ALLOMIT: BOOSTING ALLOSTERIC VIRTUAL SCREENING VIA MULTI-INSTANCE PARAMETER-EFFICIENT TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Allosteric modulation offers opportunities to alleviate target crowding and broaden the scope of drug discovery by targeting sites beyond canonical active sites. However, mainstream virtual screening methods largely focus on orthosteric screening, where a binding pocket is specified in advance. In contrast, allosteric sites may be unknown and require ligand-induced conformational changes, making allosteric screening more challenging. Moreover, the limited availability of allosteric activity data poses an additional challenge for training effective screening models. To this end, we introduce AlloMIT, a multi-instance parameter-efficient tuning strategy for allosteric hit discovery under limited activity data. AlloMIT integrates with pre-trained virtual screening alignment models in a plug-and-play manner, adapting them with only a small number of trainable parameters. Rather than relying on a single pre-defined binding pocket, AlloMIT represents each protein structure as a set of candidate pockets and learns target-ligand interactions in a weakly supervised, multi-instance manner, without requiring pocket-level annotations. To account for ligand-dependent pocket contributions, its Contextual Sparse Pocket Aggregation (CSPA) module considers potential pocket-ligand interactions within proteins and outputs the final pocket score via sparse aggregation. Additionally, we construct AlloVSBench to support parameter-efficient model adaptation and benchmark allosteric hit discovery performance across screening methods and parameter-efficient tuning methods. Ultimately, AlloMIT achieves the highest allosteric BEDROC score among all evaluated methods. Compared with adapted HypSeek, AlloMIT improves BEDROC from 17.09% to 23.25% and lifts $EF_{0.5\%}$ from 20.79 to 41.70.

1 INTRODUCTION

Orthosteric binding typically occurs at the site used by a protein's endogenous ligands or substrates, whereas allosteric binding occurs at a topographically distinct site on the same protein (Figure 1(a)) (Proudfoot et al., 2011). Binding at an allosteric site can alter protein conformation and thereby modulate its activity (Figure 1(b)) (Kruse et al., 2013; Chen et al., 2016). Allosteric modulation has therefore emerged as an important strategy for therapeutic intervention. Recent successes, including sotorasib targeting $KRAS^{G12C}$ and asciminib targeting the myristoyl pocket of BCR-ABL1, highlight the potential of allosteric ligands to expand the pharmacological space for both established and emerging therapeutic targets (Skoulidis et al., 2021; Hughes et al., 2019).

However, discovering allosteric hits remains challenging for two major reasons. First, allosteric activity data remain scarce, restricting model training. Second, allosteric binding sites are often unknown and may vary in both location and conformation. Although pocket-based virtual screening has achieved substantial progress in recent years, many existing methods still require a pre-defined binding site (Gao et al., 2023; Feng et al., 2025). A natural strategy for allosteric screening is therefore to first identify an allosteric pocket and then perform virtual screening against it. However, ligand-induced conformational changes can make allosteric pocket identification particularly difficult. Pocket detection from a single ligand-free static structure may fail to capture ligand-relevant conformations (Hollingsworth et al., 2019), whereas sampling these conformations through molecular dynamics

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

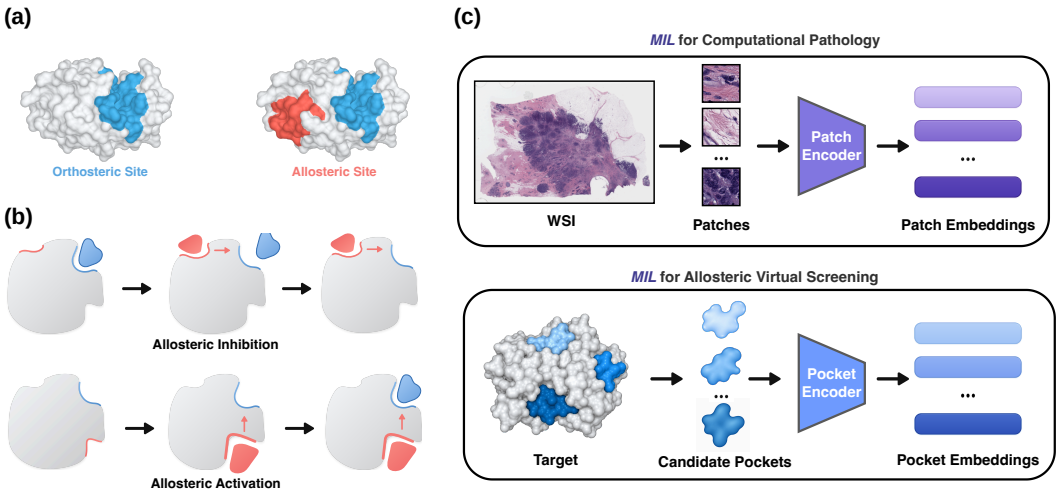


Figure 1: **Allosteric regulation and multi-instance learning.** **(a)** Orthosteric (blue) and allosteric (coral) sites occupy distinct regions of a protein. **(b)** Schematic examples show how allosteric binding changes the orthosteric site to inhibit (top) or facilitate (bottom) ligand binding. **(c)** Multi-instance learning analogy between computational pathology and allosteric screening. In computational pathology, a whole-slide image (WSI) is represented as a bag of tissue patches with slide-level supervision. Analogously, in allosteric screening, a protein target is represented as a bag of candidate pockets, while activity labels supervise target-ligand pairs without specifying which pocket mediates binding.

simulations can incur substantial computational costs (Meller et al., 2023). Moreover, erroneous pocket identification can mislead subsequent virtual screening.

Therefore, we introduce AlloMIT, a multi-instance parameter-efficient tuning framework that adapts pre-trained virtual screening alignment models for allosteric hit discovery under limited activity data. Rather than relying on a single pre-defined binding pocket, AlloMIT represents each protein target as a set of candidate pockets. Target-ligand activity labels provide weak supervision for the candidate-pocket bag without specifying which pocket mediates binding. Inspired by slide-level weak supervision in computational pathology (Lu et al., 2021) (Figure 1(c)), we formulate the candidate pockets as instances within a multi-instance learning framework. A shared pocket encoder extracts representations for all candidates, which are matched against the ligand representation. Our Contextual Sparse Pocket Aggregation (CSPA) module combines candidate-pocket context with pocket-ligand compatibility to sparsely aggregate matching scores across candidate pockets. This avoids committing screening to a single potentially misidentified pocket, while allowing the contributions of candidate pockets to vary across ligands. Together, these components extend pre-trained alignment models from orthosteric to allosteric virtual screening. This expansion preserves the full pre-trained backbone and only requires lightweight modules and selective parameter updates.

To support model adaptation and evaluate allosteric hit discovery, we construct AlloVSBench, which combines curated allosteric activity data with apo or predicted protein structures. The benchmark not only provides data for tuning but also evaluates both allosteric hit discovery and the retention of general screening capability after adaptation. Applied to HypSeek, AlloMIT achieves the highest allosteric BEDROC among a series of screening methods and outperforms other PEFT methods on the same metric. Compared with adapted HypSeek, AlloMIT increases allosteric BEDROC from 17.09% to 23.25% and $EF_{0.5\%}$ from 20.79 to 41.70. Further experiments, including ablation studies, provide complementary evidence for the effectiveness of AlloMIT in allosteric hit discovery.

Our main contributions are as follows:

- We introduce AlloMIT, a parameter-efficient tuning framework that adapts pre-trained alignment-based virtual screening methods to allosteric hit discovery without requiring a pre-defined binding pocket.
- We reformulate allosteric hit discovery as a multi-instance learning problem and develop a corresponding tuning framework with a Contextual Sparse Pocket Aggregation module that

jointly considers and sparsely aggregates pocket-ligand matching evidence across multiple candidate pockets, accommodating ligand-dependent variation in potential binding sites.

- We construct AlloVSBench to support both model adaptation and screening evaluation. The benchmark assesses allosteric hit discovery and the retention of general screening capability after adaptation.
- AlloMIT achieves the highest BEDROC, $\mathrm{EF}_{0.5\%}$, and $\mathrm{EF}_{1\%}$ among the evaluated virtual screening and parameter-efficient tuning methods. Extensive experiments further provide systematic insights into effective screening paradigms for allosteric hit discovery and the key limitations of existing approaches.

2 METHOD

2.1 PRELIMINARY

Alignment-based virtual screening methods score protein-ligand pairs by matching their representations. HypSeek (Wang et al., 2026), a three-branch network, uses structure encoders for pockets and ligands and a protein language model for sequences. An elegant design is that these representations are further mapped to Lorentz hyperbolic space, helping capture hierarchical relationships and fine-grained affinity differences. The ligand representation is then matched separately with the pocket and sequence representations, and the two matching scores are summed as:

$$\mathbf{S}_{\text{base}} = \phi_c(a_{\text{poc}} E_{\text{poc}}(P))^\top \phi_c(a_{\text{lig}} E_{\text{lig}}(L)) + \phi_c(a_{\text{seq}} E_{\text{seq}}(S))^\top \phi_c(a_{\text{lig}} E_{\text{lig}}(L)) \,. \tag{1}$$

Here, P and S contain the corresponding pocket and sequence inputs for each target, and L contains the ligands. The subscript $m \in \{\text{poc}, \text{seq}, \text{lig}\}$ indexes the pocket, sequence, and ligand branches. Each E_m includes an encoder and projection head and stacks d-dimensional features as columns, with $d = 128$. The shared mapping ϕ_c returns the spatial coordinates of the Lorentz exponential map, while a_m denotes the branch-specific scaling factor. Each row of $\mathbf{S}_{\text{base}}$ provides the ligand-ranking scores for one target.

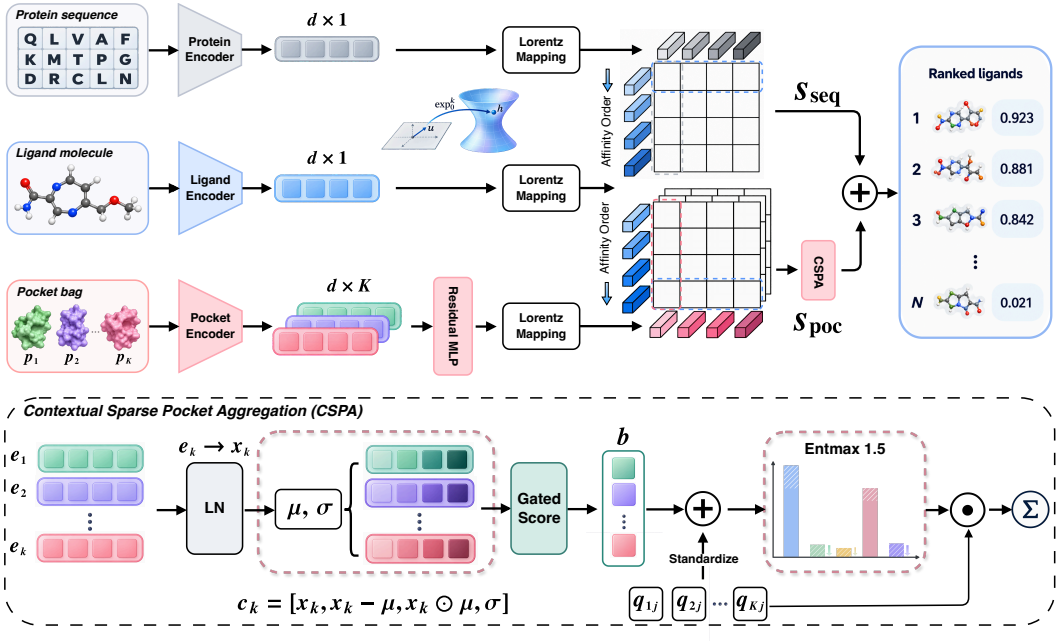


Figure 2: **Overview of AlloMIT. Top**, we apply AlloMIT to HypSeek while retaining its pocket, protein sequence, and ligand inputs. Notably, AlloMIT modifies the pocket branch by replacing the single pre-defined pocket with a set of candidate pockets and introducing a residual MLP and Contextual Sparse Pocket Aggregation (CSPA). **Bottom**, candidate-pocket context characterizes each pocket relative to the others, while pocket-ligand matching scores reflect its compatibility with the query ligand. CSPA combines these two signals to determine sparse weights and uses them to aggregate the matching scores into a final pocket score.

2.2 OVERVIEW

As shown in Figure 2, we illustrate AlloMIT using HypSeek. Given a protein target and a ligand ℓ_j from a screening library, AlloMIT replaces the original pocket P with a set of candidate pockets $\mathcal{P} = \{p_k\}_{k=1}^{K}$ extracted from the target structure using Fpocket. We introduce two lightweight learnable modules, a residual MLP and Contextual Sparse Pocket Aggregation (CSPA).

The shared residual MLP calibrates each candidate-pocket feature $\mathbf{e}_k = E_{\mathrm{poc}}(p_k)$ as $\widetilde{\mathbf{e}}_k = \mathbf{e}_k + r_\phi(\mathbf{e}_k)$. Each calibrated feature is then matched to the ligand representation $\mathbf{z}_j = \phi_c(a_{\mathrm{lig}} E_{\mathrm{lig}}(\ell_j))$, yielding a pocket-ligand score $q_{kj} = \phi_c(a_{\mathrm{poc}}\widetilde{\mathbf{e}}_k)^\top \mathbf{z}_j$. CSPA aggregates these scores using ligand-dependent weights w_{kj}. The aggregated pocket score is added to the sequence-ligand similarity. Ligands are ranked in descending order of the resulting target-ligand score:

$$s_j = \sum_{k=1}^{K} w_{kj} q_{kj} + \left[\phi_c(a_{\mathrm{seq}} E_{\mathrm{seq}}(S))\right]^\top \mathbf{z}_j. \tag{2}$$

As shown in Figure 3, AlloMIT follows a two-stage tuning strategy comprising warm-up tuning and joint tuning. We first train the residual MLP and CSPA on fixed pre-trained representations to learn pocket-feature calibration and score aggregation. We then fix these modules and jointly tune selected parameters in the pocket, sequence, and ligand branches to adapt their representations to allosteric screening. We describe CSPA and the two-stage tuning procedure in the following sections.

2.3 CONTEXTUAL SPARSE POCKET AGGREGATION

Replacing a single pocket with multiple candidates yields a set of pocket-ligand matching scores for each query ligand. These scores reflect individual pocket-ligand compatibility but do not explicitly capture each candidate's structural context within the pocket set. We introduce CSPA to combine these two sources of information through sparse aggregation. Specifically, CSPA first computes a contextual structural score for each candidate, then combines it with ligand-specific compatibility to determine sparse aggregation weights. These weights aggregate the original matching scores, allowing the contributing subset of pockets to vary across ligands.

Contextual structural scores. We normalize the uncorrected pocket features as $\mathbf{x}_k = \mathrm{LN}(\mathbf{e}_k)$ and summarize the candidate set using its feature-wise mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. Each candidate is represented relative to this shared context as:

$$\mathbf{c}_k = [\mathbf{x}_k,\, \mathbf{x}_k - \boldsymbol{\mu},\, \mathbf{x}_k \odot \boldsymbol{\mu},\, \boldsymbol{\sigma}], \qquad b_k = g_\psi(\mathbf{c}_k), \tag{3}$$

where brackets denote concatenation, $\odot$ denotes element-wise multiplication, and g_ψ is a gated scorer. The resulting b_k is shared across ligands and serves as a target-specific structural prior. The contextual scores are computed from the uncorrected pocket features, whereas the residual-corrected features determine pocket-ligand matching scores.

Ligand-conditioned sparse aggregation. Raw matching scores can differ in scale across ligands. We therefore standardize $\mathbf{q}_j = [q_{1j}, \ldots, q_{Kj}]^\top$ over the candidate pockets, obtaining the relative compatibility vector $\widehat{\mathbf{q}}_j$. With $\mathbf{b} = (b_1, \ldots, b_K)^\top$, CSPA combines this ligand-specific evidence with the structural prior to compute:

$$\mathbf{w}_j = \mathrm{entmax}_{1.5}\left(\frac{\mathbf{b} + \widehat{\mathbf{q}}_j}{\tau}\right), \qquad s_j^{\mathrm{poc}} = \sum_{k=1}^{K} w_{kj} q_{kj}. \tag{4}$$

where $\tau > 0$ controls the concentration of the weights. The resulting weights are nonnegative and sum to one, and entmax (Peters et al., 2019) allows candidates to receive exactly zero weight. The standardized scores determine the weights, whereas the final aggregation uses the original matching scores q_{kj}.

2.4 WARM-UP TUNING AND JOINT TUNING

As shown in Figure 3, we adopt a two-stage tuning strategy. We first tune the newly introduced residual MLP and CSPA, then jointly tune selected parameters in the pocket, sequence, and ligand branches.

The encoders and projection heads inherit pre-trained parameters from HypSeek, whereas the newly introduced residual MLP and CSPA have not been pre-trained. During warm-up tuning (Stage 1), we therefore freeze all three encoders and their projection heads and tune only the residual MLP and CSPA under target-ligand activity supervision. This stage learns pocket-feature calibration and score aggregation on fixed pre-trained representations.

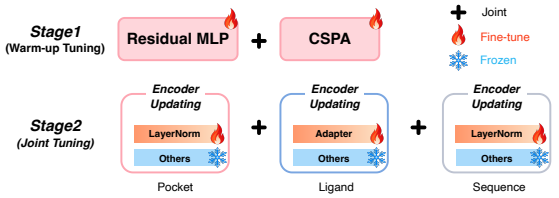


Figure 3: **Warm-up tuning and joint tuning.** Warm-up trains the residual MLP and CSPA with frozen encoders. Joint tuning then updates selected encoder parameters while keeping these modules fixed.

After warm-up tuning, we conduct joint tuning (Stage 2) to adapt the model to allosteric screening. We jointly tune the LayerNorm parameters in the pocket and sequence encoders and the bottleneck adapters in the final two layers of the ligand encoder. The parameters of the residual MLP and CSPA, together with all other pre-trained parameters, remain frozen.

Both tuning stages follow the training objective of HypSeek (Wang et al., 2026), with the pocket loss applied to aggregated matching scores and geometric regularization applied to sequence-ligand representations:

$$\mathcal{L} = \mathcal{L}_{\text{align}}^{\text{poc}} + \mathcal{L}_{\text{align}}^{\text{seq}} + \gamma \mathcal{L}_{\text{cone}} + \mathcal{L}_{\text{reg}}. \tag{5}$$

Each alignment term combines contrastive matching with within-assay activity ranking. The remaining terms provide geometric regularization, with γ controlling the contribution of the cone loss.

3 AlloVSBench: A Benchmark for Allosteric Virtual Screening

We construct AlloVSBench to support model adaptation and evaluate allosteric hit discovery performance. As shown in Figure 4, the construction pipeline includes data harmonization, leakage-controlled partitioning, and apo structural reconstruction.

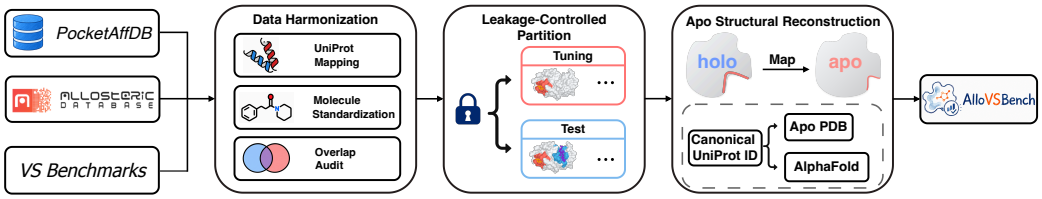


Figure 4: **Construction of AlloVSBench.** Source records are harmonized through UniProt mapping and molecular standardization, then audited for overlap to form target-disjoint tuning and test sets. Structural inputs for all targets are standardized to apo structures sourced from PDB or AlphaFold predictions.

We construct AlloVSBench using allosteric activity data from the hit-to-lead (H2L) subset of ASD2023, covering 82 protein targets (He et al., 2024). For consistent evaluation before and after adaptation, we align our target partition with the original virtual screening split. PocketAffDB defines the candidate tuning targets (Feng et al., 2025), while DUD-E, DEKOIS 2.0, and LIT-PCBA serve as the test benchmarks (Mysinger et al., 2012; Bauer et al., 2013; Tran-Nguyen et al., 2020). Specifically, mapping the ASD2023 targets to canonical UniProt accessions identifies 19 targets shared with the test benchmarks, which we reserve for testing and exclude from tuning. The remaining targets represented in PocketAffDB form the candidate pool for tuning. Within this pool, each target must pass structural and reference-site checks and have a qualified ligand-free structural input. After molecular standardization and ligand-label reconciliation, it must also retain at least two distinct allosteric compounds with $pActivity \geq 5$. These criteria yield a tuning set of 28 targets, with no targets shared between the tuning and test sets. Figure 5 summarizes the target overlap across these sources.

The resulting tuning set contains 1,243 activity measurements across 28 targets, while the test set comprises 794,668 target-ligand pairs across 19 targets. We use decoys and experimentally confirmed inactive compounds from the source benchmarks as negatives. Each test library retains these negatives and combines benchmark-derived orthosteric hits with allosteric hits from ASD2023. We distinguish these two hit categories to assess the retention of general screening capability after tuning and the ability to discover allosteric hits, respectively.

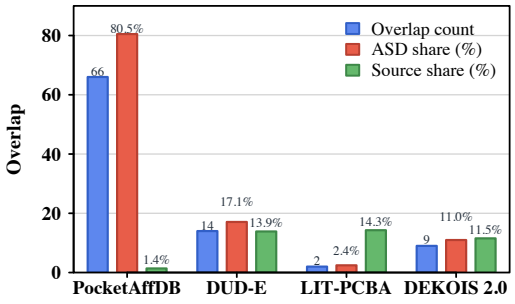


Figure 5: **Target overlap with ASD2023 H2L.**

We standardize target and ligand identities across both the tuning and test sets. Targets are mapped to canonical UniProt accessions, and molecules are reduced to their principal parent fragments and represented as canonical isomeric SMILES. Duplicate records are consolidated within compatible assay contexts after label conflicts are resolved. For each target, we search PDB for an independent apo structure, excluding the source holo structure. Candidate structures must also pass reference-site coverage and local ligand-absence checks. When no experimental structure meets these criteria, we use an AlphaFold prediction (Jumper et al., 2021). Reference-site mappings are used only for structural curation and are not supplied for candidate generation or screening.

We generate candidate pockets from each selected structure using fpocket (Le Guilloux et al., 2009). Detections from three chain-aware configurations are merged and deduplicated based on residue overlap and pocket-center proximity. We rank the remaining candidates using a composite score that combines fpocket score, druggability score, pocket volume, and support across detection configurations. The top 30 candidates form the pocket set $\mathcal{P}_i$ for each target, with all candidates retained when fewer than 30 are available. Detailed construction procedures, partition audits, and site-coverage analyses are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

We assess both general and allosteric screening performance on AlloVSBench. General screening treats both orthosteric and allosteric active compounds as hits. For allosteric screening, we exclude orthosteric hits from the screening library and evaluate how well each method retrieves allosteric hits. Following HypSeek (Wang et al., 2026), we report BEDROC and enrichment factors (EF) at the top 0.5%, 1%, and 5% (Truchon & Bayly, 2007). We compute each metric separately for each target and report the mean across targets. BEDROC is expressed as a percentage, while EF values retain their original scale. Additional results, including AUROC, are provided in Appendix D.

We compare AlloMIT with Uni-Dock (Yu et al., 2023), DeepDTA (Öztürk et al., 2018), GraphDTA (Nguyen et al., 2021), DrugCLIP (Gao et al., 2023), BindCLIP (Qiao et al., 2026), LigUnity (Feng et al., 2025), AANet (Zhu et al., 2025), and HypSeek (Wang et al., 2026) under zero-shot and tuned settings. The zero-shot setting evaluates methods without allosteric tuning. In the tuned setting, we adapt DrugCLIP, LigUnity, AANet, and HypSeek on the same AlloVSBench tuning set. Our tuning experiments cover LoRA (Hu et al., 2022), adapters (Houlsby et al., 2019), LayerNorm tuning, BitFit (Ben Zaken et al., 2022), prefix tuning (Li & Liang, 2021), projection-head tuning, and last-layer fine-tuning. For each baseline, we report the best-performing tuning configuration. We also evaluate pocket-only and sequence-only variants of HypSeek in both settings to assess pocket-based and sequence-based screening separately.

We initialize the pocket, sequence, and ligand encoders from pre-trained HypSeek (Wang et al., 2026). During warm-up tuning (the first stage of AlloMIT), we tune the residual pocket MLP and contextual scorer with a learning rate of 10^{-4}. During joint tuning (Stage 2), we tune the ligand adapters at a learning rate of 10^{-4} and the pocket and sequence LayerNorm parameters at 5×10^{-6}, while keeping all other parameters frozen. This stage updates 123,392 parameters. We train both stages for 20 epochs with a batch size of 4 and retain the checkpoint with the lowest training loss in each stage.

4.2 SCREENING PERFORMANCE

Table 1: **Screening performance on AlloVSBench.** Metrics are macro-averaged over 19 targets. BEDROC is reported as a percentage. Bold values indicate column-wise best results.

		General				Allosteric			
Model	Setting	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$
Uni-Dock	Zero-shot	4.76	2.12	2.21	1.83	0.47	0.00	0.00	1.58
DeepDTA	Zero-shot	4.16	2.88	1.92	1.58	6.62	3.43	7.00	2.98
GraphDTA	Zero-shot	22.01	15.22	11.90	5.23	0.04	0.00	0.00	0.00
BindCLIP	Zero-shot	15.04	10.20	8.08	3.62	5.70	13.90	6.99	1.93
DrugCLIP	Zero-shot	19.36	16.12	12.89	4.58	3.34	0.00	5.26	2.11
DrugCLIP*	Tuned	26.97	22.43	16.96	6.10	6.06	10.48	5.24	4.21
LigUnity	Zero-shot	23.91	15.16	12.97	5.39	2.69	0.00	0.00	3.15
LigUnity*	Tuned	28.03	17.66	16.35	5.93	6.12	10.48	5.24	2.10
AANet	Zero-shot	21.29	14.85	12.52	4.18	0.67	0.00	0.00	0.88
AANet†	Tuned	4.36	2.68	3.16	1.62	2.54	5.19	2.62	1.05
$HypSeek_{poc}$	Zero-shot	44.44	27.68	24.41	10.20	12.38	20.89	15.72	6.31
$HypSeek_{poc}$*	Tuned	40.05	24.04	22.19	9.48	15.55	31.21	15.72	4.21
$HypSeek_{seq}$	Zero-shot	47.22	30.51	26.68	10.37	12.27	20.79	10.46	7.36
$HypSeek_{seq}$*	Tuned	45.87	30.00	26.01	9.80	15.87	20.79	15.72	7.36
HypSeek	Zero-shot	48.80	30.31	26.94	10.98	14.94	20.89	20.95	6.84
HypSeek*	Tuned	47.36	28.89	26.80	10.56	17.09	20.79	20.94	**7.89**
+ AlloMIT (Ours)	Tuned	**51.22**	**31.78**	**28.90**	**11.17**	**23.25**	**41.70**	**26.19**	7.87

AlloMIT improves early enrichment of allosteric hits. Table 1 compares AlloMIT on HypSeek with zero-shot and tuned screening baselines. AlloMIT achieves the highest allosteric BEDROC, increasing from 17.09% for tuned HypSeek to 23.25%. Its allosteric $EF_{0.5\%}$ more than doubles from 20.79 to 41.70, placing more allosteric ligands near the top of the ranked libraries without requiring a pre-defined binding pocket. This stronger early enrichment is particularly useful when only a small fraction of the library can be tested experimentally. AlloMIT also raises general BEDROC from 47.36% to 51.22% on HypSeek and achieves the highest general enrichment factors at all three cutoffs. These results support improved allosteric hit discovery while maintaining strong general screening performance.

Plug-and-play performance. AlloMIT is a plug-and-play PEFT framework for allosteric screening. We also apply AlloMIT to DrugCLIP and LigUnity using the same configuration as for HypSeek. A slight difference is that these two methods include only a pocket branch and a ligand branch. AlloMIT raises allosteric $EF_{0.5\%}$ from 10.48 to 28.13 on DrugCLIP and from 10.48 to 15.71 on LigUnity relative to conventional tuning (Figure 6). Together with the results on HypSeek, these gains support AlloMIT as a plug-and-play strategy for adapting pre-trained models to allosteric screening.

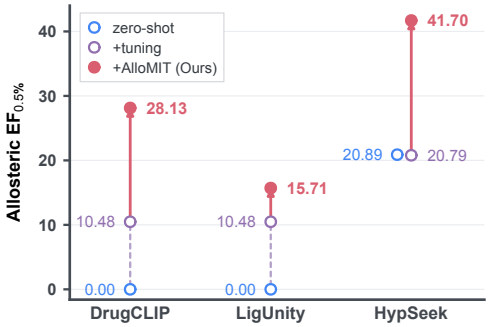


Figure 6: **Plug-and-play performance of AlloMIT.**

†Following AANet's tuning strategy, we adapt its released checkpoint on the AlloVSBench tuning set.

*We systematically compare representative PEFT methods for each marked baseline and report the tuned results obtained with the best-performing method.

4.3 BUILDING ALLOMIT STEP BY STEP

We examine the design choices behind AlloMIT on HypSeek, from pocket aggregation and feature calibration to joint tuning. Figure 7 also compares alternatives to the two-stage tuning strategy.

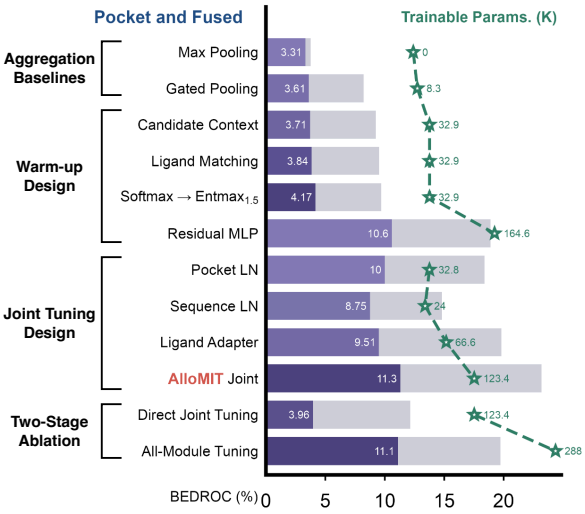


Figure 7: **Building AlloMIT on HypSeek.** Dark and light bars show allosteric BEDROC (%) using pocket scores alone and combined pocket-sequence scores, respectively. Green stars indicate trainable parameters per stage, in thousands, on a separate scale.

Warm-up Design. We first use candidate-pocket context to determine aggregation weights. Adding standardized pocket-ligand matching scores to the weighting logits raises pocket-only allosteric BEDROC to 3.84%. Replacing Softmax with $\text{Entmax}_{1.5}$ yields the highest pocket-only and fused BEDROC among these aggregation variants, at 4.17% and 9.71%. Building on the design, a shared residual MLP calibrates pocket features, further raising pocket-only and fused BEDROC to 10.6% and 18.9%, respectively.

Joint Tuning Design. Following warm-up, we fix the residual MLP and CSPA and compare tuning individual branches with jointly tuning the pocket, sequence, and ligand branches. Joint tuning achieves a fused allosteric BEDROC of 23.25%, compared with 19.8% for ligand-only tuning. This configuration updates 123,392 parameters during the joint tuning stage.

Two-Stage Ablation. Direct Joint Tuning skips warm-up and reduces fused allosteric BEDROC from 23.25% to 12.15%. All-Module Tuning jointly updates the residual MLP, CSPA, and encoder branches from the original pre-trained model, yielding 19.73% with 288.0K trainable parameters. These results suggest that joint tuning benefits from first learning pocket-feature calibration and score aggregation during warm-up.

4.4 DISCUSSION

	BEDROC (%)	$\text{EF}_{0.5\%}$
Max pooling	14.94	20.89
Weighted pooling	16.87	20.79
Top-1 selection	13.67	20.79
AlloMIT	**23.25**	**41.70**

Table 2: **Allosteric screening performance on AlloVSBench.** Using HypSeek, we compare max pooling, weighted pooling, and top-1 pocket selection before ligand matching with AlloMIT's score aggregation after matching.

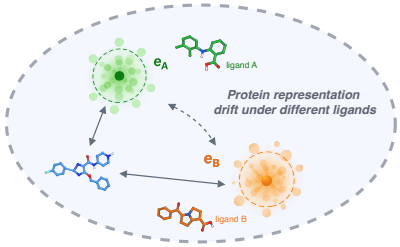


Figure 8: **Anchor drift.** Ligand-conditioned aggregation produces representations $\mathbf{e}_A$ and $\mathbf{e}_B$ of the same protein for ligands A and B. Their different distances to ligand C illustrate how drift can affect score comparability.

Aggregate First or Match First? Inspired by multi-instance learning in pathology, we explored using max pooling to aggregate candidate-pocket representations into a single target representation, which was then matched with the ligand representation. However, max pooling retains only the largest value in each feature dimension, potentially discarding information relevant to pocket-ligand matching. AlloMIT instead matches each candidate pocket with the ligand before aggregating the resulting scores, allowing aggregation to operate on ligand-specific matching evidence. As shown in Table 2, AlloMIT achieves an allosteric BEDROC of 23.25%, compared with 14.94% for the aggregate-first variant.

Can We Trust Early Pocket Selection? Inspired by AANet (Zhu et al., 2025), we also tested two variants using its ligand-conditioned attention mechanism. The first uses the embedding of the pocket with the highest attention score. The second forms the target representation as an attention-weighted average of pocket embeddings. As shown in Table 2, their allosteric BEDROC values are 13.67% and 16.87%, respectively, compared with 23.25% for AlloMIT. Both yield an allosteric $EF_{0.5\%}$ of 20.79, compared with 41.70 for AlloMIT. Predicting the allosteric pocket remains uncertain, and selecting an incorrect pocket can misdirect subsequent ligand screening. Although weighted pooling can reduce the risk of committing to a single pocket, its ligand-dependent weights cause the overall protein representation to drift with the query ligand (Figure 8). Different ligands are therefore scored against different representations of the same target, potentially affecting score comparability. Our AlloMIT instead scores each pocket-ligand pair independently before CSPA determines its contribution to the final prediction. Each candidate pocket retains a fixed representation across query ligands, providing a consistent reference for pocket-ligand scoring.

4.5 Case Study on AKT1

AKT1 is a serine-threonine kinase involved in cell survival and proliferation, making it a target of interest in cancer research. The conserved ATP-binding sites of AKT isoforms make selective inhibition challenging. **However, allosteric hits have the potential to expand the druggability of such targets.** Specifically, Akt-I-1 selectively inhibits AKT1 through a pleckstrin homology (PH)-domain-dependent mechanism, with a reported IC_{50} of 4.6 μM. Its selectivity over AKT2 and AKT3 provides an example of how allosteric inhibition can distinguish closely related kinase isoforms (Barnett et al., 2005). Here, we examined whether AlloMIT could prioritize this known allosteric inhibitor using an AlphaFold-predicted AKT1 structure without a predefined binding pocket.

The complete AKT1 screening library in AlloVSBench contains 17,957 compounds, including orthosteric hits, allosteric hits, and negatives. Zero-shot HypSeek ranked Akt-I-1 at position 1,490 (top 8.30%), whereas AlloMIT ranked it at position 294 (top 1.64%) in the same library.

Moreover, Figure 9 shows the CSPA weights for Akt-I-1. The known allosteric pocket P7 receives the largest weight (34.74%), whereas the orthosteric ATP-binding pocket P10 receives no weight. AlloMIT thus favors an alternative pocket hypothesis without explicit pocket supervision. However, P20 and P18 retain substantial weights of 25.84% and 19.24%. This illustrates a limitation of activity-supervised multi-instance learning. Target-ligand labels guide ligand ranking but do not uniquely identify the binding site.

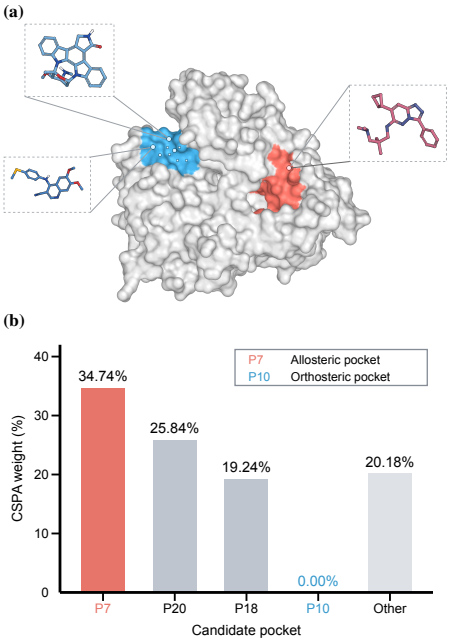


Figure 9: **Pocket contributions for Akt-I-1.** The allosteric inhibitor Akt-I-1 demonstrates **isoform selectivity** for AKT1 over AKT2 and AKT3 despite their conserved ATP-binding sites (Barnett et al., 2005). **(a)** AKT1 with the allosteric and orthosteric regions shown in coral and blue, respectively. **(b)** CSPA weights for Akt-I-1. P7 and P10 correspond to the allosteric and orthosteric pockets; Other combines the remaining candidates.

5 Conclusion

We propose AlloMIT, a parameter-efficient tuning framework that adapts pre-trained virtual screening models for allosteric hit discovery. AlloMIT keeps candidate pockets as competing hypotheses, evaluates their compatibility with each ligand, and then sparsely aggregates the resulting scores. From limited target-ligand activity data, it learns ligand-dependent pocket contributions and subsequently tunes selected representation parameters. We also built AlloVSBench to assess allosteric hit discovery and general screening performance using apo or predicted target structures. AlloMIT improves both the early enrichment of allosteric hits and the general screening performance of HypSeek. These results highlight its potential to accelerate allosteric hit discovery and broaden the search for selective modulators of protein function.

AI USE STATEMENT

Generative AI tools were used to assist with code debugging, language editing, and figure-layout development. All AI-assisted outputs were reviewed by the authors. Reported measurements were produced by the authors' experimental pipeline, and the authors take full responsibility for the final text, figures, code, and scientific claims.

ETHICS STATEMENT

This study uses public molecular and structural data and does not involve human participants, private personal data, or animal experiments. AlloVSBench is intended for methodological evaluation rather than direct clinical decision making. Dataset provenance, filtering, leakage controls, and evaluation exclusions are documented in the paper and appendix.

REPRODUCIBILITY STATEMENT

Section 2 defines the task, model, and adaptation procedure. Section 4.1 provides additional implementation details. Appendix B provides the mathematical formulation, derivations, and implementation details. Appendix C records data construction and leakage audits, and Appendix D reports additional screening results. Checkpoints are selected using training loss alone, without validation or test feedback.

REFERENCES

Stanley F. Barnett, Deborah Defeo-Jones, Sheng Fu, Paula J. Hancock, Kathleen M. Haskell, Raymond E. Jones, Jason A. Kahana, Astrid M. Kral, Karen Leander, Ling L. Lee, John Malinowski, Elizabeth M. McAvoy, Debbie D. Nahas, Ronald G. Robinson, and Hans E. Huber. Identification and characterization of pleckstrin-homology-domain-dependent and isoenzyme-specific Akt inhibitors. *Biochemical Journal*, 385(2):399–408, 2005. doi: 10.1042/BJ20041140.

Matthias R. Bauer, Tamer M. Ibrahim, Simon M. Vogel, and Frank M. Boeckler. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0: A public library of challenging docking benchmark sets. *Journal of Chemical Information and Modeling*, 53(6):1447–1462, 2013. doi: 10.1021/ci400115b.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1.

Ying-Nan P. Chen, Matthew J. LaMarche, Ho Man Chan, Peter Fekkes, Jorge Garcia-Fortanet, Michael G. Acker, Brandon Antonakos, Christine Hiu-Tung Chen, Zhouliang Chen, Vesselina G. Cooke, Jason R. Dobson, Zhan Deng, Feng Fei, Brant Firestone, Michelle Fodor, Cary Fridrich, Hui Gao, Denise Grunenfelder, Huai-Xiang Hao, Jaison Jacob, Samuel Ho, Kathy Hsiao, Zhao B. Kang, Rajesh Karki, Mitsunori Kato, Jay Larrow, Laura R. La Bonte, Francois Lenoir, Gang Liu, Shumei Liu, Dyuti Majumdar, Matthew J. Meyer, Mark Palermo, Lawrence Perez, Minying Pu, Edmund Price, Christopher Quinn, Subarna Shakya, Michael D. Shultz, Joanna Slisz, Kavitha Venkatesan, Ping Wang, Markus Warmuth, Sarah Williams, Guizhi Yang, Jing Yuan, Ji-Hu Zhang, Ping Zhu, Timothy Ramsey, Nicholas J. Keen, William R. Sellers, Travis Stams, and Pascal D. Fortin. Allosteric inhibition of SHP2 phosphatase inhibits cancers driven by receptor tyrosine kinases. *Nature*, 535(7610):148–152, 2016. doi: 10.1038/nature18621.

Bin Feng, Zijing Liu, Hao Li, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. Hierarchical affinity landscape navigation through learning a shared pocket–ligand space. *Patterns*, 6(10):101371, 2025. doi: 10.1016/j.patter.2025.101371.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. DrugCLIP: Contrastive protein–molecule representation learning for virtual screening. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44595–44614, 2023. doi: 10.52202/075280-1932.

Jixiao He, Xinyi Liu, Chunhao Zhu, Jinyin Zha, Qian Li, Mingzhu Zhao, Jiacheng Wei, et al. ASD2023: Towards the integrating landscapes of allosteric knowledgebase. *Nucleic Acids Research*, 52(D1):D376–D383, 2024. doi: 10.1093/nar/gkad915.

Scott A. Hollingsworth, Brendan Kelly, Celine Valant, Jordan Arthur Michaelis, Olivia Mastromihalis, Geoff Thompson, A. J. Venkatakrishnan, Samuel Hertig, Peter J. Scammells, Patrick M. Sexton, Christian C. Felder, Arthur Christopoulos, and Ron O. Dror. Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic GPCRs. *Nature Communications*, 10(1):3289, 2019. doi: 10.1038/s41467-019-11062-7.

James R. Horn and Brian K. Shoichet. Allosteric inhibition through core disruption. *Journal of Molecular Biology*, 336(5):1283–1291, 2004. doi: 10.1016/j.jmb.2003.12.068.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Timothy P. Hughes, Michael J. Mauro, Jorge E. Cortes, Hironobu Minami, Delphine Rea, Daniel J. DeAngelo, Massimo Breccia, Yeow-Tee Goh, Moshe Talpaz, Andreas Hochhaus, Philipp le Coutre, Oliver Ottmann, Michael C. Heinrich, Juan L. Steegmann, Michael W.N. Deininger, Jeroen J.W.M. Janssen, Francois-Xavier Mahon, Yosuke Minami, David Yeung, David M. Ross, Martin S. Tallman, Jae H. Park, Brian J. Druker, David Hynds, Yuyan Duan, Christophe Meille, Florence Hourcade-Potelleret, K. Gary Vanasse, Fabian Lang, and Dong-Wook Kim. Asciminib in chronic myeloid leukemia after ABL kinase inhibitor failure. *New England Journal of Medicine*, 381(24): 2315–2326, 2019. doi: 10.1056/nejmoa1902328.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2127–2136, 2018.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Andrew C. Kruse, Aaron M. Ring, Aashish Manglik, Jianxin Hu, Kelly Hu, Katrin Eitel, Harald Hübner, Els Pardon, Celine Valant, Patrick M. Sexton, Arthur Christopoulos, Christian C. Felder, Peter Gmeiner, Jan Steyaert, William I. Weis, K. Christopher Garcia, Jürgen Wess, and Brian K. Kobilka. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature*, 504 (7478):101–106, 2013. doi: 10.1038/nature12735.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168, 2009. doi: 10.1186/1471-2105-10-168.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

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*, 13:43, 2021. doi: 10.1186/s13321-021-00522-2.

Artur Meller, Michael Ward, Jonathan Borowsky, Meghana Kshirsagar, Jeffrey M. Lotthammer, Felipe Oviedo, Juan Lavista Ferres, and Gregory R. Bowman. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nature Communications*, 14(1):1177, 2023. doi: 10.1038/s41467-023-36699-3.

Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012. doi: 10.1021/jm300687e.

Thin Nguyen, Hang Le, Thomas P. Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021. doi: 10.1093/bioinformatics/btaa921.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. doi: 10.1093/bioinformatics/bty593.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146/.

John Proudfoot, Olivier Nosjean, Jan Blanchard, John Wang, Dominique Besson, Denis Crankshaw, Günter Gauglitz, Robert Hertzberg, Carol Homon, Lyndon Llewellyn, Richard Neubig, Larry Walker, and Pascal Villa. Glossary of terms used in biomolecular screening (IUPAC recommendations 2011). *Pure and Applied Chemistry*, 83(5):1129–1158, 2011. doi: 10.1351/pac-rec-09-05-03.

Anjie Qiao, Zhen Wang, Yaliang Li, Jiahua Rao, and Yuedong Yang. BindCLIP: A unified contrastive-generative representation learning framework for virtual screening. *arXiv preprint arXiv:2602.15236*, 2026. doi: 10.48550/arXiv.2602.15236.

Ferdinandos Skoulidis, Bob T. Li, Grace K. Dy, Timothy J. Price, Gerald S. Falchook, Jürgen Wolf, Antoine Italiano, Martin Schuler, Hossein Borghaei, Fabrice Barlesi, Terufumi Kato, Alessandra Curioni-Fontecedro, Adrian Sacher, Alexander Spira, Suresh S. Ramalingam, Toshiaki Takahashi, Benjamin Besse, Abraham Anderson, Agnes Ang, Qui Tran, Omar Mather, Haby Henary, Gataree Ngarmchamnanrith, Gregory Friberg, Vamsidhar Velcheti, and Ramaswamy Govindan. Sotorasib for lung cancers with KRAS p.G12C mutation. *New England Journal of Medicine*, 384(25): 2371–2381, 2021. doi: 10.1056/nejmoa2103695.

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. LIT-PCBA: An unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling*, 60(9): 4263–4273, 2020. doi: 10.1021/acs.jcim.0c00155.

Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi: 10.1002/jcc.21334.

Jean-François Truchon and Christopher I. Bayly. Evaluating virtual screening methods: Good and bad metrics for the early recognition problem. *Journal of Chemical Information and Modeling*, 47 (2):488–508, 2007. doi: 10.1021/ci600426e.

Jianhui Wang, Wenyu Zhu, Bowen Gao, Xin Hong, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Learning protein–ligand binding in hyperbolic space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):1150–1158, 2026. doi: 10.1609/aaai.v40i2.37086.

Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-Dock: GPU-accelerated docking enables ultralarge virtual screening. *Journal of Chemical Theory and Computation*, 19(11):3336–3345, 2023. doi: 10.1021/acs.jctc.2c01145.

Wenyu Zhu, Jianhui Wang, Bowen Gao, Yinjun Jia, Haichuan Tan, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. AANet: Virtual screening under structural uncertainty via alignment and aggregation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146613–146646, 2025. doi: 10.52202/085713-4405.

A Related Work

Virtual Screening. Structure-based virtual screening prioritizes compounds using three-dimensional information about a protein binding site. Molecular docking methods such as AutoDock Vina search for plausible ligand poses and rank them using scoring functions (Trott & Olson, 2010). Uni-Dock accelerates this process through GPU parallelization, enabling screening of ultralarge compound libraries (Yu et al., 2023). GNINA incorporates convolutional neural network scoring into the docking workflow to evaluate protein-ligand poses (McNutt et al., 2021). Representation alignment provides another approach by encoding pockets and ligands separately and learning a shared space for matching. Ligand representations can then be precomputed and reused across screening tasks, reducing the computation required to rank large libraries (Gao et al., 2023).

DrugCLIP formulates virtual screening as contrastive pocket-ligand retrieval in a shared Euclidean space, without requiring explicit affinity values for each training pair (Gao et al., 2023). To capture finer differences in binding strength, LigUnity extends this retrieval formulation by combining active-inactive discrimination with relative affinity ranking (Feng et al., 2025). HypSeek further explores the geometry of the shared representation space, embedding pockets, ligands, and protein sequences in hyperbolic space to capture hierarchical binding relationships and fine-grained affinity differences (Wang et al., 2026). Across these formulations, pocket-based prediction still depends on the binding region supplied for matching. For apo or predicted structures, the relevant site may be unknown, and geometric pocket detection can produce several plausible candidates. An inaccurate pocket definition can therefore affect ligand ranking even when a model performs well with an annotated binding site (Zhu et al., 2025).

AANet addresses this uncertainty through cavity alignment and ligand-conditioned aggregation (Zhu et al., 2025). It first aligns representations of detected cavities, ligand-bound pockets, and ligands through contrastive learning. A cross-attention module then uses the ligand representation to weight candidate cavity embeddings and form a target representation for matching. Aggregation training also uses auxiliary attention targets derived from structural annotations or pre-trained cavity scores, supporting learning from activity measurements without precise pocket annotations. The resulting target representation varies with the queried ligand. In the allosteric setting studied here, however, target-ligand activity labels do not identify the binding pocket. An inaccurate pocket preference can shift this representation toward an irrelevant candidate before final matching, causing subsequent scores to inherit this bias.

AlloMIT retains separate candidate-pocket representations through ligand matching and uses CSPA to aggregate the resulting scores. For a fixed trained model, each candidate's matching representation is shared across query ligands, while its score and aggregation weight depend on the ligand.

Multiple-Instance Learning. Multiple-instance learning (MIL) is a weakly supervised learning framework in which each sample consists of a bag of instances. Labels are provided for the bag as a whole, without specifying the label of each instance. The model predicts the bag-level outcome by combining information from these instances. In computational pathology, for example, a whole-slide image is represented as a bag of tissue patches, and a slide-level diagnosis provides supervision. Attention-based MIL learns to aggregate instance representations through attention weights for bag-level prediction (Ilse et al., 2018). CLAM combines attention-based aggregation with instance-level clustering to learn discriminative features under slide-level supervision (Lu et al., 2021).

In allosteric screening, candidate pockets form a bag for each target, but activity measurements supervise target-ligand pairs rather than the target alone. Each prediction therefore depends on both the pocket bag and the queried ligand. AlloMIT uses this conditional MIL formulation to learn ligand-dependent contributions from candidate pockets without pocket-level binding annotations. The resulting weights describe contributions to ligand ranking and do not establish physical binding-site identities.

Parameter-Efficient Adaptation. Parameter-efficient fine-tuning (PEFT) adapts pre-trained models by optimizing a small set of trainable parameters while keeping most pre-trained weights fixed. This can be achieved by introducing lightweight modules or selectively updating existing parameters. Adapters introduce trainable bottleneck modules (Houlsby et al., 2019), and LoRA represents weight updates through low-rank matrices (Hu et al., 2022). Prefix tuning learns continuous task-specific

prefixes while freezing the pre-trained model (Li & Liang, 2021). BitFit instead restricts updates to existing bias terms or a subset of them (Ben Zaken et al., 2022).

AlloMIT combines parameter-efficient adaptation with learning to aggregate candidate-pocket evidence under target-ligand supervision. In the HypSeek configuration, warm-up tuning first trains the residual pocket MLP and CSPA with frozen encoders. Joint tuning then updates selected encoder parameters while keeping the learned calibration and aggregation parameters fixed.

B AlloMIT: Mathematical Formulation and Implementation

Target-ligand activity measurements do not specify which candidate pocket mediates binding. AlloMIT addresses this uncertainty by matching each candidate to the query ligand and learning how to combine the resulting scores. We develop this formulation using HypSeek as the pre-trained backbone.

B.1 Representations and Conditional Matching

We fix one target with candidate pockets $\{p_k\}_{k=1}^{K}$ and omit its index. The indices k and j denote a candidate pocket and a query ligand, respectively.

The encoders and projection heads produce pocket features $\mathbf{e}_k = E_{\mathrm{poc}}(p_k)$, ligand features $\mathbf{e}_j^{\mathrm{lig}} = E_{\mathrm{lig}}(\ell_j)$, and a sequence feature $\mathbf{e}^{\mathrm{seq}} = E_{\mathrm{seq}}(S)$ in $\mathbb{R}^d$, with $d = 128$. HypSeek (Wang et al., 2026) scales these features before applying the Lorentz exponential map. For curvature $-c$, where $c > 0$, its spatial component is:

$$\psi_c(\mathbf{v}) = \frac{\sinh(\sqrt{c}\|\mathbf{v}\|_2)}{\sqrt{c}\|\mathbf{v}\|_2}\mathbf{v}, \qquad \psi_c(\mathbf{0}) = \mathbf{0}. \tag{6}$$

For branch $m \in \{\mathrm{poc}, \mathrm{lig}, \mathrm{seq}\}$, write $\phi_m(\mathbf{e}) = \psi_c(a_m\mathbf{e})$, where $a_m > 0$ is the inherited branch-specific scale.

AlloMIT calibrates each pocket feature using the shared residual MLP r_ϕ defined in the main text. The calibrated features and matching representations are:

$$\widetilde{\mathbf{e}}_k = \mathbf{e}_k + r_\phi(\mathbf{e}_k), \qquad \mathbf{h}_k = \phi_{\mathrm{poc}}(\widetilde{\mathbf{e}}_k), \qquad \mathbf{z}_j = \phi_{\mathrm{lig}}(\mathbf{e}_j^{\mathrm{lig}}). \tag{7}$$

Similarly, $\mathbf{z}^{\mathrm{seq}} = \phi_{\mathrm{seq}}(\mathbf{e}^{\mathrm{seq}})$. The pocket-ligand score is $q_{kj} = \mathbf{h}_k^\top \mathbf{z}_j$. These vectors contain the d spatial coordinates of the mapped representations. Matching uses their ordinary dot product, without the time-coordinate term of a full Lorentz inner product.

We learn ligand-dependent weights $\mathbf{w}_j$ from target-ligand activity labels without pocket-level supervision. Their components satisfy $w_{kj} \geq 0$ and $\sum_{k=1}^{K} w_{kj} = 1$. The screening score is:

$$s_j^{\mathrm{poc}} = \sum_{k=1}^{K} w_{kj} q_{kj}, \qquad s_j = s_j^{\mathrm{poc}} + (\mathbf{z}^{\mathrm{seq}})^\top \mathbf{z}_j. \tag{8}$$

For a fixed trained model, the candidate representations $\mathbf{h}_k$ are shared across query ligands. Their scores and aggregation weights depend on the ligand. The weights describe contributions to ranking, not calibrated probabilities of physical binding-site identity.

Aggregation order and matching anchors. The candidate-pocket representations $\mathbf{h}_k$ are shared across query ligands, whereas their aggregation weights are ligand-dependent. Since matching uses an ordinary dot product, the pocket score admits an equivalent representation:

$$s_j^{\mathrm{poc}} = \sum_{k=1}^{K} w_{kj} \mathbf{h}_k^\top \mathbf{z}_j = \bar{\mathbf{h}}_j^\top \mathbf{z}_j, \qquad \bar{\mathbf{h}}_j = \sum_{k=1}^{K} w_{kj} \mathbf{h}_k. \tag{9}$$

This identity holds even when the weights depend on the matching scores themselves. Score aggregation therefore admits an equivalent aggregate representation $\bar{\mathbf{h}}_j$, which can vary across ligands. Each candidate-specific matching anchor $\mathbf{h}_k$, however, is shared across query ligands, while

its score and contribution to the final prediction are ligand-dependent. The consistent reference discussed in the main text refers to these candidate-specific matching representations, rather than a query-independent aggregate.

The distinction becomes consequential when aggregation precedes the nonlinear pocket mapping. To isolate the effect of this ordering, consider the same calibrated pocket features $\widetilde{\mathbf{e}}_k$, ligand representation $\mathbf{z}_j$, and normalized weights $\mathbf{w}_j$ in both paths. Writing $F = \phi_{\mathrm{poc}}$, the corresponding matching anchors are

$$\mathbf{v}_j = F\left(\sum_{k=1}^{K} w_{kj}\widetilde{\mathbf{e}}_k\right), \qquad \bar{\mathbf{h}}_j = \sum_{k=1}^{K} w_{kj}F(\widetilde{\mathbf{e}}_k). \tag{10}$$

Here, $\mathbf{v}_j$ results from aggregation before mapping, while $\bar{\mathbf{h}}_j$ results from mapping each candidate separately before aggregation. The nonlinear mapping generally prevents these two operations from commuting.

For two query ligands i and j, the first path exhibits a change in its matching anchor whenever their weighted pocket features differ. Specifically, the mapping F preserves direction and transforms the input norm through a strictly increasing function,

$$\|F(\mathbf{x})\|_2 = \frac{\sinh(\sqrt{c}\, a_{\mathrm{poc}}\|\mathbf{x}\|_2)}{\sqrt{c}}. \tag{11}$$

It is therefore injective, giving

$$\mathbf{v}_i \neq \mathbf{v}_j \quad \Longleftrightarrow \quad \sum_{k=1}^{K}(w_{ki} - w_{kj})\widetilde{\mathbf{e}}_k \neq \mathbf{0}. \tag{12}$$

In early aggregation, ligand-dependent weights construct a feature aggregate before the nonlinear mapping F. We refer to the resulting query-dependent change in the pre-matching target representation $\mathbf{v}_j$ as *anchor drift*. Different weight vectors need not produce drift if they yield the same aggregate feature.

The effect of aggregation order on scoring can be characterized directly. Define

$$\boldsymbol{\delta}_j = \mathbf{v}_j - \bar{\mathbf{h}}_j, \qquad s_j^{\mathrm{early}} = \mathbf{v}_j^\top \mathbf{z}_j, \qquad s_j^{\mathrm{late}} = \bar{\mathbf{h}}_j^\top \mathbf{z}_j. \tag{13}$$

Their score difference is exactly

$$s_j^{\mathrm{early}} - s_j^{\mathrm{late}} = \boldsymbol{\delta}_j^\top \mathbf{z}_j \equiv \epsilon_j. \tag{14}$$

Consequently, a representation difference affects the score only through its component along the ligand representation. For two ligands, the corresponding ranking margin changes as

$$s_i^{\mathrm{early}} - s_j^{\mathrm{early}} = s_i^{\mathrm{late}} - s_j^{\mathrm{late}} + \epsilon_i - \epsilon_j. \tag{15}$$

An identical sequence-branch contribution can be added to both paths without changing these difference identities. The ranking-margin identity specifies how aggregation order can affect comparisons across ligands: differential score changes can alter their relative ranking. Anchor variation alone does not establish that scores are incomparable, and an ordering change need not reduce screening accuracy; its effect depends on the activity labels. AlloMIT retains a shared set of candidate-specific matching anchors across query ligands. Ligand-dependent aggregation weights determine how their individual matching scores contribute to the final prediction, preserving the consistent pocket-level reference described in the main text.

B.2 Contextual Sparse Pocket Aggregation

To determine each candidate's contribution, CSPA combines its context within the pocket set with its matching score for the query ligand. For the contextual branch, we normalize the uncorrected pocket features as $\mathbf{x}_k = \mathrm{LN}(\mathbf{e}_k)$, without trainable affine parameters. For feature dimension $r = 1, \ldots, d$, let x_{kr} denote the corresponding component of $\mathbf{x}_k$. We compute the mean and standard deviation across candidate pockets as:

$$\mu_r = \frac{1}{K}\sum_{k=1}^{K} x_{kr}, \qquad \sigma_r = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_{kr} - \mu_r)^2 + \epsilon}. \tag{16}$$

The vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ collect these feature-wise statistics, with $\epsilon = 10^{-5}$ for numerical stability. The shared gated scorer g_ψ, defined in Appendix B.4, receives a $4d$-dimensional context vector:

$$\mathbf{c}_k = [\mathbf{x}_k,\ \mathbf{x}_k - \boldsymbol{\mu},\ \mathbf{x}_k \odot \boldsymbol{\mu},\ \boldsymbol{\sigma}], \qquad b_k = g_\psi(\mathbf{c}_k). \tag{17}$$

Brackets denote concatenation and $\odot$ denotes element-wise multiplication. The structural logits $\mathbf{b} = (b_1, \ldots, b_K)^\top$ are shared across ligands. They are unconstrained scores, rather than normalized probabilities.

For each ligand, we standardize its matching scores across candidate pockets:

$$\bar{q}_j = \frac{1}{K}\sum_{k=1}^{K} q_{kj}, \qquad \sigma_{q,j} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(q_{kj} - \bar{q}_j)^2 + \epsilon}, \qquad \widehat{q}_{kj} = \frac{q_{kj} - \bar{q}_j}{\sigma_{q,j}}. \tag{18}$$

Standardized scores determine the weights, while the aggregation retains the original matching scores.

The weighting logits combine the two signals at a fixed temperature $\tau > 0$:

$$\mathbf{u}_j = \frac{\mathbf{b} + \widehat{\mathbf{q}}_j}{\tau}, \qquad \mathbf{w}_j = \mathrm{entmax}_{1.5}(\mathbf{u}_j). \tag{19}$$

The 1.5-entmax operator (Peters et al., 2019) produces sparse simplex weights with the threshold form:

$$w_{kj} = \left[\frac{u_{kj} - \lambda_j}{2}\right]_+^2, \qquad \sum_{k=1}^{K} w_{kj} = 1. \tag{20}$$

Here $[a]_+ = \max(a, 0)$, and λ_j is chosen to satisfy the unit-sum constraint. A candidate receives zero weight when $u_{kj} \le \lambda_j$. It then contributes no direct term to s_j^{poc}, but can still affect other weights through the full-bag score normalization and contextual statistics. The number of nonzero weights depends on the logits, and all candidates can receive nonzero weights. At deterministic inference, permuting the candidates permutes their weights and leaves the aggregated score unchanged.

B.3 Training Objective

The alignment objective combines target-ligand matching with activity ranking within each assay. It supervises the aggregated pocket-ligand scores and the sequence-ligand scores in separate branches, as specified in Equation (5). We additionally apply the distance and angular penalties from HypSeek (Wang et al., 2026) to sequence-ligand representations.

A batch contains B sampled target-assay groups, with n_i ligands and activity values a_{ij} in group i. The sampler redraws ligand subsets and their order at each epoch. The index $j = 1, \ldots, n_i$ denotes a ligand's position in this sampled order, rather than its activity rank. For either branch, $t_{i,aj}$ denotes the score between target i and ligand j from group a, multiplied by the inherited, frozen logit scale.

Let $\mathcal{N}_i$ contain the cross-group ligand negatives for target i, and let $\mathcal{T}_{ij}$ contain its own target i and the target negatives for ligand (i, j). Known conflicting pairs are excluded from the negatives. The target-to-ligand anchors $\mathcal{A}_i$ satisfy $a_{ij} \ge 5$ in groups with multiple ligands, while singleton groups retain their sole anchor. The two contrastive directions are:

$$\begin{aligned}
\mathcal{L}_{\mathrm{lig}\to\mathrm{tar}} &= -\sum_i \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \log \frac{\exp(t_{i,ij})}{\sum_{a\in\mathcal{T}_{ij}} \exp(t_{a,ij})}, \\
\mathcal{L}_{\mathrm{tar}\to\mathrm{lig}} &= -\sum_i \frac{1}{\sqrt{n_i}} \sum_{j\in\mathcal{A}_i} \log \frac{\exp(t_{i,ij})}{\exp(t_{i,ij}) + \sum_{(a,r)\in\mathcal{N}_i} \exp(t_{i,ar})}.
\end{aligned} \tag{21}$$

Within each group, activity ranking compares an anchor with ligands whose activity is lower by more than $\log_{10} 3$ units. With $\mathcal{R}_{ij} = \{j\} \cup \{r : a_{ir} < a_{ij} - \log_{10} 3\}$, the ranking term is:

$$\mathcal{L}_{\mathrm{rank}} = -\sum_{i:n_i>2} \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i-1} \frac{1}{\log(j+1)} \log \frac{\exp(t_{i,ij})}{\sum_{r\in\mathcal{R}_{ij}} \exp(t_{i,ir})}. \tag{22}$$

The factor $1/\log(j+1)$ weights sampled positions, while activity values define the comparison set $\mathcal{R}_{ij}$. The final sampled ligand is omitted as a ranking anchor but can still serve as a comparison ligand. Each branch combines the matching and ranking terms as:

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{lig}\to\mathrm{tar}} + \mathcal{L}_{\mathrm{tar}\to\mathrm{lig}} + \mathcal{L}_{\mathrm{rank}}. \tag{23}$$

B.4 Module Implementation

The residual pocket MLP calibrates each candidate feature before matching. It computes the correction as:

$$r_\phi(\mathbf{e}) = W_2 \,\mathrm{Dropout}(\mathrm{ReLU}(W_1\mathbf{e} + \mathbf{a}_1)) + \mathbf{a}_2. \tag{24}$$

The hidden width is 512 and the dropout probability is 0.05. We initialize W_2 and $\mathbf{a}_2$ to zero so that the initial correction leaves the pocket features unchanged.

The contextual scorer maps each candidate's context vector to a scalar score through two hidden branches of width 32:

$$g_\psi(\mathbf{c}) = \mathbf{v}^\top[\tanh(W_v\mathbf{c} + \mathbf{a}_v) \odot \mathrm{sigmoid}(W_g\mathbf{c} + \mathbf{a}_g)]\,. \tag{25}$$

We initialize $\mathbf{v}$ to zero, making the initial contextual scores zero. The initial aggregation weights therefore depend only on relative matching scores and need not be uniform.

C AlloVSBench Construction and Audits

Combining allosteric activity data with existing screening libraries requires consistent target and ligand identities. We harmonize data from these sources to construct target-disjoint tuning and test sets, then select ligand-free structural inputs and generate candidate-pocket sets.

C.1 Data Sources and Target Partition

We begin with 82 protein targets from the H2L subset of ASD2023, which provides allosteric compounds and their measured activities (He et al., 2024). We map target identifiers to canonical UniProt accessions to determine their overlap with PocketAffDB (Feng et al., 2025) and the DUD-E, DEKOIS 2.0, and LIT-PCBA screening benchmarks (Mysinger et al., 2012; Bauer et al., 2013; Tran-Nguyen et al., 2020).

Based on this target overlap, we select 19 targets shared with the screening benchmarks for the test set. The corresponding benchmark libraries provide the negatives and benchmark-derived hits, which are supplemented with allosteric compounds from ASD2023. Among the remaining targets, those represented in PocketAffDB form the candidate pool for tuning. These candidates undergo further structural, reference-site, and activity checks before inclusion in the tuning set. This target-level partition keeps the tuning and test targets disjoint.

C.2 Tuning Set Construction

Within this candidate pool, we retain targets that pass structural and reference-site checks and have a qualified ligand-free structural input. Each target must also retain at least two distinct allosteric compounds with $pActivity \geq 5$ after ligand-label reconciliation. These criteria yield a tuning set of 28 targets.

The resulting tuning set comprises 1,243 activity measurements across 29 assays. The activity threshold determines target eligibility, while the measured activity values are retained as continuous supervision. Accordingly, the 31 measurements below the threshold remain in the tuning set.

C.3 Test Set Construction and Overlap Analysis

For the held-out test targets, we assemble screening libraries from 25 source panels in the LigUnity benchmark collection, comprising 14 DUD-E, nine DEKOIS, and two LIT-PCBA panels across 19 targets. The benchmark libraries provide the negatives and benchmark-derived hits, which are supplemented with representative allosteric compounds from ASD2023 H2L.

Before combining these libraries, we reduce molecules to their principal organic parent fragments and represent them using canonical isomeric SMILES. Formal charge and stereochemistry are retained, and tautomers are not normalized. Within each target, duplicates with the same label are merged while retaining source provenance. Benchmark active labels take precedence over negative labels. ASD representatives overlapping with the negatives receive allosteric labels, while those already

labeled as benchmark actives retain their benchmark category. These rules define mutually exclusive evaluation labels.

The resulting test set contains 548,542 unique molecules and 794,668 target-ligand pairs. These pairs comprise 3,742 benchmark-derived orthosteric positives, 25 allosteric positives, and 790,901 negative pairs (Table 3). Each target has one to three allosteric positives.

After harmonization, the tuning and test sets share no UniProt targets or identical canonical protein sequences. Seventeen molecules occur in both sets across different targets, accounting for 0.0031% of the unique test molecules, but no target-ligand pair is shared. The split separates exact target identities without imposing homology, scaffold, or molecule disjointness.

Table 3: **Composition of the tuning and test sets.**

Statistic	Tuning	Test
Targets	28	19
Total samples	1,243	794,668
Orthosteric hits	—	3,742
Allosteric hits	1,212	25
Negatives	—	790,901

C.4 Ligand-Free Structure Selection

We select ligand-free structural inputs that preserve the reference binding region, using qualified experimental structures where available and AlphaFold predictions otherwise. For tuning targets, reference residues are derived from allosteric holo structures and mapped to UniProt. For test targets, benchmark pocket heavy atoms are mapped to source-structure residues and then to UniProt. A test reference is retained when at least 85% of its source atoms are recovered and at least three target residues are identified. All 25 source sites pass these checks and are consolidated by target.

Using these reference residues, we require candidate structures to cover at least 80% of each reference site and at least three mapped residues. For test structures, both conditions are checked before and after residue transfer. The ligand-absence checks differ between the two sets. Tuning structures are rejected if they contain organic ligands, whereas test structures are rejected if any organic-ligand heavy atom lies within 6 Å of the mapped site atoms. The test source holo structures are excluded, including versions with their ligands removed. Qualified test candidates are ranked by site coverage, covered-residue count, and resolution. Selected inputs retain only protein heavy atoms.

This procedure provides independent experimental structures for 17 test targets. The remaining two, P10275 and P31749, use AlphaFold Protein Structure Database predictions (model version 6) because no experimental candidate passes the selection criteria. Reference-site information is used for structural curation and subsequent coverage evaluation, but is not supplied to candidate pocket generation or the screening model.

C.5 Candidate-Pocket Bags and Target-Coverage Audit

Each selected experimental or predicted structure is processed with three chain-aware fpocket configurations. Their candidates are merged within each canonical UniProt target and greedily deduplicated. Two candidates are assigned to the same cluster only when their residue-set Jaccard similarity is at least 0.70 and their center distance is at most 4 Å. The cluster representative is selected by fpocket score, followed by druggability score and a deterministic source-rank tie-break.

For each remaining candidate, we compute target-wise empirical percentiles for fpocket score, druggability score, pocket volume, and the number of supporting fpocket configurations. The composite score is the equal-weight sum of these four percentiles. Candidates are ranked deterministically within each target, and the top K candidates form its pocket bag, with all candidates retained when fewer are available. No reference-site annotation or ligand information is accessed during candidate generation, deduplication, or composite ranking.

For the offline coverage audit, a candidate pocket matches a reference site when they share at least three UniProt-indexed residues, these shared residues account for at least 25% of each residue set, and their centers are at most 8 Å apart. The candidate center is the mean of the protein heavy-atom coordinates belonging to its fpocket residues. The reference-site center is the mean of its Cα coordinates on the candidate chain. A pair is treated as unmatched when this reference center cannot be computed. Residue-set Jaccard similarity is recorded for diagnostics but is not part of the site matching decision.

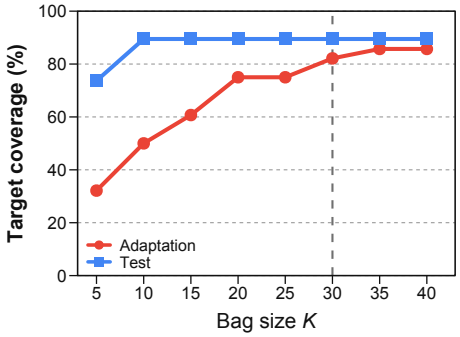


Figure 10: **Target coverage versus bag size.**

We report target coverage: the proportion of targets for which at least one reference site is matched by a candidate among the top K pockets. Each target is counted once, irrespective of how many reference sites are matched. Adaptation targets are evaluated against allosteric reference sites; test targets are evaluated against both orthosteric and allosteric reference sites. Thus, test target coverage indicates recovery of at least one site of either type, rather than coverage of every site or of allosteric sites alone. Reference-site information is accessed only after the candidate ranking has been frozen.

At $K = 30$, the candidate bags cover 23 of 28 adaptation targets (82.1%) and 17 of 19 test targets (89.5%). Increasing the bag size to $K = 40$ covers one additional adaptation target and no additional test target.

D Additional Screening Results

Table 4 complements the general and allosteric results in Table 1 with orthosteric screening results. The model configurations and checkpoints follow the main table. Allosteric hits are excluded from this evaluation, and metrics are macro-averaged over the 19 test targets. BEDROC uses $\alpha = 80.5$.

Table 4: **Orthosteric screening performance on AlloVSBench.** AUROC and BEDROC are reported as percentages. Bold values indicate column-wise best results.

		Orthosteric				
Model	Setting	AUROC	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$
Uni-Dock	Zero-shot	58.54	4.78	2.15	2.24	1.85
DeepDTA	Zero-shot	52.20	4.05	2.82	1.86	1.57
GraphDTA	Zero-shot	66.90	22.10	15.34	12.00	5.28
BindCLIP	Zero-shot	55.65	15.03	10.21	8.13	3.67
DrugCLIP	Zero-shot	63.52	19.45	16.26	13.01	4.62
DrugCLIP*	Tuned	68.90	26.98	22.60	17.08	6.12
LigUnity	Zero-shot	70.19	23.97	15.28	13.08	5.46
LigUnity*	Tuned	76.62	28.11	17.82	16.53	6.02
AANet	Zero-shot	59.57	21.42	15.09	12.72	4.27
AANet†	Tuned	50.80	4.34	2.65	3.16	1.62
$HypSeek_{poc}$	Zero-shot	84.10	44.53	27.87	24.58	10.32
$HypSeek_{poc}$*	Tuned	78.68	40.10	24.18	22.34	9.60
$HypSeek_{seq}$	Zero-shot	83.59	47.33	30.77	26.91	10.46
$HypSeek_{seq}$*	Tuned	79.84	45.96	30.27	26.20	9.87
HypSeek	Zero-shot	**85.36**	48.89	30.56	27.15	11.10
HypSeek*	Tuned	81.11	47.43	29.12	27.00	10.64
+ AlloMIT (Ours)	Tuned	80.67	**51.29**	**31.99**	**29.10**	**11.29**

Note: Superscripts follow Table 1.

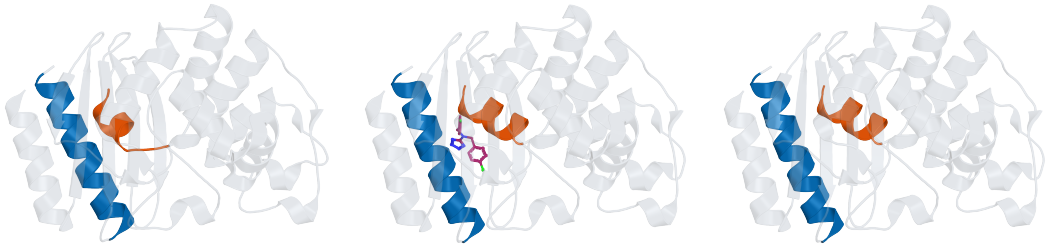
Table 5: **Experimental results for HypSeek and its variants.**

		General				Allosteric			
Strategy	Tuned	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$	BEDROC	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{5\%}$
$\text{HypSeek}_{\text{poc}}$									
Zero-shot	—	44.44	27.68	24.41	**10.20**	12.38	20.89	**15.72**	6.31
LoRA	L	**44.81**	**28.67**	**24.92**	10.19	14.21	**31.22**	**15.72**	6.31
Adapter	L	44.52	28.02	24.73	10.06	14.55	20.89	**15.72**	**7.35**
LayerNorm	L	44.44	27.68	24.41	10.20	12.48	20.89	**15.72**	6.31
Projection head	L	40.05	24.04	22.19	9.48	**15.55**	31.21	15.72	4.21
Joint	$L_{\text{H}} + P_{\text{H}}$	37.51	21.61	20.66	9.01	12.99	20.79	15.72	4.21
Sequential	$L_{\text{H}} \rightarrow P_{\text{H}}$	37.43	22.76	20.67	8.92	14.35	31.21	15.72	4.21
$\text{HypSeek}_{\text{seq}}$									
Zero-shot	—	47.22	30.51	26.68	**10.37**	12.27	20.79	10.46	7.36
LoRA	L	47.16	30.65	26.51	10.32	13.58	20.79	10.46	8.41
Adapter	L	**47.77**	**30.70**	**27.44**	10.30	13.66	20.79	10.46	**8.92**
LayerNorm	S	47.22	30.54	26.68	**10.37**	12.30	20.79	10.46	7.36
Projection head	S	46.70	30.15	26.25	9.93	15.55	**31.25**	**15.72**	7.36
Joint	$L_{\text{A}} + S_{\text{H}}$	45.87	30.00	26.01	9.80	**15.87**	20.79	**15.72**	7.36
Sequential	$S_{\text{H}} \rightarrow L_{\text{A}}$	45.24	29.66	25.69	9.75	14.85	20.79	**15.72**	7.36
HypSeek									
Zero-shot	—	48.80	30.31	26.94	**10.98**	14.94	20.89	**20.95**	6.84
LoRA	L	**48.90**	30.05	**27.30**	10.94	16.76	10.47	**20.95**	7.89
Adapter	P	48.42	**30.65**	26.98	10.86	15.41	**31.21**	15.72	7.89
LayerNorm	L	48.80	30.31	26.90	10.98	14.97	20.89	**20.95**	6.84
Projection head	S	48.51	29.77	26.83	10.98	16.53	10.47	**20.95**	**8.41**
Joint	$L_{\text{A}} + P_{\text{H}} + S_{\text{H}}$	47.36	28.89	26.80	10.56	**17.09**	20.79	20.94	7.89
Sequential	$S_{\text{H}} \rightarrow P_{\text{H}} \rightarrow L_{\text{A}}$	47.77	29.09	27.02	10.67	17.04	20.79	20.94	6.83

Note: L, P, and S denote the ligand, pocket, and sequence branches. H and A denote projection-head tuning and adapters. Arrows indicate sequential tuning order. Given the limited allosteric training data, we use the entire tuning set for adaptation without holding out a validation subset. For each experimental configuration, we select the checkpoint with the lowest training loss.

E Apo versus Holo Structural Inputs

The choice of protein structure determines which binding-site information is available when screening begins. A holo complex records a protein conformation observed with a bound ligand. Removing that ligand deletes its coordinates, but leaves the protein in the conformation resolved in the complex. If a screening pocket is then extracted around the original ligand position, the input also specifies where binding occurred. Pocket geometry and site identity thus provide distinct sources of prior information, and deleting the ligand atoms alone does not remove either cue.



(a) **Apo structure.** (b) **Holo structure.** (c) **Ligand-removed holo.**

Figure 11: **Removing the ligand preserves the open holo pocket. (a)** The apo structure of TEM-1 β-lactamase (PDB 1JWP) presents a closed α11-α12 interface. **(b)** The CBT-bound holo structure (PDB 1PZO) presents an open interface, with the larger displacement occurring in α11. **(c)** Removing CBT from the holo structure leaves this open geometry intact. The protein is gray, α11 is orange, and α12 is blue. CBT is shown as sticks with magenta carbon atoms in (b). All three views use the same orientation.

What ligand removal retains. The three views of TEM-1 β-lactamase in Figure 11 make this distinction concrete (Horn & Shoichet, 2004). The apo structure has a closed α11-α12 interface, whereas the CBT-bound holo structure has an open interface, with the larger displacement occurring in α11. Panel (c) shows the holo structure after CBT removal while retaining the protein coordinates. The ligand is absent, but the open pocket remains. This operation changes which atoms are supplied to the model without reconstructing the unbound conformation. The comparison does not establish whether CBT induces the open state or stabilizes a state that already exists in the protein ensemble. In either case, the holo structure records a conformation observed to support binding.

That retained conformation changes the information available for screening. A model given the ligand-removed holo structure can evaluate compounds against a pocket geometry already observed in a bound complex. When the pocket is also centered on the original ligand, the model does not need to resolve its location among alternative sites. Both cues can remain available even when the co-crystallized ligand is hidden and differs from every screened compound. Their presence therefore concerns the structural information inherited from a known complex, rather than direct access to the screened compounds' bound poses or activity labels.

Structural inputs in AlloVSBench. To reduce dependence on this information, AlloVSBench excludes the source holo structures used by the test benchmarks, including copies with their ligands removed. Each target is instead assigned an independent experimental structure that passes the ligand-absence and site-coverage checks, or an AlphaFold prediction when no experimental candidate qualifies. Reference sites are used for these structural checks and for a separate coverage audit after the candidate ranking is fixed. Candidate-pocket generation and ranking operate on the selected protein structure without access to those references. Neither the reference-site annotations nor the co-crystallized ligand positions are supplied to the screening model.

This separation leaves the contribution of each candidate pocket unresolved when ligand ranking begins. It does not guarantee that the selected structure contains every binding-competent conformation. A static apo or predicted structure may omit a transient pocket, as the closed TEM-1 interface illustrates. AlloVSBench therefore evaluates screening from these available structural inputs under candidate-pocket uncertainty, while keeping the conformational limitations of those inputs explicit.

F Computational Cost

We report screening wall time and the one-time cost of adapting AlloMIT. The screening comparison with Uni-Dock covers the same 19 targets and 794,668 target-ligand pairs.

F.1 Inference Cost

The timed AlloMIT run starts from prepared ligand and candidate-pocket inputs and protein sequences, with no precomputed neural embeddings. Ligand, pocket, and sequence representations are computed during the run, including the forward pass through the sequence encoder, whose LayerNorm parameters are updated during joint tuning. The recorded interval includes model initialization, checkpoint loading, encoding, CSPA aggregation, scoring, and prediction output. It excludes upstream structure preparation and pocket detection, as well as downstream metric computation and quality checks.

Uni-Dock v1.2.0 uses Vinardo scoring in fast mode with an exhaustiveness of 128. Its timing starts from prepared receptor and ligand PDBQT files and sums the wall times of 1,563 active batches, including docking and associated batch handling. AlloMIT's wall time is estimated from inference-log timestamps. These records give 3.06 minutes for AlloMIT and 7.39 hours for Uni-Dock, corresponding to an approximately 145-fold difference in observed throughput (Table 6). The runs do not establish a controlled comparison under identical hardware and software settings, and the historical Uni-Dock GPU model was not recorded. These timings therefore provide an empirical runtime reference for the recorded configurations.

F.2 ADAPTATION COST

The timed AlloMIT configuration uses ligand adapters and both pocket and sequence LayerNorm tuning. Its tuning log records an NVIDIA GeForce RTX 4080 GPU. Warm-up tuning updates CSPA and the residual pocket MLP with fixed encoders and takes 8.41 minutes. Joint tuning then updates the ligand adapters and the pocket and sequence LayerNorm parameters while keeping CSPA and the residual pocket MLP fixed, taking a further 20.25 minutes (Table 7).

The two stages require 28.66 minutes in total. Adding the first screening run on the 19-target panel gives 31.72 minutes for adaptation and screening from prepared inputs. This adaptation cost is incurred once per adapted checkpoint and can be amortized over subsequent screening runs that reuse it. Those runs require only inference, with wall time depending on the screening workload and execution settings.

Table 6: **Recorded screening cost.** Throughput counts target-ligand pairs per second.

Method	Wall time	Pairs/s
AlloMIT	3.06 min	4,328
Uni-Dock	7.39 h	29.9

Table 7: **One-time adaptation cost.** Times correspond to the configuration evaluated above.

Stage	Time (min)
Warm-up tuning	8.41
Joint tuning	20.25
Total	28.66