FlashBind: Towards Accurate and Efficient Structure-based Virtual Screening

Songlin Jiang[1], Yifan Chen[1], Aarti Krishnan[2,3,4,5], Yu Zhang[2,3,4], Wengong Jin[1,2,†]

Abstract

Accurate prediction of protein-ligand interactions is central to computational drug discovery. Recent foundation models such as Boltz-2 have achieved remarkable accuracy in binding affinity prediction, yet their prohibitive computational cost remains a major barrier to large-scale virtual screening. Here we introduce **FlashBind**, a lightweight structure-based model that achieves a **50× speedup** over Boltz-2 at inference time by replacing expensive structure prediction with a fast docking model and substituting costly PairFormer modules with a streamlined EGNN architecture. FlashBind matches Boltz-2 on standard virtual screening benchmarks and demonstrates superior generalization to enzyme-substrate specificity prediction. To evaluate real-world applicability, we apply FlashBind to target-based antibiotic screening against the essential bacterial proteins in *E. coli* and show that FlashBind substantially outperforms Boltz-2 and other virtual screening baselines. Notably, several top-ranked candidates exhibit potent inhibition of DnaG and effective bacterial growth inhibition against E. coli in wet-lab validation. Together, these results demonstrate that FlashBind bridges the gap between accuracy and efficiency, enabling ultra-fast, high-fidelity screening of massive chemical libraries for drug discovery.

Keywords: virtual screening, drug discovery, deep learning, artificial intelligence, graph neural networks

1 Introduction

The discovery of novel bioactive small molecules is a fundamental pursuit in pharmaceutical science, yet it remains hindered by the vast chemical space, which is estimated to contain more than 10^{60} drug-like compounds [1]. High-throughput virtual screening serves as the critical filter in this process, aiming to identify potential binders from massive chemical libraries before experimental validation. Currently, structure-based virtual screening methods are divided into two categories. Physics-based docking approaches, such as AutoDock Vina [2], GNINA [3], and Glide [4], are computationally expensive and struggle to scale to ultra-large chemical libraries. On the other hand, deep learning models offer greater throughput but often suffer from limited generalizability across diverse targets.

In recent years, this lack of generalizability has been reshaped by the emergence of foundation models trained on immense biological datasets. Models such as Boltz-2 [5] have achieved remarkable accuracy in predicting protein-ligand complex structures and binding affinities. However, these gains come with prohibitive computational cost. The intricate architecture of such models, often relying on expensive recycling mechanisms and PairFormer modules, poses a major barrier to their deployment in large-scale virtual screening campaigns. For instance, processing a single protein-ligand complex with Boltz-2 requires approximately 35 seconds, a timeframe that renders the screening of billion-scale libraries computationally intractable. Therefore, a critical gap remains: the field lacks a solution that can match the predictive fidelity of foundation models while maintaining the throughput required for industrial-scale discovery.

To address this challenge, we introduce **FlashBind**, a lightweight geometric deep learning framework designed to bridge the gap between accuracy and efficiency. FlashBind achieves the accuracy-efficiency trade-off by structurally decoupling the screening process into two streamlined stages: rapid structure generation and geometric scoring. Instead of relying on computationally intensive diffusion-based generation or end-to-end folding, our framework utilizes a fast docking model (FABind+) [6] to provide a physically plausible structural prior. This structure is then

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA; [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3]Institute for Medical Engineering & Science and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; [4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA; [5]Whitehead Institute for Biomedical Research, Cambridge, MA, USA; [†]Correspondence to: w.jin@northeastern.edu

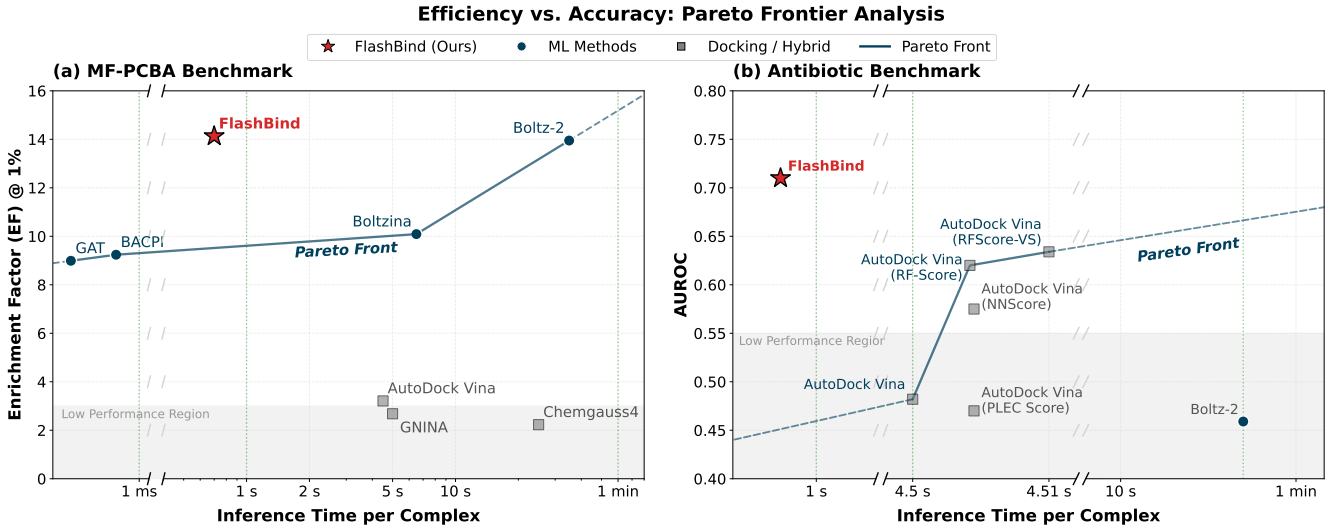

Fig. 1: **Computational efficiency vs. screening accuracy.** FlashBind (star) occupies the optimal efficiency-accuracy trade-off, matching Boltz-2 with 50× faster inference (0.7s vs. 35s). The inference time is measured on a single NVIDIA L40S GPU.

processed by a highly efficient $E(3)$-equivariant Graph Neural Network (EGNN) [7], which replaces heavy attention mechanisms to capture local physical interactions within the binding pocket.

The impact of this architectural shift is transformative. By replacing expensive structure prediction modules with a simplified docking-then-scoring workflow, FlashBind achieves a **50-fold speedup** at inference time compared to Boltz-2, reducing the processing time per complex to approximately 700 milliseconds. We demonstrate that this speed does not come at the cost of accuracy. On the standard MF-PCBA benchmark [8], FlashBind matches the early enrichment performance of Boltz-2, significantly outperforming both traditional physics-based scoring and sequence-based baselines.

Furthermore, we show that FlashBind possesses superior generalization capabilities beyond virtual screening. It effectively deciphers fine-grained enzyme-substrate specificity, matching the performance of enzyme-specific methods such as EZSpecificity [9]. Specifically, under the highly challenging "unknown enzyme and substrate" setting on the ESIBank benchmark, FlashBind achieves an overall AUROC of 0.7229, significantly outperforming sequence-based baselines like ESP. More notably, across diverse and data-scarce enzyme families, our general-purpose model consistently outperforms foundation models like Boltz-2, demonstrating that its learned geometric features successfully capture complex functional catalytic patterns without relying on heavy, task-specific architectures.

Lastly, we demonstrate real-world applicability of FlashBind via a structure-based antibiotic screening campaign targeting essential *E. coli* proteins [10], where Boltz-2 yield nearly random accuracy (AUROC = 0.46), and FlashBind achieved state-of-the-art performance (AUROC = 0.71). To further validate real-world utility, we conducted a prospective wet-lab screening campaign against *E. coli* DNA primase (DnaG), scoring 9,289 compounds with FlashBind and selecting 136 structurally diverse, PAINS-free candidates for experimental assay. Of these, 10 were confirmed as active inhibitors, achieving a hit rate of 7.4%, substantially higher than the typical of unguided screening. Moreover, 4 of the confirmed hits demonstrated whole-cell antibacterial activity against *E. coli*, indicating that FlashBind can identify compounds capable of not only inhibiting the target but also suppressing bacterial growth. These results highlight the robustness of our lightweight geometric encoder in capturing the structural determinants of bacterial target inhibition. In summary, FlashBind bridges the gap between accuracy and efficiency, providing a scalable tool for ultra-fast virtual screening of massive chemical libraries.

2 Results

We present FlashBind, a geometric deep learning framework designed to enable structure-based virtual screening at the scale of massive chemical libraries. We begin by detailing our rigorous data curation pipeline (Fig. 2a) and the inference architecture (Fig. 2b), which collectively ensure robust training signals and resolve the accuracy-efficiency trade-off by decoupling structure generation from geometric scoring (Section 2.1). To validate its industrial relevance, we first demonstrate its state-of-the-art performance in ultra-fast virtual screening on the MF-PCBA benchmark [8], where it matches foundation models in early enrichment while offering a 50-fold speedup (Section 2.2). Moving beyond

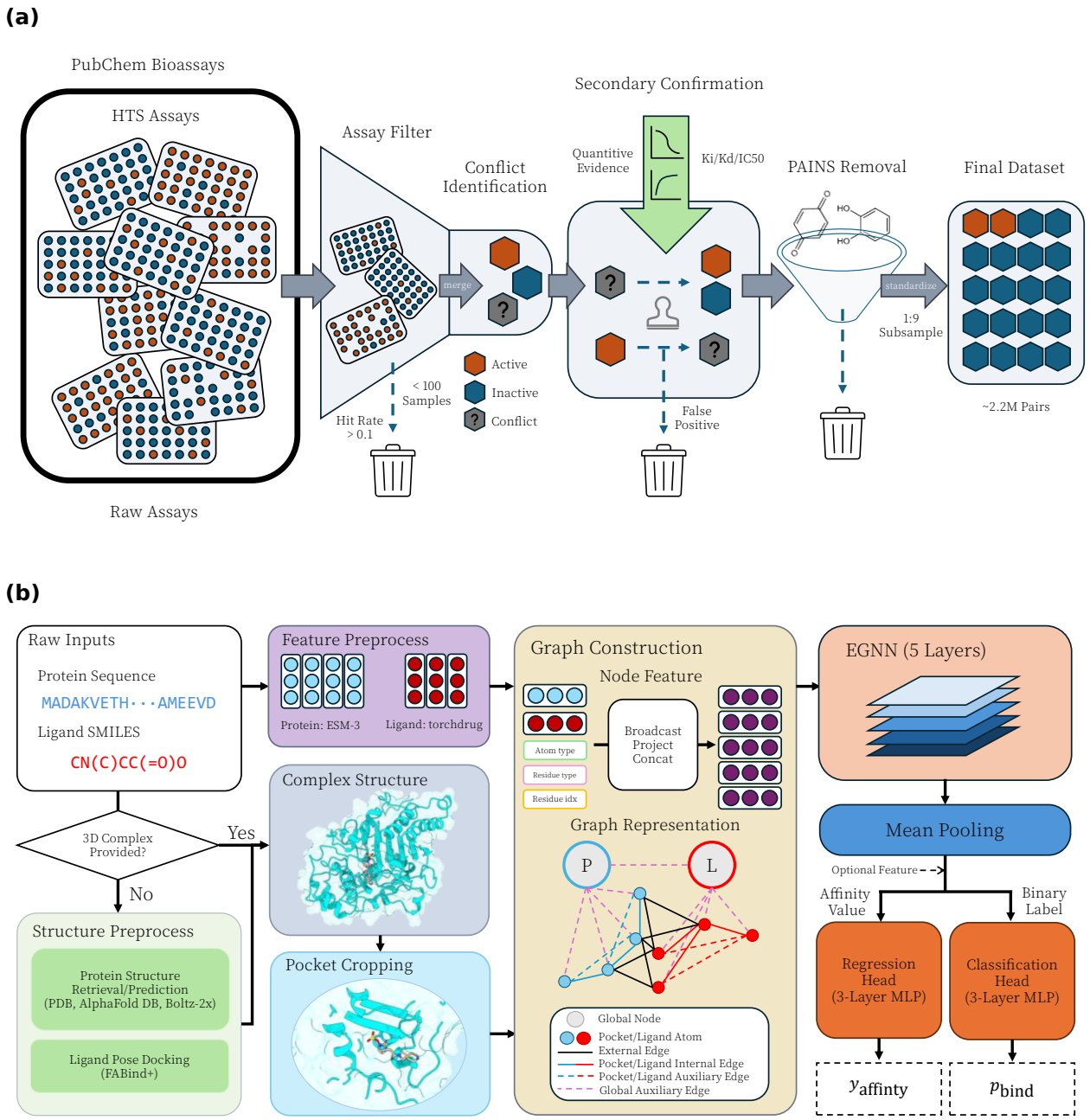

Fig. 2: **The FlashBind framework. (a)** Training data curation. To mitigate experimental noise, we constructed a multi-stage filtration pipeline with assay consistency checks, secondary confirmation, and PAINS removal. **(b)** FlashBind model architecture. The pipeline decouples structure generation from scoring. Inputs are mapped to a 3D complex using FABind+, then scored by an EGNN to predict binding probability.

general protein-ligand binding, we extend its capabilities to enzyme-substrate specificity prediction (Section 2.3). Finally, we showcase the model's real-world utility in a challenging antibiotic discovery campaign against E. coli, culminating in the experimental validation of selected compounds (Section 2.4).

2.1 FlashBind Framework

The FlashBind framework is designed to function as a high-throughput geometric encoder. Formally, given a protein amino acid sequence $\mathcal{S}_p$ and a ligand SMILES string $\mathcal{S}_l$, our primary objective is to learn a mapping function $\mathcal{F} : (\mathcal{S}_p, \mathcal{S}_l) \rightarrow p_{\text{bind}} \in [0, 1]$, where p_{bind} represents the probability of a ligand binding a protein.

A critical prerequisite for learning robust binding signals is the construction of a high-quality training set from inherently noisy experimental sources. As illustrated in Fig. 2a, we designed a multi-stage data curation pipeline based on Boltz-2 [5] to address the high false-positive rates endemic to High-Throughput Screening (HTS) assays from PubChem BioAssays [11]. The pipeline begins with an assay-level filter, retaining only confirmatory and primary screens with over 100 compounds and hit rates below 10%. Assays targeting the same protein (identified via UniProt ID [12]) were merged to maximize chemical diversity and reduce redundancy. We then applied a *secondary confirmation* mechanism at the compound level: active labels were retained only if corroborated by quantitative evidence (e.g., K_d, K_i, or IC_{50}) in the database, while conflicting entries lacking such confirmation were discarded. Finally, Pan-Assay Interference Compounds (PAINS [13]) were removed to eliminate frequent hitters. After balancing to a 1:9 binder-to-decoy ratio, the resulting dataset comprises approximately 2.2 million protein-ligand pairs.

The inference pipeline proceeds in three stages as shown in Fig. 2b. First, to circumvent the computational bottleneck of conformational sampling, we map the raw 1D inputs to a 3D structural complex $(\mathbf{X}, \mathbf{H})$ using a fast regression-based docking model (FABind+) [6]. Here, $\mathbf{X} \in \mathbb{R}^{N\times 3}$ represents atomic coordinates and $\mathbf{H} \in \mathbb{R}^{N\times d}$ represents atomic features. Crucially, this step provides a physically plausible structural prior without the prohibitive cost of diffusion-based generation, ensuring the scalability required for million-level libraries.

Second, to efficiently encode the physicochemical environment, we apply an adaptive cropping function $\mathcal{F}_{crop}$ that isolates the binding interface based on spatial proximity. This operation reduces the input space to a biologically relevant subset, filtering out noise from distant solvent-exposed regions. The cropped structure is then transformed into a multi-relational geometric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{h}, \mathbf{x})$. The node set $\mathcal{V}$ comprises both protein and ligand atoms, while the edge set $\mathcal{E}$ encodes interactions at multiple scales: internal edges representing intra-molecular topology, external edges capturing non-covalent contacts at the binding interface, and auxiliary edges incorporating global structural context.

Finally, the graph $\mathcal{G}$ is processed by an E(3)-equivariant graph neural network (EGNN) [7]. By enforcing E(3)-equivariance, the network updates node coordinates and features robustly regardless of the arbitrary orientation of the docked pose, eliminating the need for data augmentation. The network outputs a global graph representation $z_{\mathcal{G}}$, which is projected by a Multi-Layer Perceptron (MLP) classification head to predict the binary probability p_{bind}. We note that this architecture is task-agnostic regarding the final projection; for continuous affinity prediction, as detailed in the Supplementary Information A, we simply replace the classification head with a regression head to predict continuous affinity values (y_{affinity}).

2.2 Ultra-fast Virtual Screening

To assess the model's capability in identifying active compounds from vast chemical spaces, we benchmarked FlashBind on the MF-PCBA dataset [8], a standard benchmark adopted by Boltz-2 for evaluating virtual screening performance. The primary metric is the enrichment factor (EF), the ability to rank true binders at the very top of a prioritized list.

FlashBind demonstrates exceptional performance in this critical metric. As illustrated in Fig. 3a, our model achieves an Enrichment Factor at the top 1% (EF@1%) of 14.13, significantly outperforming traditional physics-based scoring functions like Chemgauss4 [14] (EF@1% = 2.23) and sequence-based deep learning baselines [15]. Most notably, FlashBind rivals the performance of the computationally intensive foundation model, Boltz-2 [5], which achieves an EF@1% of 13.95 on the same test set. This indicates that our lightweight geometric encoder effectively distills the complex structural signals required for hit identification.

To rigorously validate the sources of our efficiency and accuracy, we conducted a step-wise ablation study on a representative subset of the benchmark (Fig. 3b), constructed by sampling one-tenth of the compounds per target proportionally to the hit rate, as running ablation variants and traditional docking baselines on the full 500k-compound library is prohibitively expensive. Our analysis supports three key design premises. First, we confirmed that a fast docking oracle provides a sufficient structural foundation. The Boltz-2(FABind+) variant, which utilizes pre-computed poses from FABind+, maintains robust performance compared to the original Boltz-2 (EF@1%: 11.12 vs. 10.45). This conclusion is further supported by Boltzina [16], which substitutes the diffusion module with AutoDock Vina [2] and similarly achieves comparable results (EF@1%: 10.09). Together, these findings suggest that precise conformational sampling from diffusion models is not strictly necessary if a high-quality docked pose is available.

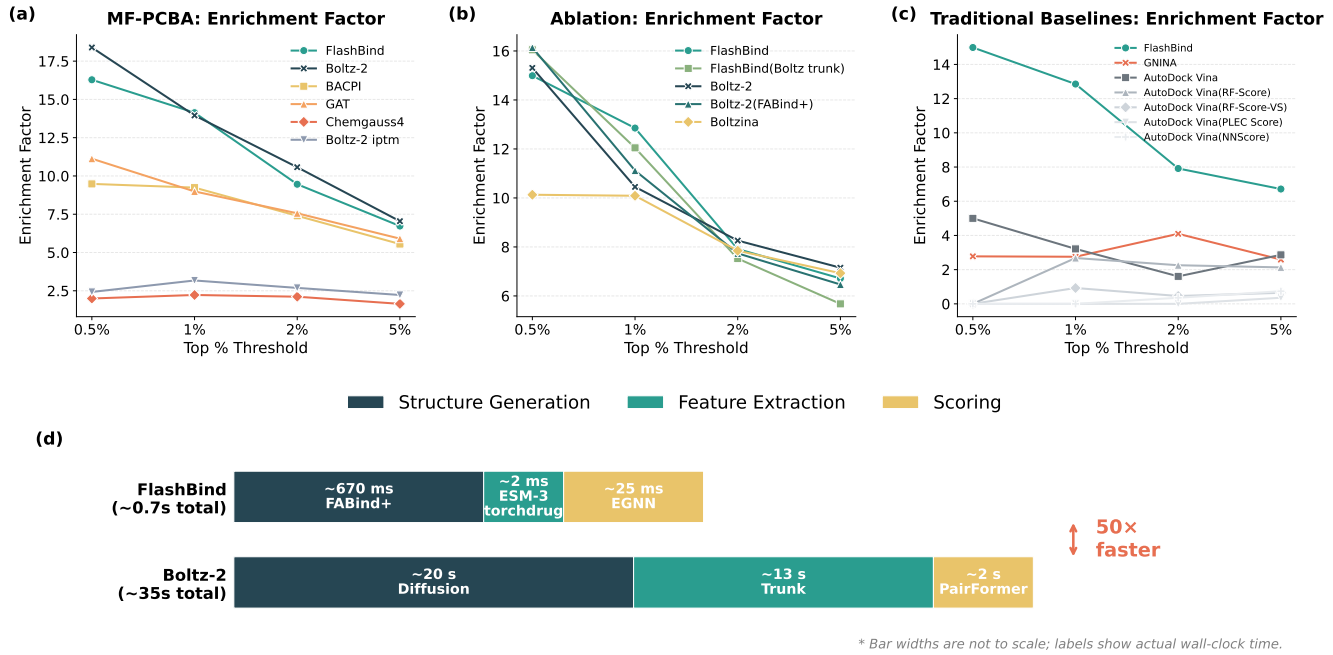

Fig. 3: **Enrichment Factor (EF) on the MF-PCBA benchmark and per-complex inference time breakdown. (a)** Comparison with deep learning and physics-based scoring baselines on the full benchmark. FlashBind rivals the foundation model Boltz-2 while substantially outperforming traditional methods. **(b)** Step-wise ablation study on a representative subset, validating each efficiency-oriented design choice. **(c)** Comparison with traditional docking and rescoring pipelines on the same subset. **(d)** Per-complex inference time breakdown. Each bar is decomposed into three pipeline stages: structure generation, feature extraction, and scoring. FlashBind replaces Boltz-2's diffusion sampling with FABind+, its deep trunk with pre-trained ESM-3/torchdrug embeddings, and its PairFormer scorer with a lightweight EGNN, achieving a cumulative **50-fold speedup** on a single NVIDIA L40S GPU. Bar widths are not to scale.

Second, we assessed the necessity of heavy-weight scoring architectures. By comparing Boltz-2(FABind+) with a variant of our model using Boltz-2's latent representations (FlashBind(Boltz trunk)), we observe that replacing the massive PairFormer module with our lightweight EGNN results in no significant performance loss (EF@1%: 12.05 vs. 11.12). This confirms that a streamlined equivariant graph network is sufficient to capture critical protein-ligand interactions.

Third, we validated the use of efficient pre-trained embeddings. The standard FlashBind model, which utilizes accessible ESM-3 [17] and torchdrug [18] features, achieves performance fully comparable to the variant relying on computationally expensive Boltz-2 trunk outputs (EF@1%: 12.85 vs. 12.05), effectively decoupling our framework from the foundation model.

Beyond internal validation, we benchmarked FlashBind against established traditional docking and rescoring methods on the same subset (Fig. 3c). Our geometric deep learning approach substantially outperforms standard AutoDock Vina scoring (EF@1% = 3.21) as well as random forest and neural network-based rescoring functions (e.g., RF-Score, GNINA [3, 19–22]).

The cumulative effect of these architectural optimizations is a decisive improvement in practical throughput, as visualized by the per-complex time breakdown in Fig. 3d. For Boltz-2, each complex requires ∼35 s of wall-clock time: ∼20 s for iterative diffusion-based structure generation, ∼13 s for computing trunk representations through the deep PairFormer, and ∼2 s for final scoring. FlashBind systematically replaces every expensive component with a lightweight counterpart. FABind+ generates a docked pose in ∼0.67 s; ESM-3 protein embeddings are computed once per target and amortized across all ligands, while torchdrug molecular features, based on lightweight RDKit descriptors, are generated at a throughput exceeding 800 molecules per CPU-second, together contributing negligible per-complex cost; and the equivariant EGNN scores each complex in only ∼25 ms (Table S5). The resulting end-to-end latency is ∼0.7 s, a **50-fold speedup**, with the docking oracle as the sole remaining bottleneck. In a practical rescoring scenario where a pre-docked pose library is already available, the scoring step alone enables evaluation of over 140,000 complexes per GPU-hour, making million-scale campaigns tractable on modest hardware. Fig. 1a

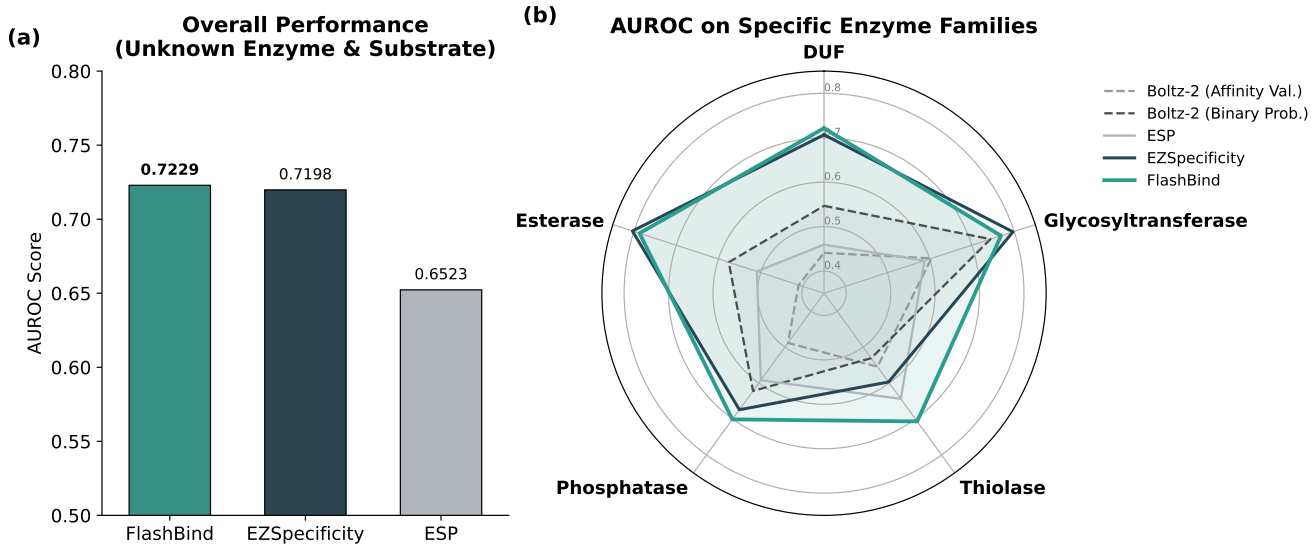

Fig. 4: **Evaluation of enzyme-substrate specificity prediction on the ESIBank benchmark. (a)** Overall performance comparison on the "unknown enzyme & substrate" split (4-fold cross-validation). FlashBind achieves an AUROC of 0.7229, performing comparably to the specialized state-of-the-art model EZSpecificity (0.7198) and significantly outperforming sequence-based baselines like ESP. **(b)** Fine-grained analysis on representative enzyme families (Thiolase, Glycosyltransferase, DUF, Phosphatase, Esterase). FlashBind demonstrates robust generalization in these specific categories, outperforming the foundation model Boltz-2 and matching the specialized architecture of EZSpecificity, indicating that our geometric encoder effectively captures functional catalytic patterns.

further contextualizes this gain as a Pareto frontier of enrichment versus inference cost: FlashBind occupies the optimal region, matching foundation-model-level enrichment at a fraction of the computational budget.

2.3 Enzyme-Substrate Interaction

In addition to virtual screening, another application of the protein-ligand binding predictor is the decoding enzyme-substrate specificity, a functional property governed by precise catalytic alignment rather than thermodynamic stability alone. To evaluate FlashBind's ability in this fine-grained regime, we conducted a case study using the ESIBank data set [9], a comprehensive benchmark for enzyme specificity prediction, on which FlashBind was retrained from scratch.

We rigorously evaluated our model under the "unknown enzyme & substrate" split, the most challenging setting, where neither the protein nor the small molecule have been seen during training. Following the protocol of the state-of-the-art method EZSpecificity [9], we performed 4-fold cross-validation to ensure statistical robustness. Recognizing that enzyme specificity is often governed by subtle chemical functional group interactions that pure geometric scoring might miss, we adopted a similar strategy to the EZSpecificity framework by augmenting our geometric encoder with explicit chemical descriptors (UniMol embeddings [23] and Morgan fingerprints [24]).

As shown in Fig. 4a, FlashBind achieves an overall AUROC of 0.7229, demonstrating performance fully comparable to the specialized EZSpecificity model (AUROC = 0.7198) and significantly outperforming the sequence-based baseline ESP [25] (AUROC = 0.6523). This result is particularly notable given that FlashBind utilizes a general-purpose geometric encoder, whereas EZSpecificity employs a heavy, task-specific cross-attention architecture designed exclusively for this problem.

We further investigated performance across specific enzyme families, including data-scarce categories like Thiolases and Domain of Unknown Function (DUF) proteins (Fig. 4b). In these specific regimes, FlashBind consistently outperforms the foundation model Boltz-2 and the sequence-based ESP, while maintaining parity with EZSpecificity. For instance, in the Glycosyltransferase family, our model effectively captures the subtle structural determinants required for sugar transfer, a task where pure sequence-based methods often falter. The ability to match a specialized SOTA model on its own benchmark serves as strong validation of our architectural soundness, suggesting that the geometric features learned by FlashBind are not limited to binding affinity but are transferable to complex functional prediction tasks.

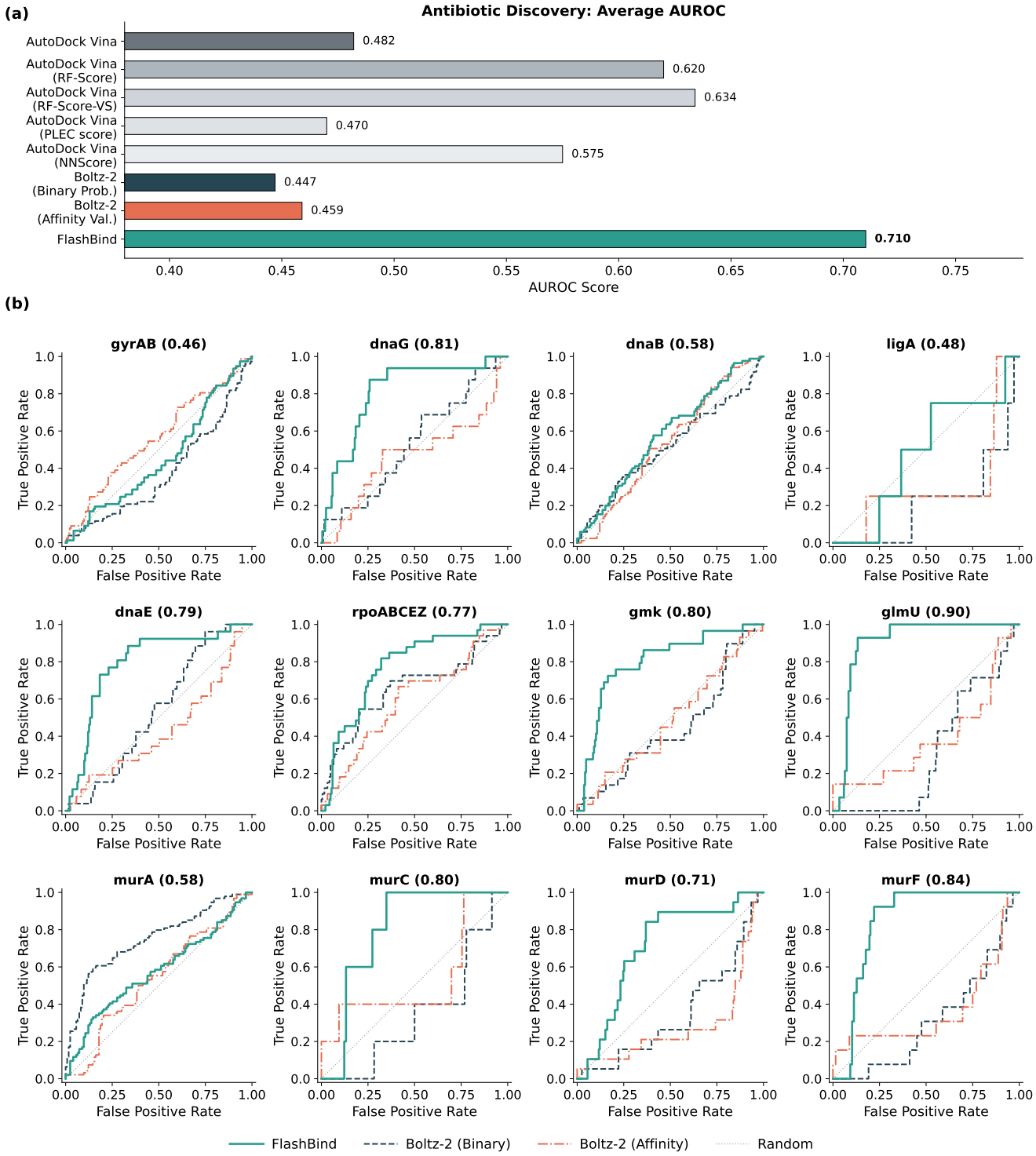

Fig. 5: **Benchmarking performance on the antibiotic discovery task. (a)** Comparison of average AUROC across 12 essential *E. coli* targets. FlashBind (0.710) demonstrates practical ranking capability, significantly outperforming standard docking (AutoDock Vina) and machine-learning rescoring functions. **(b)** Representative ROC curves for key targets including guanylate kinase (*gmk*) and bifunctional acetyltransferase (*glmU*). FlashBind (green) maintains higher true positive rates compared to baselines.

2.4 Antibiotic Discovery and Experimental Validation

While the previous section established FlashBind as an ultra-fast filter for large-scale screening benchmarks, validating its utility in real-world discovery campaigns is the ultimate test. To this end, we apply FlashBind to a structure-based antibiotic discovery task [10]. This task is particularly challenging, as a model must not only identify compounds that bind bacterial targets but also inhibit bacterial growth, and must do so within a chemical space distinct from standard training sets.

The benchmark dataset [10] comprises 218 active antibacterial compounds and 100 inactive compounds docked to essential *E. coli* proteins, whose structures are predicted by AlphaFold2. Ground-truth labels are derived from *in vitro* enzymatic inhibition assays across 12 essential proteins of *E. coli* (e.g., DNA gyrase, MurA). A compound is labeled positive if it shows more than 50% dnaG inhibition in both replicates. Model performance is evaluated via the area under the receiver operating characteristic curve (AUROC), averaged across all 12 target proteins.

We compared FlashBind with Boltz-2, standard molecular docking tools such as AutoDock Vina [2], and various machine-learning scoring functions (Fig. 5). To evaluate the zero-shot performance of FlashBind, we tested all models without finetuning. We found that traditional physics-based docking struggles to distinguish actives from decoys (Vina Avg. AUROC ≈ 0.48). Similarly, Boltz-2 does not transfer effectively to this dataset, with performance comparable to random guessing (AUROC ≈ 0.45).

In contrast, FlashBind demonstrates practical utility in this regime, achieving a mean AUROC of 0.710. As shown in the target-specific ROC curves (Fig. 5b), our model consistently retrieves active scaffolds for critical targets such as *gmk* and *glmU*. The performance gap compared to baselines suggests that our approach is sufficiently robust to prioritize candidates even in complex biological assays, making it a viable tool for the initial stages of antibiotic discovery campaigns.

Furthermore, as illustrated in the Pareto frontier analysis (Fig. 1b), FlashBind lies above the Pareto front, achieving superior accuracy while remaining over 6× faster than AutoDock Vina and other rescore method (0.7s vs. 4.5s per complex). This positions FlashBind as the only method simultaneously surpassing both the accuracy and efficiency of existing baselines in this benchmark.

To assess whether FlashBind's predictions translate into real-world success, we conducted a prospective virtual screening campaign targeting *E. coli* DNA primase (dnaG). We scored all 9,289 compounds from the Broad Institute compound collection using FlashBind (Fig. 6a). From the top-ranked candidates, we applied two filters to ensure chemical quality and diversity: compounds containing pan-assay interference (PAINS) substructures were excluded, and redundant scaffolds were removed by retaining only compounds with pairwise Tanimoto similarity below 0.5. This yielded a final selection of 136 compounds for experimental testing.

Following Wong et al. [10], each compound was assayed against DnaG in two independent replicates at a concentration of 100μM. Normalized inhibition scores were averaged to produce a final activity score, where higher scores indicate stronger DnaG inhibition, and a compound was classified as active if it induced more than 50% inhibition of DnaG. As shown in Fig. 6a, 10 of the 136 compounds tested were confirmed to be active, corresponding to a hit rate of 7.4%. This substantially exceeds the typical hit rates observed in unguided random screening campaigns, demonstrating that FlashBind's rankings provide actionable enrichment in a prospective setting.

We further evaluated each dnaG inhibitor for whole-cell antibacterial activity using two *E. coli* inhibition assays at a concentration of 128μg/mL (Fig. 6b). The first assay measures *E. coli* cell viability upon compound treatment alone, while the second assay co-administers a sub-inhibitory concentration of polymyxin B (PMB) to permeabilize the outer membrane. The goal of the second assay is to isolate intracellular target inhibition from membrane permeability, since the screening process did not take into account permeability. Among the 10 confirmed hits, 4 compounds exhibited strong EC+PMB inhibition ($\geq 50\%$) alongside high DnaG inhibition, demonstrating that FlashBind can effectively prioritize compounds with genuine whole-cell antibacterial potential. All 10 confirmed hits span structurally diverse chemotypes (Fig. 6c); for example, BRD-K00003155 combines potent DnaG inhibition with whole-cell activity within a compact, drug-like scaffold, illustrating that FlashBind captures genuine binding signals rather than overfitting to a narrow chemical series.

3 Discussion

The central dichotomy in modern AI-driven drug discovery has long been the trade-off between computational efficiency and predictive accuracy. While foundation models have demonstrated remarkable capacity for capturing molecular interactions, their prohibitive computational costs effectively restrict their utility to lead optimization stages, leaving the vast chemical space of initial screening to faster, yet less accurate, physics-based scoring functions. FlashBind reconciles this conflict. By occupying the optimal region on the efficiency-accuracy Pareto frontier, our framework demonstrates that the early enrichment capabilities of large-scale foundation models can be achieved with

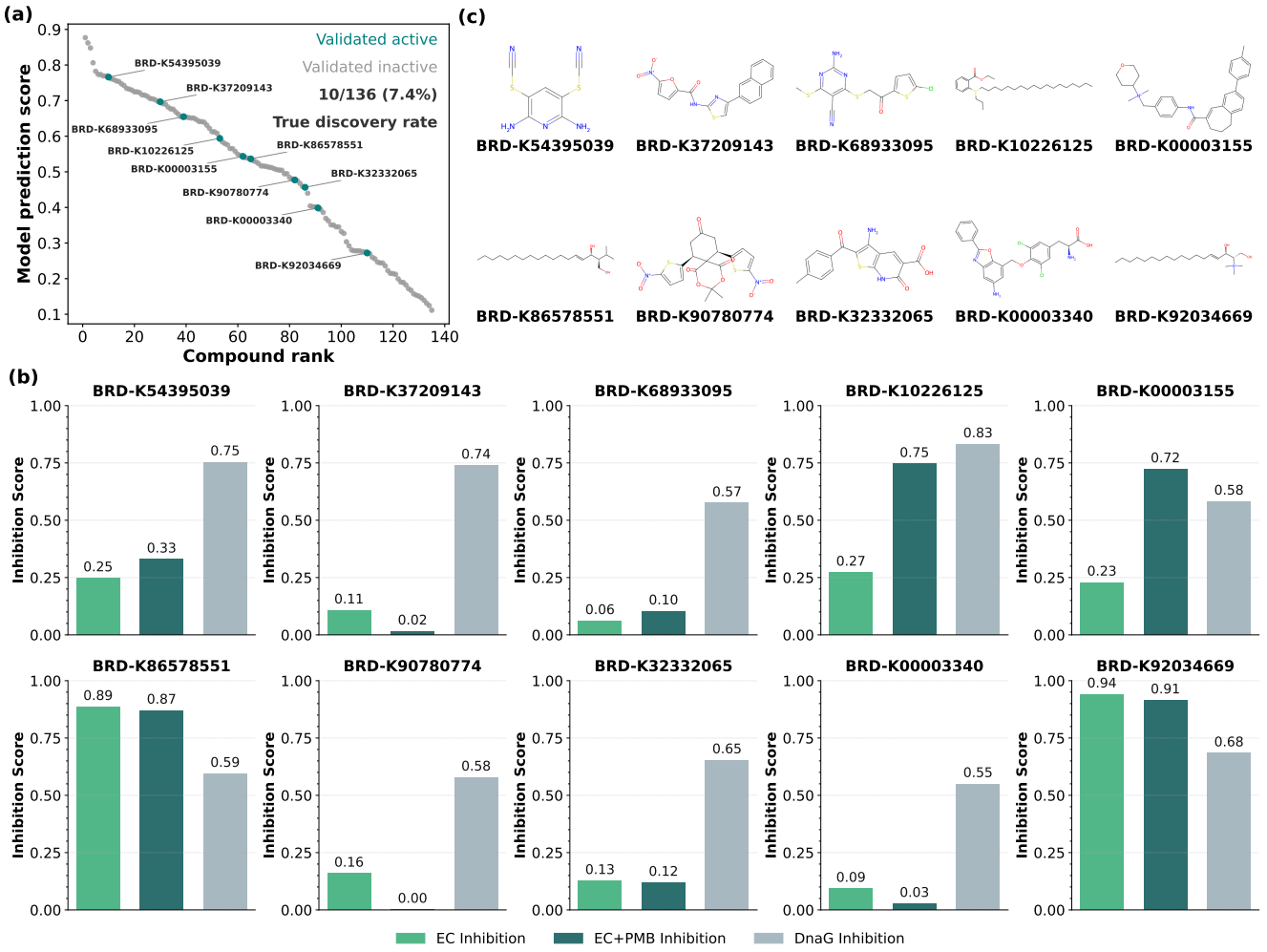

Fig. 6: **Prospective wet-lab validation of FlashBind on DnaG. (a)** Rank-ordered FlashBind prediction scores for the 136 selected compounds screened against *E. coli* DNA primase (dnaG). Teal points indicate experimentally confirmed actives (DnaG inhibition $\geq 50\%$), yielding a hit rate of 10/136 (7.4%).**(b)** Experimental profiles of all 10 confirmed hits, showing EC inhibition, EC+PMB inhibition, and DnaG inhibition for each compound. Higher values indicate stronger inhibitory activity. **(c)** Chemical structures of all 10 confirmed hits identified by FlashBind.

a 50-fold reduction in inference latency. This "best-of-both-worlds" capability is pivotal for the democratization of high-fidelity virtual screening, enabling academic laboratories and smaller biotech entities to interrogate billion-scale chemical libraries that were previously accessible only via immense computational resources.

Our findings challenge the prevailing assumption that end-to-end, heavy-weight structure generation is a prerequisite for high-accuracy scoring. The success of FlashBind suggests that precise conformational sampling via expensive diffusion models is not strictly necessary for hit identification. We observe that a lightweight E(3)-equivariant geometric encoder can effectively catch the protein-ligand interactions in fast docking priors (e.g., FABind+ [6]), focusing on robust geometric equivariants, such as heavy-atom contact maps and spatial proximity patterns, rather than relying on precise atomic coordinates that require explicit hydrogen or solvent modeling. Furthermore, our study underscores the critical importance of rigorous data curation in the era of large-scale learning. While we significantly expanded the training scale compared to previous sequence-based methods, data quantity alone is insufficient; the expansion of scale must be rigorously coupled with strict filtration pipelines to mitigate experimental noise and assay artifacts. This suggests that for biological learning, simultaneously pursuing both data quantity and quality control is essential to prevent model hallucinations.

Crucially, the geometric representations learned by our framework are versatile and not limited to binary classification. As detailed in the A, the FlashBind architecture is inherently task-agnostic; by substituting the classification head with a regression module, we extended the model to quantitative affinity prediction. While the model achieves competitive performance superior to traditional baselines and demonstrates comparable results on

the recent CASP16 benchmark [26], it does not consistently outperform the foundation model Boltz-2 [5] across all datasets (e.g., OpenFE [27]). We hypothesize that this performance gap stems largely from data provenance rather than architectural limitations, a conclusion further supported by our ablation study in Appendix A.3; whereas Boltz-2 leverages a massive, rigorously self-curated dataset with optimized structures, our current regression training relied on the SAIR database [28]. This highlights that while our lightweight architecture is theoretically sound, its peak performance in regression tasks is heavily contingent on the granularity and quality of the underlying training data distributions.

The generalization capability of FlashBind is further evidenced by its performance on other biological tasks. In the antibiotic discovery and enzyme specificity campaigns, where foundation models frequently suffered from negative transfer between stable crystal structures and complex enzymatic assays, our approach maintained high predictive validity. This suggests that by prioritizing explicit local geometric constraints, FlashBind mitigates the risk of overfitting to the specific protein families or eukaryotic targets that dominate standard training sets, thereby offering reliable predictions in novel, data-scarce biological regimes.

Despite these advances, limitations remain. As a modular structure-based method, the system's predictive upper bound is inherently constrained by the fidelity of the upstream docking oracle. Although our pipeline can dynamically identify binding pockets, in scenarios involving significant conformational plasticity or cryptic binding sites where FABind+ fails to generate a physically plausible pose, the downstream EGNN [7] may inevitably process incorrect geometric signals, leading to false predictions. Future iterations will aim to mitigate this dependency by integrating lightweight flexible-docking modules to better handle backbone dynamics. Furthermore, given the deliberately streamlined design of our current encoder, there is significant potential to scale the architecture; we plan to explore more expressive geometric networks to capture higher-order many-body interactions and expand the training horizon to encompass broader bioactivity datasets. In summary, FlashBind presents a scalable, accurate, and physically grounded framework that redefines the practical limits of structure-based virtual screening.

4 Methods

4.1 Data Curation

To ensure rigorous evaluation across diverse drug discovery tasks, we curated three distinct datasets focusing on high-throughput screening, enzymatic specificity, and binding affinity quantification.

Virtual Screening Dataset

The construction of our virtual screening dataset from PubChem BioAssays [11], including the multi-stage filtration pipeline with secondary confirmation and PAINS removal, is described in detail in Section 2.1. The final training set contains **2,237,058 protein-ligand pairs** spanning **451 protein targets** and **368,812 ligands**.

For validation, we utilized a curated subset of the LIT-PCBA benchmark [29], a challenging virtual screening dataset derived from PubChem. Our validation set consists of 43,492 protein-ligand pairs, quantitatively sampled from each of LIT-PCBA's 15 targets to ensure that the hit rate for every target remained above 0.5%.

Enzyme-Substrate Interaction Dataset

For the fine-grained enzyme-substrate interaction task, we utilized the standard ESIBank benchmark [9] dataset directly. This dataset comprises **323,783 pairs**, covering **8,124 enzymes** and **34,417 ligands**, representing a challenging functional space where distinguishing subtle substrate specificity is required. Consistent with the training protocol of *EZSpecificity*, we adhered to the standard unknow enzyme&substrate splits defined within the ESIBank benchmark itself to ensure a fair comparison.

Affinity Value Dataset

For the auxiliary task of continuous affinity prediction, we curated a high-quality subset of the SAIR dataset [28]. We filtered for assays with distinct chemical series and sufficient dynamic range to ensure learnable structure-activity relationships. The resulting dataset contains **796,557 pairs**, covering **28,821 proteins** and **403,124 unique ligands**. We constructed the validation set by holding out a portion of the SAIR training data, created by randomly selecting 20 distinct assays to assess model generalization across different experimental conditions.

Data Partitioning and Leakage Prevention.

To ensure a rigorous evaluation of generalization capabilities, we implemented stringent protocols to prevent data leakage for the curated datasets.

- **Protein-level:** We removed any protein from the training set sharing $\geq 90\%$ sequence identity with proteins in the validation or test sets. This exclusion was determined by clustering the complete protein set using MMseqs2 [30].
- **Ligand-level:** To prevent the model from exploiting learned chemical scaffolds, we filtered the training data based on chemical similarity. Any ligand in the training set exhibiting a Tanimoto similarity > 0.4 [31] to any *active* ligand in the test set was discarded.

4.2 Structure Generation

FlashBind operates as a scoring function that relies on 3D structural inputs. For entries defined only by protein sequence and ligand SMILES, we employ a hierarchical structure generation pipeline designed to prioritize experimental accuracy while maintaining coverage.

Protein Structure Retrieval

We first query the Protein Data Bank (PDB) [32] for experimental structures matching the target sequence with 100% identity and high resolution. If no exact experimental match is found, we search the AlphaFold Protein Structure Database [33] for predicted structures and select the model with the highest mean pLDDT confidence score. In rare cases where neither is available, we generate the structure de novo using Boltz-2x [5]. This hierarchy ensures that the model utilizes the most reliable structural prior available. In practice, we find that over 60% of target structures can be directly retrieved from PDB or the AlphaFold Database, substantially reducing the computational overhead associated with de novo structure prediction.

Ligand Docking and Cropping

Ligand poses are generated using FABind+ [6], a regression-based docking model selected for its inference speed (< 0.7s per complex). To focus the geometric encoder on the relevant binding interface, we apply a budget-constrained cropping algorithm. We consider only protein residues with atoms falling within 20Å of the ligand center. From this subset, residues are selected greedily based on proximity until an atom budget ($B_a = 2048$) or residue budget ($B_r = 512$) is reached.

4.3 Geometric Graph Architecture

The cropped complex is transformed into a multi-relational geometric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, h, x)$. Node features h are constructed by concatenating embeddings from the ESM-3 [17] protein language model with torchdrug [18] ligand chemical descriptors generated via RDKit [34] (detailed in B.1). The edge set $\mathcal{E}$ encodes interactions at multiple scales, including internal covalent bonds, external protein-ligand contacts (< 10Å), and auxiliary global connections.

The core encoder is an $E(3)$-Equivariant Graph Neural Network (EGNN) [7] comprising $L = 5$ layers with a hidden dimension of 192. This architecture ensures that the learned representations and predicted scores are invariant to arbitrary rotations and translations of the input docked poses. The final graph representation is pooled and passed to task-specific Multi-Layer Perceptrons (MLP), optionally concatenated with auxiliary feature vectors depending on the task.

4.4 Training Strategy

We employ a group-based mini-batch sampling strategy trained on a cluster of NVIDIA L40S GPUs.

Virtual Screening

To optimize for early enrichment, we train the model using the Focal Loss ($\gamma = 1, \alpha = 0.7$) [35]. We utilize a targeted sampling strategy where each training batch consists of groups sampled from a single assay, enforcing a strict 1:4 binder-to-decoy ratio.

Enzyme-Substrate Interaction

For the enzyme specificity task, we strictly adhere to the evaluation protocol of the state-of-the-art benchmark, *EZSpecificity*. We perform **4-fold cross-validation** on the "unknown enzyme & substrate" split. The model employs the same Focal Loss as the screening module. To ensure a rigorous comparison and align with the input modalities of the baseline architecture, the geometric representation is concatenated with UniMol embeddings[23] and Morgan fingerprints [24] specifically for this task.

Affinity Value Prediction

For the regression task, the model is trained to minimize a composite objective function combining pairwise ranking accuracy and absolute error [36]:

$$\mathcal{L}_{\text{affinity}} = 0.9 \cdot \mathcal{L}_{\text{Huber}}(y_i - y_j, \hat{y}_i - \hat{y}_j) + 0.1 \cdot \mathcal{L}_{\text{Huber}}(y_i, \hat{y}_i) \tag{1}$$

where the first term penalizes errors in the pairwise difference between compounds within the same assay, and the second term penalizes absolute prediction error.

4.5 Inference and Ensembling

We employ an ensembling technique, selecting the top $k = 2$ checkpoints based on validation performance.

- **Classification:** For screening and enzyme tasks, the final probability is the average of the ensemble outputs.
- **Affinity Prediction:** We apply a post-hoc polynomial correction to the averaged predictions based on the ligand's molecular weight (MW). This step corrects for systematic size-related scoring biases common in docking functions.

4.6 Experimental validation

***E. coli* dnaG inhibition assay**. Inhibition of *E. coli* DNA primase (dnaG) was assessed using an *in vitro* assay developed by ProFoldin (Hudson, MA), following the manufacturer's instructions. The assay is based on the measurement of the RNA primers synthesized by DNA primase in the presence of DNA template and NTPs. For screening experiments, reactions were performed using 40μl of reaction mixture including 24μl ultrapure Milli-Q water, 4μl of 10x assay buffer, 4μl of 10x DNA template, 4μl of 10x enzyme, and 4μl of 10x NTP mix, resulting in final concentrations of 10mM HEPES (pH 7.5), 5mM magnesium sulfate, 0.5mM dithiothreitol, 0.003% Brij-35, 100nM DNA, 0.5mM NTPs, and 100nM enzyme. 36μl of diluted buffer containing enzyme and NTP mix was plated into standard black 384-well plates (Corning 3575). Where applicable, 0.8μl of test compound (or DMSO as a negative control) was added, and plates were incubated at room temperature for at least 5min. Four μl of 10x DNA template was then added to each reaction. For generating standard curves, the amount of substrate (DNA template) added was decreased in proportion to activity. Plates were incubated at 37°C for 2h. The provided 10× fluorescence dye was diluted 10-fold with ultrapure Milli-Q water. After incubation, 60μl of 1 dye was added to each reaction, and mixtures were incubated at room temperature for 5min. The fluorescence excitation/emission at 485/535nm was then measured using a SpectraMax M3 plate reader. For each sample, activity was determined by linear interpolation with respect to the standard curves provided that the resulting fluorescence intensity value fell within the standard curve range. Otherwise, fluorescence intensity values below that of the zero standard were clipped to that of the zero standard, and fluorescence intensity values above that of the highest standard were linearly extrapolated with respect to that of the highest standard. For subsequent validation dose–response experiments, half the indicated reaction volumes—that is, 20μl for each reaction mixture—was used, and 40μl of 1x dye was added to each reaction.

***E. coli* inhibition assay**. To test the antibacterial activity of each compound, we grow *E. coli* cells overnight in 3 mL LB medium and diluted 1/10,000 into fresh LB. In 96-well flat-bottom plates (Corning), cells are then introduced to compound at an initial concentration of 128μg/mL, either mixed or not mixed with 32μg/mL polymyxin B nonapeptide. The plates are then incubated at 37°C without shaking until untreated control cultures reach stationary phase, at which time plates were read at 600 nm using a SpectraMax M3 plate reader. Cell viability values are normalized by the mean of two DMSO controls.

Compound preparation. Compounds with high purity were procured from the Broad Institute Center for the Development of Therapeutics.

Declarations

Code Availability

The source code for FlashBind, including model architecture, training scripts, and inference pipeline, is publicly available at https://github.com/AIDD-Lab/FlashBind.

Data Availability

All datasets used in this study, including the curated virtual screening dataset, the MF-PCBA benchmark splits, the ESIBank enzyme-substrate interaction dataset, the antibiotic discovery benchmark, and the SAIR affinity dataset, are publicly available at https://huggingface.co/datasets/clorf6/FlashBind.

Appendix

Appendix A Binding Affinity Prediction

While the main text focuses on the binary classification task for high-throughput virtual screening, the geometric encoder of FlashBind is inherently task-agnostic. To adapt the framework for quantitative binding affinity prediction (regression), we replaced the binary classification head with an architecturally identical Multi-Layer Perceptron (MLP) regression head, simply removing the final sigmoid activation function to produce a continuous scalar output. In this section, we demonstrate the model's capability in this regime and investigate the impact of encoder architecture on performance.

A.1 Data Curation and Filtering

To train the model for continuous affinity prediction (e.g., K_i, K_d, IC_{50}), we utilized the **SAIR** dataset [28], a large-scale synthetic structural dataset integrating data from ChEMBL [37] and BindingDB [38]. Our curated training set comprises 796,557 protein-ligand pairs, covering 28,821 proteins and 403,124 unique ligands.

Unlike the binary classification task where assays were merged to maximize chemical diversity, for the regression task, we strictly grouped entries by their unique assay identifier defined by the combination of `protein` (UniProt ID) [12], `source` (ChEMBL or BindingDB), and `description`. This ensures that all data points within a group originate from a single assay with a consistent experimental setup, as continuous affinity values are highly dependent on specific experimental contexts.

We applied a rigorous filtration pipeline focusing on both label reliability and structural quality:

Label Quality Filtering

Following the protocol of Boltz-2, we applied the following filters to each assay group:

- All affinity values were standardized to a common logarithmic scale relative to a 1 μM baseline.
- We discarded assays where the mean pairwise Tanimoto similarity (using ECFP4 Morgan fingerprints) [24] between compounds was below 0.25. Such assays lack a discernible chemical series, making them uninformative for learning structure-activity relationships.
- We removed assays with fewer than 10 data points, fewer than 10 unique affinity values, or a unique-to-total ratio below 0.2.
- We discarded assays where the standard deviation of internal affinity values was below 0.25 (log scale), as a narrow activity range provides insufficient signal for regression.
- Assays containing extreme affinity values (less than 10^{-6} μM) were discarded to prevent artifacts from unit inconsistencies.

Structural Quality Filtering

For each protein-ligand complex, SAIR provides candidate structures predicted by Boltz-1x [39]. We filtered individual data points based on structural fidelity:

- We discarded entries where the mean ipTM of the predicted structure was ≤ 0.5, indicating low confidence in the binding interface.
- We discarded entries where the PoseBusters [40] pass rate across the five predicted poses was ≤ 0.5 (i.e., fewer than three poses passed consistency checks).

The threshold of 0.5 was chosen as a deliberate trade-off between rigor and data retention, ensuring that the retained structures are physically plausible while maintaining a dataset scale sufficient for training.

Validation and Test Sets

To prevent data leakage, we removed any protein from the training set sharing $\geq 90\%$ sequence identity with proteins in the validation or test sets, and any ligand with a Tanimoto similarity > 0.4 to active ligands in the test set.

- **Validation**: Constructed by randomly holding out all data points from 20 diverse assays in the SAIR dataset.

Table S1: **Comprehensive performance comparison on affinity value prediction benchmarks.** Metrics are averaged per-assay. "non-cent." and "cent." denote metrics computed on raw and centered predictions, respectively. PW1/PW2 refer to the percentage of predictions within 1 and 2 kcal/mol of the experimental value. Baseline data sourced from Boltz-2.

Dataset	Method	Pearson R ↑	Kendall τ ↑	PMAE ↓	MAE ↓		PW1 ↑		PW2 ↑	
					non-cent.	cent.	non-cent.	cent.	non-cent.	cent.
OpenFE	Boltz-2	0.62	0.46	0.93	1.22	0.64	0.49	0.80	0.82	0.96
	BACPI	0.29	0.19	1.21	1.44	0.85	0.40	0.67	0.74	0.94
	GAT	0.28	0.20	1.30	1.42	0.91	0.40	0.64	0.75	0.92
	FlashBind	0.44	0.33	1.13	1.46	0.79	0.39	0.71	0.71	0.95
FEP+ (4 targets)	Boltz-2	0.66	0.48	0.85	0.75	0.59	0.69	0.83	0.97	0.98
	BACPI	0.14	0.09	1.18	1.40	0.82	0.43	0.62	0.73	1.00
	GAT	0.40	0.28	1.07	1.19	0.71	0.43	0.72	0.86	0.95
	FlashBind	0.53	0.38	1.10	1.44	0.76	0.39	0.71	0.74	0.95
CASP16	Boltz-2	0.65	0.45	1.36	1.28	0.95	0.48	0.61	0.81	0.90
	BACPI	0.41	0.31	1.55	1.25	1.10	0.45	0.51	0.81	0.89
	GAT	0.50	0.35	1.58	1.28	1.13	0.44	0.49	0.79	0.84
	FlashBind	0.65	0.51	1.14	1.21	0.88	0.29	0.68	0.94	0.94

- **Test**: We evaluated performance on three benchmarks identical to those used in Boltz-2: the **OpenFE subset** [27], a **4-target subset (CDK2, TYK2, JNK1, P38)** of the FEP+ benchmark [41], and the blind **CASP16** benchmark [26].

A.2 Benchmarking Results

We compared our method FlashBind against the state-of-the-art foundation model Boltz-2, as well as sequence-based (BACPI) [15] and ligand-only (GAT) baselines.

The comprehensive results are summarized in Table S1. FlashBind consistently and significantly outperforms the non-structural baselines across all datasets. For instance, on the FEP+ 4-target benchmark, our model achieves a Pearson's R of 0.53, a substantial improvement over the sequence-based BACPI (0.14). This performance gap underscores the effectiveness of our geometric encoder in leveraging 3D structural information for affinity prediction.

When compared to the computationally intensive Boltz-2 model, our lightweight approach achieves competitive performance, particularly in ranking metrics. This is most evident on the blind CASP16 benchmark, where FlashBind's rank correlation matches and slightly exceeds that of Boltz-2 (Kendall's τ of 0.51 vs. 0.45), highlighting its strong generalization capability in challenging, blind evaluation settings. While Boltz-2 generally yields lower absolute errors on the OpenFE and FEP+ benchmarks, FlashBind's performance remains comparable, validating its utility as an efficient alternative for rapid affinity estimation.

A.3 Ablation of Encoder Architecture

To further investigate the performance gap between FlashBind and Boltz-2 observed in the affinity prediction benchmarks, we conducted an additional ablation study focusing on the model's encoder capacity. A potential hypothesis for the performance difference is that the $E(3)$-equivariant Graph Neural Network (EGNN) [7] used in FlashBind might lack the representational power of the computationally heavier PairFormer architecture employed by Boltz-2.

To test this hypothesis, we developed a variant of our model, denoted as **FlashBind (PairFormer)**. In this variant, we replaced the EGNN encoder with a PairFormer module with hyperparameters similar to the Boltz-2 affinity head, while keeping the rest of the pipeline identical (i.e., using FABind+ generated structures as input and the same MLP prediction head).

The results are summarized in Table S2. Contrary to the expectation that a more complex architecture would yield significant gains, the PairFormer variant demonstrated negligible performance improvements compared to the standard EGNN-based FlashBind. For instance, on the OpenFE dataset, the Pearson's R only marginally fluctuated (from 0.44 to 0.43), and on the FEP+ 4 benchmark, the performance remained statistically comparable.

Consequently, this result validates our architectural choice: the EGNN provides a much more favorable trade-off, offering comparable accuracy to a Transformer-based architecture at a fraction of the computational and memory cost. Furthermore, this finding implies that the remaining performance gap between FlashBind and Boltz-2 on affinity regression tasks is unlikely to originate from architectural differences. A more plausible explanation lies in the

disparity of training data: Boltz-2 is trained on a proprietary, rigorously curated dataset that is approximately 1.5x larger than our training set derived from SAIR [28], and likely contains substantially fewer noisy labels. This data advantage, rather than the sophistication of the scoring network, is the more probable driver of Boltz-2's stronger affinity prediction performance.

Table S2: **Ablation study on encoder architecture.** Detailed performance comparison between the proposed EGNN-based scoring module and a computationally heavier PairFormer-based variant across three benchmarks. The results indicate that increasing the encoder complexity yields negligible performance gains.

Dataset	Method	Pearson R ↑	Kendall tau ↑	PMAE ↓	MAE ↓		PW1 ↑		PW2 ↑	
					non-cent.	cent.	non-cent.	cent.	non-cent.	cent.
OpenFE	FlashBind (EGNN)	0.44	0.33	1.13	1.46	0.79	0.39	0.71	0.71	0.95
	FlashBind (PairFormer)	0.43	0.32	1.11	0.84	0.76	0.32	0.61	0.74	0.95
FEP+ 4 targets	FlashBind (EGNN)	0.53	0.38	1.10	1.44	0.76	0.39	0.71	0.74	0.95
	FlashBind (PairFormer)	0.53	0.39	1.13	1.39	0.78	0.47	0.68	0.79	0.97
CASP16	FlashBind (EGNN)	0.65	0.51	1.14	1.21	0.88	0.29	0.68	0.94	0.94
	FlashBind (PairFormer)	0.71	0.53	1.19	1.28	0.89	0.68	0.71	0.94	0.97

Appendix B Details about the Methods

This section supplements the "Methods" section by providing specific hyperparameters, feature definitions, and algorithmic logic used for graph construction and model training.

B.1 Node Featurization

To construct the node representations for the EGNN, we generate initial features for proteins and ligands using distinct pipelines before projecting them into a unified space.

Protein Features

We utilize the **ESM-3** [17] language model (`esm3_sm_open_v1`) to capture sequence-based semantics. We extract the representation from the final hidden layer for each residue. To map these residue-level embeddings to the atomic graph, we broadcast the embedding of a residue to all its constituent atoms. These embeddings serve as the initial sequence features and are computationally efficient to generate, given the limited number of unique protein targets in screening tasks.

Ligand Features

We adopt a featurization pipeline same to **torchdrug** [18] using **RDKit** [34]. For each ligand atom, we extract a feature vector consisting of the following one-hot encoded chemical properties:

- Atom symbol (e.g., C, N, O, S, F, Cl, etc.).
- Atom degree (number of heavy-atom neighbors).
- Total number of attached hydrogen atoms.
- Implicit valence.
- Formal charge.
- Aromaticity (boolean flag).

This pipeline is highly optimized for CPU execution, processing over 800 ligands per second, which allows for on-the-fly feature generation during training.

Unified Representation

Protein and ligand features are projected into a common hidden dimension via separate linear layers and concatenated to form a base embedding. This base embedding is further augmented by concatenating: (1) One-hot encodings for atom type and residue type; (2) A binary indicator for molecule type (protein vs. ligand); (3) Positional encodings representing the residue's sequence index.

B.2 Pocket Cropping

We apply a deterministic, budget-constrained cropping algorithm to isolate the binding interface. The algorithm operates in a greedy manner to select a subset of protein residues $\mathcal{P}$ based on spatial proximity to the ligand, subject to an atom budget $B_a = 2048$ and a residue budget $B_r = 512$.

The procedure is as follows:

1. **Initialization**: Calculate the effective budgets for the protein by subtracting the number of ligand atoms: $B'_a = B_a - |\mathcal{A}_L|$ and $B'_r = B_r - |\mathcal{A}_L|$.
2. **Distance Calculation**: For every protein residue, calculate the minimum Euclidean distance between any of its atoms and ligand center.
3. **Sorting**: Sort all protein residues in ascending order of this distance.
4. **Greedy Selection**: Iterate through the sorted list. Add a residue to the pocket candidates $\mathcal{P}_{budget}$ if adding it does not exceed B'_a or B'_r.
5. **Distance Filtering**: Filter $\mathcal{P}_{budget}$ to retain only residues within a cutoff $d_{\max} = 20.0$ Å, forming the set $\mathcal{P}_{dist}$.
6. **Robustness Fallback**: If $|\mathcal{P}_{dist}| < k_{\min}$ (where $k_{\min} = 100$), ignore the distance cutoff and return the top $k_{\min}$ closest residues from $\mathcal{P}_{budget}$ to prevent creating overly sparse graphs.

B.3 Edge Construction

Inspired by MEAN [42], edges are constructed to capture interactions at multiple scales. All edges are directed. If the total number of edges exceeds the budget $B_e = 16384$, the edge list is truncated based on the priority scheme described below.

Edge Categories

- **Internal Edges ($\mathcal{E}_{\text{internal}}$)**:
  - *Ligand Covalent*: Single, double, triple, and aromatic bonds derived from the molecular graph.
  - *Protein Covalent*: Intra-residue bonds based on standard amino acid templates.
  - *Protein Sequential*: Connections between Cα atoms of adjacent residues (k to $k+1$).
- **External Edges ($\mathcal{E}_{\text{external}}$)**:
  - *Protein-Ligand Proximity*: Connected if Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\| < d_{\text{cross}}$ (10.0 Å).
- **Auxiliary Edges ($\mathcal{E}_{\text{aux}}$)**:
  - *Global*: Connections from global virtual nodes ($v_{\text{P}}, v_{\text{L}}$) to all respective atoms, and between v_{P} and v_{L}.
  - *Protein Spatial*: Connections between atoms of different residues if distance $< d_{\text{protein}}$ (4.0 Å).
  - *Ligand LAS (Local Atomic Structure)*: Virtual edges between ligand atoms within a 2-hop covalent distance or within the same ring.

Priority Scheme

To ensure critical interactions are preserved under the edge budget, edges are added in the following strict priority order (highest to lowest):

External (Protein-Ligand) > Global > Ligand Covalent > Protein Sequential > Protein Covalent > Ligand LAS > Protein Spatial.

B.4 Group-Based Sampling

To optimize training stability across diverse datasets, we employ a group-based mini-batch sampling strategy. Each training batch ($B = 20$) consists of multiple independent groups ($N = 5$ samples per group), where all samples in a group originate from the same experimental assay.

The sampling procedure is performed as follows:

1. **Assay Filtering**: Prior to training, we discard invalid assays. Binary assays must contain at least one binder and one decoy. Affinity assays must contain at least two ligands.
2. **Dataset Selection**: For each batch, a dataset is selected based on predefined probabilities.
3. **Group Population**:
   - *For Binary Datasets*: An assay is sampled uniformly at random. A group is formed by sampling 1 binder and 4 decoys (1:4 ratio) from that assay.

Table S3: Performance comparison on the MF-PCBA benchmark for binder classification. Our model demonstrates competitive performance, particularly in enrichment factors, compared to established baselines. All baseline data is taken from original Boltz-2 publication. AUROC is computed globally across all targets.

Method	AUROC ↑	EF at 0.5% ↑	EF at 1% ↑	EF at 2% ↑	EF at 5% ↑
Boltz-2	0.8056	18.3916	13.9540	10.5706	7.0448
BACPI	0.7205	9.4818	9.2397	7.3983	5.5533
GAT	0.7867	11.1279	8.9897	7.5630	5.9055
Chemgauss4	0.5706	1.9969	2.2257	2.1136	1.6462
Boltz-2 iptm	0.6134	2.4242	3.1728	2.6881	2.2263
FlashBind	0.7826	16.2832	14.1343	9.4587	6.7300

- *For Affinity Datasets*: Assays are sampled with probability proportional to the Interquartile Range (IQR) of their affinity values. A group is formed by uniformly sampling N ligands from the selected assay.

4. **Replacement**: Sampling is performed without replacement by default, automatically switching to replacement if the pool size is insufficient.

This strategy ensures that the model learns from consistent experimental contexts while balancing class distribution for screening tasks and prioritizing informative dynamic ranges for regression tasks.

B.5 Optimizer Strategy

The models were trained with different optimization strategies to maximize performance on their respective tasks.

For the virtual screening task, we employed a hybrid optimization strategy by mixing two different optimizers. Specifically, for the parameters within the EGNN hidden layers, we used the **Muon** optimizer [43]. This allowed us to apply a high learning rate of 2×10^{-3} and a weight decay of 0.01 to accelerate the convergence of the model's core representation learning component. All other model parameters (e.g., embedding layers, prediction head) were optimized using a standard **AdamW** optimizer [44] with more conservative hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 1×10^{-3}).

However, for the more sensitive affinity value regression task, this aggressive, hybrid strategy led to training instability. We therefore opted for a simpler and more stable approach, using only the standard **AdamW** optimizer for all model parameters.

Finally, for the enzyme-substrate interaction task, we strictly adhered to the hyperparameter settings of the baseline method, EZSpecificity, to ensure a rigorous and fair comparison. Consequently, we employed the standard **AdamW** optimizer with a learning rate of 3×10^{-4} for all model parameters.

Appendix C Detailed Results

This section provides the detailed numerical data corresponding to the figures presented in the main text, including specific Enrichment Factor (EF) values for virtual screening, precise inference speed measurements, and fine-grained performance breakdowns for enzyme families.

C.1 Virtual Screening Performance

We present the comprehensive metrics for the MF-PCBA benchmark (Fig. 3 in main text). Table S3 details the performance of FlashBind against primary baselines, and Table S4 provides the detailed results for the ablation study and comparison against various docking and rescoring strategies.

C.2 Computational Efficiency

To quantify the efficiency gains, we benchmarked the inference speed of various methods. Table S5 reports the average inference time per protein-ligand complex. Measurements were conducted on a single **NVIDIA L40S GPU**, averaged over 100 randomly selected samples to account for variance in protein size and graph complexity.

Table S4: **Ablation study and extended comparison on MF-PCBA subset.** This table details the performance of various architectural variants and traditional docking methods. "Boltz-2 (FABind+)" indicates Boltz-2 scoring using FABind+ poses. "FlashBind (Boltz trunk)" uses the Boltz-2 trunk for feature extraction. AUROC is computed globally across all targets.

Method	**AUROC** ↑	**EF at 0.5%** ↑	**EF at 1%** ↑	**EF at 2%** ↑	**EF at 5%** ↑
Boltz-2	0.7788	15.3087	10.4519	8.2599	7.1524
Boltz-2 (FABind+)	0.7852	16.1420	11.1186	7.7377	6.4675
Boltzina	0.7699	10.1308	10.0894	7.8458	6.9278
AutoDock Vina	0.5956	5.0000	3.2143	1.6071	2.8661
AutoDock Vina (RF-Score)	0.4899	0.0000	2.6800	2.2581	2.1325
AutoDock Vina (RF-Score-VS)	0.4668	0.0000	0.9271	0.4589	0.6463
AutoDock Vina (PLEC Score)	0.4817	0.0000	0.0000	0.0000	0.3568
AutoDock Vina (NNScore)	0.5062	0.0000	0.0000	0.3571	0.7389
GNINA	0.5592	2.7767	2.7543	4.0957	2.5896
FlashBind (Boltz trunk)	0.7553	16.0468	12.0458	7.5255	5.6767
FlashBind	0.7778	14.9905	12.8531	7.9154	6.7143

Table S5: Inference time comparison per protein-ligand pair on a single NVIDIA L40S GPU. Our full pipeline offers a 50x speedup over Boltz-2.

Method	Boltz-2	BACPI	GAT	Chemgauss4	AutoDock Vina	GNINA	Boltzina	FlashBind (Full Pipeline)	FlashBind (Scoring Only)
Avg. Time (s)	35	0.00065	0.00028	25	4.5	5	6.5	0.7	0.025

Table S6: **AUROC performance on specific enzyme families (ESIBank).** FlashBind consistently matches or outperforms baselines across diverse enzyme categories, including those with sparse data (e.g., DUF).

Method	**Enzyme Family (AUROC)**				
	DUF	**Glycosyltransferase**	**Thiolase**	**Phosphatase**	**Esterase**
Boltz-2 (Binary Prob.)	0.5468	0.7440	0.5307	0.6219	0.5755
Boltz-2 (Affinity Val.)	0.4409	0.6042	0.5545	0.4882	0.4101
ESP	0.4594	0.5878	0.6439	0.5921	0.5091
EZSpecificity	0.7067	**0.7972**	0.5970	0.6741	**0.8035**
FlashBind	**0.7216**	0.7688	**0.7067**	**0.7010**	0.7869

C.3 Enzyme Specificity Results

Table S6 provides the numerical breakdown of the enzyme-substrate specificity performance (AUROC) across five distinct enzyme families, corresponding to the radar plot in Fig. 4(b) of the main text. FlashBind demonstrates robust generalization, particularly in the DUF and Esterase families, outperforming the foundation model Boltz-2.

References

[1] Reymond, J.-L.: The chemical space project. Accounts of Chemical Research **48**(3), 722–730 (2015) https://doi.org/10.1021/ar500432k https://doi.org/10.1021/ar500432k. PMID: 25687211 (Cited on page 1)

[2] Trott, O., Olson, A.J.: Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry **31** (2009) (Cited on pages 1, 4 and 8)

[3] McNutt, A.T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., Koes, D.R.: Gnina 1.0: molecular docking with deep learning. Journal of cheminformatics **13**(1), 43 (2021) (Cited on pages 1 and 5)

[4] Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T., Banks, J.L.: Glide: A new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. Journal of Medicinal Chemistry **47**(7), 1750–1759 (2004) https://doi.org/10.1021/jm030644s https://doi.org/10.1021/jm030644s. PMID: 15027866 (Cited on page 1)

[5] Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V.R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., Barzilay, R.: Boltz-2: Towards accurate and efficient binding affinity prediction. bioRxiv (2025) https://doi.org/10.1101/2025.06.14.659707 https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707.full.pdf (Cited on pages 1, 4, 10 and 11)

[6] Gao, K., Pei, Q., Zhang, G., Zhu, J., He, K., Wu, L.: Fabind+: Enhancing molecular docking through improved pocket prediction and pose generation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1. KDD '25, pp. 330–341. ACM, ??? (2025). https://doi.org/10.1145/3690624.3709253 . http://dx.doi.org/10.1145/3690624.3709253 (Cited on pages 1, 4, 9 and 11)

[7] Satorras, V.G., Hoogeboom, E., Welling, M.: E(n) Equivariant Graph Neural Networks (2022). https://arxiv.org/abs/2102.09844 (Cited on pages 2, 4, 10, 11 and 15)

[8] Buterez, D., Janet, J.P., Kiddle, S.J., Liò, P.: Mf-pcba: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. Journal of Chemical Information and Modeling **63**(9), 2667–2678 (2023) https://doi.org/10.1021/acs.jcim.2c01569 https://doi.org/10.1021/acs.jcim.2c01569. PMID: 37058588 (Cited on pages 2 and 4)

[9] Cui, H., Su, Y., Dean, T.J., Yu, T., Zhang, Z., Peng, J., Shukla, D., Zhao, H.: Enzyme specificity prediction using cross-attention graph neural networks. Nature **647**(8090), 639–647 (2025) https://doi.org/10.1038/s41586-025-09697-2 (Cited on pages 2, 6 and 10)

[10] Wong, F., Krishnan, A., Zheng, E.J., Stärk, H., Manson, A.L., Earl, A.M., Jaakkola, T., Collins, J.J.: Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. Molecular Systems Biology **18**(9), 202211081 (2022) https://doi.org/10.15252/msb.202211081 (Cited on pages 2 and 8)

[11] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: Pubchem 2023 update. Nucleic Acids Research **51**(D1), 1373–1380 (2022) https://doi.org/10.1093/nar/gkac956 https://academic.oup.com/nar/article-pdf/51/D1/D1373/48441598/gkac956.pdf (Cited on pages 4 and 10)

[12] Consortium, T.U.: Uniprot: the universal protein knowledgebase in 2025. Nucleic Acids Research **53**(D1), 609–617 (2024) https://doi.org/10.1093/nar/gkae1010 https://academic.oup.com/nar/article-pdf/53/D1/D609/60719276/gkae1010.pdf (Cited on pages 4 and 14)

[13] Baell, J.B., Holloway, G.A.: New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. Journal of medicinal chemistry **53 7**, 2719–40 (2010) (Cited on page 4)

[14] Nishimoto, Y., Fedorov, D.G.: The fragment molecular orbital method combined with density-functional tight-binding and the polarizable continuum model. Phys. Chem. Chem. Phys. **18**, 22047–22061 (2016) https:

//doi.org/10.1039/C6CP02186G (Cited on page 4)

[15] Li, M., Lu, Z., Wu, Y., Li, Y.: Bacpi: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction. Bioinformatics **38** (2022) https://doi.org/10.1093/bioinformatics/btac035 (Cited on pages 4 and 15)

[16] Furui, K., Ohue, M.: Boltzina: Efficient and Accurate Virtual Screening via Docking-Guided Binding Prediction with Boltz-2 (2025). https://arxiv.org/abs/2508.17555 (Cited on page 4)

[17] Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R.S., Thomas, N., Khan, Y.A., Mishra, C., Kim, C., Bartie, L.J., Nemeth, M., Hsu, P.D., Sercu, T., Candido, S., Rives, A.: Simulating 500 million years of evolution with a language model. Science **387**(6736), 850–858 (2025) https://doi.org/10.1126/science.ads0018 https://www.science.org/doi/pdf/10.1126/science.ads0018 (Cited on pages 5, 11 and 16)

[18] Zhu, Z., Shi, C., Zhang, Z., Liu, S., Xu, M., Yuan, X., Zhang, Y., Chen, J., Cai, H., Lu, J., Ma, C., Liu, R., Xhonneux, L.-P., Qu, M., Tang, J.: TorchDrug: A Powerful and Flexible Machine Learning Platform for Drug Discovery (2022). https://arxiv.org/abs/2202.08320 (Cited on pages 5, 11 and 16)

[19] Ballester, P.J., Mitchell, J.B.: A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics **26**(9), 1169–1175 (2010) (Cited on page 5)

[20] Wójcikowski, M., Ballester, P.J., Siedlecki, P.: Performance of machine-learning scoring functions in structure-based virtual screening. Scientific Reports **7**(1), 46710 (2017)

[21] Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M.M., Siedlecki, P.: Development of a protein–ligand extended connectivity (plec) fingerprint and its application for binding affinity predictions. Bioinformatics **35**(8), 1334–1341 (2019)

[22] Durrant, J.D., McCammon, J.A.: Nnscore: a neural-network-based scoring function for the characterization of protein- ligand complexes. Journal of chemical information and modeling **50**(10), 1865–1871 (2010) (Cited on page 5)

[23] Ji, X., Wang, Z., Gao, Z., Zheng, H., Zhang, L., Ke, G., et al.: Uni-mol2: Exploring molecular pretraining model at scale. arXiv preprint arXiv:2406.14969 (2024) (Cited on pages 6 and 11)

[24] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling **50**(5), 742–754 (2010) (Cited on pages 6, 11 and 14)

[25] Kroll, A., Ranjan, S., Engqvist, M.K., Lercher, M.J.: A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nature communications **14**(1), 2787 (2023) (Cited on page 6)

[26] Gilson, M.K., Eberhardt, J., Škrinjar, P., Durairaj, J., Robin, X., Kryshtafovych, A.: Assessment of pharmaceutical protein-ligand pose and affinity predictions in casp16 (2025) https://doi.org/10.22541/au.174562565.51283311/v1 (Cited on pages 10 and 15)

[27] Gowers, R.J., Alibay, I., Swenson, D.W.H., Henry, M.M., Ries, B., Baumann, H.M., Eastwood, J.R.B.: The Open Free Energy Library. https://doi.org/10.5281/zenodo.8344248 . https://doi.org/10.5281/zenodo.8344248 (Cited on pages 10 and 15)

[28] Lemos, P., Beckwith, Z., Bandi, S., Damme, M., Crivelli-Decker, J., Shields, B.J., Merth, T., Jha, P.K., De Mitri, N., Callahan, T.J., Nish, A., Abruzzo, P., Salomon-Ferrer, R., Ganahl, M.: Sair: Enabling deep learning for protein-ligand interactions with a synthetic structural dataset. bioRxiv (2025) https://doi.org/10.1101/2025.06.17.660168 https://www.biorxiv.org/content/early/2025/06/21/2025.06.17.660168.full.pdf (Cited on pages 10, 14 and 16)

[29] Tran-Nguyen, V.-K., Jacquemard, C., Rognan, D.: Lit-pcba: An unbiased dataset for machine learning and virtual screening. Journal of Chemical Information and Modeling (2020) https://doi.org/10.1021/acs.jcim.0c00155 (Cited on page 10)

[30] Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. **35**(11), 1026–1028 (2017) (Cited on page 11)

[31] Butina, D.: Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. Journal of Chemical Information and Computer Sciences **39**(4), 747–750 (1999) (Cited on page 11)

[32] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Res. **28**(1), 235–242 (2000) (Cited on page 11)

[33] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S.: Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research **50**(D1), 439–444 (2021) https://doi.org/10.1093/nar/gkab1061 https://academic.oup.com/nar/article-pdf/50/D1/D439/43502749/gkab1061.pdf (Cited on page 11)

[34] Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., sriniker, Gedeck, P., Jones, G., Kawashima, E., NadineSchneider, Nealschneider, D., Dalke, A., tadhurst-cdd, Swain, M., Cole, B., Turk, S., Savelev, A., Vaucher, A., Wójcikowski, M., Faara, H., Take, I., Walker, R., Scalfani, V.F., Maeder, N., Probst, D., Ujihara, K., Pahl, A., godin, Lehtivarjo, J.: Rdkit/rdkit: 2025_09_1 (Q3 2025) Beta Release. https://doi.org/10.5281/zenodo.17193272 . https://doi.org/10.5281/zenodo.17193272 (Cited on pages 11 and 16)

[35] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection (2018). https://arxiv.org/abs/1708.02002 (Cited on page 11)

[36] Huber, P.J.: In: Kotz, S., Johnson, N.L. (eds.) Robust Estimation of a Location Parameter, pp. 492–518. Springer, New York, NY (1992). https://doi.org/10.1007/978-1-4612-4380-9_35 . https://doi.org/10.1007/978-1-4612-4380-9$_3$5 (Cited on page 12)

[37] Zdrazil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D.M., Mosquera, J., Magarinos, M., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A., Adasme, M., Monecke, P., Landrum, G., Leach, A.: The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research **52**(D1), 1180–1192 (2023) https://doi.org/10.1093/nar/gkad1004 (Cited on page 14)

[38] Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K.: Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Research **35**, 198–201 (2006) https://doi.org/10.1093/nar/gkl999 (Cited on page 14)

[39] Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., Barzilay, R.: Boltz-1 democratizing biomolecular interaction modeling. bioRxiv (2024) https://doi.org/10.1101/2024.11.19.624167 https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167.full.pdf (Cited on page 14)

[40] Buttenschoen, M., Morris, G.M., Deane, C.M.: Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chemical Science **15**, 3130–3139 (2023) (Cited on page 14)

[41] Hahn, D.F., Bayly, C.I., Boby, M.L., Bruce Macdonald, H.E., Chodera, J.D., Gapsys, V., Mey, A.S.J.S., Mobley, D.L., Benito, L.P., Schindler, C.E.M., Tresadern, G., Warren, G.L.: Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks [article v1.0]. Living Journal of Computational Molecular Science **4**(1) (2022) https://doi.org/10.33011/livecoms.4.1.1497 (Cited on page 15)

[42] Kong, X., Huang, W., Liu, Y.: Conditional Antibody Design as 3D Equivariant Graph Translation (2023). https://arxiv.org/abs/2208.06073 (Cited on page 17)

[43] Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for

hidden layers in neural networks (2024). https://kellerjordan.github.io/posts/muon/ (Cited on page 18)

[44] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2017). https://arxiv.org/abs/1412.6980 (Cited on page 18)