or IC_{50} and are included as pK values (pK = –log(K_i/K_d/IC_{50})). To further demonstrate that the observed performance drop is not merely a consequence of reduced training set size, we include figures illustrating performance evolution on both the CASF benchmark and our test sets (Figure 3 and Supplementary Figure S2). These figures clearly show that CASF performance, while slightly affected, remains very high and consistently above the performance on novel test targets.

Recognizing that a robust evaluation of scoring functions extends beyond scoring power, we assessed out-of-distribution docking and screening performance of GenScore on CASF-2016 targets from the 1NVQ cluster, using a model trained without 1NVQ cluster proteins. This cluster offers a relatively large sample of targets ($n = 50$ for docking, $n = 10$ for screening), whereas other clusters did not provide enough samples to yield stable, statistically significant results. As it represents only a single target, we report these results in Supplementary Figure S3.

ATOMICA embeddings generation and projection. To examine whether pretrained molecular representations capture properties of protein–ligand interactions that support better generalization of scoring models, we generated ATOMICA [32] graph-level embeddings for ligand–pocket complexes in PDBbind. Each embedding is a 32-dimensional vector, and we successfully obtained representations for 19,189 complexes (98.7% of the dataset, see below). For visualization, we employed t-distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction method that projects high-dimensional vectors into a low-dimensional space while preserving local similarity structure. We used two components and a perplexity of 30. This projection provides an intuitive view of how ligand–pocket complexes are arranged in the learned representation space, offering qualitative insight into the organization of ATOMICA embeddings through Figure 2 and Supplementary Figure S1.

Evaluated models. We evaluate the following four models. **(1) GEMS** [7] is a graph neural network that integrates ligand embeddings from ChemBERTa [55] and protein embeddings from ESM2 [53] and ANKH [54]. **(2) GenScore** [8] employs a graph transformer trained with a mixture density network (MDN) objective to model interatomic distance distributions, building upon the approach of RTMScore [43]. Next two methods use ATOMICA embeddings. In **(3) ATOMICA-MLP**, we train a simple multilayer perceptron using ATOMICA embeddings. In **(4) $\text{GEMS}_{\text{ATOMICA}}$**, we create a variant of the GEMS scoring function in which ChemBERTa ligand embeddings were concatenated with ATOMICA embeddings. Embedding extraction failed for 1.3% of complexes, primarily due

to ligand fragmentation issues. As a result, the training and validation sets for $\text{GEMS}_{\text{ATOMICA}}$ and ATOMICA-MLP were marginally smaller, reflecting the complexes for which embeddings could not be computed. The test sets reported in Tables 1, 2, and 3 are identical to ensure a fair comparison across all evaluated models.

Leveraging limited data. Laboratories often specialize in specific types of proteins and may obtain only a limited set of crystal structures with ligands and corresponding experimental affinity measurements for a protein of interest. Effectively leveraging such target-specific data is important to improve predictive models under realistic conditions. To study this, we hold out 25 complexes from each target-specific test set. These complexes are then used in two separate scenarios, with full consistency maintained between training, validation, and test splits, accounting for ATOMICA preprocessing errors: **(i) Validation scenario**, where the 25 complexes serve solely as a fixed validation set, while the original training data remains unchanged. Models trained in this scenario are denoted as GEMS^{VAL} and $\text{GEMS}^{\text{VAL}}_{\text{ATOMICA}}$. **(ii) Finetuning scenario**, where, starting from the model trained with the original stratified k-fold (SKF) procedure, the best model is further finetuned on the 25 held-out complexes using a small learning rate for 25 epochs. The original training and validation sets are kept unchanged to ensure comparability. These models are denoted as $\text{GEMS}^{\text{FT}-25}$ and $\text{GEMS}^{\text{FT}-25}_{\text{ATOMICA}}$ for the ATOMICA variant.

4 Results

In this section, we present a comprehensive analysis of our results. We begin by visualizing ATOMICA embeddings with t-SNE, using experimental and structural labels to provide qualitative intuition about the organization of the representation space. We then evaluate model performance on the PDBbind CleanSplit benchmark, followed by assessment on our stricter out-of-distribution (OOD) splits. Finally, we investigate strategies for leveraging limited data from the test target distribution and compare their effectiveness.

PDBbind embeddings are well organized in ATOMICA space. We first inspect the projection of ATOMICA embeddings described in Section 3. In Figure 2, we observe an affinity gradient along the larger, crescent-shaped cluster, which coincides with a gradient in molecular weight. Interestingly, the very heavy molecules appear to concentrate predominantly in the smaller, well-separated cluster. Supplementary Figure S1 shows that the three clusters used for our OOD evaluation are also localized in distinct regions of the embedding space. Notably, the 3F3E cluster, which proved particularly challenging for GenScore (see Table 2), is concentrated mostly in the smaller cluster corresponding to heavier molecules. Overall, our findings provide empirical and visual evidence confirming that the ATOMICA embedding space is well structured with respect to fundamental properties such as pocket shapes, ligand sizes, and binding affinity.

Performance on PDBbind CleanSplit. In Table 1, we evaluate the effect of enriching the GEMS scoring function with ATOMICA graph-level embeddings and compare it with the original GEMS and GenScore on the CASF benchmark after training on the full PDBbind CleanSplit. In this setup, we do not observe any advantage from incorporating the additional embeddings, with $\text{GEMS}_{\text{ATOMICA}}$ achieving marginally worse performance than the original GEMS (Table 1). Interestingly, while the simplistic ATOMICA-MLP is not competitive overall, it attains a reasonable performance on the independent subset of CASF (Table 1), highlighting that even simple models can capture some useful information from the embeddings.

Table 1: Scoring performance of different methods trained on PDBbind CleanSplit evaluated on CASF-2016 and the independent subset of CASF-2016 [7].

	Pearson correlation ↑		RMSE ↓	
Method	**CASF-2016**	**CASF-2016 Indep.**	**CASF-2016**	**CASF-2016 Indep.**
GenScore	0.807	0.737	1.280	1.580
GEMS	**0.815**	**0.828**	**1.274**	**1.347**
ATOMICA-MLP	0.583	0.704	1.965	1.960
$\text{GEMS}_{\text{ATOMICA}}$	0.808	0.822	1.292	1.361

Performance on strictly novel targets. We next evaluate model performance on the novel-target splits introduced in Section 3. Scoring previously unseen proteins proves substantially more challenging than established benchmarks such as CASF. For example, the best Pearson correlation on a novel target is 0.736 (Table 2, $\text{GEMS}_{\text{ATOMICA}}$, target 1SQA), whereas the average across all clusters is only 0.470 for GEMS and 0.498 for $\text{GEMS}_{\text{ATOMICA}}$ (Table 2). By comparison, both models achieve Pearson correlations of 0.815 and 0.808 on CASF, respectively (Table 1). Performance also varies widely across targets: while the 1SQA cluster yields relatively strong results, clusters such as 3DD0, 2P15, 3F3E, and 3O9I all result in correlations below 0.5 for every method, underscoring the difficulty imposed by strict generalization conditions.

$\text{GEMS}_{\text{ATOMICA}}$ consistently outperforms the standard GEMS across all clusters except 1NVQ (Table 2). It also achieves both the highest average and the highest minimum performance, indicating that incorporating ATOMICA embeddings improves the robustness of scoring protein–ligand interactions on unseen proteins. This demonstrates that more challenging benchmarks can reveal advantages of certain approaches that remain hidden on easier benchmarks like CASF.

There is no universal winner, however. GenScore achieves the best results on the 3DD0 and 3O9I clusters but fails completely on 3F3E, with a near-random correlation of 0.047 (Table 2). Interestingly, even the simple ATOMICA-MLP model manages to reach top performance on the 2P15 cluster (Table 2), showing that useful signal can be extracted from ATOMICA embeddings even with smaller modeling capacity.

Table 2: **Performance on strictly novel targets.** Scoring power (Pearson correlation $r \uparrow$) across different unseen protein clusters from the CASF benchmark. "Avg." column reports the average across all protein clusters and "Min." the minimum.

Method / Target	**1NVQ**	**1SQA**	**2P15**	**2VW5**	**3DD0**	**3F3E**	**3O9I**	**Avg.**	**Min.**
#Complexes	2714	736	462	207	475	391	469	—	—
GenScore	0.451	0.696	0.415	0.384	**0.481**	0.047	**0.480**	0.422	0.047
ATOMICA-MLP	0.540	0.631	**0.479**	0.524	0.346	0.149	0.400	0.438	0.149
GEMS	**0.609**	0.717	0.364	0.551	0.236	0.404	0.410	0.470	0.236
$\text{GEMS}_{\text{ATOMICA}}$	0.602	**0.736**	0.379	**0.596**	0.311	**0.434**	0.431	**0.498**	**0.311**

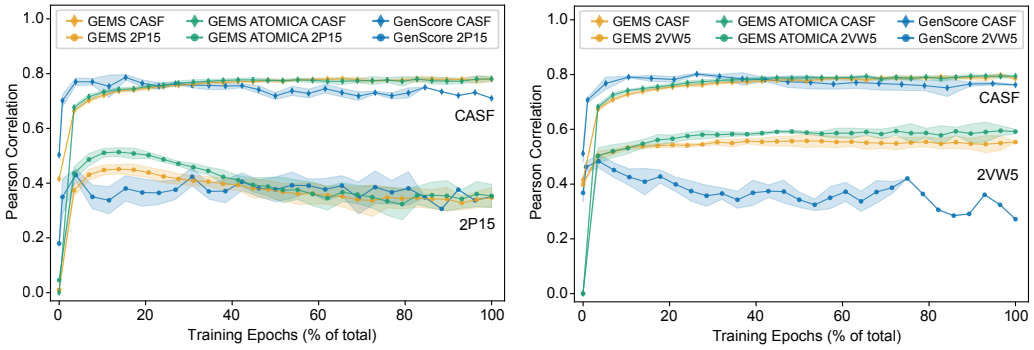


Figure 3: Evolution of the scoring power of the three scoring methods GenScore, GEMS and $\text{GEMS}_{\text{ATOMICA}}$ during model training with the original stratified k-fold splitting of the train–validation data, evaluated on CASF-2016 and the out-of-distribution (OOD) clusters 2P15 (left) and 2VW5 (right). The x-axis shows training progress as a percentage of total training epochs, while the y-axis displays the Pearson correlation coefficient $\uparrow$ between predicted and true affinity. Each line represents the mean performance across five cross-validation folds, with shaded uncertainty regions (± 1 standard deviation) indicating the variability in performance across the five training runs. Note that the uncertainty regions disappear towards higher epoch numbers due to early stopping, which results in different models completing training at different epochs. Consequently, the later portions of the curves are based on fewer than five models. The final results imply that the original stratified k-fold splitting leads to overfitted models, and that using limited test-target validation data allows for better early stopping and the selection of better models.

Exploring strategies for utilizing limited data. We next examine strategies for leveraging a small number of additional target complexes, either by using them for validation or by fine-tuning on them (Table 3). The results are mixed: both validation and fine-tuning improve performance on some clusters but degrade it on others. For instance, validation with ATOMICA embeddings improves performance on 2P15 and 3DD0, but decreases it on 3F3E. Similarly, fine-tuning boosts results on 3DD0 and 3O9I, yet reduces them on 1NVQ and 2VW5.

A motivating example for the validation strategy is shown in Figure 3. On the left, the 2P15 cluster illustrates that peak performance is reached early in training, visible as the hill-shaped curve of performance evolution. Without proper early stopping, all methods exhibit overfitting, and their performance declines to a similar level. Training dynamics varies across targets, as shown on the right side of the figure and in Supplementary Figure S2.

Despite these fluctuations, both strategies yield overall gains compared to training without access to the additional complexes (see GEMS and $\text{GEMS}_{\text{ATOMICA}}$ in Table 2). In particular, they improve not only the average but also the worst-case (minimum) performance across clusters. Among them, the fine-tuned models achieve the strongest results, with $\text{GEMS}^{\text{FT}-25}_{\text{ATOMICA}}$ delivering the highest average correlation (0.550) and $\text{GEMS}^{\text{FT}-25}$ attaining the highest minimum (0.386).

Table 3: **Performance on strictly novel targets with additional target-specific annotated data.** Scoring power (Pearson correlation $r \uparrow$) across different test protein clusters of models utilizing the additional 25 target protein complexes for validation and fine-tuning.

Method / Target	**1NVQ**	**1SQA**	**2P15**	**2VW5**	**3DD0**	**3F3E**	**3O9I**	**Avg.**	**Min.**
#Complexes	2689	711	437	182	450	366	444	—	—
GEMS^{VAL}	0.621	0.726	0.432	0.558	0.497	0.376	0.280	0.499	0.280
$\text{GEMS}^{\text{VAL}}_{\text{ATOMICA}}$	**0.626**	**0.742**	0.514	**0.602**	0.419	0.375	0.354	0.519	0.375
$\text{GEMS}^{\text{FT}-25}$	0.563	0.670	0.556	0.537	**0.620**	**0.386**	0.424	0.537	**0.386**
$\text{GEMS}^{\text{FT}-25}_{\text{ATOMICA}}$	0.561	0.717	**0.610**	0.531	0.572	0.385	**0.476**	**0.550**	0.385

5 Conclusions

We studied the challenge of training scoring functions that generalize reliably to unseen proteins. Our results reveal a substantial gap between performance on standard benchmarks such as CASF and on stricter, novel-target splits. While CASF remains a useful reference, its optimistic estimates mask the difficulty of true generalization. Achieving robustness to entirely new proteins will likely require the models to capture the underlying physical principles of protein–ligand binding. These findings highlight the need for stricter, carefully designed benchmarks to become standard practice, particularly in scenarios where realistic performance estimates are essential.

Consistent with prior reports on GEMS [7], our study suggests that recent gains in scoring may partly reflect improved data memorization rather than genuine generalization. Nevertheless, augmenting GEMS with ATOMICA embeddings consistently improved performance, showing that richer representations can enhance generalization. We also found that even limited target-specific data can narrow the generalization gap, boosting both average and worst-case performance.

Future directions. Future work should emphasize more restrictive data splits and systematic reevaluation of existing methods to identify the drivers of true generalization, extending this analysis also to docking and screening scenarios, motivated by our preliminary results showing a performance drop on the 1NVQ cluster (see Supplementary Figure S3). Expanding our exploration of strategies for leveraging sparse target data, such as quantifying how performance gains scale with available data and testing more advanced fine-tuning techniques, represents another promising avenue. Finally, extending beyond graph-level ATOMICA embeddings to block-level (i.e., residue, nucleotide or chemical motif level) or atom-level representations may further strengthen protein–ligand modeling.

References

[1] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, 2024. DOI: 10.1038/s41586-024-07487-w.

[2] R. Krishna *et al.*, "Generalized biomolecular modeling and design with RoseTTAFold all-atom," *Science*, vol. 384, no. 6693, eadl2528, 2024. DOI: 10.1126/science.adl2528.

[3] J. Wohlwend *et al.*, "Boltz-1 democratizing biomolecular interaction modeling," *bioRxiv*, p. 2024.11.19.624167, 2025. DOI: 10.1101/2024.11.19.624167.

[4] G. Corso *et al.*, "DiffDock: Diffusion steps, twists, and turns for molecular docking," *arXiv*, 2022. DOI: 10.48550/arxiv.2210.01776. eprint: 2210.01776.

[5] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction," *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, 2018, PDBbind 2016. DOI: 10.1093/bioinformatics/bty374.

[6] Y. Li *et al.*, "DeepAtom: A framework for protein-ligand binding affinity prediction," *arXiv*, 2019. DOI: 10.48550/arxiv.1912.00318. eprint: 1912.00318.

[7] D. Graber *et al.*, "GEMS – enhancing generalizable binding affinity prediction by removing data leakage and integrating language model embeddings into graph neural networks," *bioRxiv*, p. 2024.12.09.627482, 2025. DOI: 10.1101/2024.12.09.627482.

[8] C. Shen *et al.*, "A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers," *Chemical Science*, vol. 14, no. 30, pp. 8129–8146, 2023, GenScore. DOI: 10.1039/d3sc02044d.

[9] W. Walters, M. T. Stahl, and M. A. Murcko, "Virtual screening—an overview," *Drug Discovery Today*, vol. 3, no. 4, pp. 160–178, 1998. DOI: https://doi.org/10.1016/S1359-6446(97)01163-X.

[10] H. Y. I. Lam *et al.*, "Protein language models are performant in structure-free virtual screening," *Briefings in Bioinformatics*, vol. 25, no. 6, bbae480, 2024. DOI: 10.1093/bib/bbae480.

[11] T. Harren *et al.*, "Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 14, no. 3, 2024. DOI: 10.1002/wcms.1716.

[12] Valsson *et al.*, "Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data," *Communications Chemistry*, vol. 8, no. 1, p. 41, 2025. DOI: 10.1038/s42004-025-01428-y.

[13] J. Durairaj *et al.*, "PLINDER: The protein-ligand interactions dataset and evaluation resource," *bioRxiv*, p. 2024.07.17.603955, 2024. DOI: 10.1101/2024.07.17.603955.

[14] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind database: Collection of binding affinities for proteinligand complexes with known three-dimensional structures," *Journal of Medicinal Chemistry*, vol. 47, no. 12, pp. 2977–2980, 2004. DOI: 10.1021/jm0305801.

[15] Z. Liu *et al.*, "PDB-wide collection of binding data: Current status of the PDBbind database," *Bioinformatics*, vol. 31, no. 3, pp. 405–412, 2015. DOI: 10.1093/bioinformatics/btu626.

[16] H. C. Hunsberger *et al.*, "The role of apoe4 in alzheimer's disease: Strategies for future therapeutic interventions," *Neuronal Signal*, vol. 3, no. 2, NS20180203, 2019, Epub 2019 Apr 18. DOI: 10.1042/NS20180203.

[17] X. Li, D. Zhang, M. Hannink, and L. J. Beamer, "Crystal structure of the kelch domain of human keap1*," *Journal of Biological Chemistry*, vol. 279, no. 52, pp. 54 750–54 758, 2004. DOI: https://doi.org/10.1074/jbc.M410073200.

[18] S. Adinolfi *et al.*, "The keap1-nrf2 pathway: Targets for therapy and role in cancer," *Redox Biology*, vol. 63, p. 102 726, 2023, Epub 2023 Apr 29. DOI: 10.1016/j.redox.2023.102726.

[19] P. Notin *et al.*, "Machine learning for functional protein design," *Nature Biotechnology*, vol. 42, no. 2, pp. 216–228, 2024. DOI: 10.1038/s41587-024-02127-0.

[20] J. L. Watson *et al.*, "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023. DOI: 10.1038/s41586-023-06415-8.

[21] D. M. Camacho *et al.*, "Next-generation machine learning for biological networks," *Cell*, vol. 173, no. 7, pp. 1581–1592, 2018. DOI: 10.1016/j.cell.2018.05.015.

[22] J. Yang, C. Shen, and N. Huang, "Predicting or pretending: Artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets," *Frontiers in Pharmacology*, vol. 11, p. 69, 2020. DOI: 10.3389/fphar.2020.00069.

[23] M. Volkov *et al.*, "On the frustration to predict binding affinities from protein–ligand structures with deep neural networks," *Journal of Medicinal Chemistry*, vol. 65, no. 11, pp. 7946–7958, 2022, PDBbind 2019. DOI: 10.1021/acs.jmedchem.2c00487.

[24] C. Kramer and P. Gedeck, "Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets," *Journal of Chemical Information and Modeling*, vol. 50, no. 11, pp. 1961–1969, 2010. DOI: 10.1021/ci100264e.

[25] Y. Li and J. Yang, "Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions," *Journal of Chemical Information and Modeling*, vol. 57, no. 4, pp. 1007–1012, 2017. DOI: 10.1021/acs.jcim.7b00049.

[26] M. S. Sellner, M. A. Lill, and M. Smieško, "Quality matters: Deep learning-based analysis of protein-ligand interactions with focus on avoiding bias," *bioRxiv*, p. 2023.11.13.566916, 2023. DOI: 10.1101/2023.11.13.566916.

[27] M. Su *et al.*, "Comparative assessment of scoring functions: The CASF-2016 update," *Journal of Chemical Information and Modeling*, vol. 59, no. 2, pp. 895–913, 2019. DOI: 10.1021/acs.jcim.8b00545.

[28] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking," *Journal of Medicinal Chemistry*, vol. 55, no. 14, pp. 6582–6594, 2012. DOI: 10.1021/jm300687e.

[29] M. R. Bauer, T. M. Ibrahim, S. M. Vogel, and F. M. Boeckler, "Evaluation and optimization of virtual screening workflows with dekois 2.0 – a public library of challenging docking benchmark sets," *Journal of Chemical Information and Modeling*, vol. 53, no. 6, pp. 1447–1462, 2013. DOI: 10.1021/ci400115b.

[30] M. Buttenschoen and C. M Deane, "PoseBusters: AI-based Docking Methods Fail to Generate Physically Valid Poses or Generalise to Novel Sequences," *Chemical Science*, vol. 15, no. 9, pp. 3130–3139, 2024. DOI: 10.1039/D3SC04185A.

[31] G. Corso *et al.*, "Deep confident steps to new pockets: Strategies for docking generalization," in *The Twelfth International Conference on Learning Representations*, 2024.

[32] A. Fang, Z. Zhang, A. Zhou, and M. Zitnik, "ATOMICA: Learning universal representations of intermolecular interactions," *bioRxiv*, p. 2025.04.02.646906, 2025. DOI: 10.1101/2025.04.02.646906.

[33] O. Trott and A. J. Olson, "AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2010. DOI: 10.1002/jcc.21334.

[34] R. A. Friesner *et al.*, "Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, 2004. DOI: 10.1021/jm0306430.

[35] R. A. Friesner *et al.*, "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for proteinligand complexes," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6177–6196, 2006. DOI: 10.1021/jm051256o.

[36] U. Ryde and P. Söderhjelm, "Ligand-binding affinity estimates supported by quantum-mechanical methods," *Chemical Reviews*, vol. 116, no. 9, pp. 5520–5566, 2016, PMID: 27077817. DOI: 10.1021/acs.chemrev.5b00630. eprint: https://doi.org/10.1021/acs.chemrev.5b00630.

[37] A. Pecina, J. Fanfrlík, M. Lepšík, and J. Řezáč, "Sqm2.20: Semiempirical quantum-mechanical scoring function yields dft-quality protein–ligand binding affinity predictions in minutes," *Nature Communications*, vol. 15, no. 1, p. 1127, 2024. DOI: 10.1038/s41467-024-45431-8.

[38] Z. Cournia *et al.*, "Free energy methods in drug discovery—introduction," in *Free Energy Methods in Drug Discovery: Current State and Future Directions*, ser. ACS Symposium Series, vol. 1397, American Chemical Society, 2021, pp. 1–38, ISBN: 9780841298064. DOI: 10.1021/bk-2021-1397.ch001.

[39] D. S. Karlov, S. Sosnin, M. V. Fedorov, and P. Popov, "graphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes," *ACS Omega*, vol. 5, no. 10, pp. 5150–5159, 2020, PDBbind 2018. DOI: 10.1021/acsomega.9b04162.

[40] D. Jiang *et al.*, "InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions," *Journal of Medicinal Chemistry*, vol. 64, no. 24, pp. 18 209–18 232, 2021. DOI: 10.1021/acs.jmedchem.1c01830.

[41] S. Li *et al.*, "GIANT: Protein-ligand binding affinity prediction via geometry-aware interactive graph neural network," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, no. 99, pp. 1–17, 2023, PDBbind 2016 refined. DOI: 10.1109/tkde.2023.3314502.

[42] D. Cao *et al.*, "Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling," *Nature Machine Intelligence*, vol. 6, no. 6, pp. 688–700, 2024, EquiScore. DOI: 10.1038/s42256-024-00849-z.

[43] C. Shen *et al.*, "Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer," *Journal of Medicinal Chemistry*, vol. 65, no. 15, pp. 10 691–10 706, 2022, RTMScore. DOI: 10.1021/acs.jmedchem.2c00991.

[44] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021.

[45] W. Jin, C. Uhler, and N. Hacohen, "DSMBind: An unsupervised generative modeling framework for binding energy prediction," in *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[46] A. Mastropietro, G. Pasculli, and J. Bajorath, "Learning characteristics of graph neural networks predicting protein–ligand affinities," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1427–1436, 2023. DOI: 10.1038/s42256-023-00756-9.

[47] M. Brocidiacono, K. I. Popov, and A. Tropsha, *An improved metric and benchmark for assessing the performance of virtual screening models*, 2024. arXiv: 2403.10478 [q-bio.QM].

[48] J. Yang, C. Shen, and N. Huang, "Predicting or pretending: Artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets," *Frontiers in Pharmacology*, vol. Volume 11 - 2020, 2020. DOI: 10.3389/fphar.2020.00069.

[49] L. Chen *et al.*, "Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening," *PLoS ONE*, vol. 14, no. 8, e0220113, 2019. DOI: 10.1371/journal.pone.0220113.

[50] V.-K. Tran-Nguyen, C. Jacquemard, and D. Rognan, "Lit-pcba: An unbiased data set for machine learning and virtual screening," *Journal of Chemical Information and Modeling*, vol. 60, no. 9, pp. 4263–4273, 2020. DOI: 10.1021/acs.jcim.0c00155.

[51] A. Huang, I. S. Knight, and S. Naprienko, *Data leakage and redundancy in the lit-pcba benchmark*, 2025. arXiv: 2507.21404 [cs.LG].

[52] H. Cheng *et al.*, "Ecod: An evolutionary classification of protein domains," *PLoS Computational Biology*, vol. 10, no. 12, e1003926, 2014. DOI: 10.1371/journal.pcbi.1003926.

[53] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic level protein structure with a language model," 2022. DOI: 10.1101/2022.07.20.500902.

[54] A. Elnaggar *et al.*, "Ankh: Optimized protein language model unlocks general-purpose modelling," *arXiv*, 2023. DOI: 10.48550/arxiv.2301.06568. eprint: 2301.06568.

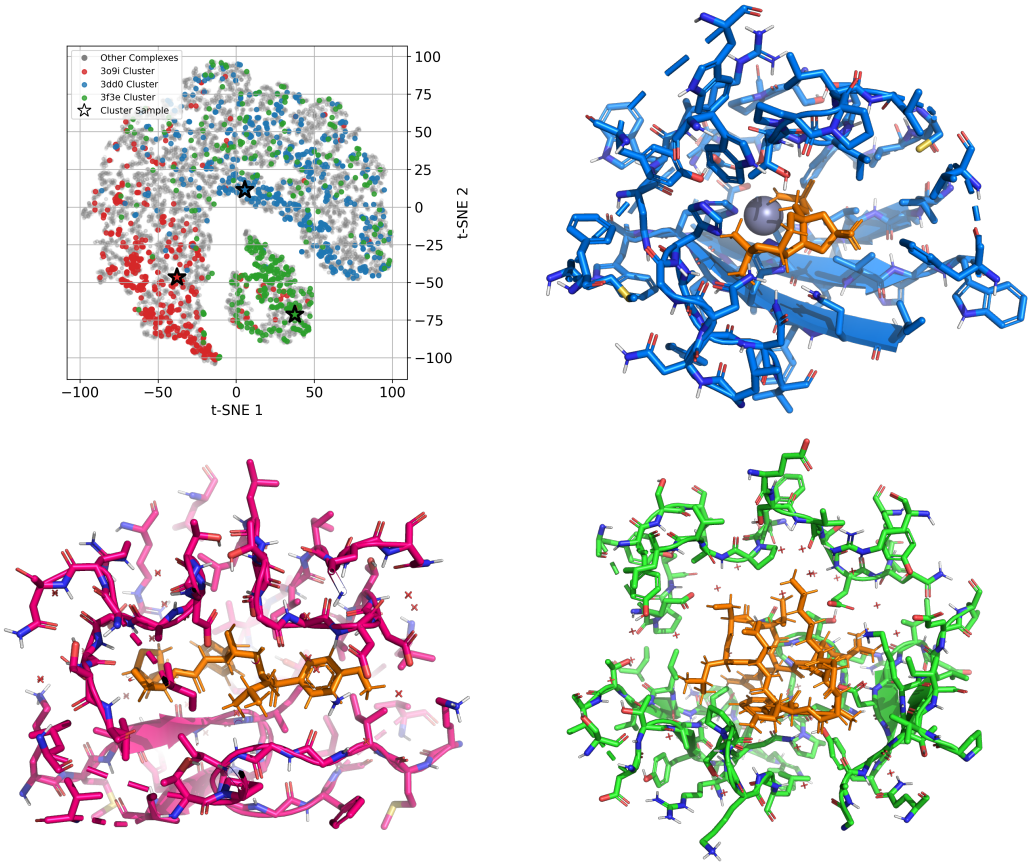
[55] W. Ahmad *et al.*, "ChemBERTa-2: Towards chemical foundation models," *arXiv*, 2022. DOI: 10.48550/arxiv.2209.01712. eprint: 2209.01712.

[56] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The cambridge structural database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, vol. 72, pp. 171–179, 2016. DOI: 10.1107/S2052520616003954.

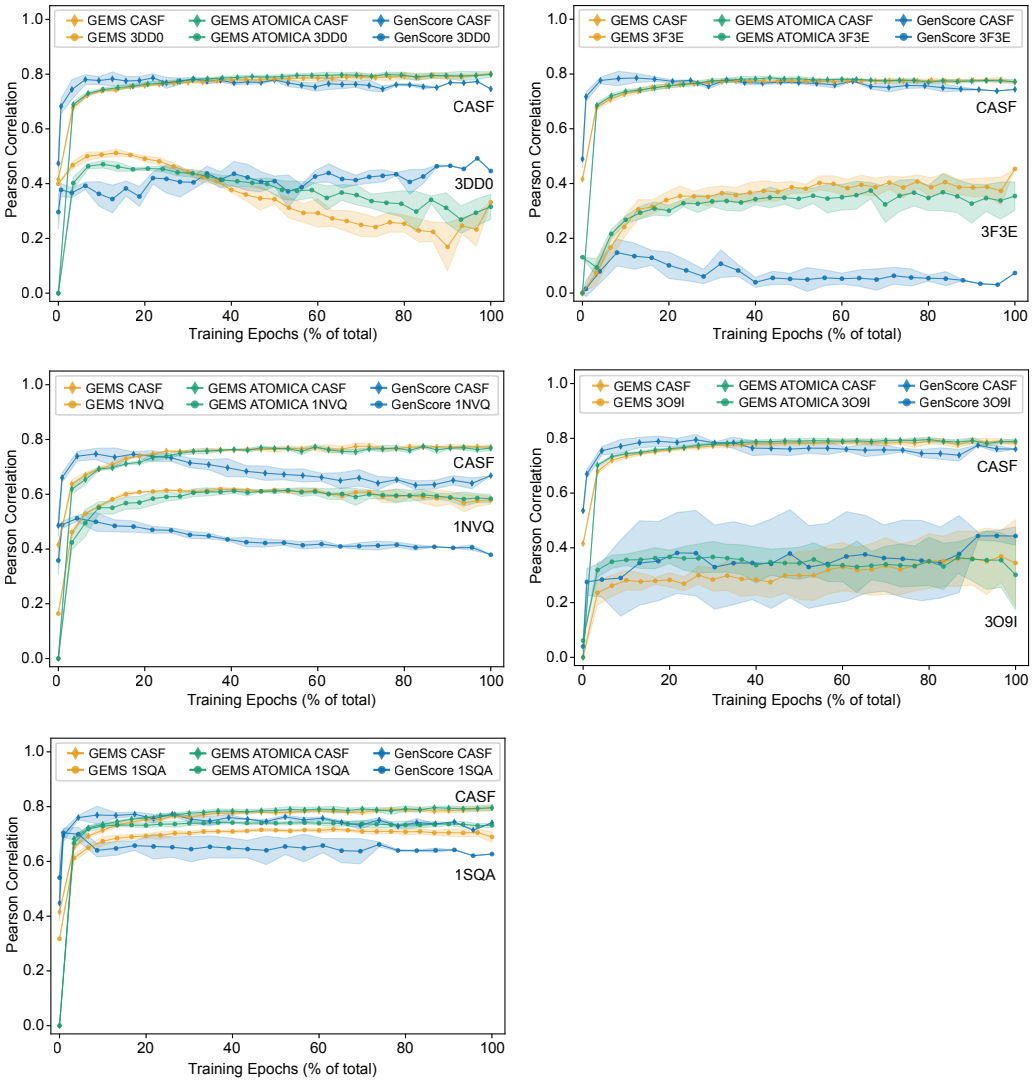
[57] J. Yang, A. Roy, and Y. Zhang, "BioLiP: A semi-manually curated database for biologically relevant ligand–protein interactions," *Nucleic Acids Research*, vol. 41, no. D1, pp. D1096–D1103, 2013. DOI: 10.1093/nar/gks966.

[58] H. Wei, W. Wang, Z. Peng, and J. Yang, "Q-biolip: A comprehensive resource for quaternary structure-based protein–ligand interactions," *Genomics, Proteomics Bioinformatics*, vol. 22, no. 1, qzae001, Jan. 2024. DOI: 10.1093/gpbjnl/qzae001. eprint: https://academic.oup.com/gpb/article-pdf/22/1/qzae001/58457634/qzae001.pdf.

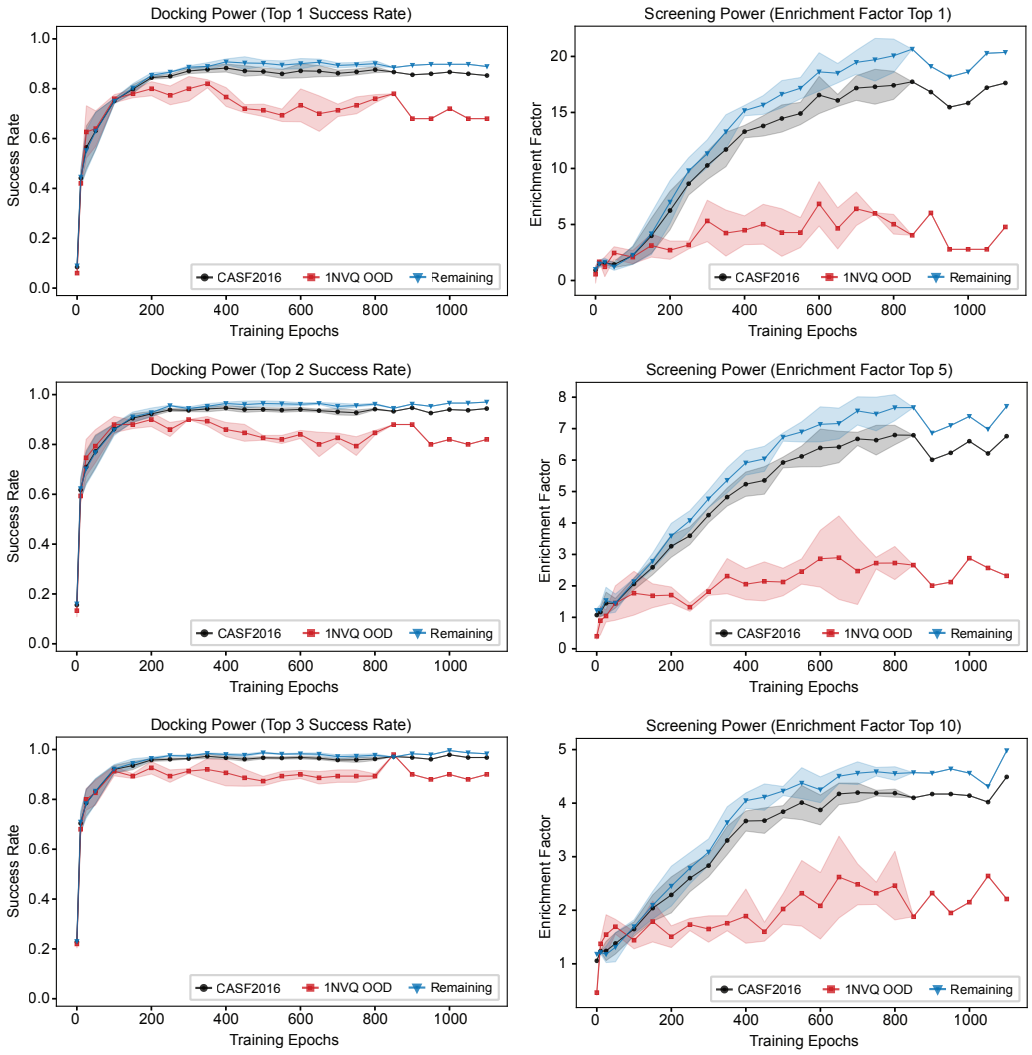
Supplementary material



Supplementary Figure S1: **Upper left**: t-SNE projections of ATOMICA embeddings of ligand–pocket interactions from PDBbind. The three colors show projected embeddings from three different test target clusters. One sample from each cluster is indicated by a star. The complexes corresponding to the three stars are shown in the remaining plots of this figure. These projections indicate that the ATOMICA space is organized by pocket shape, with some pockets forming tight regions. Considered together with Figure 2, they suggest that these regions align with specific ranges of ligand sizes and affinities; in our experiments, scoring interactions representing such regions was particularly difficult when no representatives were available during training. **Upper right**: 3HKU, example from the 3DD0 cluster. **Lower right**: 3QAA, example from the 3O9I cluster. **Lower left**: 5ITF, example from the 3F3E cluster.



Supplementary Figure S2: Evolution of the scoring power of the three scoring methods GenScore, GEMS, and $\text{GEMS}_{\text{ATOMICA}}$ during model training, evaluated on CASF-2016 and the out-of-distribution (OOD) clusters 3DD0, 3F3E, 1NVQ, 3O9I and 1SQA. The x-axis shows training progress as a percentage of total training epochs, while the y-axis displays the Pearson correlation coefficient ↑ between predicted and true affinity. Each line represents the mean performance across five cross-validation folds, with shaded uncertainty regions (± 1 standard deviation) indicating the variability in performance across the five training runs. Note that the uncertainty regions disappear towards higher epoch numbers due to early stopping, which results in different models completing training at different epochs. Consequently, the later portions of the curves are based on fewer than five models.



Supplementary Figure S3: Evolution of docking power (left column) and screening power (right column) of GenScore on the complete decoy datasets of the CASF-2016 benchmark (black), the 1NVQ cluster subset (red), and the remaining decoy datasets with the 1NVQ cluster excluded (blue). The x-axis shows the absolute number of training epochs, while the y-axis displays the success rate (for docking) and the enrichment factors (for screening). Each line represents the mean performance across three cross-validation folds, with shaded regions indicating ± 1 standard deviation around the mean, providing a measure of variability in performance across the three independent model training runs. Note that the shaded uncertainty regions disappear towards the right end of each line due to early stopping, which results in different models completing training at different epochs. Consequently, the later portions of the curves are based on fewer than three models, precluding the calculation of meaningful standard deviations.