Site4Drug: Predicting Drug-Binding Target Sites with an AI Agent

Taehan Kim[†1] **Sarrah Rose Mikhail Leung**[1] **Bharat Mekala**[1] **Jeongbin Park**[†2]

Abstract

Selecting *where* to intervene on a protein (i.e., choosing a **targetable site**) is often a more ambiguous and failure-prone bottleneck than selecting *what* binds, especially for membrane proteins where accessibility, topology, and post-translational modifications (PTMs) constrain actionable regions. We present **Site4Drug**, a modality-aware site-finding agent that outputs a ranked list of **targetable regions** with explicit constraints, evidence summaries, risk flags, and a traceable decision log.

Rather than requiring users to specify the drug modality upfront, Site4Drug can **recommend a binding modality** (e.g., antibody/peptide-like vs small-molecule) from the same evidence used for site discovery, including topology, hydropathy, PTM propensity, disulfides, domain context, and sequence. Importantly, this evidence is applied consistently across modalities, including small-molecule pocket discovery, to avoid selecting chemically plausible but biologically occluded sites.

1. Introduction

Many discovery pipelines assume a binding site is known and focus on docking, screening, or binder generation. For example, this includes models like BoltzGen (Stark et al., 2025) for binder generation and CLIP-style pocket scoring models (e.g., DrugCLIP (Gao et al., 2023), BindCLIP (Qiao et al., 2026)) that evaluate candidates given a predefined site.

In practice, teams often stall earlier: *where and with what modality should we target on the protein?* For membrane proteins this is especially difficult: only certain regions are physically accessible, topology predictions can disagree, and PTMs (notably glycosylation) can occlude or destabilize candidate epitopes. When downstream screening fails, teams may not know whether the binding model failed or the *site choice* was wrong, partly because the reasoning behind site selection is rarely logged or can be arbitrary.

Curated resources can help, but they are not absolute ground truths for therapeutic site selection. For example, IEDB (Vita et al., 2025) aggregates immune epitope evidence from highly specialized experimental settings (e.g., T cell and B cell assays) that are context-dependent and inherently incomplete; critically, these assays were not designed to exhaustively enumerate *drug-actionable* binding sites on proteins.

On the small-molecule side, "known pockets" are often operationally defined by taking residues contacting a co-crystallized ligand and padding by a fixed neighborhood. Such definitions can be brittle: a limited number of known ligand-bound structures leaves the pocket undefined (no prior discoveries). This is particularly important for cancer neoantigens and mutated viral antigens. Together, these limitations motivate epitope/pocket discovery that is grounded in feasibility constraints along with reasoning traces.

Conventional approaches such as fpocket (Le Guilloux et al., 2009) and RAPID-Net (Balytskyi et al., 2025) rely on geometric features or learned structural representations, and thus can primarily be applied to the identification of canonical binding pockets. As a result, they are limited in their ability to incorporate diverse metadata or to define sites for alternative drug modalities. Recently, large language models (LLMs) have been increasingly utilized in tasks such as protein structure prediction and drug-genomic response prediction. Moreover, their ability to flexibly integrate diverse forms of unstructured data suggests that they should be considered for adoption in site definition tasks.

Additionally, LLM-based AI agentic systems are being introduced into AI-driven scientific discovery research (e.g., AI-Descartes (Cornelio et al., 2023), Theorizer (Jansen et al., 2026), AutoDiscovery (Agarwal et al., 2026), The AI Scientist (Lu et al., 2026), AI co-scientist (Gottweis et al., 2026), The Virtual Lab (Swanson et al., 2024)), because they require repeated execution of LLMs to achieve reproducible scientific conclusions. Therefore, we explore the potential of drug site detection as an LLM agentic system.

[†]Corresponding authors. [1]UC Berkeley [2]University of Michigan. Correspondence to: Taehan Kim <terry.kim@berkeley.edu>, Jeongbin Park <jeongbp@umich.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

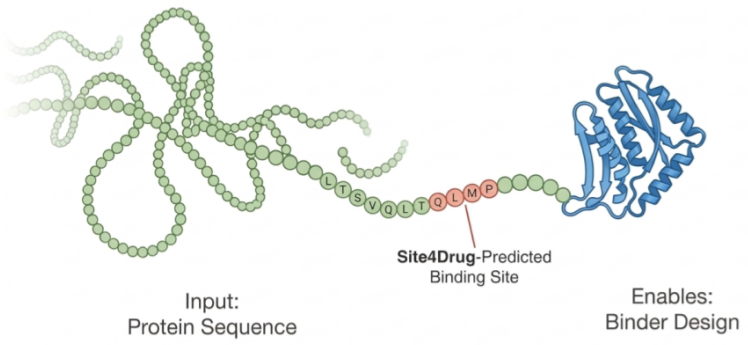


Figure 1. Site4Drug discovers a potential binding region for binder design.

Thus, Site4Drug attempts to provide a solution for the upstream bottleneck. It reframes site selection as a **constraint-first, evidence-integrated decision problem**. Given a protein sequence, it proposes candidate regions, ranks them, and emits a structured report: what constraints were applied, what evidence supported each candidate, and what risks remain. This makes the site-selection step auditable and debuggable. With these results, third-party tools can also be invoked as part of an agentic system as summarized in Figure 1.

2. Problem Setup

2.1. Inputs and Outputs

Input. A target protein specified by an amino-acid sequence $x_{1:L}$.

Output. Site4Drug returns a structured, auditable site-selection result consisting of:

- **Recommended binding modality:** a categorical recommendation $\hat{m} \in \{\text{epitope}, \text{pocket}, \text{other}\}$ with an evidence-backed rationale and uncertainty notes.[1]
- **Ranked candidates with scores:** a list of K candidates $\{r_k\}_{k=1}^{K}$ with per-candidate score $S(r_k)$:
  - **Accessibility/topology label:** a coarse label derived from hydropathy and transmembrane (TM) overlap;
  - **Evidence summary:** key signals supporting or penalizing the region (hydropathy, TM overlap, typed PTM masks, motif hits, cysteine/disulfide context);
  - **Flags:** typed annotations including *risk flags* (e.g., "TM-overlap", "PTM-overlap", "glyco-mask-overlap", "disulfide-constrained", "motif-overlap")

A representative output report is shown in the Appendix A, Fig. S1.

2.2. Method Overview: Two Modules

Site4Drug potentially offers two modules:

1. **Module 1 (Main Discovery Engine):** evidence extraction → constraint-first candidate generation → scoring/ranking with risk flags and audit log.
2. **Module 2 (Potential Design Handoff):** route top candidates to modality-appropriate downstream design/scoring tools (e.g., peptide-binder generation in epitope mode; pocket-aware ligand scoring in pocket mode).

3. Module 1: Evidence Aggregation and Region Discovery

As summarized in Figure 2, Module 1 constructs a residue- and region-level evidence summary from three signal classes: i) topology/accessibility priors, ii) PTM risk signals, and iii) motif/domain context. These signals are used to guide 1) binding site candidate proposal, 2) validate proposed spans, and 3) attach interpretable risk annotations.

3.1. Evidence Extractors

Given a protein sequence, Site4Drug computes three classes of feasibility-aware signals that are exposed to the proposal

[1] Epitope refers to where antibodies or peptide binders attach, usually on extracellular protruding regions of membrane proteins and often tied to immune responses. Pocket refers to where small molecules bind, typically in intracellular proteins or membrane channels, covering immune as well as general signaling and metabolic pathways. Other covers cases where the evidence suggests non-standard or mixed modes (e.g., ambiguous accessibility or competing feasibility constraints).

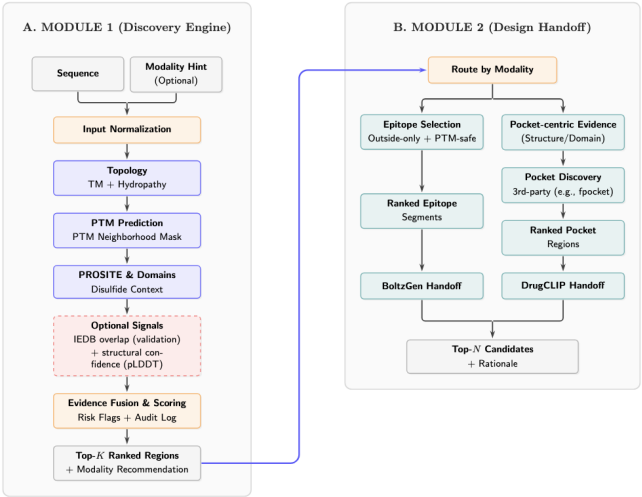


Figure 2. Two-module pipeline: constraint-first discovery (Module 1) followed by modality-specific design handoff (Module 2).

stage and retained for downstream validation.

I. Coarse topology + hydropathy. Site4Drug computes a sliding-window Kyte–Doolittle hydropathy profile (Kyte & Doolittle, 1982) together with a heuristic TM detector to derive a coarse accessibility prior from sequence alone. Regions overlapping detected TM segments are labeled `tmd/restricted`, whereas non-overlapping regions are labeled `outside/exposed`, with confidence determined by the hydropathy margin from the TM threshold.

II. PTM prediction + neighborhood masking. Site4Drug uses PTM calls as feasibility constraints of targetability. PTM sites are obtained from a MusiteDeep-backed extractor (Wang et al., 2020) with rule-based augmentation, and each site is expanded into a typed local mask. We then record whether a candidate overlaps these masks, together with overlap counts by PTM type and local PTM density. For example, if MusiteDeep predicts a phosphoserine/phosphothreonine site at residue 211, we expand it into a typed local mask (e.g., residues 208–214) and record whether a candidate region overlaps that mask.

III. Motif hits + cysteine/disulfide proxy. Site4Drug also queries ScanProsite (De Castro et al., 2006) and records motif spans as cautionary context for targetability. Overlap with a known motif yields a `motif-overlap` annotation, since such regions may be conserved or functionally constrained. We also track cysteine counts as a lightweight proxy for disulfide-constrained segments; regions containing multiple cysteines receive a `disulfide-constrained` flag.

3.2. Candidate Generation and Ranking

Conditioned on the sequence and the compact evidence summary described above, the language model proposes a ranked JSON list of candidate regions directly. After filtering invalid candidates, each is annotated with topology/accessibility evidence, PTM overlap, motif overlap, cysteine count, typed risk flags, and a heuristic score.

3.3. Reranking with Specialist

A specialist panel provides a secondary adjudication layer over the validated candidates. BioAgent, ChemAgent, and RiskAgent receive the same compact evidence summary and return claim → evidence → impact critiques. DecisionAgent then synthesizes these critiques into a final modality decision and an adjusted ranking, while remaining restricted to evidence already present in the context. Representative prompt excerpts and training examples appear in Appendix B.

3.4. Scoring, Risk Flags, and Audit Log

Conceptually, the candidate-ranking logic can be viewed as following a mode-specific heuristic of the form

$$S_0(r) = s_{\text{mode}}(r) - p_{\text{TM}}(r) - p_{\text{PTM}}(r) - p_{\text{motif}}(r),$$

where s_{mode} captures the base preference, while the penalty terms summarize constraint-aware adjustments from TM overlap, PTM-mask overlap, and motif context. In particular, epitope scoring favors polar, non-TM, PTM-light windows, whereas pocket scoring places more weight on hydrophobicity and applies weaker PTM penalties, with small motif-specific adjustments in some cases. In our inference pipeline, the LLM proposes and ranks candidates using this evidence-rich context.

We also compute a separate risk vector $g(r)$ producing typed flags such as TM-overlap, PTM-overlap, glyco-mask-overlap, PTM-dense, disulfide-constrained, hydrophobic-core, and motif-overlap. These flags are logged as interpretable annotations alongside the heuristic score. The detailed rules that determine the flags are noted in Appendix C–E.

4. Experimental Results

Evaluating AI agents for drug-target site discovery is inherently challenging because there is no widely accepted benchmark for this task. In addition, for a novel target-site discovery, there is no definitive "ground truth". We therefore evaluate Site4Drug using a comprehensive, modality-aware validation data rather than a single fixed benchmark.

Our evaluation has three goals: (i) to test whether the predicted regions overlap known *small-molecule pockets*, (ii) to assess whether predicted epitope regions agree with annotated *antibody epitope*s, and (iii) to examine whether predicted sites remain *structurally plausible* even when Site4Drug does not directly use protein structure as input.

Table 1. Overview of the curated datasets used for final reporting

Drug Modality	Count
Small molecule (S)	55
Antibody (A)	26
Antibody / Small molecule (AS)	8
Total	89

To support this evaluation, we curated a dataset spanning pocket, epitope, and mixed-modality targets. Table 1 summarizes the final reporting groups. Here, `Group S` denotes pocket-mode small-molecule validation cases, `Group A` denotes antibody epitope validation cases, and `Group AS` denotes mixed-modality targets that are known to support both pocket and epitope interpretations. This organization reflects the fact that therapeutic site selection is modality-dependent and that different sources provide different levels of supervision.

4.1. Assessment on small-molecule targets (Pockets)

We first evaluated Site4Drug on small-molecule targets for which experimentally determined target–drug co-crystal structures (from methods such as cryo-EM or X-ray crystallography) are available. In this setting, we use deposited structures from public resources (i.e., RCSB) to define approximate ground-truth pocket residues. Specifically, for each target–drug complex, the pocket is defined as the set of residues located within 4 Å of the bound ligand. These residue indices are then remapped from structure numbering to the input FASTA sequence using Needleman–Wunsch alignment. This yields a sequence-level reference site that can be directly compared against the sequence-defined outputs of Site4Drug. Fig. 3a shows examples of co-crystals of EGFR and six different small molecule ligands. This visually shows that our validation data is composed of pocket locations that do not differ significantly among small molecule drugs, making it ideal for evaluation purposes.

Following the procedure above, we initially collected 91 literature-based target–small-molecule pairs with corresponding RCSB entries and retained (`Group S` + `Group AS`) 63 after excluding cases in which a representative FASTA sequence could not be identified reliably. For example, this is the case of multi-chain ambiguity or difficult residue mapping. Examples of representative validation data are shown in Table 2.

Then, we ran Site4Drug on this data and first verified that the curated proteins were classified as pocket. A few targets, including EGFR and HER2, were classified as epitope modalities by Site4Drug. However, these were acceptable cases that are known to have both corresponding antibody and small molecule drugs.[2] Proceeding with actual pocket binding site evaluation, we compared the top-1 and top-5 sites predicted by Site4Drug against the reference pocket residues by performing hypergeometric distribution-based test. The results for cases with p-values below 0.05 are reported in Fig. 3c. We additionally compared Site4Drug against fpocket as a structure-based baseline. Structure-trained models such as RAPID-Net were excluded from the main comparison because they were trained on scPDB, which is derived from RCSB and would therefore introduce a substantial risk of data leakage in this benchmark.

Under this evaluation, Site4Drug achieved pocket-site localization performance comparable to fpocket under general conditions even without direct structure information (Fig. 3c). We also discovered that the statistically significant top-1 results were mainly kinases, which is biologically plausible because kinases often contain characteristic and recurrent small-molecule-accessible pockets (Fig. 3b). For instance, pralsetinib acts as a multi-kinase inhibitor that targets 11 kinases: DDR1, FGFR1, FGFR2, FLT3, JAK1, JAK2, KDR, NTRK1, NTRK3, PDGFRB, and RET (Gonzalez et al., 2025). As expected, fpocket applied directly to ligand-bound RCSB structures achieved near-ceiling performance because the bound ligand is already present and

[2]EGFR is well known for several small-molecule ligands, including afatinib, canertinib, erlotinib, gefitinib, lapatinib, and osimertinib, making it a pocket-mode case. At the same time, it is also targeted by antibodies such as cetuximab, panitumumab, nimotuzumab, and 806, which makes it a natural mixed-modality test case. Meanwhile, HER2 is also associated with both small-molecule drugs (e.g., lapatinib, neratinib, sapitinib, and CI-1033) and antibody drugs (e.g., margetuximab, trastuzumab, disitamab, and pertuzumab).

Table 2. Representative entries from the RCSB target–drug co-crystal dataset. `Site (FASTA)` denotes binding-site residues located within 4 Å of the co-crystallized drug, remapped onto the input FASTA coordinate.

Target	Drug	Mode	PDB ref.	Site (FASTA)	Description
CYP17A1	Abiraterone	pocket	`3RUK`	95, 96, 183, 184, 187, 191, 279, 280, 284, 288, 348, 424, 464	Decreases PSA.
AKT1	Borussertib	pocket	`6HHF`	17, 54, 79, 80, 84, 85, 210, 211, 264, 268, 271, 272, 273, 274, 290, 292, 296, 297	Covalent-allosteric.
KIF11 (Eg5)	Ispinesib	pocket	`4A5Y`	116, 117, 118, 119, 127, 130, 133, 136, 137, 160, 211, 214, 215, 217, 218, 221	Inhibits KSP.

Table 3. Representative target–antibody pairs from the ABCD database with available epitope site annotations.

Target	Antibody	Mode	Epitope site	Reference
`ADAM17`	`MEDI3622`	epitope	`PKAYYSPVGKKNIYLN; 366-381`	`ABCD_AR537`
`malE / MBP`	`alpha-MBP-scFv`	epitope	`ELAKKFEKDTGIKV; 51-61` *(not including ELA at the front)*	`ABCD_AW485`
`F`	`MPE8`	epitope	`SKGYLSALRTGWYTS + QLPLYGVIDT; 41-55, 302--311` *(neutralizes HRSV and HMPV isolates of several subgroups)*	`ABCD_AR272`

strongly constrains the pocket geometry (orientation leaks the answer). To complement the binary $p < 0.05$ criterion, we provide the full prediction outputs, reference pocket residues, and raw top-1/top-5 hypergeometric p-values for all evaluated cases in the GitHub repository, enabling readers to examine non-significant predictions in more detail. The main value of Site4Drug is that it can still propose meaningful sites when such structural context is absent, incomplete, ambiguous, or difficult to use directly.

4.2. Ablation: Sequence-only baseline

To test whether Site4Drug benefits from its explicit evidence pipeline beyond sequence-only domain-aligned prompting, we constructed a sequence-only ablation on the same included pocket benchmark used in the main small-molecule evaluation (`Group S + Group AS`, $n = 63$). We chose this setting because it is the largest and cleanest benchmark in this study and also supports direct comparison with the structure-based baseline fpocket. In this ablation, the model received only a generic target identifier and the raw amino-acid sequence. Otherwise, the prompt structure, inference setting, and output format were kept nearly identical to the main pipeline, with the explicit TM/topology, PTM, motif, and cysteine/disulfide evidence fields removed. Exact prompt details are provided in Appendix D.1. We also assessed the effect of self-consistency by comparing a single proposal ($k = 1$) against three-attempt voting ($k = 3$). The sequence-only baseline yielded significant overlap ($p < 0.05$) for 3/63 cases at both top-1 and top-5 under $k = 1$, improving to 7/63 and 6/63 under $k = 3$, respectively. These values remain well below the full Site4Drug pipeline (Fig. 3c), suggesting that Site4Drug benefits materially from its explicit evidence pipeline rather than merely exploiting generic sequence-level patterns.

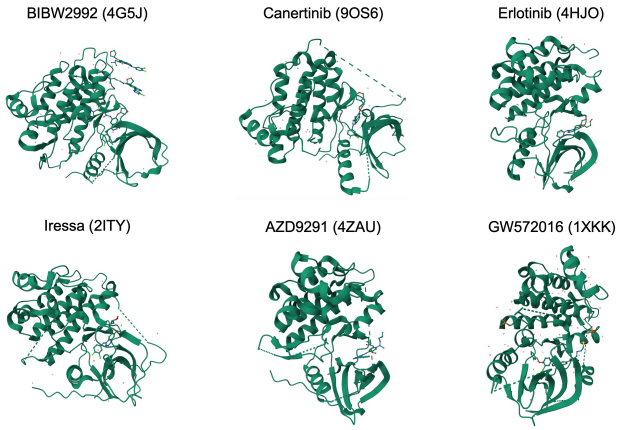
4.3. Assessment on antibody targets (Epitope)

We next evaluated Site4Drug in epitope mode using antibody–target pairs curated from the AntiBodies Chemically Defined (ABCD) database. In contrast to the pocket benchmark above, residue-level epitope annotations were much sparser in this setting: many ABCD entries identify the target and antibody but do not provide a precise sequence-resolved epitope. We therefore retain only the subset with sufficiently clear annotated epitope positions for direct overlap-based evaluation. In total, we initially curated 52 target–antibody pairs and retained 26 for final epitope benchmarking `Group A`. Representative examples are shown in Table 3.

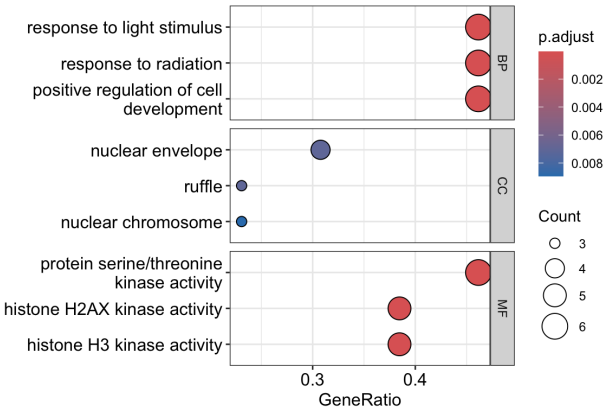
For each retained sequence, we evaluated whether the top-1 or top-5 predictions from Site4Drug significantly overlapped the annotated epitope using the same hypergeometric framework as in pocket mode. When applied in epitope mode, Site4Drug produced significant overlap ($p < 0.05$) for 8 cases under the top-1 criterion and for 11 cases under the top-5 criterion (Fig. 3d). Although this benchmark is much smaller and noisier than the small-molecule benchmark, the observed success rate indicates that Site4Drug can recover meaningful epitope-like regions from sequence-derived evidence.

4.4. Validation of Structural Reliability

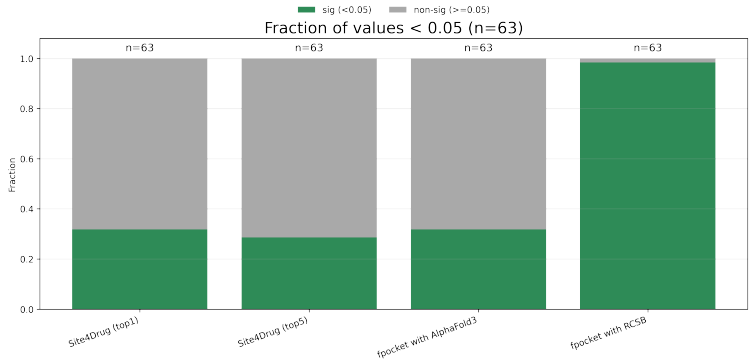
Finally, we asked whether the sites predicted by Site4Drug tend to fall in structurally credible regions even though the model does not directly ingest protein structure. For this analysis, we used AlphaFold3 (Abramson et al., 2024) to predict structures corresponding to the 63 retained pocket-mode validation target pairs and examined the pLDDT values of the residues contained in the predicted sites.



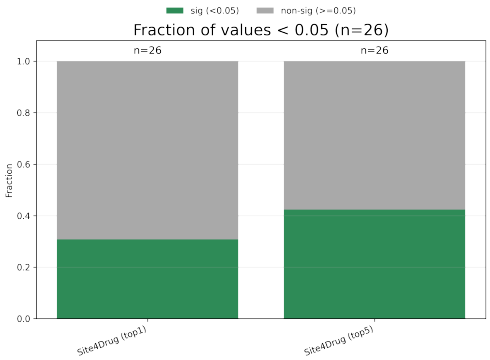
(a) Six EGFR–drug cocrystal structures with similar orientations.



(b) GO plot of targets of interest from Site4Drug.



(c) Proportion of targets with pocket site predictions at p-value < 0.05 for each baseline method.



(d) Proportion of targets with epitope site predictions at p-value < 0.05.

Figure 3. Overview of structural examples, functional enrichment results, and baseline significance ratios. (a) Six EGFR–drug cocrystal structures viewed from similar orientations along with the corresponding drugs and RCSB entries. (b) GO plot of targets for which the top-1 Site4Drug LLM prediction achieved p-value < 0.05. (c) The proportion of targets with significant site predictions (p-value < 0.05) on the pocket benchmark ($n = 63$): Site4Drug top-1 (20/63), Site4Drug top-5 (18/63), fpocket with AlphaFold3 input (20/63), and fpocket with RCSB input (62/63). Site4Drug produced comparable results to fpocket with AlphaFold3 structure, showing Site4Drug implies structural information under general conditions. Since the RCSB structure contains the bound ligand and thus adopts a bent conformation, fpocket with ligand-bound RCSB structures yielded results very close to the ground truth (98.41%) as expected. (d) Applying Site4Drug to the ABCD database with epitope mode yielded a similar proportion of targets with significant site predictions (p-value < 0.05) as in pocket mode.

Regions with lower PLDDT are generally interpreted as less structurally reliable or more conformationally flexible, whereas higher-PLDDT regions tend to correspond to more confidently modeled local structure. Although the drug-binding site does not have to lie exclusively in the highest-PLDDT region of a protein, structural confidence remains a useful secondary plausibility check, especially for novel sites that lack direct experimental annotation.

To examine whether the sites predicted by Site4Drug, despite not using structural information, still correspond to structurally reliable regions that are favorable for drug binding, we compared the PLDDT values of Site4Drug's top-1 and top-5 predictions on the AlphaFold3 structures corresponding to the 63 cases used in the pocket validation. In most cases, the top-1 site had a higher mean PLDDT than the broader top-5 set, with only 9 exceptions (Fig. 4). This suggests that the highest-ranked prediction from Site4Drug is often not only statistically meaningful with respect to known sites, but also structurally more plausible than lower-ranked alternatives. More broadly, this behavior indicates that the sequence-based evidence aggregation used by Site4Drug may implicitly recover signals correlated with structural reliability, even without direct structure input.

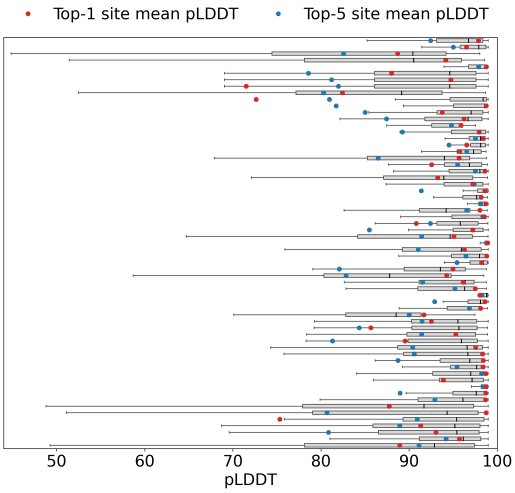


Figure 4. Horizontal box plot of per-residue pLDDT values for the corresponding AlphaFold3 structures of 63 records included in pocket-mode validation (`Group S` and `Group AS` groups). Each box plot summarizes the whole-structure pLDDT distribution for one record. The red marker denotes the mean pLDDT of the best site, while the blue marker denotes the mean pLDDT averaged over top-5 annotated sites in that record. Together, these markers show how local site confidence compares with the overall structural confidence of each predicted site.

5. Module 2: Modality-specific Design Handoff

Module 1 addresses the challenge of selecting *where* to intervene on a protein (i.e., choosing a targetable site). Thus, ranked outputs of Module 1 can now be treated by Module 2 as an interface to downstream design tools. In epitope mode, top-ranked spans together with topology/PTM risk annotations can be passed to peptide or antibody-binder design tools (e.g., BoltzGen). In pocket mode, top-ranked pocket regions lead to small-molecule design or scoring tools (e.g., DrugCLIP, BindCLIP).

Thus, we also expand and report the capability of not only providing the target site, but also end-to-end design. While the following results illustrate the potential of Module 2 for end-to-end design workflows, they should be interpreted as proof-of-concept demonstrations on a single target, EGFR, rather than as broad validation across diverse targets.

5.1. Application of Module 2 for pocket mode.

Because DrugCLIP takes as input a three-dimensional structure that can be defined as a pocket within the full protein structure, we extracted only the atoms belonging to the predicted site from the AlphaFold3 structure and used them to construct the pocket. DrugCLIP was then applied to the top-ranked site predicted by Site4Drug for EGFR (RCSB code: 1XKK; p-value $= 8.47 \times 10^{-5}$). Known EGFR inhibitors including afatinib, canertinib, erlotinib, gefitinib, lapatinib, and osimertinib were absent from the DrugCLIP ligand database and therefore could not appear in the retrieved list. However, visualization of the top-6 hits together with these reference compounds showed that the retrieved molecules were structurally similar to the EGFR inhibitors and shared structural motifs likely to bind this pocket, such as benzene and cyclohexane moieties (Tab. S4, Fig. S2). This observation was further supported by a hypergeometric test evaluating the overlap between the structures of the top-6 hits and the ground truth (lapatinib; 1XKK) site, which yielded p-value lower than 10^{-11} (almost identical). For this analysis, we used Boltz2 (Passaro et al., 2025), because AlphaFold3 does not currently support general ligand-containing predictions.

5.2. Application of Module 2 for epitope mode.

We ran BoltzGen using the top-1 to top-5 EGFR epitopes predicted by Site4Drug to generate peptide binders. Because BoltzGen tends to encounter memory issues with protein sequences longer than approximately 1,000 amino acids, we truncated the 1,210-aa EGFR precursor to retain only the first 1,000 aa from the N terminus. We then provided the epitope information predicted by Site4Drug and ran BoltzGen with the number of designs set to 50 and the peptide length constrained to 12- to 18-mers. Among the generated candidates, the binder with the best BoltzGen score was selected as the representative peptide binder for each epitope.

The peptide and EGFR sequences were then provided to AlphaFold3 to generate five multimer structures, from which we calculated the median Local Interaction Score (LIS) (Kim et al., 2024) and the median of the minimum distance between the peptide and the target (Fig. S3).

It is straightforward to understand why rank-5 received the

lowest priority in Site4Drug, as unlike the higher-priority sites, it is located intracellularly. Predicted aligned error (PAE) originally indicates higher confidence in the relative positioning of two residues when its value is lower. Using the commonly applied cutoff PAE value of 12 for LIS calculation, only rank-1 yielded a definable LIS, whereas all PAE values for the remaining extracellular sites exceeded 12. The high ranking of rank-2 is also associated with the fact that it corresponds to Domain III of EGFR, which is known to be targeted by antibody therapeutics such as petosemtamab and 7D12. For rank-3 and rank-4, a difference was observed in the minimum distance metric.

These results show that the final binder candidates generated from Site4Drug's site ranking exhibit consistency with prior literature and structural information.

6. Discussion and Limitations

SFT-centric supervision. We initially considered supervised fine-tuning (SFT) on structured demonstrations as a post-training strategy. An example is shown in Appendix F. In preliminary experiments with the Qwen backbone (Qwen3-235B Instruct), the SFT checkpoint improved output formatting but also showed shortcut behavior, including repeated selection of similar N-terminal spans. We therefore report quantitative results using the base model. This suggests that post-training for target-site discovery may require biologically grounded reward or preference signals, which we view as an important direction for future work.

Structure is sometimes gating. Especially, pocket mode depends on structural context, and some models take structural information as input. However, this not only requires a large number of tokens (e.g., more than 5,000 amino acids do not provide structure in AlphaFold3) but is also not suitable for scalable screening purposes. Therefore, it is important to use an appropriate level of structural information.

No consideration of quaternary structure. At present, the model takes a single protein sequence as input, but in reality multiple proteins can also assemble into a quaternary structure. In fact, many drugs commonly target channel proteins composed of multiple domains. It will therefore be necessary to extend the approach to handle such cases as well. Unlike fpocket and RAPID-Net, which take single amino acids as input, the LLM-based Site4Drug, capable of leveraging diverse metadata including pocket / epitope mode designation, appears well positioned to address this cochain issue with ease.

More available information. Predictions may become more accurate as broader databases and annotations are integrated. For example, curated information such as exact disulfide-bond annotations from UNIPROT is not yet used throughout the present pipeline.

Sensitivity to topology. It should be noted that, rather than inputting the full amino acid sequence, partial sequence input may also be used to obtain a more context-specific epitope or to avoid token limit constraints. However, caution is warranted in such cases, as the topology may be inferred in the reverse orientation.

Concentration dependency. Target engagement is concentration-dependent, as some compounds are inactive at lower concentrations but exhibit measurable interactions with specific targets at higher concentrations. We need to test Site4Drug on data that includes concentration context.

Scale and breadth of validation and baseline comparison. A key limitation of this study is the limited scale of validation and fair baseline comparison. Although Site4Drug was evaluated using modality-aware validation datasets, the current datasets remain modest in size, particularly for antibody-relevant target sites where reliable site-level annotations are difficult to curate. Moreover, direct comparison with existing structure-trained machine learning models is challenging because of potential data leakage from overlapping training databases. Future work should address these limitations by expanding validation to larger and more diverse targets under careful leakage control, and by incorporating fair apo-structure-based or sequence-based ML baselines that are independent of the evaluation set.

7. Conclusion

Site4Drug aims to improve the reliability and transparency of target-site selection by producing constraint-aware region proposals and audit-ready decision logs. Risks include over-reliance on incomplete evidence sources or misinterpretation of immune epitope data. We mitigate these risks through explicit uncertainty flags and traceable audit trails. By identifying targetable protein regions together with the evidence supporting each decision, Site4Drug provides an interpretable upstream interface for downstream binder or small-molecule design workflows. For example, we applied Site4Drug for diverse drug modalities: small molecules, antibodies, and peptide drugs. We also tested that our site detection can be easily combined with third-party tools, such as Diffuse Bio Sandbox, a platform for nanobody and scFv modeling (Fig. S4).

We have confirmed that the epitope location predicted by Site4Drug matched exactly with the Hydrogen-Deuterium Exchange Mass Spectrometry result for an actual target. Interestingly, antibody physicochemical properties affected target binding, and our pipeline suggested a possible mechanism (Appendix J). Additional experiments (e.g., alanine scanning) on multiple targets and high-throughput screening tests are planned for future studies. Additionally, future research may focus on a lab-in-the-loop framework

for Site4Drug, integrating real MS data and iterative experimental feedback. Furthermore, Site4Drug will be incorporated into automated target-discovery pipelines (Zhang et al., 2026), thereby enabling end-to-end applications.

Impact Statement

This work aims to advance machine learning methods for drug-target site discovery and to support downstream small-molecule, peptide, and antibody-binder design. The potential positive impact is to accelerate early-stage therapeutic research by helping prioritize candidate binding sites, particularly for targets with limited experimental information. However, because such predictions may be used together with generative design tools, responsible use requires caution. Site4Drug outputs should be treated as hypotheses for experimental follow-up, not as validated therapeutic candidates. Any downstream molecules designed from these predictions should undergo independent validation for binding, specificity, off-target effects, safety, and regulatory suitability before translational or clinical use.

Acknowledgements

We implemented the LLM-agent pipeline using the Tinker API at Thinking Machines Lab. This project was developed as an extension of coursework in CDSS 94 at UC Berkeley, and we thank Kevin Miao and Karina Nguyen for course support and guidance. We also thank Diffuse Bio for providing sandbox credits. The corresponding author, Jeongbin Park, gained expertise in in silico drug design while working at Portrai, Inc.

Code Availability

The demo for Site4Drug is available at `https://github.com/winterrykim/Site4Drug_Demo`.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Agarwal, D., Majumder, B. P., Adamson, R., Chakravorty, M., Gavireddy, S. R., Parashar, A., Surana, H., Dalvi Mishra, B., McCallum, A., Sabharwal, A., et al. Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems*, 38:25181–25219, 2026.

Balytskyi, Y., Hubenko, I., Balytska, A., and Kelly, C. V. Rapid-net: Accurate pocket identification for binding-site-agnostic docking. *Journal of Chemical Information and Modeling*, 65(22):12221–12239, 2025.

Cornelio, C., Dash, S., Austel, V., Josephson, T. R., Goncalves, J., Clarkson, K. L., Megiddo, N., El Khadir, B., and Horesh, L. Combining data and theory for derivable scientific discovery with ai-descartes. *Nature Communications*, 14(1):1777, 2023.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl_2): W362–W365, 2006.

Gao, B., Qiang, B., Tan, H., Jia, Y., Ren, M., Lu, M., Liu, J., Ma, W.-Y., and Lan, Y. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36: 44595–44614, 2023.

Gonzalez, G., Lin, X., Herath, I., Veselkov, K., Bronstein, M., and Zitnik, M. Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*, pp. 1–18, 2025.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., Myaskovsky, A., Glowaty, G., Weissenberger, F., Orlandi, A., Popovici, D., et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Jansen, P., Clark, P., Downey, D., and Weld, D. S. Generating literature-driven scientific theories at scale. *arXiv preprint arXiv:2601.16282*, 2026.

Kim, A.-R., Hu, Y., Comjean, A., Rodiger, J., Mohr, S. E., and Perrimon, N. Enhanced protein-protein interaction discovery via alphafold-multimer. *BioRxiv*, 2024.

Kyte, J. and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):168, 2009.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., and Clune, J. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Qiao, A., Wang, Z., Li, Y., Rao, J., and Yang, Y. Bindclip: A unified contrastive-generative representation learning framework for virtual screening. *arXiv preprint arXiv:2602.15236*, 2026.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Swanson, K., Wu, W., Bulaong, N., Pak, J., and Zou, J. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. biorxiv, 2024. *bioRxiv*, 11, 2024.

Vita, R., Blazeska, N., Marrama, D., Marcus, I. C. T. M. S. D. Z. L. F. G. Z. L. C. K. R. B. F. S. B. M. C. S. S. C. E. L. M. C. E., Duesing, S., Bennett, J., Greenbaum, J., De Almeida Mendes, M., Mahita, J., Wheeler, D. K., et al. The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443, 2025.

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., Li, J., and Xu, D. Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, 48 (W1):W140–W146, 2020.

Zhang, H. G., Eckmann, P., Miao, J., Mahon, A. B., and Zou, J. The virtual biotech: A multi-agent ai framework for therapeutic discovery and development. *bioRxiv*, pp. 2026–02, 2026.

A. Appendix: Site4Drug Report

SITE4DRUG

Prediction Report

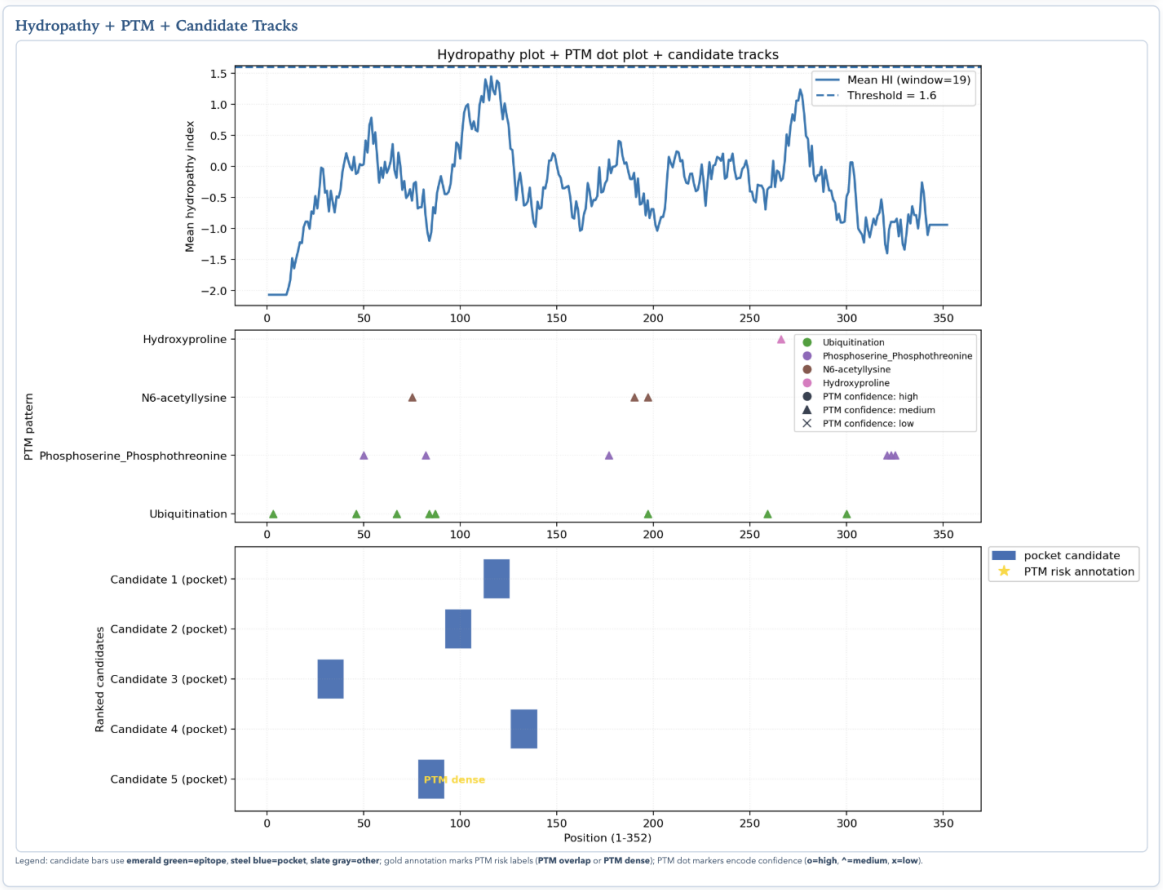
RUN ID	RUN STATUS	UNIPROT / LABEL
20260324_142742_SEQ_idx201.seq5	OK	SEQ_idx201.seq5

REQUESTED MODE	RECOMMENDED MODALITY	CANDIDATE SOURCE
auto	pocket (0.90)	llm_propose

Ranked Candidates

Rank	ID	Mode	Peptide	Position	Confidence	Score	Source	Flags	Reason
1	L_C0005	pocket	LTSTVQLITQLMPF	112-125	High	0.82	llm_self_consistency_vote	motif-overlap, self_consistency_votes_2_of_3, span-repaired-from-peptide	High hydrophobicity (0.571) and no PTM overlap support strong pocket formation; overlaps PS50011 motif but no direct PTM interference.
2	L_C0003	pocket	EAYVMASVDNPHVC	92-105	High	0.75	llm_self_consistency_vote	motif-overlap, self_consistency_votes_3_of_3, span-repaired-from-peptide	Moderate hydrophobicity (0.5) and no PTM overlap; overlaps PS50011 motif, which may constrain flexibility.
3	L_C0001	pocket	GEAPNQALLRILKE	26-39	Moderate	0.70	llm_self_consistency_vote	self_consistency_votes_3_of_3, span-repaired-from-peptide	No PTM or motif overlap and moderate hydrophobicity (0.5); low polar fraction may limit solvation.
4	L_C0004	pocket	GCLLDYVREHKDNI	126-139	Moderate	0.60	llm_self_consistency_vote	fuzzy_consensus_clustered, motif-overlap, self_consistency_votes_3_of_3, span-repaired-from-peptide	Low hydrophobicity (0.286) and motif-overlap reduce pocket plausibility despite no PTM interference.
5	L_C0002	pocket	REATSPKANKEILD	78-91	Low	0.35	llm_self_consistency_vote	PTM-clustered, PTM-dense, PTM-overlap, motif-overlap, self_consistency_votes_3_of_3, span-repaired-from-peptide	PTM-dense region with overlapping phospho/ubiquitin sites at positions 82, 84, 87 disrupts pocket integrity.

(a) Structured output report with recommended modality, ranked candidates, confidence, flags, and rationale.



(b) Sequence-level evidence view showing hydropathy, PTM annotations, and candidate tracks.

Figure S1. Example Site4Drug prediction interface for Epidermal Growth Factor Receptor (EGFR). (a) Structured output report returned by Site4Drug, including recommended modality, ranked candidates, confidence, flags, and rationale. (b) Corresponding evidence view used to support candidate ranking and risk annotation. Pocket-mode prediction is supported by the observation of lower hydropathy scores. Ground-truth small-molecule binding-site residues are L48, V56, A73, K75, M96, C105, R106, L107, L118, T120, Q121, L122, M123, G126, C127, L129, D130, R133, L174, T184, D185, F186, and L188.

B. Appendix: Prompt Templates

Inference system prompt.

```
You are Site4Drug, an expert system for constraint-first targetable site discovery on
    proteins.
You must decide target modality (epitope vs pocket) unless explicitly fixed.
Epitope mode predicts where antibodies or peptide binders attach, usually on
    extracellular protruding regions of membrane proteins, often tied to immune
    responses.
Pocket mode predicts where small molecules bind, typically in intracellular proteins or
     membrane channels, covering immune as well as general signaling and metabolic
    pathways.
Follow a constraint hierarchy: topology/TM -> PTM masks -> motif-functional caveats ->
    disulfide/context.
Generate or rank targetable candidates with evidence-grounded rationale.
Always output strict JSON.
```

Full-sequence inference template. Here, `<target_id>` denotes a generic target identifier. In benchmark evaluation this field was typically a neutral sequence label (e.g., `SEQ_idx001.seq0`).

The default proposal prompt combines the full target sequence with PTM and motif summaries:

```
Objective:
Identify targetable site regions using constraint-first analysis and return strict JSON.


Modality policy:
<mode_instruction>

<auto_block_if_mode_is_auto>Constraint hierarchy:
1) TM/topology constraints
2) PTM mask constraints (typed)
3) Motif-functional caveats
4) Disulfide/context risks

Target: UniProt <uniprot>
Length: <len(sequence)> aa
TM regions: <seq_summary.tm_regions>
Cysteines: <len(seq_summary.cysteine_positions)>

<ptm_text>

<motif_text>

Antigen sequence:
<sequence>

Output request:
Return Top-<k> candidates.
<schema_instruction>
```

If auto mode is enabled, the prompt additionally inserts a deterministic policy block of the form:

```
Auto mode policy (deterministic project rule):
- Non-membrane protein -> pocket
- Membrane protein -> pocket or epitope
- This target policy decision: <mode> (n_tm_regions=<n>, channel_like=<bool>)
- Policy reason: <reason>
Set recommended_modality to this policy decision.
```

```
When emitting ranked_candidates, prefer row modes that match this policy decision.
```

Candidate-ranking inference template. When candidate preprocessing is available, the system can be configured to ask the model to rank a provided candidate table:

```
Objective:
Rank provided candidate regions using constraint-first evidence and return strict JSON.

Modality policy:
<mode_instruction>
- <chunk_note>

<auto_block_if_mode_is_auto>Constraint hierarchy:
1) TM/topology constraints
2) PTM mask constraints (typed)
3) Motif-functional caveats
4) Disulfide/context risks

Target: Sequence ID <target_id>
Length: <seq_summary.sequence_length> aa
TM regions: <seq_summary.tm_regions>
Cysteine count: <len(seq_summary.cysteine_positions)>

<ptm_text>

<motif_text>

Candidate table:
<candidate_table>

Output request:
Return exactly Top-<k>.
<schema_instruction>
```

Example instantiated evidence snippets. The PTM and motif summaries are rendered in the following format:

```
<ptm_text> example:
PTM summary: total=3, by_type=N-linked_glycosylation:2, Phosphoserine_Phosphothreonine
    :1
PTM sites (typed):
- N-linked_glycosylation @ 82 (mask 77-87, confidence=high)
- N-linked_glycosylation @ 144 (mask 139-149, confidence=high)
- Phosphoserine_Phosphothreonine @ 211 (mask 208-214, confidence=medium)

<motif_text> example:
Motif summary: total=2, by_name=EGF_2:1, N-glyco_site:1
Motif hits:
- EGF_2 (PS01186) @ 128-165 [remote_scanprosite_biopython]
- N-glyco_site (PS00001) @ 141-143 [remote_scanprosite_biopython]
```

Raw proposal JSON schema. The top-level proposer is asked to emit a compact JSON object with the following structure:

```
{
  "recommended_modality": "epitope|pocket|other",
  "modality_confidence": 0.0,
  "ranked_candidates": [
```

```
    {
      "rank": 1,
      "start": 1,
      "end": 15,
      "peptide": "AAAA",
      "mode": "epitope|pocket|other",
      "confidence_score": 0.0,
      "reason": "..."
    }
  ]
}
```

The schema instruction also requires that ranked candidates remain span-consistent, that PTM/motif caveats be mentioned when present, and that the output begin with { and end with } with no extra prose.

Final normalized output artifact. After parsing, validation, enrichment, and optional agent-panel adjudication, the run artifact is normalized into a richer schema with the following high-level fields:

```
schema_version
recommended_modality
modality_confidence
ranked_candidates[{rank, candidate_id, start, end, peptide, mode, confidence,
    confidence_score, flags, reason, mean_hydropathy, overlaps_tm, overlaps_ptm_mask,
    ptm_overlap_by_type, motif_hit_count, cysteine_count}]
candidate_evidence
risk_flags
agent_traces
feature_provenance
token_strategy_used
ptm_summary
motif_summary
orchestrator_trace
audit_log
```

Specialist-agent critique format. Each specialist agent is prompted to return compact JSON of the following form:

```
{
  "agent": "...",
  "modality_votes": {"epitope": 0.0, "pocket": 0.0, "other": 0.0},
  "candidate_adjustments": [
    {"candidate_id": "...", "delta": 0.0, "reason": "...", "evidence": ["..."]}
  ],
  "risk_flags": ["..."],
  "summary": "..."
}
```

Decision-agent format. The decision agent uses a different output contract:

```
{
  "recommended_modality": "epitope|pocket|other",
  "modality_confidence": 0.0,
  "ranking": [
    {"rank": 1, "candidate_id": "...", "reason": "...", "confidence_score": 0.0, "
    confidence_reason": "..."}
  ],
  "global_risks": ["..."]
}
```

C. Appendix: Hydropathy Profile (Reference Implementation)

Reference implementation used by the sequence feature engine to build a center-aligned hydropathy profile.

```
HI = {
  "A":1.8, "C":2.5, "D":-3.5, "E":-3.5, "F":2.8,
  "G":-0.4, "H":-3.2, "I":4.5, "K":-3.9, "L":3.8,
  "M":1.9, "N":-3.5, "P":-1.6, "Q":-3.5, "R":-4.5,
  "S":-0.8, "T":-0.7, "V":4.2, "W":-0.9, "Y":-1.3,
}

def sliding_hydropathy(sequence, window=19):
  if not sequence:
    return []
  vals = [HI.get(a, 0.0) for a in sequence]
  if len(vals) < window:
    mean_val = sum(vals) / max(len(vals), 1)
    return [mean_val] * len(vals)

  means = []
  running_sum = sum(vals[:window])
  means.append(running_sum / window)
  for i in range(1, len(vals) - window + 1):
    running_sum += vals[i + window - 1] - vals[i - 1]
    means.append(running_sum / window)

  pad_left = window // 2
  pad_right = len(vals) - len(means) - pad_left
  return [means[0]] * pad_left + means + [means[-1]] * pad_right
```

D. Appendix: Sequence-only Ablation

Additional prompt and evaluation details for the sequence-only ablation benchmark.

D.1. Sequence-only Ablation Prompt

Exact prompt template used in the sequence-only ablation setting.

To isolate the contribution of the explicit evidence pipeline, we constructed a sequence-only ablation using a prompt that was kept nearly identical to the main full-sequence inference prompt. The same system instruction, `auto` modality setting, and JSON output contract were retained. The only intentional prompt-level change was to remove the structured evidence fields, including TM/topology summaries, cysteine counts, PTM summaries and typed PTM sites, motif summaries and motif hits, and the deterministic auto-policy block. Here, `<target_id>` denotes a generic target identifier; in benchmark evaluation this field was typically a neutral sequence label such as `SEQ_idx001.seq0`.

```
System:
You are Site4Drug, an expert system for constraint-first targetable site discovery on
    proteins.
You must decide target modality (epitope vs pocket) unless explicitly fixed.
Epitope mode predicts where antibodies or peptide binders attach, usually on
    extracellular protruding regions of membrane proteins, often tied to immune
    responses.
Pocket mode predicts where small molecules bind, typically in intracellular proteins or
     membrane channels, covering immune as well as general signaling and metabolic
    pathways.
Follow a constraint hierarchy: topology/TM -> PTM masks -> motif-functional caveats ->
    disulfide/context.
Generate or rank targetable candidates with evidence-grounded rationale.
Always output strict JSON.
```

```
User:
Objective:
Identify targetable site regions using constraint-first analysis and return strict JSON.


Modality policy:
Choose modality automatically. Keep each candidate mode consistent with your final
    recommended_modality.

Constraint hierarchy:
1) TM/topology constraints
2) PTM mask constraints (typed)
3) Motif-functional caveats
4) Disulfide/context risks

Target: Sequence ID <target_id>
Length: <len(sequence)> aa

Antigen sequence:
<sequence>

Output request:
Return Top-<k> candidates.
<schema_instruction>
```

D.2. Sequence-only Ablation Results

Supplementary results for the sequence-only ablation on the included pocket benchmark.

Table S1. Sequence-only ablation on the included pocket benchmark (`Group S + Group AS`, $n = 63$).

Setting	Predicted pocket	Top-1 sig.	Top-5 sig.
Sequence-only ($k = 1$)	$23/63$ (36.5%)	$3/63$ (4.8%)	$3/63$ (4.8%)
Sequence-only ($k = 3$)	$22/63$ (34.9%)	$7/63$ (11.1%)	$6/63$ (9.5%)

E. Appendix: Reference Candidate Features and Risk Annotations

In the current default `llm_propose` inference path, the LLM proposes and ranks candidate spans from the evidence-rich prompt context. The deterministic feature engine is then applied post hoc to each accepted span for candidate enrichment, typed risk annotation, audit logging, constraint checking. The same features are also used directly for ranking.

Candidate-level reference features. For a candidate span r, we compute mean hydropathy $h(r)$, amino-acid composition fractions, TM overlap, PTM-mask overlap, motif overlap, and cysteine count. Table S2 summarizes the main derived quantities used by the current reference implementation.

Reference heuristic score. Conceptually, these features support a mode-specific preference of the form

$$S_0(r) = s_{\text{mode}}(r) - p_{\text{TM}}(r) - p_{\text{PTM}}(r) - p_{\text{motif}}(r),$$

although the current implementation uses different exact formulas by mode. Under the default `tiered` PTM policy, epitope mode favors mildly hydrophilic, polar, non-TM, PTM-light spans, whereas pocket mode favors more hydrophobic spans and penalizes PTM overlap less aggressively.

Typed risk annotations. In parallel with scalar score computation, we emit typed risk annotations that preserve *why* a candidate may be unfavorable. These annotations are not treated as interchangeable penalties. Table S3 summarizes the current rule triggers.

Table S2. Candidate-level deterministic features used for post hoc enrichment and, in deterministic mode, heuristic ranking.

Feature	Definition
$h(r)$	mean hydropathy over span r
$\phi_{\text{hydrophobic}}(r)$	hydrophobic residue fraction in r
$\phi_{\text{polar}}(r)$	polar residue fraction in r
$\phi_{+}(r)$	positively charged residue fraction in r
$\phi_{-}(r)$	negatively charged residue fraction in r
$I_{\text{TM}}(r)$	indicator that r overlaps a detected TM segment
$c_{\text{motif}}(r)$	number of overlapping PROSITE/ScanProsite motif matches
$n_{\text{glyco}}(r)$	number of overlapping N-linked glycosylation masks
$n_{\text{other}}(r)$	number of overlapping non-glycosylation PTM masks
$d_{\text{PTM}}(r)$	$\big(n_{\text{glyco}}(r) + n_{\text{other}}(r)\big)/\lvert r\rvert$
$c_{\text{cys}}(r)$	number of cysteines in r

Table S3. Reference typed risk annotations and their current trigger conditions.

Risk flag	Trigger condition
`TM-overlap`	$I_{\text{TM}}(r) = 1$
`PTM-overlap`	$n_{\text{glyco}}(r) + n_{\text{other}}(r) > 0$
`glyco-mask-overlap`	$n_{\text{glyco}}(r) > 0$
`PTM-dense`	$d_{\text{PTM}}(r) \geq 0.25$ or total overlapping PTMs ≥ 3
`PTM-clustered`	$d_{\text{PTM}}(r) \geq 0.25$ or total overlapping PTMs ≥ 3
`disulfide-constrained`	$c_{\text{cys}}(r) \geq 2$
`hydrophobic-core`	$h(r) > 1.0$
`motif-overlap`	$c_{\text{motif}}(r) > 0$

Additional motif-specific flags may be emitted when motif names contain particular substrings; for example, zinc- or DNA-related motif names produce extra cautionary annotations in epitope mode. Because explicit disulfide bond pairing is not modeled, `disulfide-constrained` should be interpreted as a cysteine-based structural-risk proxy rather than a direct bond-level call.

F. Appendix: Illustrative Behavior Shift After Post-Training

Quantitative summary. For this target, both models returned 10 ranked candidates. In the raw top-level output, the mean candidate-level reason length decreased from 20.6 words in the base model to 5.5 words in the SFT checkpoint. It gives a more concise formatted "reason" for each candidate.

Original base-model raw output (abridged).


```
"recommended_modality": "pocket",
"modality_confidence": 1.0,
"ranked_candidates": [
  {
    "rank": 1,
    "start": 619,
    "end": 633,
    "peptide": "TPHIDNLEVANSFAV",
    "mode": "pocket",
    "confidence_score": 0.91,
    "reason": "High-confidence phosphosite cluster at 624 and 628; no TM or motif
   conflicts; region is solvent-accessible and flexible, favorable for small molecule
   binding; PTM mask 619-633 reduces risk of interference"
  },
  {
    "rank": 2,
    "start": 724,
    "end": 738,
    "peptide": "DYDDSLVSKFIERWS",
```

```
    "mode": "pocket",
    "confidence_score": 0.89,
    "reason": "Dense phosphorylation cluster (729, 733, 737) with medium confidence;
    part of regulatory region; PTM mask 724-738 suggests dynamic regulation; suitable
    for allosteric pocket targeting"
  },
  ...
]
```

SFT checkpoint raw output (abridged).

```
{
  "recommended_modality": "pocket",
  "modality_confidence": 0.60,
  "ranked_candidates": [
    {
      "rank": 1,
      "start": 1,
      "end": 15,
      "peptide": "NSIQIGGLFPRGADQ",
      "mode": "epitope",
      "confidence_score": 0.40,
      "reason": "PTM-overlap (Phosphoserine_Phosphothreonine at 2, Methylarginine at
    59)"
    },
    {
      "rank": 2,
      "start": 16,
      "end": 30,
      "peptide": "EYSAFRVGMVQFSTS",
      "mode": "epitope",
      "confidence_score": 0.40,
      "reason": "PTM-overlap (Phosphoserine_Phosphothreonine at 29)"
    },
    ...
  ]
}
```

Original base-model final returned output (abridged).

```
Rank 1 | pocket | TPHIDNLEVANSFAV | 35--49
Reason: High hydrophobic fraction (0.533) and no PTM overlap support strong pocket
    binding potential.

Rank 2 | pocket | QFGGAEVSGFQIVDY | 227--241
Reason: Favorable mean hydropathy (0.2) and no PTM interference enhance pocket
    suitability.
```

SFT checkpoint final returned output (abridged).

```
Rank 1 | epitope | NSIQIGGLFPRGADQ | 1--15
Reason: PTM-overlap (Phosphoserine\_Phosphothreonine at 2, Methylarginine at 59).

Rank 2 | epitope | EYSAFRVGMVQFSTS | 16--30
Reason: PTM-overlap (Phosphoserine\_Phosphothreonine at 29).
```

G. Appendix: Module 2 DrugCLIP Results

Table S4. Top-6 DrugCLIP hits represented as SMILES strings.

Rank	SMILES
Rank-1	CCCN(C)c1nc(SCc2ccc(C(=O)N3CCN(C(C)=O)CC3)cc2)nc(C)c1C
Rank-2	CCCc1c(C#N)c(N)nc(SCCCN2CCN(c3cccc(Cl)c3)CC2)c1C#N
Rank-3	CCCCNC(=O)CSc1cc(N2CCN(c3ccccc3OC)CC2)ncn1
Rank-4	COc1ccc(N2CCN(c3cc(SCC(=O)NCCC(C)C)ncn3)CC2)cc1
Rank-5	Cc1sc2nc(CCc3ccccc3)nc(N3CCN(C(=O)Nc4cccc(C#N)c4)CC3)c2c1C
Rank-6	CCCCc1ccc(NC(=O)CSc2cc(N3CCN(c4ccccc4OC)CC3)ncn2)cc1

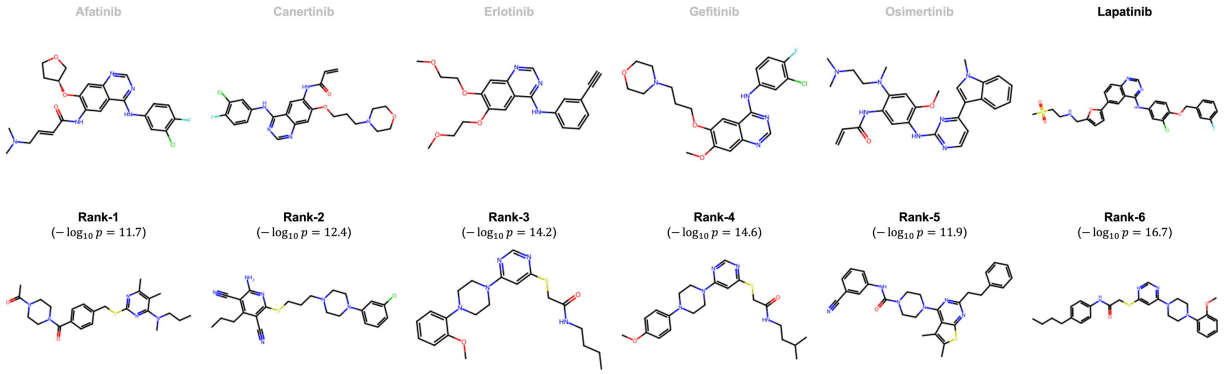


Figure S2. Comparison of reference EGFR inhibitors and the top-6 compounds retrieved by DrugCLIP for the EGFR pocket. The top row displays the structures of afatinib, canertinib, erlotinib, gefitinib, osimertinib, and lapatinib, and the bottom row displays the top-6 DrugCLIP hits. The input FASTA sequence was derived from RCSB 1XKK (EGFR bound to lapatinib). Despite the absence of the known EGFR inhibitors from the DrugCLIP ligand database, the top-ranked compounds exhibit structurally related chemical motifs, indicating that DrugCLIP captures pocket-compatible structural patterns. Figures were drawn by rdkit.

H. Appendix: Module 2 BoltzGen Results

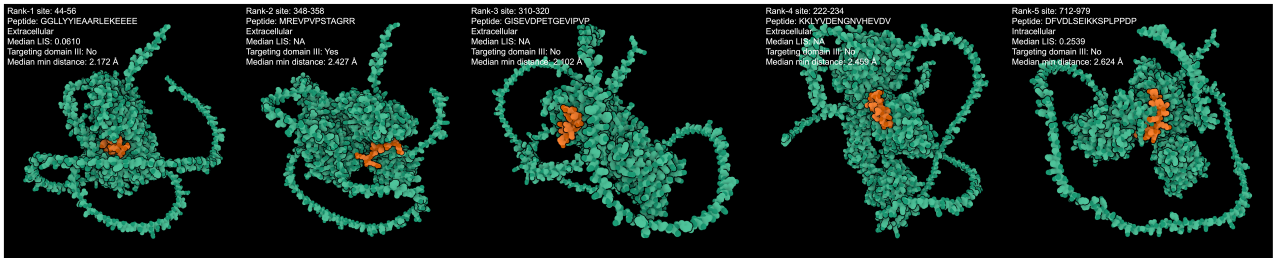
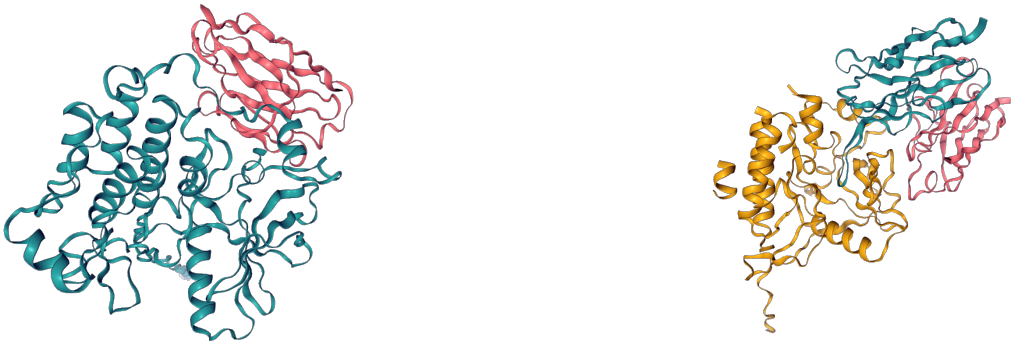


Figure S3. Peptide binder results generated from the top-5 epitopes recommended by Site4Drug in the EGFR precursor. EGFR is initially translated as a 1,210-aa precursor, after which an approximately 24-aa N-terminal signal peptide is removed to produce the 1,186-aa mature protein. Based on the mature protein sequence, the extracellular region consists of domain I (1-165 aa), domain II (166-310 aa), domain III (311-480 aa), and domain IV (481-620 aa), while the intracellular region begins at residue 643. The sites recommended by Site4Drug are consistent with the ranking of the final binder epitopes as assessed by the membrane topology of epitope, the median LIS, the domain III relatedness, and the median minimum distance, in order. EGFR and peptide binders are shown in green and orange.

I. Appendix: Integration with Diffuse Bio Sandbox



(a) Nanobody binder designed in Diffuse Bio Sandbox.

(b) scFv binder designed in Diffuse Bio Sandbox.

Figure S4. Integration with Diffuse Bio Sandbox for structure-conditioned binder generation using the Site4Drug epitope candidate on EGFR (RCSB code: 1XKK). The selected epitope residues are 34–36, 41, 55–60, 70–75, 92, 105–110, and 115–121, which are aligned with top-5 sites by Site4Drug: 112-125, 92-105, 26-39, 126-139, and 78-91. (a) Nanobody design generated against the specified epitope. Green and red parts correspond to EGFR and designed nanobody drug, respectively. (b) scFv design generated against the same epitope. Yellow and the other parts correspond to EGFR protein and designed scFv drug, respectively. These examples illustrate how Site4Drug-selected target regions can be passed downstream to external binder-generation frameworks. Figures were generated by Diffuse Bio Sandbox visualization tool.

J. Appendix: Audit Logs as a Tool for Identifying Binding Failure Modes

The following example summarizes how the Site4Drug audit log can support interpretation of a binding failure mode.

```
Rank 1 | epitope | candidate region

Reason: The candidate was ranked first because it showed no PTM mask overlap and low
    PTM density, suggesting a PTM-free target region with high initial confidence.
    However, the region was also highly hydrophobic, with mean_hydropathy = 0.9733 and
    hydrophobic_fraction = 0.6. ChemAgent flagged this high hydrophobicity as a
    potential risk because it may reduce solubility and epitope accessibility.

Interpretation: Although the predicted epitope appears favorable from a PTM perspective,
     its high hydrophobicity may indicate limited exposure in the native protein or
    cell-surface context. Therefore, the candidate may be detectable as a linear
    sequence-based epitope but less accessible in the folded protein.

Mechanistic hypothesis: After antibody mutation, a subtle change in paratope geometry
    may have reduced tolerance to this hydrophobic and partially exposed epitope. The
    WT antibody may still recognize the region weakly, whereas the mutant antibody may
    no longer bind it stably.

Additional context: The candidate may not function as an isolated linear epitope, but
    rather as part of a conformational epitope shaped by neighboring structural
    features. Full-length sequence analysis identified multiple risk flags, including
    PTM-dense regions, glyco-mask overlap, and disulfide-constrained regions,
    suggesting that folding, glycosylation, and structural context may influence
    binding.

Conclusion: The binding loss after antibody mutation is therefore more consistent with
    reduced recognition of a partially exposed or structure-dependent epitope than with
     an incorrect epitope sequence prediction.

Suggested follow-up: Compare WT and mutant antibodies using peptide ELISA, recombinant
    extracellular domain binding, and cell-based binding assays. Binding in peptide
    ELISA but not in full-length or cell-based assays would support an accessibility or
     conformational epitope issue. Loss of binding even in peptide ELISA would instead
    suggest that the paratope mutation disrupted affinity for the linear epitope itself.
```