Scaling Pocket Docking with Data Augmentation and Heterogeneous Equivariant Graph Attention

Huyen Nguyen [*1] **Tuan Le** [*1] **Loc Nguyen** [1] **Yuxing Peng** [1] **Emine Kucukbenli** [1] **Hien Mai** [1] **Steven Truong** [1] **Ha Tang** [1]

Abstract

Accurate pocket-level molecular docking is limited by narrow training distributions, costly all-atom modeling, and ranking functions that often fail to select the best pose from high-quality generated ensembles. We introduce a scalable docking pipeline that combines large-scale data augmentation, efficient all-atom equivariant modeling, and improved pose ranking. Our diffusion model is trained on curated PLINDER complexes augmented with SAIR synthetic structures, using leakage removal, structural quality filtering, and dataset-specific pocket cutoffs to improve generalization. For ranking, we replace fully connected tensor-product convolutions in the confidence model with Heterogeneous Equivariant Graph Attention (HeteroEGA), enabling interaction-specific attention across ligand, receptor-residue, and receptor-atom graphs. We also evaluate an independent physics-based refinement track using Vina minimization followed by GNINA reranking. On PoseBusters-308, our confidence-ranking pipeline achieves 81.85% Top-1 success rate and 94.51% Oracle success rate, surpassing SigmaDock and DiffDock-RL++. The proposed HeteroEGA confidence model slightly outperforms the post-refinement track while ranking poses substantially faster. These results show that combining broader training data with attention-based equivariant ranking can close much of the gap between Top-1 and Oracle docking accuracy.

[*]Equal contribution [1]NVIDIA. Correspondence to: Huyen Nguyen <huyenn@nvidia.com>, Tuan Le <tvle@nvidia.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Molecular docking serves as a fundamental cornerstone in modern drug discovery, enabling high-throughput virtual screening and the prioritization of lead compounds for experimental validation (Trott & Olson, 2010). By predicting the preferred orientation of a ligand within a protein binding site, docking provides the structural insights necessary for lead optimization. However, traditional search-based methods, such as AutoDock Vina, often struggle to navigate the vast, high-dimensional conformational search space efficiently (Eberhardt et al., 2021). Furthermore, their reliance on empirical scoring functions frequently leads to inaccuracies when modeling complex interactions, limiting their predictive reliability in novel chemical spaces.

To overcome these search-space limitations, the field has transitioned from classical algorithms to deep learning-based frameworks. Early architectures treated docking primarily as a regression task, such as EquiBind (Stärk et al., 2022) and TankBind (Lu et al., 2022). While these models offer significant speed advantages, they often converge to local minima and struggle with significant structural occlusions. A paradigm shift occurred with the emergence of generative models, most notably DiffDock (Corso et al., 2023), which reformulated docking as a reverse-diffusion process over the $SE(3)$ manifold. This approach treats ligand placement as a continuous transformation, significantly improving the diversity and quality of sampled poses.

More recently, there has been a marked shift toward pocket-specific docking, which focuses computational resources on all-atom environments with known binding sites to achieve higher precision. Early efforts in this domain, such as DiffDock-Pocket (Plainer et al., 2023), adapted the original DiffDock baseline to predict ligand poses within specific sites while supporting receptor flexibility through side-chain torsion modeling. Subsequently, ArtiDock (Voitsitskyi et al., 2024) introduced a lightweight Graph Neural Network (GNN) architecture—drawing inspiration from TankBind (Lu et al., 2022)—to encode the pocket and ligand as interacting graphs; it specifically leveraged extensive synthetic data to enhance generalization. Emerging trends have further modified the core diffusion architec-

ture; for instance, DTMol (Teng et al., 2025) replaced traditional message-passing layers with Diffusion Transformers (DiTs) to better capture the intricate, long-range dependencies between ligand atoms and pocket residues. Current state-of-the-art research has since branched into specialized domains: DiffDock-RL++ (Broster et al., 2026) prioritizes physical validity by integrating reinforcement learning (RL) with relative reward mechanisms, while SigmaDock (Prat et al., 2025) utilizes a fragment-based diffusion approach to navigate high-dimensional conformational complexity efficiently.

Despite these advancements for docking, four critical challenges remain unaddressed. First, a data-scarcity bottleneck persists; most models are trained on relatively small, homogeneous datasets, limiting generalization. Second, the scaling–precision trade-off hinders real-world utility; all-atom pocket models often incur high computational costs, resulting in prohibitive inference latencies. Third, many frameworks rely on suboptimal scoring metrics, where high geometric accuracy (low RMSD) does not always translate to physically valid binding poses. Finally, the significant performance gap between Top-1 and Oracle success rates highlights the limitations of current ranking strategies, which often fail to correctly identify the most physically plausible pose from a generated ensemble.

To overcome these challenges, we present a high-performance docking pipeline that integrates large-scale data curation, structural optimization, and physics-informed refinement. Our work introduces four primary advancements:

- **Enhanced Data Robustness:** We expand the training distribution by curating a massive, heterogeneous dataset from PLINDER (Durairaj et al., 2024) and SAIR (Lemos et al., 2025), utilizing dataset-specific preprocessing and optimized pocket-cutoff adjustments to improve data quality.
- **Computational Efficiency via cuEquivariance:** We leverage hardware-accelerated cuEquivariance (NVIDIA, 2024) kernels to optimize our all-atom pocket docking model, significantly improving inference speed.
- **Refinement Strategies:** We apply a hybrid post-refinement stage—utilizing Vina-minimization for structural relaxation and GNINA (McNutt et al., 2025) re-ranking to prioritize physically plausible poses.
- **EGA-based Confidence Model:** Most significantly, we introduce a novel architecture for the confidence model that achieves performance competitive with state-of-the-art refinement strategies. By replacing traditional Fully Connected Tensor Product Convolutions with Equivariant Graph Attention (EGA), we improve the model's ability to rank poses based on complex geometric features.

Our all-atom scoring model achieves a state-of-the-art 81.85% Top-1 success rate using the improved EGA-based confidence model, slightly surpassing the 81.60% achieved when utilizing the physics-based post-refinement pipeline on the PoseBusters-308 benchmark. These results demonstrate that our EGA-based ranking model successfully captures distinct attention patterns for various types of molecular interactions, potentially leading to a promising exploration into various transformer-based architectures with attention mechanisms for boosting docking accuracy. Also, the integration of high-fidelity datasets, hardware acceleration, and physical refinement proves essential for the next generation of molecular docking frameworks.

2. Methodology

2.1. Data Curation and Preprocessing Pipeline

2.1.1. Multi-Source Dataset Integration

To enhance the model's generalization across chemical and biological space, we combine two massive, complementary data sources: PLINDER (Durairaj et al., 2024) and SAIR (Lemos et al., 2025). By integrating these sources, we leverage the high-quality, non-redundant protein diversity of PLINDER alongside the extensive structural ligand ensembles provided by SAIR.

PLINDER (Durairaj et al., 2024) is a curated dataset of 449,383 systems designed specifically to minimize data leakage, which features high-quality structures, and offers specialized splits. We adopt the PLINDER-TIME split, a time-based partition that evaluates model performance on future, unseen data. This subset contains 106,745 unique protein systems and 35,255 unique ligand SMILES, providing a significantly larger and more chemically robust training signal than traditional dataset like PDBbind (Wang et al., 2004).

SAIR - Structurally Augmented IC_{50} Repository. (Lemos et al., 2025) To supplement the experimental structures, we incorporate SAIR, a massive open-access collection of 1,048,857 protein-ligand systems derived from ChEMBL and BindingDB. For each system, Boltz-1 (Wohlwend et al., 2025) was used to generate 5 structures $\{\text{model_}i\}_{i=0}^{4}$, resulting in a total of 5,244,285 samples. While encompassing 5,149 unique protein sequences, SAIR's primary contribution is its immense ligand pose diversity, which allows our models to learn from a vastly expanded conformational space.

By aligning these sources, our dataset maximizes both the

diversity of protein binding pockets and the density of sampled ligand orientations, facilitating superior performance in real-world docking scenarios.

2.1.2. Data Leakage Prevention and Structural Quality Filtering

To guarantee unbiased evaluation and robust model training, we implemented a strict data curation pipeline focused on leakage prevention and conformational quality control.

Data Leakage Prevention: Fair benchmarking requires strict separation between training data and downstream evaluation sets. While PLINDER natively excludes targets present in the PoseBusters benchmark (Durairaj et al., 2024), the SAIR dataset required explicit deduplication. We applied a two-tier filter to SAIR eliminating exact PDB/CCD code matches and homologous protein sequences—which successfully reduced the data size from over 1 million to approximately 875,000 systems.

Structural Quality Filtering: Consistent with the DiffDock pipeline (Corso et al., 2023; Plainer et al., 2023), our training routine generates initial 3D ligand conformers directly from SMILES strings using RDKit. To ensure the physical reliability of these inputs, we enforce a strict quality threshold: any sample where the RDKit-generated conformer deviates from the ground truth by an RMSD greater than 2Å is discarded. This aligns our inclusion criteria with standard benchmarking success thresholds, ensuring the model learns exclusively from structurally viable data.

2.1.3. Dataset-Specific Pocket Cutoff Optimization

Standard docking pipelines typically employ a fixed pocket extraction radius for simplicity. However, this "one-size-fits-all" approach fails to account for the diverse geometries of different data sources, where pockets vary significantly in depth and spatial extent. A rigid threshold often leads to either the exclusion of key residues or the inclusion of non-informative structural noise.

To address this, we employ a dataset-specific tuning approach rather than a universal cutoff. We investigate the impact of varying extraction radius across different data sources and found that optimal performance is achieved by tuning the threshold for each specific source. This strategy ensures that the pocket representation is precisely aligned with the characteristic binding-site geometry of the dataset, maximizing the signal-to-noise ratio and improving the quality of structural inputs for the docking model.

2.2. Model Architecture:

Our framework builds on the diffusion-based paradigm of DiffDock (Corso et al., 2023) and DiffDock-Pocket (Plainer et al., 2023), utilizing a sequential pipeline of a score model for pose generation and a confidence model for structural ranking.

2.2.1. Score Model: All-Atom Representation

The score model learns the score functions over the ligand's degrees of freedom—translation ($\mathbb{T}^3$), rotation ($SO(3)$), and torsion ($SO(2)^k$)—to sample candidate poses $\{(\tilde{x}^{(j)}, \tilde{y}^{(j)})\}_{j=1}^{\mathbb{N}}$. We refine the protein representation by utilizing an all-atom graph. This provides a richer geometric context than the C_α-only approach of DiffDock (Corso et al., 2023), while remaining more computationally streamlined than DiffDock-Pocket (Plainer et al., 2023) by excluding flexible sidechain atoms. This balance captures essential pocket topography without the overhead of modeling sidechain degrees of freedom.

2.2.2. Confidence Model: Architectural Refinement and Soft-Labeling

The confidence model serves as the critical ranking component of the pipeline, estimating the probability that a generated pose is a successful docking event. We introduce two major enhancements: a transition to specialized equivariant attention and a continuous training objective.

From Convolution to Heterogeneous Attention

In DiffDock (Corso et al., 2023), confidence model operates on a heterogeneous graph with three node types — ligand atoms $\mathcal{L}$, receptor residues $\mathcal{R}$, and receptor atoms $\mathcal{A}$ — connected by 9 edge types per layer (3 intra-entity: $(\mathcal{L}\to\mathcal{L})$, $(\mathcal{R}\to\mathcal{R})$, $(\mathcal{A}\to\mathcal{A})$; 6 cross-entity: all pairwise combinations). Each edge type is handled by a separate `FullyConnectedTensorProductConv` (TP-conv), which performs SO(3)-equivariant message passing via spherical harmonics but weights all neighbors equally within a cutoff — limiting the model's ability to distinguish interactions of different chemical importance.

To advance confidence models capturing highly complex, anisotropic geometric relationships—an area where attention mechanisms have shown immense potential. The success of attention in structural biology is exemplified by AlphaFold2 (Jumper et al., 2021), which utilizes Triangular Attention within its Evoformer block to enforce geometric consistency by explicitly modeling spatial relationships between residue pairs. Building on spatial concepts, EquiformerV3 (Liao et al., 2026) recently introduced Equivariant Graph Attention (EGA), which utilizes depth-wise tensor products to modulate attention weights based on the relative 3D orientation of $SE(3)$-equivariant irreducible representations. Inspired by Equiformer-V3, we introduce HeteroEGA (Heterogeneous Equivariant Graph Attention), which upgrades every TP-conv in DiffDock's confidence

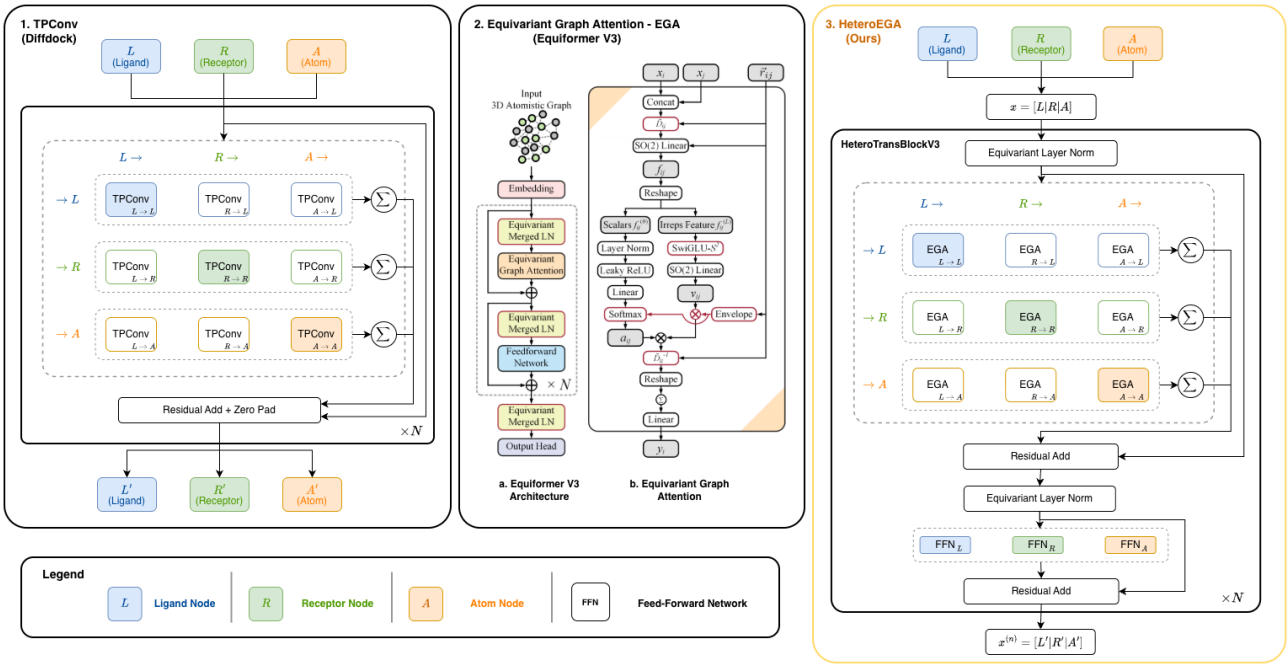

Figure 1. Architecture overview. (1) DiffDock's TPConv backbone uses 9 tensor product convolutions per layer across three node types (ligand, receptor residue, receptor atom) with uniform neighbor aggregation. (2) Equiformer V3's Equivariant Graph Attention (EGA) module computes SO(2)-equivariant attention weights via scalar softmax modulated by a polynomial envelope. (3) Our HeteroEGA replaces each TPConv with an independent EGA module, preserving the 9-way heterogeneous interaction topology while introducing learned attention weighting, per-entity-type feed-forward networks, and pre-norm residual connections within a unified equivariant feature tensor. $\mathbf{x}^{(n)}$ denoted as output after N `HeteroTransBlockV3` layers.

model to an SO(2)-equivariant graph attention module while preserving the full 9-way heterogeneous topology (Figure 1.3).

Concretely, each `HeteroTransBlockV3` layer contains 9 independent `EquivariantGraphAttention` (EGA) modules — one per interaction type — each with its own learned query, key, and value projections, radial weighting functions, and SO(2) linear layers. This enables the network to learn fundamentally different attention patterns for each type of molecular interaction: intra-ligand covalent connectivity, receptor backbone contacts, long-range ligand-receptor electrostatics, and binding-pocket atom contacts are each attended to by a dedicated, specialized module. Node features are maintained in a unified SO(3)-equivariant tensor $\mathbf{x} \in \mathbb{R}^{N \times (\ell_{\max}+1)^2 \times C}$, where N is the total number nodes, $(l_{\max}+1)^2$ spherical harmonic coefficients encode geometric information up to degree $l_{\max}$, C is the hidden dimension. At the layer i^{th}, it applies:

$$\mathbf{x}'^{(i)} = \mathbf{x}^{(i-1)} + \sum_{(s,t)\in\mathcal{E}} \mathrm{EGA}_{s\to t}\big(\mathrm{LN}(\mathbf{x}^{(i-1)})\big) \quad (1)$$

then,

$$\begin{aligned}\mathbf{x}^{(i)} = \mathbf{x}'^{(i)} + \big[&\mathrm{FFN}_{\mathcal{L}}\big(\mathrm{LN}(\mathbf{x}'^{(i)}_{\mathcal{L}})\big);\\ &\mathrm{FFN}_{\mathcal{R}}\big(\mathrm{LN}(\mathbf{x}'^{(i)}_{\mathcal{R}})\big);\\ &\mathrm{FFN}_{\mathcal{A}}\big(\mathrm{LN}(\mathbf{x}'^{(i)}_{\mathcal{A}})\big)\big]\end{aligned} \quad (2)$$

where $\mathrm{EGA}_{s\to t}$ is the specific Equivariant Graph Attention module for the source node type s and target node type t, with $\mathcal{E} \in \{\mathcal{L}, \mathcal{R}, \mathcal{A}\}$; LN denotes Equivariant Layer Normalization and (FFN_{τ}) are per-entity-type Feed-Forward Networks that allow the model to learn distinct nonlinear transformations for each chemical domain. The final layer applies only the 3 ligand-targeting attention modules, and confidence is predicted via learned attention-weighted pooling over ligand $(\ell{=}0)$ scalars followed by a 3-layer MLP.

Soft-Label Training.

To select the highest-quality poses, the confidence model predicts the likelihood of a successful docking event. While standard models utilize a hard RMSD threshold θ (e.g., 2 Å) for binary classification (Corso et al., 2023), this ignores the

nuances of near-threshold samples.

We implement a *"soft-labeling"* technique, converting RMSD labels into a continuous probability p via a logistic function:

$$p = \frac{1}{1 + e^{k(\mathrm{RMSD} - \theta)}} \tag{3}$$

By replacing rigid labels with a smooth gradient, the model accounts for uncertainty near the boundary and better utilizes "near-miss" data. The resulting output provides a more reliable confidence logit for practitioners to assess prediction accuracy in the absence of ground truth.

2.3. Post-Refinement & Ranking

Based on the findings of (Buttenschoen et al., 2024) and recent benchmarks from (Truong et al., 2026), which highlight the intrinsic lack of physical realism in end-to-end machine learning docking, we implement a refinement pipeline to reconcile diffusion-generated poses with established biophysical force fields: Vina minimization + GNINA (McNutt et al., 2021) Reranking.

We evaluated post-hoc minimization strategies using SMINA (Koes et al., 2013), a specialized fork designed for high-precision energy minimization. SMINA utilizes refined minimization algorithms that converge more precisely on local optima. Following the minimization step, we employ GNINA (McNutt et al., 2021) for final pose re-ranking, utilizing its integrated convolutional neural networks (CNNs) to evaluate protein–ligand complementarity and prioritize chemically accurate interactions through learned scoring. This strategy effectively recovers performance loss typically caused by structural inconsistencies, achieving a Top-1 success rate of 81.6% on the PoseBusters benchmark.

3. Experiments

3.1. Experiment Setup

3.1.1. Datasets

Score Model The training set comprises 940,000 samples, merging 80,000 complexes from PLINDER-time with 860,000 synthetic samples from SAIR *model_0* to enhance interaction accuracy and structural robustness. Validation is performed on the PLINDER-time split ($\sim$11,000 samples) for model selection.

Confidence Model To generate training data for the confidence model, we utilize the trained score model to perform inference on the PLINDER dataset, producing a set of candidate poses for each training example. We map the Root Mean Square Deviation (RMSD) between a predicted pose and its ground-truth pose to a continuous soft label $p \in [0, 1]$, as previously defined in Section 2.2.2, with $\theta = 2$ Å.

Benchmark Benchmarking is conducted on PoseBusters v2 (Buttenschoen et al., 2024), which includes 308 protein-ligand complexes with unseen protein sequences released post-2021. Rigorous verification was performed to ensure zero data leakage between the training and benchmark sets.

3.1.2. Training Configuration

Models are trained on NVIDIA A100 (80GB) GPUs using PyTorch, with cuEquivariance (NVIDIA, 2024) accelerating SE(3)-equivariant operations. We utilize the AdamW optimizer with an initial learning rate of 10^{-3} and a `ReduceLROnPlateau` scheduler, which dynamically reduces the learning rate when the validation metric plateaus. Training is performed with a batch size of 16 per GPU.

3.1.3. Inference Protocol

For each protein–ligand complex in PoseBusters-308, the model generates $N = 40$ candidate binding poses using the reverse diffusion process. We evaluate Top-1 and Oracle success rates (RMSD $<$2Å) by two ranking tracks as below:

(1) Confidence Ranking. Poses are ranked by the confidence model score, following a strategy similar to DiffDock, without physic-based minimization & ranking.

(2) Refinement & Final Ranking. Each candidate pose is minimized using SMINA with Vina scoring function (as described in section 2.3) to resolve minor geometric inconsistencies and improve local interactions. The refined pose is then re-ranked using the GNINA scoring function.

3.1.4. Evaluation Metrics

Pose quality is assessed using a 2.0 Å RMSD threshold between the predicted ligand pose and the ground truth. The Top-k success rate denotes the fraction of complexes where at least one of Top-k highest-ranked poses (from 40 generated poses per complex) meets this criterion. To ensure statistical robustness against sampling stochasticity, all metrics are averaged over 20 independent runs.

3.2. Main Results

3.2.1. Comparative Analysis with Baseline Methods

Table 1 shows our model consistently outperforms existing deep-learning methods on the PoseBusters-308 benchmark. We achieve an 81.85% Top-1 success rate and 94.51% Oracle performance, surpassing the state-of-the-art SigmaDock by +1.35% and +2.51%, respectively. To our knowledge, this is the first work to utilize the synthetic SAIR dataset for pocket docking, proving that synthetic data can effectively drive generalization. Furthermore, we are the first to

Table 1. Performance benchmark on the PoseBusters-308 test set against prior methods. Top-1 and Oracle denote success rate with RMSD < 2 Å, respectively. Best results are highlighted in **bold**.

Method	Type	Top-1 (%) ↑	Oracle (%) ↑
EqBind (Buttenschoen et al., 2024)	Blind	2.0	–
TankBind (Buttenschoen et al., 2024)	Blind	16.0	–
DiffDock (Buttenschoen et al., 2024)	Blind / Conf	38.0	–
ArtiDock (Voitsitskyi et al., 2024)	Pocket-AA	78.0	–
SigmaDock (Prat et al., 2025)	Pocket-AA / Physics	80.5	92.0
DiffDock-RL (Broster et al., 2026)	Pocket-AA / RL / Conf	69.0	84.8
DiffDock-RL++ (Broster et al., 2026)	Pocket-AA/ RL / Physics	80.2	88.5
Ours (Confidence Ranking)	Pocket-AA / Conf	**81.85**	**94.51**
Ours (Post-refinement + GNINA)	Pocket-AA / Physics	**81.60**	**94.17**

bring attention to the potential of confidence model, showing it can slightly outperform traditional post-refinement re-ranking (SMINA/GNINA) and achieve superior results without computationally intensive physics-based minimization.

3.2.2. Impact of Ranking Strategy

Table 2. Comparison of ranking strategies. Evaluated on identical generated poses, the public confidence model (Corso et al., 2023) is compared against our internal confidence model and a SMINA/GNINA post-refinement pipeline. Best results are highlighted in **bold**.

Ranking Strategy	Top-1 (%) ↑	Oracle (%) ↑
Confidence Model (Public)	77.39 ± 1.60	94.59 ± 0.71
Ours (Confidence Model)	**81.85 ± 1.62**	**94.51 ± 0.93**
Ours (GNINA only)	79.62 ± 1.85	94.51 ± 0.93
Ours (SMINA + GNINA)	81.60 ± 1.41	94.17 ± 0.69

While our model achieves a robust Oracle success rate of 94.59%, a significant gap exists at Top-1, which reaches only 77.39% when using a standard public checkpoint similar to DiffDock (Corso et al., 2023). This 17.2% discrepancy underscores the critical need for improved ranking strategies. Table 2 highlights our contribution: our proposed confidence model with HeteroEGA achieves 81.85%, representing a 4.46% improvement compared to public checkpoint and slightly outperforming traditional post-refinement re-ranking with SMINA and GNINA (81.60%). Notably, we find that GNINA re-ranking without structural minimization does not enhance performance significantly.

3.2.3. Ablation of Large-Scale Data Augmentation

To evaluate the impact of data augmentation, we conducted an ablation study using various training sets, in Table 3, employing a public DiffDock confidence checkpoint (Corso et al., 2023) for fair ranking across all tests. While PDBBind (∼20,000 samples) is a standard baseline in prior studies, training on PLINDER (∼80,000 samples) improved Top-1 success rate by 9.1% and Oracle performance by 7.6%, suggesting the model is significantly data-constrained. Incorporating PLINDER with the synthetic SAIR dataset as an augment further boosted both Top-1 and Oracle metrics by 5.74%. Notably, SAIR is most effective as an augmentation; it underperforms when used as a standalone training set compared to PDBBind or PLINDER.

Table 3. Impact of multi-source data augmentation. Performance is compared across models trained on the standard PDBBind baseline, synthetic (SAIR), curated experimental (PLINDER), and combined datasets. All configurations utilize the public confidence model (Corso et al., 2023) for pose ranking. Best results are highlighted in **bold**.

Training Dataset	Top-1 (%) ↑	Oracle (%) ↑
PDBBind (Baseline)	62.55 ± 1.86	82.94 ± 1.01
SAIR (Synthetic)	45.44 ± 2.40	77.64 ± 1.02
PLINDER	71.65 ± 1.59	90.54 ± 0.86
PLINDER + SAIR (Ours)	**77.39 ± 1.60**	**94.59 ± 0.71**

Table 4. Effect of dataset-specific pocket extraction radii. We compare model performance using different extraction cutoffs across the PDBBind, PLINDER, and hybrid datasets. For the hybrid configuration, cutoffs are denoted as (PLINDER / SAIR), respectively. All models utilize the public confidence model (Corso et al., 2023) for pose ranking. Best results within each dataset group are highlighted in **bold**.

Dataset	Cutoff (Å)	Top-1 (%) ↑	Oracle (%) ↑
PDBBind	5	62.55 ± 1.86	82.94 ± 1.01
	6	61.27 ± 1.23	**83.76 ± 1.04**
	7	56.81 ± 1.77	79.98 ± 1.18
PLINDER	5	71.44 ± 2.19	89.51 ± 0.95
	6	**71.65 ± 1.59**	**90.54 ± 0.86**
Hybrid	5 / 5	76.60 ± 1.30	93.73 ± 0.98
(PLINDER / SAIR)	**6 / 5**	**77.39 ± 1.60**	**94.59 ± 0.71**

3.2.4. Ablation of Pocket Cutoff Sensitivity Analysis

While pocket-specific docking typically uses a fixed cutoff threshold of 5Å we investigated adaptive cutoffs in Table 4. On PDBBind, a 6Å cutoff outperformed both 5Å and 7Å settings for Oracle success rate. Since PLINDER and PDBBind share the same data source, we evaluated PLINDER only at 5Å and 6Å, finding that 6Å provided slightly better results. This advantage was maintained when augmenting with SAIR (which uses a default 5Å cutoff). Consequently, we adopted the combination of a 6Å PLINDER cutoff and a 5Å SAIR cutoff as our optimal configuration for all experiments.

3.3. Inference speed:

Given that all-atom models are often computationally intensive, inference latency is a critical consideration. We compares our pipeline against high-performance baselines: SigmaDock (Prat et al., 2025) and DiffDock-RL++ (Broster et al., 2026), in Table 5. While SigmaDock scales to 22.8s for 40 poses (0.57s/pose), our reproduction of DiffDock-RL++[1] averaged 24.49 ± 0.11s. In contrast, by leveraging NVIDIA's cuEquivariance (NVIDIA, 2024) library to highly optimize our equivariant tensor product operations, our generation model requires only 11.26 ± 3.99s, achieving a nearly two-fold speedup. Furthermore, our EGA-based confidence model reduces ranking time to just 0.49 ± 0.19s—roughly 7 times faster than the traditional 3.28 ± 0.39s required for Vina and GNINA refinement—while maintaining comparable accuracy as detailed in Section 3.2.

Table 5. Inference time comparison. Generation time is measured for generating an ensemble of 40 poses per complex. Ranking time represents the overhead of evaluating and selecting the top pose. The performance metrics for *SigmaDock* are obtained from (Prat et al., 2025). Best results are highlighted in **bold**.

Method	Generation (s) ↓	Ranking (s) ↓
SigmaDock	22.80	–
DiffDock-RL++	24.49 ± 0.11	3.28 ± 0.39
Ours (Post-ref.)	$\mathbf{11.26 \pm 3.99}$	3.28 ± 0.39
Ours (Confidence)	$\mathbf{11.26 \pm 3.99}$	$\mathbf{0.49 \pm 0.19}$

4. Conclusion

In this work, we demonstrate that leveraging large-scale data—combining experimental and synthetic datasets—alongside an adaptive pocket cutoff significantly enhances model generalization. Specifically, we observed a 14.84% improvement in Top-1 success rate and an 11.65% increase in Oracle performance when transitioning from a PDBBind (5Å cutoff) baseline to our combined PLINDER (6Å cutoff) and SAIR (5Å cutoff) approach. These results, obtained using a consistent ranking strategy for fairness, suggest that synthetic data is a vital resource for addressing current challenges in pocket-specific docking.

Furthermore, we show that the ranking strategy is critical for narrowing the gap between Top-1 and Oracle success rates. While the Oracle performance reaches 94.59%, the public checkpoint achieves only 77.39%, highlighting a significant opportunity for ranking optimization. A key contribution of our work is the integration of Equivariant Graph Attention (EGA) into the confidence model. Our EGA-based model yields a 4.46% improvement over the public DiffDock confidence model and even slightly outperforms traditional physics-based methods, such as SMINA (Vina) minimization followed by GNINA ranking. This demonstrates that attention mechanisms are highly effective for ranking and provide a computationally efficient alternative to traditional refinement.

Our results highlight that the EGA-based architecture opens a promising direction for future research, offering a path to deliver even more robust performance within deep-learning-based pocket docking pipelines. Due to time constraints, we have not yet explored the full SAIR dataset or investigated the complete range of data features that could further enhance performance. However, our research confirms that synthetic data significantly benefits pocket docking, and the exploration of diverse synthetic resources remains a high priority. In future work, we will further validate our model using PoseBusters (PB-valid) metrics to confirm that our all-atom (AA) approach maintains high physical plausibility and structural integrity.

[1] https://github.com/oxpig/RLDiff

References

Broster, J. H., Popovic, B., Kondinskaia, D., Deane, C. M., and Imrie, F. Teaching diffusion models physics: Reinforcement learning for physically valid diffusion-based docking. *bioRxiv*, pp. 2026–03, 2026.

Buttenschoen, M., Morris, G. M., and Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.

Durairaj, J., Adeshina, Y., Cao, Z., Zhang, X., Oleinikovas, V., Duignan, T., McClure, Z., Robin, X., Studer, G., Kovtun, D., et al. Plinder: The protein-ligand interactions

dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

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 interactions with a synthetic structural dataset. *bioRxiv*, pp. 2025–06, 2025.

Liao, Y.-L., Hoffman, A. J., Shen, S. C., Duval, A., Norwood, S. W., and Smidt, T. Equiformerv3: Scaling efficient, expressive, and general se (3)-equivariant graph attention transformers. *arXiv preprint arXiv:2604.09130*, 2026.

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., and Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

McNutt, A. T., Li, Y., Meli, R., Aggarwal, R., and Koes, D. R. Gnina 1.3: the next increment in molecular docking with deep learning. *Journal of Cheminformatics*, 17(1): 28, 2025.

NVIDIA. cuequivariance: A math library for equivariant neural networks. https://github.com/nvidia/cuequivariance, 2024.

Plainer, M., Toth, M., Dobers, S., Stark, H., Corso, G., Marquet, C., and Barzilay, R. Diffdock-pocket: Diffusion for pocket-level docking with sidechain flexibility. 2023.

Prat, A., Zhang, L., Deane, C. M., Teh, Y. W., and Morris, G. M. Sigmadock: Untwisting molecular docking with fragment-based se (3) diffusion. *arXiv preprint arXiv:2511.04854*, 2025.

Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning (ICML)*, 2022.

Teng, H., Wang, R., Shen, Y., Yuan, Y., and Kingsford, C. Dtmol: pocket-based molecular docking using diffusion transformers. *bioRxiv*, pp. 2025–04, 2025.

Trott, O. and 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.

Truong, C.-M., Ballester, P. J., Taboureau, O., Tran-Nguyen, V.-K., et al. Nexttopdocker: the largest-to-date docking power benchmark reveals that deep learning performs generally much worse than logistic regression models. 2026.

Voitsitskyi, T., Yesylevskyy, S., Bdzhola, V., Stratiichuk, R., Koleiev, I., Ostrovsky, Z., Vozniak, V., Khropachov, I., Henitsoi, P., Popryho, L., Zhytar, R., Nafiiev, A., and Starosyla, S. Artidock: fast and accurate machine learning approach to protein-ligand docking based on multimodal data augmentation. *bioRxiv*, 2024. doi: 10.1101/2024.03.14.585019.

Wang, R., Fang, X., Lu, Y., and Wang, S. The pdbbind database: collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025.