FlashBind: Towards Accurate and Efficient Structure-based Virtual Screening

Anonymous authors
Paper under double-blind review

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.

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 (Reymond, 2015). 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 (Trott & Olson, 2009), GNINA (McNutt et al., 2021), and Glide (Halgren et al., 2004), 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 (Passaro et al., 2025) 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

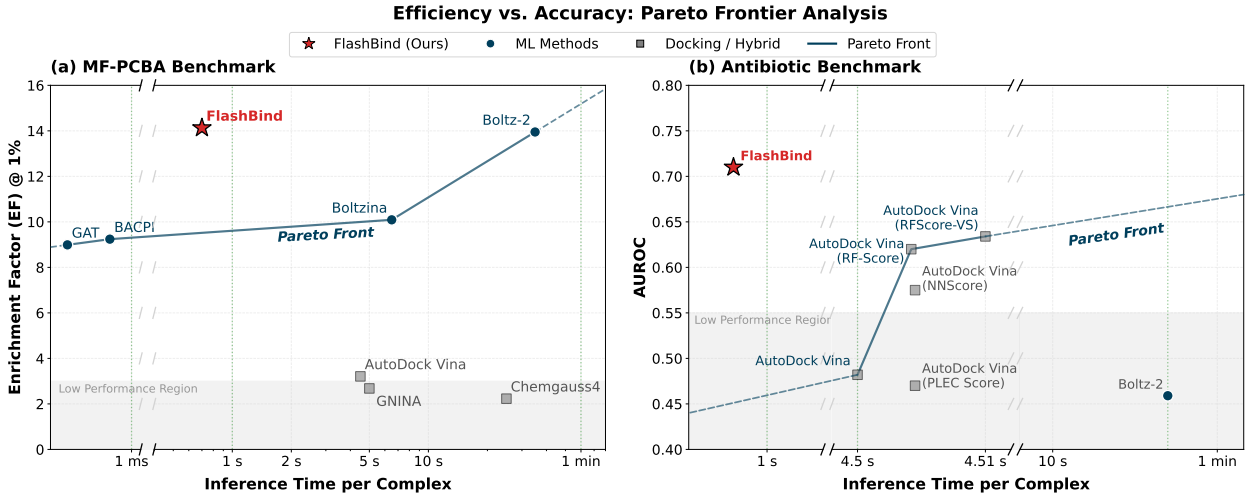

Figure 1: **Computational efficiency vs. screening accuracy.** FlashBind (star) occupies the optimal efficiency-accuracy trade-off, matching Boltz-2 with 50× faster inference (0.7s vs. 35s). The inference time is measured on a single NVIDIA L40S GPU.

folding, our framework utilizes a fast docking model (FABind+) (Gao et al., 2025) to provide a physically plausible structural prior. This structure is then processed by a highly efficient $E(3)$-equivariant Graph Neural Network (EGNN) (Satorras et al., 2022), which replaces heavy attention mechanisms to capture local physical interactions within the binding pocket.

This simple design proves remarkably effective: across virtual screening, enzyme-substrate specificity, and a prospective antibiotic discovery campaign, FlashBind matches or exceeds far heavier foundation models at a fraction of the computational cost. We summarize our main contributions as follows:

- **A lightweight, ultra-fast screening framework.** We replace expensive diffusion-based structure generation and PairFormer-based scoring with a fast docking model (FABind+) (Gao et al., 2025) followed by an $E(3)$-equivariant graph neural network (Satorras et al., 2022), yielding a **50-fold inference speedup** over Boltz-2 (∼0.7 s vs. ∼35 s per complex on a single NVIDIA L40S GPU).

- **State-of-the-art enrichment at a fraction of the cost.** On the MF-PCBA benchmark (Buterez et al., 2023), FlashBind matches the foundation model Boltz-2 in early enrichment (EF@1% of 14.13 vs. 13.95) while substantially outperforming physics-based and sequence-based baselines. Ablation studies confirm that neither diffusion sampling nor a heavy PairFormer trunk is necessary to attain this accuracy.

- **Generalization to enzyme-substrate specificity.** Under the challenging "unknown enzyme & substrate" setting of the ESIBank benchmark (Cui et al., 2025), FlashBind reaches an AUROC of 0.7229, on par with the specialized EZSpecificity (0.7198) and well above the sequence-based ESP (Kroll et al., 2023), while surpassing Boltz-2 on data-scarce enzyme families.

- **Prospective experimental validation.** On a structure-based antibiotic benchmark against essential *E. coli* proteins (Wong et al., 2022), FlashBind attains an AUROC of 0.71 where Boltz-2 is near-random (0.46). In a prospective campaign against *E. coli* DNA primase (DnaG), we screened 9,289 compounds, experimentally assayed 136, and confirmed 10 active inhibitors (a 7.4% hit rate), 4 of which further exhibited whole-cell antibacterial activity.

Together, these results establish FlashBind as a scalable and accurate framework for ultra-fast virtual screening of massive chemical libraries.

2 Related Work

Structure-based Virtual Screening. Virtual screening seeks to identify, from chemical libraries now reaching the billion-compound scale, the small subset of molecules likely to bind a given target. The classical workhorse is molecular docking (Trott & Olson, 2009; Halgren et al., 2004), which predicts a ligand's binding pose and energy within a pocket. Docking is, however, computationally intensive, making exhaustive screening of modern libraries impractical. Supervised learning offers a faster alternative: discriminative models trained to classify protein-ligand pairs as binding or non-binding trade physical interpretability for throughput, but often generalize poorly to unseen targets. A more recent paradigm reframes screening as dense retrieval (Radford et al., 2021): methods such as DrugCLIP (Gao et al., 2023) contrastively align separate protein and ligand encoders into a shared embedding space, enabling library embeddings to be precomputed and screening to reduce to fast similarity search. Geometric deep learning instead operates directly on 3D complex structures for higher fidelity, typically by representing the complex as an atomic graph and enforcing geometric symmetries (Satorras et al., 2022); this is the regime in which FlashBind operates, with the key distinction that we decouple a fast docking prior from a lightweight equivariant scorer to retain structural accuracy without per-compound conformational sampling.

Enzyme-Substrate Specificity. Predicting which substrates an enzyme acts upon is a functional task governed by catalytic alignment rather than thermodynamic stability alone, and it has historically been treated separately from binding prediction. Sequence-based approaches such as ESP (Kroll et al., 2023) pair enzyme representations with molecular fingerprints to classify enzyme-substrate pairs, but discard the 3D geometry of the active site. More recent structure-aware methods exploit predicted complex geometry: EZSpecificity (Cui et al., 2025) introduces a task-specific cross-attention architecture together with the ESIBank benchmark, establishing the current state of the art under stringent unknown-enzyme-and-substrate splits. Our work shows that a general-purpose equivariant encoder matches this specialized model, suggesting that the geometric features useful for binding transfer to catalytic specificity.

Machine Learning for Antibiotic Discovery. Machine learning has emerged as a powerful tool against the slowing pace of antibiotic discovery. Early efforts were predominantly ligand-based: deep classifiers trained on phenotypic growth-inhibition data prioritize compounds by predicted antibacterial activity directly from molecular structure (Stokes et al., 2020). Such phenotype-driven models are agnostic to mechanism and offer little insight into the molecular target. Structure-based formulations address this by scoring compounds against specific essential bacterial proteins; the benchmark of Wong et al. (Wong et al., 2022) couples docking against a panel of essential *E. coli* targets with experimental inhibition assays, providing a realistic and mechanistically grounded testbed. We adopt this structure-based setting and further close the loop with a fully prospective wet-lab campaign.

Bridging the Accuracy-Speed Gap. A central obstacle in structure-based screening is the trade-off between accuracy and throughput. Boltz-2 (Passaro et al., 2025) marked a turning point, reportedly approaching the accuracy of free-energy perturbation for protein-ligand affinity while being orders of magnitude cheaper, making high-fidelity prediction feasible in discovery settings. Even so, its per-complex inference cost remains a bottleneck for routine large-library screening, motivating efforts to accelerate the architecture. Derivative methods carry their own constraints: Boltzina (Furui & Ohue, 2025), for instance, requires predefined pocket information and retains the original Boltz trunk, leaving substantial headroom on speed. FlashBind takes a more aggressive stance, replacing both the diffusion-based structure generator and the heavy trunk with a docking oracle and an equivariant graph network, which removes the trunk bottleneck entirely while preserving early-enrichment accuracy.

3 Method

3.1 Problem Formulation

We address structure-based virtual screening: identifying the small subset of compounds in a large chemical library that bind a given protein target. This hit-discovery setting prioritizes ranking active compounds

(binders) above inactive decoys over precisely quantifying binding strength, and we therefore formulate it as a **binary classification** task. Formally, given a protein amino acid sequence $\mathcal{S}_p$ and a ligand SMILES string $\mathcal{S}_l$, FlashBind learns a mapping $\mathcal{F} : (\mathcal{S}_p, \mathcal{S}_l) \to p_{\text{bind}} \in [0, 1]$, where p_{bind} is the predicted probability that the ligand binds the target. As described below, $\mathcal{F}$ is realized by mapping the raw inputs to a 3D complex $(\mathbf{X}, \mathbf{H})$ via fast docking, cropping it to the binding interface, encoding the result as a geometric graph $\mathcal{G}$, and scoring $\mathcal{G}$ with an $E(3)$-equivariant network. The same encoder is task-agnostic: replacing the classification head with a regression head yields continuous affinity prediction (y_{affinity}), which we treat as an auxiliary task and detail in Appendix A.

3.2 The FlashBind Architecture

FlashBind resolves the accuracy-efficiency trade-off by structurally decoupling screening into two streamlined stages: rapid structure generation followed by geometric scoring (Fig. 2b). This avoids both diffusion-based conformational sampling and end-to-end folding, the two dominant bottlenecks in foundation-model pipelines.

Structure generation. For inputs given only as a sequence and a SMILES string, we obtain a protein structure through a hierarchical retrieval cascade that prioritizes experimental fidelity: we first query the PDB (Berman et al., 2000) for an experimental structure matching the target at 100% identity, otherwise select the highest-pLDDT model from the AlphaFold Database (Varadi et al., 2021), and only as a last resort predict the structure de novo with Boltz-2x (Passaro et al., 2025). In practice, over 60% of targets are resolved directly from the PDB or AlphaFold, sharply reducing prediction overhead. The ligand is then docked into this structure with FABind+ (Gao et al., 2025), a regression-based docking model chosen for its speed (< 0.7 s per complex), producing a 3D complex $(\mathbf{X}, \mathbf{H})$ with atomic coordinates $\mathbf{X} \in \mathbb{R}^{N\times 3}$ and features $\mathbf{H} \in \mathbb{R}^{N\times d}$. Crucially, this step supplies a physically plausible structural prior at a fraction of the cost of diffusion-based generation.

Graph construction. To focus the encoder on the binding interface, an adaptive cropping function $\mathcal{F}_{\text{crop}}$ isolates the local pocket: residues are added greedily by proximity until an atom budget ($B_a = 2048$) or residue budget ($B_r = 512$) is reached, after which residues beyond 20Å are removed. The cropped complex is converted into a multi-relational geometric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{h}, \mathbf{x})$, whose nodes $\mathcal{V}$ are protein and ligand atoms. Node features $\mathbf{h}$ concatenate pre-trained ESM-3 (Hayes et al., 2025) protein embeddings with TorchDrug (Zhu et al., 2022) ligand descriptors computed via RDKit (Landrum et al., 2025). Following MEAN (Kong et al., 2023), the edge set $\mathcal{E}$ captures interactions at multiple scales: **internal edges** for intra-molecular covalent topology, **external edges** for non-covalent protein-ligand contacts (< 10Å), and **auxiliary edges** that inject global structural context through global nodes. Full node featurization, the budget-constrained cropping algorithm, and the complete edge taxonomy with its priority scheme are detailed in Appendix B.1, B.2, and B.3, respectively.

Equivariant scoring network. The graph $\mathcal{G}$ is processed by an $E(3)$-equivariant graph neural network (EGNN) (Satorras et al., 2022) of $L = 5$ layers with hidden dimension 192. Enforcing $E(3)$-equivariance makes the prediction invariant to the arbitrary rotation and translation of the docked pose, removing any need for orientation augmentation. The network produces a pooled graph representation $z_{\mathcal{G}}$, which a task-specific MLP head maps to the binding probability p_{bind} (and, for the auxiliary regression task, to y_{affinity}).

3.3 Data Curation

A reliable binding signal can only be learned from high-quality supervision, yet high-throughput screening (HTS) data are notoriously noisy. We therefore curate the training data with a multi-stage filtration pipeline (Fig. 2a).

Virtual screening dataset. We build our screening corpus from PubChem BioAssays (Kim et al., 2022), explicitly counteracting the high false-positive rates endemic to HTS. The pipeline begins with an *assay-level* filter that retains only confirmatory and primary screens containing more than 100 compounds with hit rates below 10%; assays targeting the same protein (matched by UniProt ID (Consortium, 2024)) are merged to

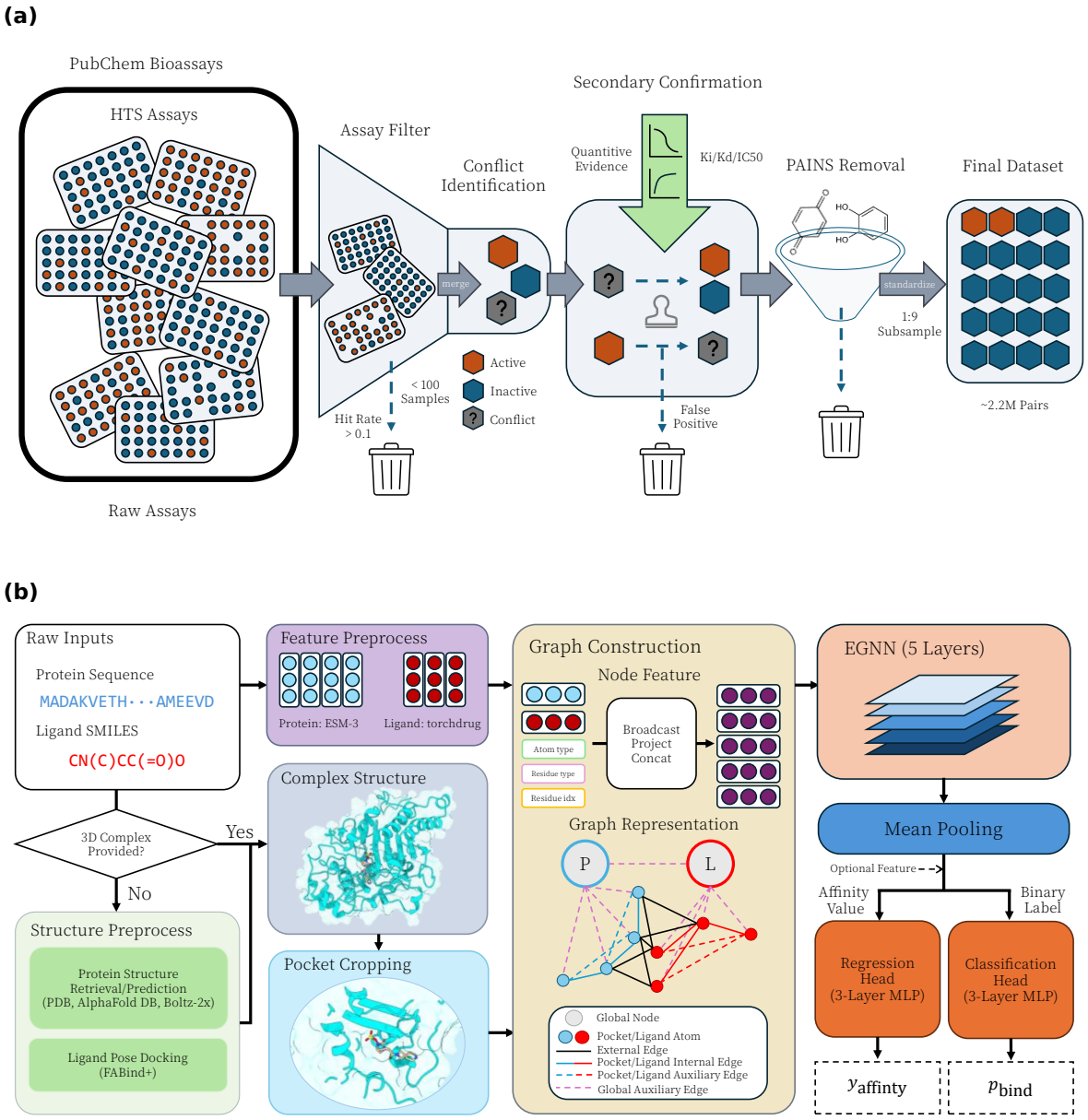

Figure 2: **The FlashBind framework. (a)** Training data curation. To mitigate experimental noise, we construct a multi-stage filtration pipeline with assay-consistency checks, secondary confirmation, and PAINS removal. **(b)** Inference pipeline. FlashBind decouples structure generation from scoring: raw inputs are mapped to a 3D complex with FABind+, cropped to the binding interface, and scored by an EGNN to predict the binding probability.

maximize chemical diversity. We then apply a *compound-level* secondary-confirmation step: an active label is kept only if corroborated by quantitative evidence (K_d, K_i, or IC_{50}), and conflicting entries lacking such confirmation are discarded. Finally, Pan-Assay Interference Compounds (PAINS) (Baell & Holloway, 2010) are removed to eliminate frequent hitters. After balancing to a 1:9 binder-to-decoy ratio, the final training set contains **2,237,058 protein-ligand pairs** spanning **451 protein targets** and **368,812 ligands**. For

validation we use a curated subset of LIT-PCBA (Tran-Nguyen et al., 2020) comprising 43,492 pairs sampled across its 15 targets such that every target retains a hit rate above 0.5%.

Enzyme-substrate specificity dataset. For the fine-grained enzyme-substrate task, we use the ESIBank benchmark (Cui et al., 2025) directly. It comprises **323,783 pairs** covering **8,124 enzymes** and **34,417 ligands**. To ensure a fair comparison, we adhere to the standard "unknown enzyme & substrate" splits defined within the benchmark, matching the protocol of EZSpecificity (Cui et al., 2025).

Leakage prevention. To rigorously assess generalization, we enforce strict train/test separation at two levels. At the *protein* level, we cluster all sequences with MMseqs2 (Steinegger & Söding, 2017) and remove any training protein sharing $\geq 90\%$ sequence identity with a validation or test protein. At the *ligand* level, we discard any training ligand whose Tanimoto similarity (Butina, 1999) to an *active* test ligand exceeds 0.4, preventing the model from exploiting memorized scaffolds.

3.4 Training and Inference

All models are trained on a cluster of NVIDIA L40S GPUs using a group-based mini-batch sampling strategy (Appendix B.4), in which every batch is drawn from a single experimental assay so that the loss compares compounds tested under identical conditions.

Virtual screening. To emphasize early enrichment, we optimize a Focal Loss (Lin et al., 2018) ($\gamma = 1$, $\alpha = 0.7$) under a per-assay sampling scheme that enforces a 1:4 binder-to-decoy ratio within each batch.

Enzyme-substrate specificity. We follow the EZSpecificity protocol exactly, performing 4-fold cross-validation on the "unknown enzyme & substrate" split with the same Focal Loss. Because catalytic specificity hinges on subtle functional-group chemistry that pure geometry can miss, we augment the geometric representation with UniMol embeddings (Ji et al., 2024) and Morgan fingerprints (Rogers & Hahn, 2010) for this task only, matching the input modalities of the baseline.

Inference and ensembling. At inference we ensemble the top $k = 2$ checkpoints selected on validation performance, averaging their outputs for classification. Protein embeddings are computed once per target and amortized across all candidate ligands, and ligand descriptors are generated by lightweight RDKit featurization, so the per-complex cost is dominated by docking and EGNN scoring (analyzed in Section 4.1).

4 Experiments

We evaluate FlashBind across three increasingly demanding settings that probe complementary aspects of structure-based prediction. We first establish its core competency in ultra-fast virtual screening on the MF-PCBA benchmark (Buterez et al., 2023), where it matches foundation models in early enrichment at a 50-fold lower inference cost, and ablate each design choice to isolate the sources of this efficiency (Section 4.1). We then test generalization beyond thermodynamic binding to enzyme-substrate specificity, a fine-grained functional task governed by catalytic alignment (Section 4.2). Finally, we assess real-world utility in antibiotic discovery: FlashBind substantially outperforms both physics-based docking and foundation models on a retrospective *E. coli* benchmark (Section 4.3), and in a fully prospective campaign against DNA primase (DnaG) we experimentally confirm 10 active inhibitors among 136 tested compounds, 4 of which show whole-cell antibacterial activity (Section 4.4).

4.1 Ultra-fast Virtual Screening

Enrichment performance. To assess the model's capability in identifying active compounds from vast chemical spaces, we benchmarked FlashBind on the MF-PCBA dataset (Buterez et al., 2023), 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.

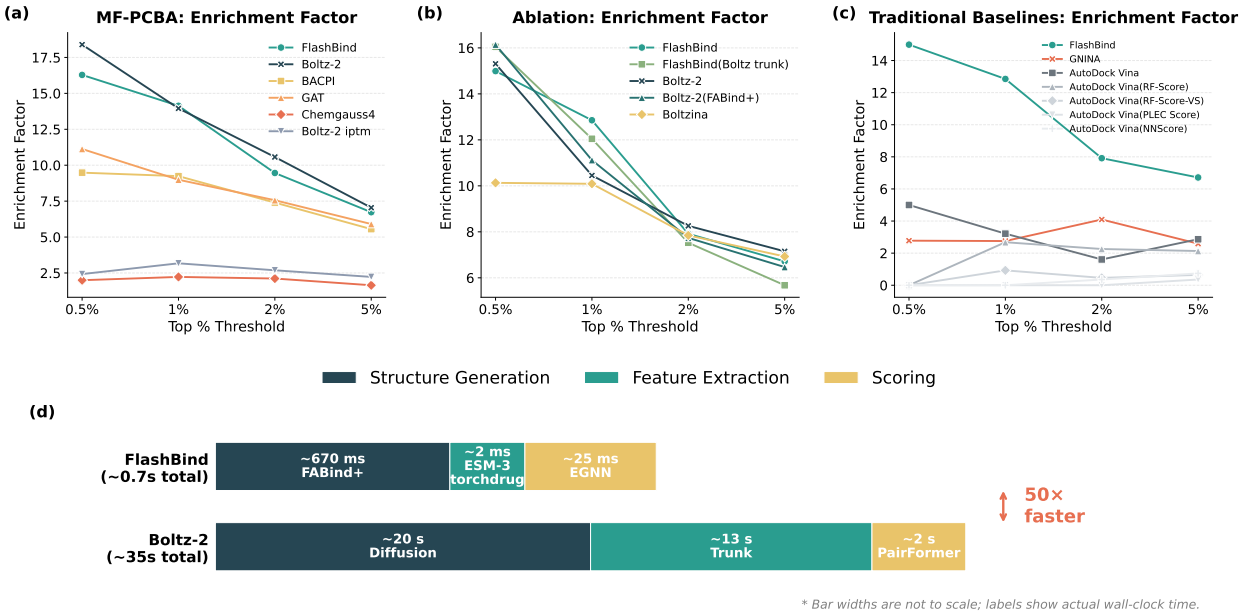

Figure 3: **Enrichment Factor (EF) on the MF-PCBA benchmark and per-complex inference time breakdown. (a)** Comparison with deep learning and physics-based scoring baselines on the full benchmark. FlashBind rivals the foundation model Boltz-2 while substantially outperforming traditional methods. **(b)** 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.

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 (Nishimoto & Fedorov, 2016) (EF@1% = 2.23) and sequence-based deep learning baselines (Li et al., 2022). Most notably, FlashBind rivals the performance of the computationally intensive foundation model, Boltz-2 (Passaro et al., 2025), 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.

Ablation study. 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 (Furui & Ohue, 2025), which substitutes the diffusion module with AutoDock Vina (Trott & Olson, 2009) 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.

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 (Hayes et al., 2025) and torchdrug (Zhu et al., 2022) 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.

Comparison with traditional pipelines. 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 (Ballester & Mitchell, 2010; Wójcikowski et al., 2017; 2019; Durrant & McCammon, 2010; McNutt et al., 2021)).

Inference efficiency. 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 5). 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 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.

4.2 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 (Cui et al., 2025), 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 (Cui et al., 2025), 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 (Ji et al., 2024) and Morgan fingerprints (Rogers & Hahn, 2010)).

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 (Kroll et al., 2023) (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.

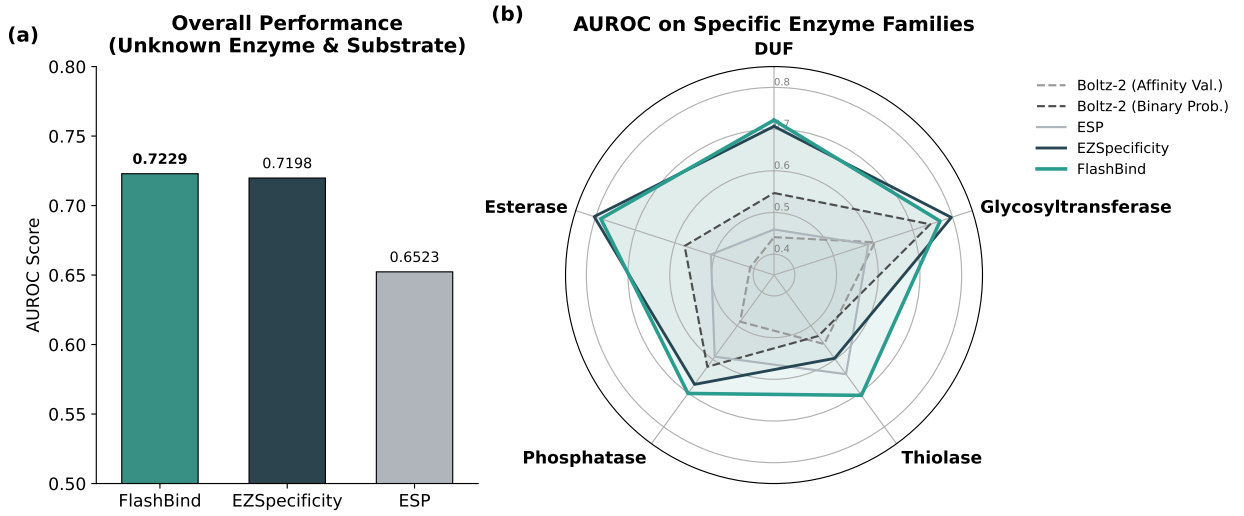

Figure 4: **Evaluation of enzyme-substrate specificity prediction on the ESIBank benchmark. (a)** Overall performance comparison on the "unknown enzyme & substrate" split (4-fold cross-validation). FlashBind achieves an AUROC of 0.7229, performing comparably to the specialized state-of-the-art model 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.

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.

4.3 Antibiotic Discovery Benchmark

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 (Wong et al., 2022). 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 (Wong et al., 2022) 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% 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 (Trott & Olson, 2009), 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

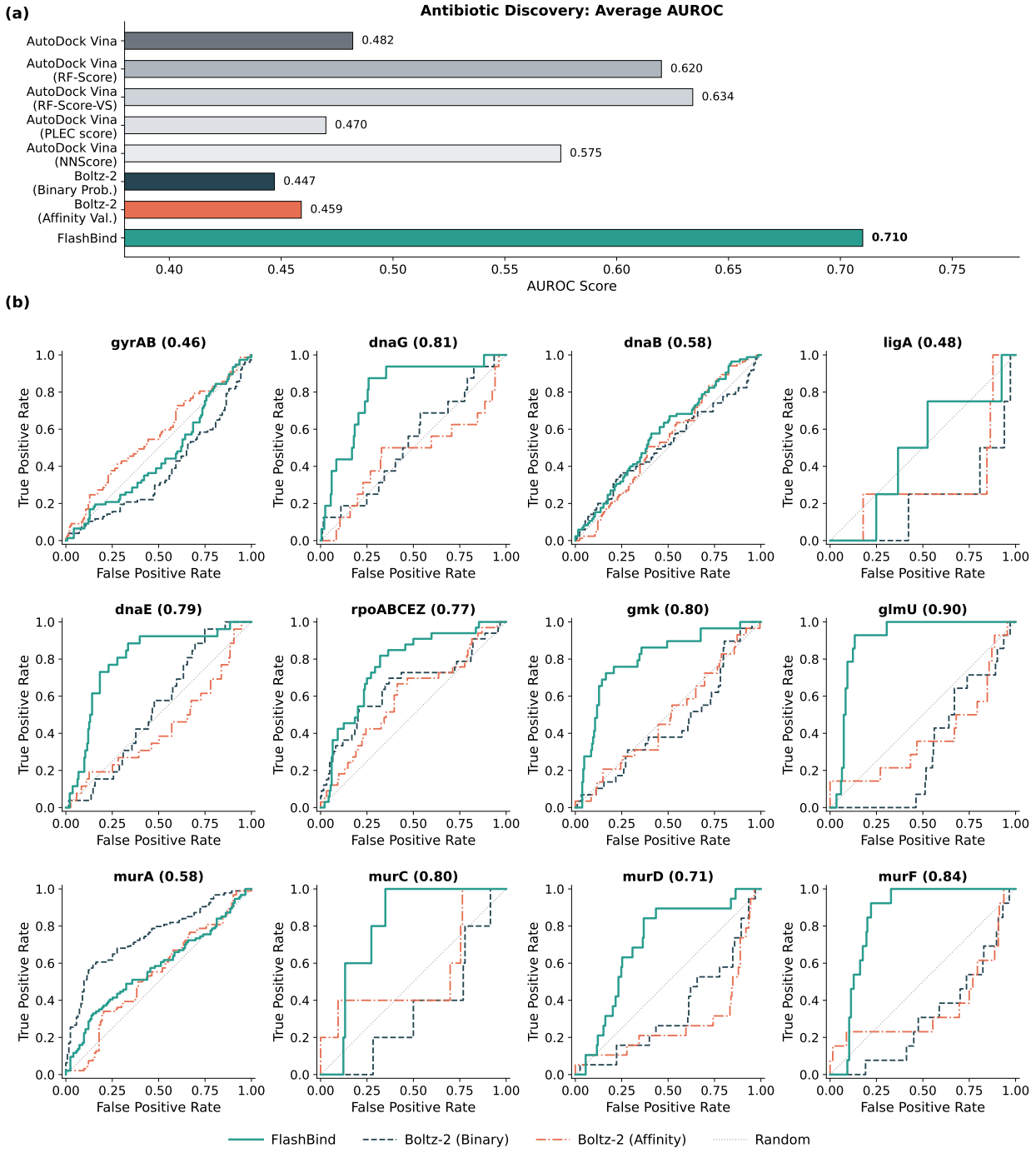

Figure 5: **Benchmarking performance on the antibiotic discovery task. (a)** Comparison of average AUROC across 12 essential *E. coli* targets. FlashBind (0.710) demonstrates practical ranking capability, significantly outperforming standard docking (AutoDock Vina) and machine-learning rescoring functions. **(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.

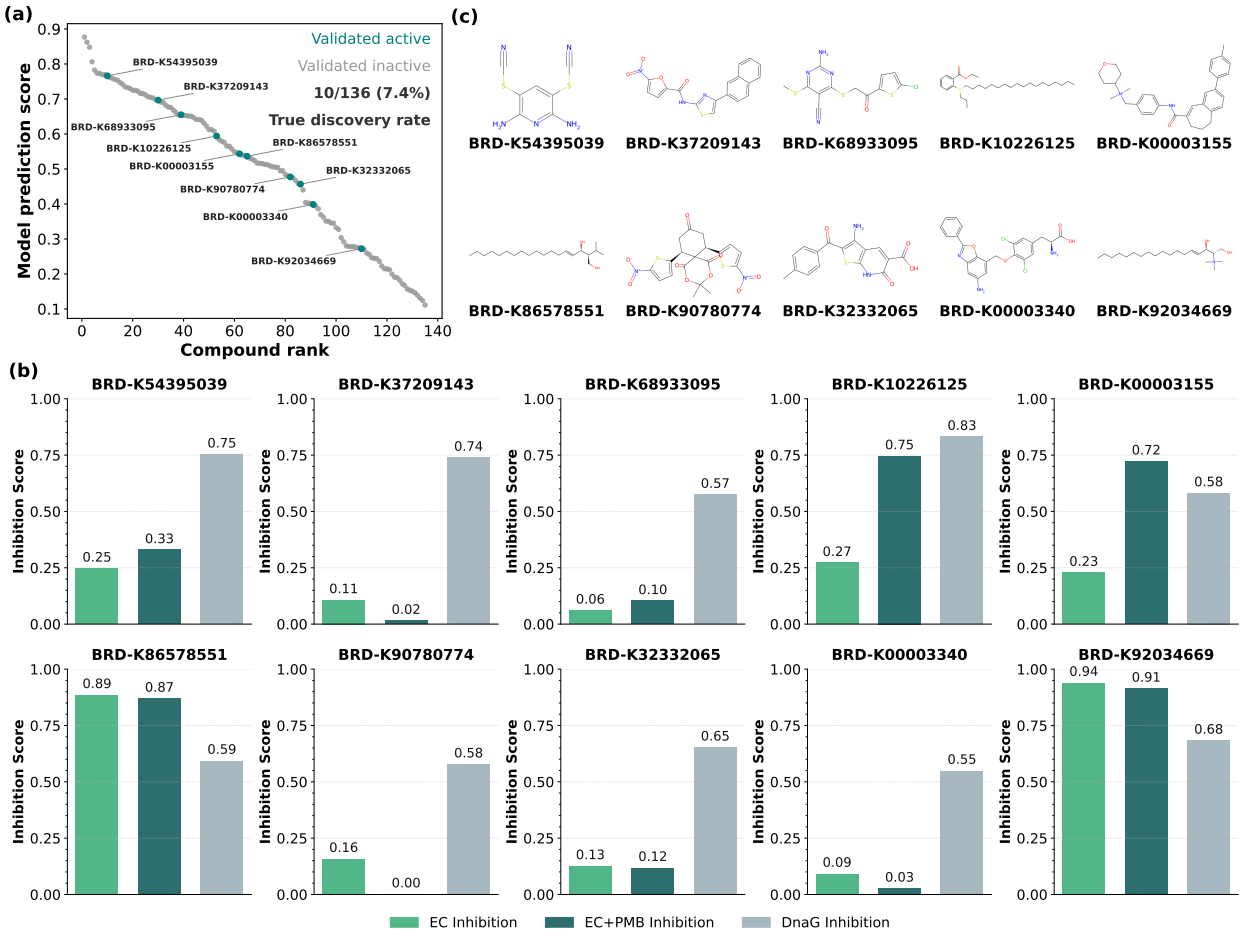

Figure 6: **Prospective wet-lab validation of FlashBind on DnaG. (a)** Rank-ordered FlashBind prediction scores for the 136 selected compounds screened against *E. coli* DNA primase (dnaG). Teal points indicate experimentally confirmed actives (DnaG inhibition $\geq 50\%$), yielding a hit rate of 10/136 (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.

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.

4.4 Prospective Wet-Lab Validation

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. (Wong et al., 2022), each compound was assayed against DnaG in two independent replicates at a concentration of 100μM. Full assay conditions are provided in Appendix C. 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.

5 Discussion and Conclusion

We presented FlashBind, a lightweight geometric framework that resolves the long-standing efficiency-accuracy trade-off in structure-based virtual screening. By occupying the optimal region of the Pareto frontier, it attains the early-enrichment capability of large-scale foundation models with a 50-fold reduction in inference latency, making high-fidelity screening of massive libraries accessible to laboratories without immense computational resources.

Our findings challenge the assumption that heavy-weight, end-to-end structure generation is a prerequisite for high-accuracy scoring. The success of FlashBind indicates that precise conformational sampling via expensive diffusion models is not strictly necessary for hit identification: a lightweight $E(3)$-equivariant encoder can capture the protein-ligand interactions present in fast docking priors such as FABind+ (Gao et al., 2025), relying on robust geometric features like heavy-atom contact patterns rather than precise atomic coordinates that demand explicit hydrogen or solvent modeling. Equally, scale alone is insufficient: our expanded training set was only effective when coupled with rigorous filtration to suppress experimental noise, suggesting that data quantity and quality control must be pursued together.

This robustness extends to 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, FlashBind maintained high predictive validity; by prioritizing explicit local geometric constraints, it mitigates overfitting to the protein families and eukaryotic targets that dominate standard training sets. The same task-agnostic encoder also extends to affinity regression (Appendix A), where it surpasses traditional baselines but does not consistently match Boltz-2 (Passaro et al., 2025), a gap our ablation (Appendix A.4) attributes to training-data provenance rather than architecture.

The modular design underlying FlashBind's efficiency also defines its principal limitation: its predictive ceiling is bounded by the fidelity of the upstream docking oracle. Where FABind+ fails to produce a plausible pose, under significant conformational plasticity or at cryptic binding sites, the downstream EGNN (Satorras et al., 2022) may process incorrect geometric signals. Future work will reduce this dependency via lightweight flexible-docking modules and more expressive geometric networks that capture higher-order many-body interactions over broader bioactivity datasets. Taken together, FlashBind offers a scalable, accurate, and physically grounded framework that redefines the practical limits of structure-based virtual screening.

References

Jonathan Bayldon Baell and Georgina A. Holloway. 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. URL https://api.semanticscholar.org/CorpusID:18795270. (Cited on page 5)

Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010. (Cited on page 8)

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000. (Cited on page 4)

David Buterez, Jon Paul Janet, Steven J. Kiddle, and Pietro Liò. Mf-pcba: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *Journal of Chemical Information and Modeling*, 63(9):2667–2678, 2023. doi: 10.1021/acs.jcim.2c01569. URL https://doi.org/10.1021/acs.jcim.2c01569. PMID: 37058588. (Cited on pages 2 and 6)

Darko Butina. 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 6)

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130 – 3139, 2023. URL https://api.semanticscholar.org/CorpusID:260865809. (Cited on page 17)

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010. (Cited on pages 4 and 16)

Haiyang Cui, Yufeng Su, Tanner J. Dean, Tianhao Yu, Zhengyi Zhang, Jian Peng, Diwakar Shukla, and Huimin Zhao. Enzyme specificity prediction using cross-attention graph neural networks. *Nature*, 647 (8090):639–647, Nov 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09697-2. URL https://doi.org/10.1038/s41586-025-09697-2. (Cited on pages 2, 3, 6 and 8)

Jacob D Durrant and J Andrew McCammon. 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 8)

Kairi Furui and Masahito Ohue. Boltzina: Efficient and accurate virtual screening via docking-guided binding prediction with boltz-2, 2025. URL https://arxiv.org/abs/2508.17555. (Cited on pages 3 and 7)

Bowen Gao, Bo Qiang, Haichuan Tan, Minsi Ren, Yinjun Jia, Minsi Lu, Jingjing Liu, Weiying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening, 2023. URL https://arxiv.org/abs/2310.06367. (Cited on page 3)

Kaiyuan Gao, Qizhi Pei, Gongbo Zhang, Jinhua Zhu, Kun He, and Lijun Wu. 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, July 2025. doi: 10.1145/3690624.3709253. URL http://dx.doi.org/10.1145/3690624.3709253. (Cited on pages 2, 4 and 12)

M. K. Gilson, J. Eberhardt, P. Škrinjar, J. Durairaj, X. Robin, and A. Kryshtafovych. Assessment of pharmaceutical protein-ligand pose and affinity predictions in casp16. 2025. doi: 10.22541/au.174562565.51283311/v1. (Cited on page 17)

Richard J. Gowers, Irfan Alibay, David W.H. Swenson, Michael M. Henry, Benjamin Ries, Hannah M. Baumann, and James R. B. Eastwood. The open free energy library, September 2023. URL https://doi.org/10.5281/zenodo.8344248. (Cited on page 17)

David F. Hahn, Christopher I. Bayly, Melissa L Boby, Hannah E. Bruce Macdonald, John D. Chodera, Vytautas Gapsys, Antonia S. J. S. Mey, David L. Mobley, Laura Perez Benito, Christina E. M. Schindler, Gary Tresadern, and Gregory L. Warren. Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks [article v1.0]. *Living Journal of Computational Molecular Science*, 4(1), 2022. ISSN 2575-6524. doi: 10.33011/livecoms.4.1.1497. URL http://dx.doi.org/10.33011/livecoms.4.1.1497. (Cited on page 17)

Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. 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. doi: 10.1021/jm030644s. URL https://doi.org/10.1021/jm030644s. PMID: 15027866. (Cited on pages 1 and 3)

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018. (Cited on pages 4, 8 and 19)

Peter J. Huber. *Robust Estimation of a Location Parameter*, pp. 492–518. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_35. URL https://doi.org/10.1007/978-1-4612-4380-9_35. (Cited on page 17)

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, et al. Uni-mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*, 2024. (Cited on pages 6 and 8)

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. (Cited on page 21)

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac956. URL https://doi.org/10.1093/nar/gkac956. (Cited on page 4)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980. (Cited on page 21)

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation, 2023. URL https://arxiv.org/abs/2208.06073. (Cited on pages 4 and 20)

Alexander Kroll, Sahasra Ranjan, Martin KM Engqvist, and Martin J Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature communications*, 14(1): 2787, 2023. (Cited on pages 2, 3 and 8)

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, Eisuke Kawashima, NadineSchneider, Dan Nealschneider, Andrew Dalke, tadhurst cdd, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Hussein Faara, Ichiru Take, Rachel Walker, Vincent F. Scalfani, Niels Maeder, Daniel Probst, Kazuya Ujihara, Axel Pahl, guillaume godin, and Juuso Lehtivarjo. rdkit/rdkit: 2025_09_1 (q3 2025) beta release, September 2025. URL https://doi.org/10.5281/zenodo.17193272. (Cited on pages 4 and 19)

Pablo Lemos, Zane Beckwith, Sasaank Bandi, Maarten van Damme, Jordan Crivelli-Decker, Benjamin J. Shields, Thomas Merth, Punit K. Jha, Nicola De Mitri, Tiffany J. Callahan, AJ Nish, Paul Abruzzo, Romelia Salomon-Ferrer, and Martin Ganahl. Sair: Enabling deep learning for protein-ligand interactions with a synthetic structural dataset. *bioRxiv*, 2025. doi: 10.1101/2025.06.17.660168. URL https://www.biorxiv.org/content/early/2025/06/21/2025.06.17.660168. (Cited on pages 16 and 19)

Min Li, Zhangli Lu, Yifan Wu, and Yaohang Li. Bacpi: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction. *Bioinformatics*, 38, 01 2022. doi: 10.1093/bioinformatics/btac035. (Cited on pages 7 and 18)

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002. (Cited on page 6)

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35:D198–D201, 12 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl999. URL https://doi.org/10.1093/nar/gkl999. (Cited on page 16)

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021. (Cited on pages 1 and 8)

Yoshio Nishimoto and Dmitri G. Fedorov. The fragment molecular orbital method combined with density-functional tight-binding and the polarizable continuum model. *Phys. Chem. Chem. Phys.*, 18:22047–22061, 2016. doi: 10.1039/C6CP02186G. URL http://dx.doi.org/10.1039/C6CP02186G. (Cited on page 7)

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707. (Cited on pages 1, 3, 4, 7 and 12)

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020. (Cited on page 3)

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

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. (Cited on pages 6, 8 and 17)

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022. URL https://arxiv.org/abs/2102.09844. (Cited on pages 2, 3, 4, 12 and 18)

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, November 2017. (Cited on page 6)

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020. (Cited on page 3)

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: An unbiased dataset for machine learning and virtual screening. *Journal of Chemical Information and Modeling*, 04 2020. doi: 10.1021/acs.jcim.0c00155. (Cited on page 6)

Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31, 2009. URL https://api.semanticscholar.org/CorpusID:30245244. (Cited on pages 1, 3, 7 and 9)

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis,

and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061. (Cited on page 4)

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167. URL https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167. (Cited on page 17)

Maciej Wójcikowski, Pedro J Ballester, and Pawel Siedlecki. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 7(1):46710, 2017. (Cited on page 8)

Maciej Wójcikowski, Michał Kukiełka, Marta M Stepniewska-Dziubinska, and Pawel Siedlecki. Development of a protein–ligand extended connectivity (plec) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341, 2019. (Cited on page 8)

Felix Wong, Aarti Krishnan, Erica J. Zheng, Hannes Stärk, Abigail L. Manson, Ashlee M. Earl, Tommi Jaakkola, and James J. Collins. Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):MSB202211081, Sep 2022. ISSN 1744-4292. doi: 10.15252/msb.202211081. URL https://doi.org/10.15252/msb.202211081. (Cited on pages 2, 3, 9 and 12)

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1004. URL https://doi.org/10.1093/nar/gkad1004. (Cited on page 16)

Zhaocheng Zhu, Chence Shi, Zuobai Zhang, Shengchao Liu, Minghao Xu, Xinyu Yuan, Yangtian Zhang, Junkun Chen, Huiyu Cai, Jiarui Lu, Chang Ma, Runcheng Liu, Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. Torchdrug: A powerful and flexible machine learning platform for drug discovery, 2022. URL https://arxiv.org/abs/2202.08320. (Cited on pages 4, 8 and 19)

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 (Lemos et al., 2025), a large-scale synthetic structural dataset integrating data from ChEMBL (Zdrazil et al., 2023) and BindingDB (Liu et al., 2006). 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) (Consortium, 2024), `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) (Rogers & Hahn, 2010) 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 (Wohlwend et al., 2024). 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 (Buttenschoen et al., 2023) 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.
- **Test**: We evaluated performance on three benchmarks identical to those used in Boltz-2: the **OpenFE subset** (Gowers et al., 2023), a **4-target subset (CDK2, TYK2, JNK1, P38)** of the FEP+ benchmark (Hahn et al., 2022), and the blind **CASP16** benchmark (Gilson et al., 2025).

A.2 Training Objective and Inference

Composite ranking objective. Because binding affinity is reported on assay-specific scales, absolute values are far less transferable across experiments than the *relative* ordering of compounds within a single assay. We therefore train the regression head with a composite objective that weights pairwise ranking accuracy over absolute error (Huber, 1992):

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

where (i, j) index two compounds drawn from the same assay group (Appendix B.4). The first term penalizes errors in the predicted pairwise difference and dominates the loss, while the second term anchors the predictions to the absolute scale. This emphasis on intra-assay ranking aligns the training signal with the lead-optimization use case, where prioritizing compounds within a series matters more than exact K_d recovery.

Table 1: **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

Molecular-weight bias correction. Structure-based scoring functions are known to exhibit a systematic bias toward larger ligands, as additional atoms tend to inflate raw interaction scores irrespective of true binding strength. To mitigate this, we apply a post-hoc polynomial correction to the ensemble-averaged predictions as a function of ligand molecular weight (MW), fitted on the validation set and applied at inference. This calibration removes size-correlated artifacts without retraining and is applied only to the regression task; the classification setting, which operates on within-assay ranking of p_{bind}, does not use it.

A.3 Benchmarking Results

We compared our method FlashBind against the state-of-the-art foundation model Boltz-2, as well as sequence-based (BACPI) (Li et al., 2022) and ligand-only (GAT) baselines.

The comprehensive results are summarized in Table 1. 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.4 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) (Satorras et al., 2022) 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 2. 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 (Lemos et al., 2025), 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 2: **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

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** (Hayes et al., 2025) 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** (Zhu et al., 2022) using **RDKit** (Landrum et al., 2025). 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 (Kong et al., 2023), 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.
   - *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 (Jordan et al., 2024). 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 (Kingma & Ba, 2017) 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.

C Experimental Protocols

This section provides the full experimental conditions for the prospective wet-lab validation, covering the *E. coli* DnaG inhibition assay, the whole-cell antibacterial assay, and compound sourcing.

C.1 *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.

C.2 *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.

C.3 Compound Preparation

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

D 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.

D.1 Virtual Screening Performance

We present the comprehensive metrics for the MF-PCBA benchmark (Fig. 3 in main text). Table 3 details the performance of FlashBind against primary baselines, and Table 4 provides the detailed results for the ablation study and comparison against various docking and rescoring strategies.

Table 3: 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

Table 4: **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

D.2 Computational Efficiency

To quantify the efficiency gains, we benchmarked the inference speed of various methods. Table 5 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.

D.3 Enzyme Specificity Results

Table 6 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.

Table 5: 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 6: **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