CoPocket: Pocket Identification Without Receptor Structures Using Co-Folding Representations

Anonymous authors
Paper under double-blind review

Abstract

Locating ligandable pockets without a known ligand is an early step in drug discovery. However, most pocket predictors need a receptor structure and are benchmarked on ligand-bound conformations that do not exist before a binder is found. Sequence-based methods avoid this requirement, although UniSite's sequence-only variant underperforms its structure-informed counterpart. Benchmarks also conflate new structures of known sites with newly discovered pockets, and compute residue-level and pocket-level metrics from different outputs. CoPocket predicts pockets from sequence and multiple sequence alignment, without receptor structures or a ligand. Its set-prediction decoder maps frozen co-folding residue and pair representations to scored masks, evaluated at both pocket and residue levels. We also build the ChronoPocket dataset, which consolidates over 230,000 PDB ligand contacts into 37,269 pockets and splits them by the date each pocket was first observed. Its ChronoPocket-Discovery subset holds out proteins with previously known sites and additional pockets first observed after the training cutoff. On the temporal test set, CoPocket reaches a pocket average precision of 0.430 at an intersection-over-union threshold of 0.5, compared with 0.230 for an ESM C-based sequence-only baseline. On UniSite test proteins absent from our training data, it outperforms the structure-based UniSite-3D (0.347 vs. 0.303). On ChronoPocket-Discovery, it recovers 20.5% of later-observed pockets in its top five predictions, a challenging reference for pocket discovery.

1 Introduction

Identifying ligandable pockets, protein regions capable of accommodating small molecules, supports molecular docking, virtual screening, and fragment-based drug discovery (Krivák & Hoksza, 2018). Protein–ligand co-folding models such as AlphaFold 3 predict how a specified ligand binds (Abramson et al., 2024). Pocket predictors instead ask where a small molecule could bind, so they can prioritize sites before suitable binders are known (Krivák & Hoksza, 2018; Gazizov et al., 2026).

However, most pocket predictors require a receptor structure. Geometric and learned methods such as fpocket, P2Rank, and DeepPocket detect cavities or score surface points of an input structure (Le Guilloux et al., 2009; Krivák & Hoksza, 2018; Aggarwal et al., 2022), and residue-level methods such as IF-SitePred score the residues of an input structure (Carbery et al., 2024). Such methods are commonly developed and benchmarked on ligand-bound (holo) structures (Krivák & Hoksza, 2018; Aggarwal et al., 2022), and accuracy on holo structures does not establish reliability on unbound (apo) or predicted conformations (Clark et al., 2020; Carbery et al., 2024). Yet a holo structure exists only after a binder has been found. Sequence-based prediction avoids this dependence but has been less accurate: UniSite's sequence-only variant trails its structure-informed counterpart on the accompanying UniSite-DS dataset (Fan et al., 2025).

Co-folding models suggest a way around this trade-off. They are trained to predict protein–ligand complexes from a protein sequence and its multiple sequence alignment (MSA) across many experimental structures (Abramson et al., 2024; Wohlwend et al., 2024; ByteDance AML AI4Science Team et al., 2025), so their internal representations of a protein may already encode where ligands bind. AF2BIND supports this view for structure predictors, showing that the pair representation of

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.

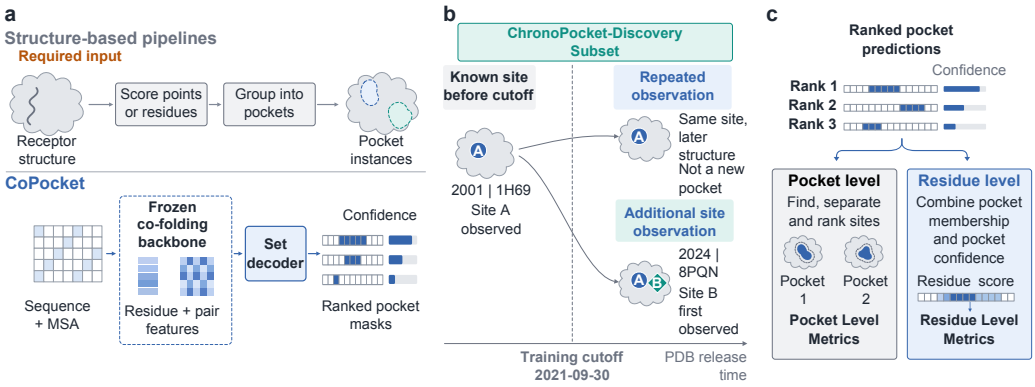


Figure 1: Overview of CoPocket and ChronoPocket. **a,** CoPocket replaces receptor-structure input with sequence and MSA, decoding co-folding features into ranked pocket masks. **b,** First-observation dating distinguishes repeated observations from additional sites, illustrated by NQO1 (UniProt P15559): site A in 2001 (PDB 1H69) and site B in 2024 (PDB 8PQN). Such cross-period proteins form ChronoPocket-Discovery and are withheld from decoder training. **c,** The same predictions assess pocket separation and ranking, and binding-residue identification through confidence-weighted membership. Protein shapes are schematic.

AlphaFold 2 carries binding-site information (Gazizov et al., 2026). **We therefore ask whether the residue and pair representations of a frozen co-folding model can replace a supplied receptor structure for pocket identification.**

Answering this question also requires evaluation that matches how pockets are discovered. First, chronological benchmarks evaluate complexes released after a cutoff (Aggarwal et al., 2022; Morehead et al., 2026), but a later complex may show a site already known from earlier structures of the same protein. Recovering such a site measures rediscovery rather than discovery. Because co-folding models are trained on structures released before a fixed date, the split should also respect that date (Abramson et al., 2024; Wohlwend et al., 2024). Second, residue-level precision–recall and ROC areas (AUPRC and AUROC) assess residue ranking (Utgés & Barton, 2024), whereas pocket-level metrics assess localization or instance detection. Examples include center-to-ligand (DCA) and center-to-center (DCC) distances (Krivák & Hoksza, 2018; Jiménez et al., 2017), and average precision (AP) at residue intersection-over-union (IoU) thresholds (Fan et al., 2025). These metrics are complementary: residue ranking alone cannot establish pocket separation, nor center proximity accurate boundaries (Utgés & Barton, 2024; Fan et al., 2025).

We address these limitations with three contributions (Figure 1):

- **CoPocket identifies pockets from a protein sequence and its MSA, without receptor coordinates or a candidate ligand.** Following UniSite's set-prediction objective (Carion et al., 2020; Fan et al., 2025), a decoder predicts scored residue masks from frozen co-folding features. Queries initialized from high-scoring residues (Zhao et al., 2024) represent candidate pockets; pair features bias attention, and confidences combine pocket presence with mask quality.

- **ChronoPocket separates newly discovered pockets from new structures of known sites.** Protein-centric labels (Utgés & Barton, 2024; Fan et al., 2025) are split by each pocket's first observation. The 30 September 2021 training cutoff matches AlphaFold 3, Boltz-1, and the original Protenix (Abramson et al., 2024; Wohlwend et al., 2024; ByteDance AML AI4Science Team et al., 2025). ChronoPocket-Discovery (ChronoPocket-D) holds out proteins with known sites before the cutoff and additional sites afterward, testing later-site recovery without training on their incomplete historical labels.

Table 1: ChronoPocket splits after model-input filtering. Each column is defined by newly observed sites, but its pocket labels also retain earlier sites on those proteins. Cross-period proteins are excluded entirely from training; their earlier pockets appear only in evaluation.

	Train	Validation	Test
First observed after (exclusive)	—	2021-09-30	2024-01-01
First observed by (inclusive)	2021-09-30	2024-01-01	2025-12-31
Proteins	15,444	2,842	2,890
Pockets first observed in this period	23,800	5,574	5,739
Pockets first observed earlier	0	1,392	1,345
Total pocket labels	**23,800**	**6,966**	**7,084**
Multi-pocket proteins (%)	31.2	46.6	46.6
ChronoPocket-D proteins (subset)	0	437	308

- **We evaluate one set of predictions at both the pocket and the residue level.** Pocket AP assesses site separation and ranking; residue AUPRC and AUROC assess binding-residue identification using confidence-weighted membership.

2 Related Work

From local scores to pocket instances. Kalasanty and PUResNet segment binding regions in voxelized receptor structures (Stepniewska-Dziubinska et al., 2020; Kandel et al., 2021); DeepSurf scores surface-centered grids (Mylonas et al., 2021), whereas EquiPocket uses equivariant message passing over surface atoms (Zhang et al., 2024b). These methods learn spatial binding propensities, with pocket extraction separate from local scoring. UniSite instead learns a set of residue masks with one-to-one matching (Fan et al., 2025). CoPocket retains this instance-level objective, but conditions its decoder on co-folding features and initializes queries from the target protein.

Transferring pretrained binding information. bindEmbed21DL predicts binding residues from sequence embeddings (Littmann et al., 2021). Hybrid models combine such embeddings with structure-derived residue graphs; Gamouh et al. (2025) show that the benefit of these graphs depends on the underlying language model. AF2BIND instead extracts binding information from AlphaFold 2 pair features, using a backbone template and amino-acid baits (Gazizov et al., 2026). These studies motivate separating the source of pretrained information from the inputs required at inference. CoPocket tests frozen co-folding residue and pair features for pocket-instance prediction from sequence and MSA, without supplying receptor coordinates or a ligand.

Annotation and evaluation units. BioLiP2 curates biologically relevant ligand contacts (Zhang et al., 2024a), while LIGYSIS and UniSite-DS consolidate evidence across structures of a protein (Utgés & Barton, 2024; Fan et al., 2025). For docking, PLINDER provides similarity measures for proteins, pockets, interactions, and ligands to construct task-specific evaluation splits (Durairaj et al., 2024). ChronoPocket addresses a complementary question: whether a later structure reveals an additional site on the same protein, rather than another observation of a known site. It therefore dates consolidated pocket instances by their first observation and evaluates later sites on held-out proteins. Appendix A details input requirements, annotation choices, and the distinction between residue- and pocket-level evaluation.

3 ChronoPocket: Pocket Labels and Temporal Evaluation

ChronoPocket combines repeated structural observations into protein-level pocket labels and dates each site by its first observation.

3.1 Predicting a Set of Pockets

Given a sequence $x = (x_1, \ldots, x_L)$ and its MSA A, the task predicts distinct ligandable sites without receptor coordinates or a candidate ligand. Reference sites form a set of binary residue masks, $\mathcal{Y} = \{\mathbf{y}_m\}_{m=1}^{M}$; overlapping sites retain separate identities. Following UniSite (Fan et al., 2025),

predictions comprise K masks $\mathbf{p}_k \in [0,1]^L$ and confidences $c_k \in [0,1]$, where p_{ki} scores residue membership and c_k ranks pocket k. The number of candidates K need not equal the reference count M.

3.2 Building Protein-Centric Pocket Labels

Different structures can show different ligands occupying the same site. Counting each complex as a separate pocket would duplicate these sites and fragment their binding evidence. We therefore consolidate observations for each UniProt protein (Utgés & Barton, 2024; Fan et al., 2025).

First, we extract ligand-contact residues from filtered X-ray and electron microscopy structures. One ligand contacting one protein chain defines a *pocket observation*, with contacts measured within 5 Å. Second, sequence alignment maps these contacts onto the canonical UniProt sequence, giving different structures a shared residue index. Third, we align observations to a protein-only predicted reference and group nearby ligand centroids. Each group becomes one *pocket instance*; its mask contains the union of the observed contact residues. Mapping retains 232,891 of 233,128 observations, and consolidation produces 37,269 pocket instances. These labels summarize observed binding evidence, not every potentially ligandable site. Filtering, alignment, clustering, and model-window projection are detailed in Appendix B.

3.3 Temporal Splits and the Discovery Subset

We date sites, not repeated structures of the same site. A pocket's earliest recorded PDB release determines its period (Table 1); subsequent observations do not make it new again. A protein enters validation or test when at least one of its pockets is first observed in that period. Its labels include all sites first observed by that period's end, preserving earlier sites alongside the new ones.

Proteins with both pre-cutoff and later-observed pockets are excluded entirely from decoder training. Neither their earlier nor their later pocket labels enter training. These proteins form *ChronoPocket-Discovery* (ChronoPocket-D), which tests recovery of additional sites on historically characterized targets (Figure 1b). "Known" here means documented in earlier structures, not seen during decoder training. For a held-out protein with sites first observed in training, validation, and test periods, validation labels include the first two sites and test labels all three. Training is protein-disjoint from evaluation, but validation and test share 152 proteins and some labels; their pocket counts are not additive. Projection and filtering details accompany Table S1.

The temporal cutoff selects pocket instances, not individual contacts: each site's residue mask combines all its observations, including later structures. The benchmark is therefore retrospective. No sequence-identity cutoff is imposed; split overlap and backbone exposure are examined in Appendix B.3.

4 CoPocket

CoPocket combines frozen co-folding features with a trainable pocket decoder. The decoder produces scored residue masks, which are merged across sequence windows to form full-length pocket predictions (Figure 2).

4.1 Adapting Co-Folding Representations

The frozen IntelliFold V2-Flash backbone (Qiao et al., 2026) encodes a sequence segment and its MSA using an AlphaFold 3-style Pairformer (Abramson et al., 2024). It returns residue features $S^{(0)}$ and pair features $Z^{(0)}$, with one vector per residue and per residue pair, respectively. We use them directly, without running the diffusion module to generate receptor coordinates. Two trainable Pairformer blocks jointly refine both tensors for pocket prediction, producing S and Z. The residue features s_i supply the decoder's input and mask readout; the pair features z_{ij} supply relational context.

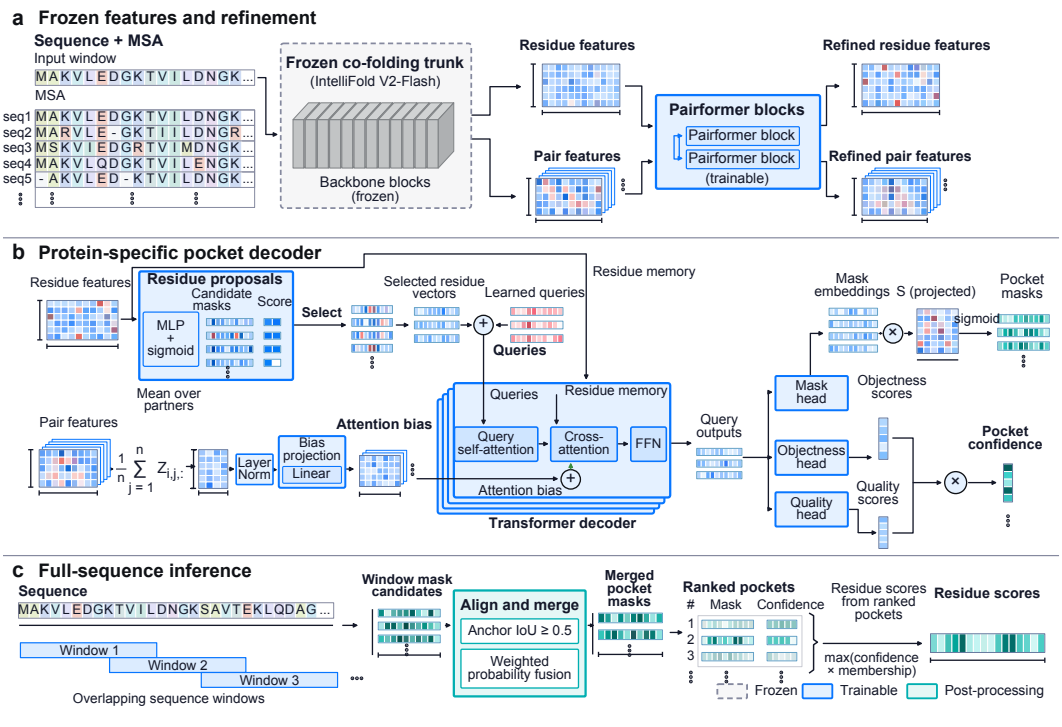


Figure 2: CoPocket architecture and full-sequence inference. (a) Frozen residue and pair features undergo trainable Pairformer refinement. (b) Residue-initialized queries and pair-derived attention biases guide pocket-mask prediction; objectness and mask quality determine confidence. (c) Overlapping-window predictions are fused into pockets and residue scores. Gray dashed, blue, and teal outlines denote frozen, trainable, and post-processing modules, respectively. Feature patterns are illustrative; settings are in Appendix C.1 and Table S2.

4.2 Predicting Pocket Instances

Each query is a vector representing one candidate pocket. To give these candidates protein-specific starting points, we adapt the proposal-based initialization of RT-DETR (Zhao et al., 2024). Each residue seeds a provisional mask, ranked by predicted presence and mask quality. Within each input segment, features from the $K = 32$ highest-ranked residues are projected and added to learned query embeddings. These queries, rather than the provisional masks, enter the decoder.

Decoder layers update queries by attending to one another and to the residue features S. Pair context modifies the attention paid to each residue: we average z_{ij} over valid partner residues $\mathcal{V}$ and project this summary into an additive attention bias,

$$\bar{z}_i = \mathrm{LN}\left(\frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}} z_{ij}\right), \qquad b_{ki}^{(\ell,h)} = w_{\ell hk}^{\top}\bar{z}_i + \beta_{\ell hk}, \tag{1}$$

where LN is layer normalization and the learned readout varies with layer ℓ, attention head h, and query slot k. The bias is added before the attention softmax. Unlike the preceding Pairformer blocks, the decoder uses a per-residue summary rather than the full pair tensor; this summary is fixed across decoder layers and covers only the input segment.

Each final query is compared with the residue features to predict membership probabilities p_{ki}. Two further outputs assess the candidate: objectness o_k predicts whether it corresponds to an annotated pocket, and quality u_k estimates its mask overlap with that pocket. We rank candidates by $c_k = o_k u_k$, combining evidence for pocket presence and mask accuracy. This confidence is a ranking score, not a probability of ligandability. The readout equations and key architecture settings are in Appendix C.1.

4.3 Training With One-to-One Matching

Hungarian matching (Carion et al., 2020; Fan et al., 2025) assigns unordered pocket labels to predictions using mask agreement and objectness. Matched queries receive positive objectness targets; unmatched queries receive negative targets. Matched masks use binary cross-entropy (BCE) and Dice losses, while quality BCE targets their soft IoU with gradients stopped. Including auxiliary proposal supervision, the objective is

$$\mathcal{L} = \mathcal{L}_{\text{obj}} + 2\mathcal{L}_{\text{mask}} + 2\mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{quality}} + 0.5\mathcal{L}_{\text{proposal}}. \tag{2}$$

Training crops come only from training-split proteins and contain at most 384 residues. They are spatial neighbourhoods with probability 0.75, preferentially centred on annotated binding residues, and contiguous segments otherwise; reference masks retain only visible residues. Structures guide crop selection during training, not window placement at inference. Training updates the two refinement blocks, proposal modules, decoder, and prediction heads; the backbone remains frozen. Loss definitions and training settings are in Appendices C.2 and C.3.

4.4 Full-Sequence Pockets and Residue Scores

Inference uses 384-residue windows at stride 192 without structures or pocket labels. Masks are mapped to full-sequence indices and grouped in descending confidence order. Each candidate joins its highest-IoU anchor if binary-mask IoU is at least 0.5; otherwise, it becomes a new fixed anchor. Within group g, confidence-weighted averaging over covering windows gives membership $\tilde{p}_{gi}$, thresholded at 0.5 for the final mask. The maximum member confidence gives the group score $\tilde{c}_g$.

The same pocket predictions also provide residue-level scores. Each residue retains its strongest confidence-weighted membership across pockets,

$$r_i = \max_g \big(\tilde{c}_g\, \tilde{p}_{gi}\big), \tag{3}$$

with zero score when no pocket remains. Mask filtering and fusion details are given in Appendix D.1.

5 Experimental Protocol

5.1 Benchmarks and Baselines

ChronoPocket test evaluation covers 2,890 proteins and 7,084 pockets. P2Rank (Krivák & Hoksza, 2018) predicts pockets and residue scores from predicted structures. A sequence-only ESM C variant (ESM Team, 2024) shares the set decoder, labels, losses, and training-step budget but changes inputs, backbone fine-tuning, and training settings. It is a system-level comparison, not a single-factor ablation (Appendix E.1).

External evaluation uses released UniSite annotations and predictions from UniSite-3D, which takes sequence and receptor structures (Fan et al., 2025). Excluding CoPocket's training UniProt identifiers leaves 537 proteins and 822 pockets, without ensuring homology or backbone-pretraining disjointness. This compares systems with different inputs and training data.

5.2 Pocket, Residue, and Discovery Metrics

Pocket detection. Pocket AP uses residue-set IoU thresholds 0.3 and 0.5, with one-to-one matching within proteins and detections pooled across proteins. ChronoPocket matches each prediction to the best *unmatched* reference. UniSite's official evaluator chooses the best reference among *all* sites and counts repeated assignments as false positives without fallback. Each benchmark applies the same rule to all compared methods.

Residue identification. ChronoPocket uses Equation 3 for CoPocket and ESM C, and native residue probabilities for P2Rank. UniSite uses a common confidence-weighted binary-mask projection (Appendix E.2). Macro AUPRC/AUROC averages per-protein values; micro pools residues. AUPRC is non-interpolated average precision. Both retain proteins with both label classes: 2,888 on ChronoPocket and 537 on UniSite.

Table 2: Overall performance (higher is better; best values bold). Compare methods within panels: cohorts, matching, and residue scoring differ (Section 5.2; Appendix E.2). Residue metrics use 2,888 proteins in A and 537 in B. One seed per training run.

		Pocket AP ↑		Residue AUPRC ↑	
Method	Input	IoU 0.3	IoU 0.5	Macro	Micro
A. ChronoPocket temporal test: 2,890 proteins; 7,084 pockets					
P2Rank	predicted structure	0.227	0.074	0.498	0.346
ESM C variant	sequence	0.377	0.230	0.609	0.558
CoPocket	sequence + MSA	**0.548**	**0.430**	**0.707**	**0.645**
B. UniSite test: 537 training-ID-excluded proteins; 822 pockets					
UniSite-3D (released)	sequence + structure	0.459	0.303	0.517	0.566
CoPocket	sequence + MSA	**0.492**	**0.347**	**0.564**	**0.572**

Discovery. On ChronoPocket-D, we report the fraction of new reference sites recovered by the top 1, 5, 10, or all predictions, using one-to-one matching at IoU 0.5. Earlier sites are not supplied to the model. Predictions matching only earlier sites receive no credit and count as false positives for new-site-only AP. Novelty filters and scoring details are in Appendix D.3.

5.3 Model Selection, Controls, and Uncertainty

CoPocket checkpoint selection maximizes validation AP at IoU 0.3 under binary-union merging at IoU 0.3. Final reporting uses the same weights with the later probability-fusion protocol at merge IoU 0.5. Checkpoint selection is described in Appendix C.3, and post-processing sensitivity in Appendix D.2.

The pair-bias ablation retains Pairformer refinement and the training and selection recipe. Ranking controls reorder fixed masks; fusion controls change grouping and aggregation of fixed window outputs. Discovery reuses the checkpoint with MSA-sampling seeds reset per protein. Each training run uses seed 42. Discovery intervals use 10,000 paired protein bootstrap resamples (Appendix D.3), measuring cohort-sampling uncertainty, not variation across training seeds or inference runs.

6 Results and Analysis

6.1 Overall Performance

On ChronoPocket test, CoPocket reaches pocket AP at IoU 0.5 of 0.430, versus 0.230 for ESM C and 0.074 for P2Rank (Table 2). Its residue macro AUPRC is 0.707 and macro AUROC 0.923 (Table S4). The ESM C comparison favours CoPocket as a system, not co-folding representations in isolation (Section 5.1).

On the external UniSite subset excluding our training UniProt identifiers, CoPocket reaches AP at IoU 0.5 of 0.347 versus 0.303 for released UniSite-3D predictions, with residue macro AUPRC of 0.564 versus 0.517. Inputs, training data, and pretraining exposure differ; training-seed uncertainty has not been measured.

6.2 Recovery of Later-Observed Pockets

ChronoPocket-D tests recovery of sites first observed in the test period on proteins with earlier known pockets. All 308 test proteins are excluded entirely from decoder training, including their earlier pocket labels; 63 also appear in discovery validation. Without historical labels as input, CoPocket recovers 108 of 527 new sites (20.5%) within its top five predictions, including at least one site on 95 proteins (Table 3). Recall rises from 8.7% at top 1 to 24.5% at top 10 and 29.8% over all candidates (Figure 3a).

To distinguish recovery from re-detection of earlier sites, we examine the 392 residue-novel sites on 239 proteins. These sites have IoU below 0.3 with every earlier pocket and at least half their residues outside the earlier union. Top-five recall is 20.2% (79 sites), falling to 19.4% (76 sites) when the

Table 3: Top-five recovery on the ChronoPocket-D test subset at IoU 0.5. Low historical overlap additionally requires predicted-mask IoU below 0.3 with every earlier site, applied after top-five selection without replacement. Residue-novel sites are a subset of period-new sites; denominators include incompletely covered sites. Intervals use 10,000 protein bootstrap resamples of fixed predictions (Appendix D.3).

Matching condition	Sites hit (n/N)	Recall (%) [95% CI]	Proteins hit (n/N)
Period-new: 308 proteins; 527 sites			
Standard	108/527	20.5 [16.7, 24.4]	95/308
Low historical overlap	82/527	15.6 [12.2, 19.1]	73/308
Residue-novel: 239 proteins; 392 sites			
Standard	79/392	20.2 [15.8, 25.0]	70/239
Low historical overlap	76/392	19.4 [15.1, 24.2]	67/239

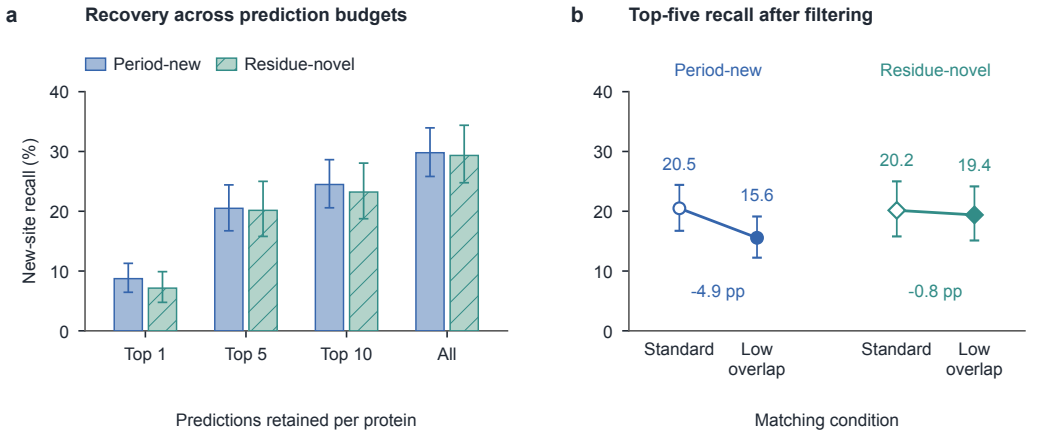


Figure 3: ChronoPocket-D discovery: period-new (308 proteins, 527 sites) and residue-novel (239 proteins, 392 sites). (a) Recall at IoU 0.5 across prediction budgets; "All" includes every candidate. (b) Top-five recall before and after excluding predictions with IoU ≥ 0.3 to any earlier site, without replacement. Labels give recall (%) and changes (pp). Whiskers show pointwise 95% recall intervals from 10,000 protein bootstrap resamples; paired-change intervals are in Appendix D.3.

matched prediction must also have low historical overlap (Figure 3b). Thus, these recoveries are not solely matches to earlier pockets. NQO1 provides an illustrative success (Figure 1b): its 2024 site overlaps the 2001 site at IoU 0.27 and is recovered at rank 1 with IoU 0.89 (Table S5). This case was selected retrospectively, not through prospective validation.

Increasing the prediction budget reveals both ranking and recovery limitations. On the residue-novel subset, the full candidate list recovers 115 of 392 sites, compared with 79 in the top five (Figure 3a). Thus, 36 recoveries lie beyond the top-five budget, while 277 sites remain unmatched at IoU 0.5 under the same one-to-one protocol. On the period-new subset, new-site-only AP is 0.028, with predictions matching only earlier sites counted as false positives. New-site prioritization therefore remains limited; no baseline comparison on this discovery subset is yet available.

6.3 Ablation and Post-Processing Analysis

Pair-derived attention bias. Removing additive pair bias produces no consistent AP change (Table 4); one run per condition cannot establish equivalence. Pair-aware refinement remains, so this does not test all pair information.

Quality-aware ranking. Combining objectness and quality improves validation AP at IoU 0.5 to 0.327, versus 0.307 and 0.284 individually. Masks are fixed, so this tests ranking rather than quality-head training (Appendix F.1).

Table 4: Training ablation and inference controls. Compare rows within blocks: independently trained models (one seed each), re-ranked fixed masks, and re-aggregated fixed window outputs, respectively.

Comparison	Split	$AP_{0.3}$ ↑	$AP_{0.5}$ ↑
Pair-bias training ablation (test; one run per condition)			
CoPocket, full model	test	0.548	0.430
without pair-derived attention bias	test	0.544	0.434
Ranking key on fixed candidate masks (validation; union merge 0.3)			
Objectness × quality (used)	val	0.541	0.327
Objectness only	val	0.528	0.307
Quality only	val	0.447	0.284
Full-sequence fusion on fixed window outputs (validation)			
Binary union, merge 0.5	val	0.550	0.409
Probability fusion, merge 0.5 (used)	val	0.548	0.442

Full-sequence fusion. With window outputs fixed, probability fusion raises validation AP at IoU 0.5 from 0.409 to 0.442 relative to binary union. This reflects changes to both grouping and aggregation, not probability averaging alone (Appendix D.2).

7 Limitations and Conclusion

Limitations. ChronoPocket is retrospective: residue masks aggregate all structures of a site, including later ones, and the temporal split controls pocket supervision only, not the pretraining of the frozen backbone (Appendix B.3). Validation and test share 152 proteins, and splits are not sequence-identity controlled, so homology to training proteins is not bounded. All CoPocket results use one training seed, and the UniSite-3D comparison involves different inputs and training data. The contribution of the trainable pair refinement has not been isolated. Inference requires an MSA, but its search cost has not been measured (Appendix F.3); pairwise context is limited to 384-residue windows. Cross-structure supervision has not been compared with single-structure supervision under matched conditions. Unmatched predictions count as false positives, but incomplete annotations prevent distinguishing incorrect predictions from unannotated pockets.

Conclusion. CoPocket identifies pockets from co-folding representations without supplied receptor structures, outperforming the tested baselines at both pocket and residue levels. ChronoPocket distinguishes later-observed sites from repeated structures: top-five predictions recover one in five later sites, leaving substantial scope to improve retrospective pocket discovery.

AI Use Statement

Language-model assistants helped draft and edit the manuscript and appendices, write analysis and figure scripts, and check reported numbers against archived result files. All experiments and values come from the authors' code and records, and the authors reviewed the manuscript.

Ethics Statement

ChronoPocket is derived from public Protein Data Bank and UniProt records. Predicted pockets indicate where small molecules may bind; they do not establish druggability, safety, or efficacy, and carry no misuse risk beyond that of existing pocket predictors.

Reproducibility Statement

Dataset construction is detailed in Section 3 and Appendix B; architecture and training in Section 4 and Appendix C; inference and evaluation in Section 5 and Appendix D; baseline provenance in Appendix E.1. Code, model weights, and machine-readable datasets and predictions are not provided with this submission.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Rishal Aggarwal, Akash Gupta, Vineeth R. Chelur, C. V. Jawahar, and U. Deva Priyakumar. DeepPocket: Ligand binding site detection and segmentation using 3d convolutional neural networks. *Journal of Chemical Information and Modeling*, 62(21):5069–5079, 2022. doi: 10.1021/acs.jcim.1c00799.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix: Advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

Anna Carbery, Martin Buttenschoen, Rachael Skyner, Frank von Delft, and Charlotte M. Deane. Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *Journal of Cheminformatics*, 16(1):32, 2024. doi: 10.1186/s13321-024-00821-4.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Jordan J. Clark, Zachary J. Orban, and Heather A. Carlson. Predicting binding sites from unbound versus bound protein structures. *Scientific Reports*, 10:15856, 2020. doi: 10.1038/s41598-020-72906-7.

Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=UfBIxpTK10`.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Gerardo Tauriello, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. PLINDER: The protein–ligand interactions dataset and evaluation resource. *bioRxiv*, 2024. doi: 10.1101/2024.07.17.603955.

ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale blog, `https://evolutionaryscale.ai/blog/esm-cambrian`, 2024. Published 4 December 2024.

Jigang Fan, Quanlin Wu, Shengjie Luo, and Liwei Wang. UniSite: The first cross-structure dataset and learning framework for end-to-end ligand binding site detection. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2506.03237. Spotlight.

Hamza Gamouh, Marian Novotný, and David Hoksza. Hybrid protein–ligand binding residue prediction with protein language models: Does the structure matter? *Bioinformatics*, 41(8):btaf431, 2025. doi: 10.1093/bioinformatics/btaf431.

Artem Gazizov, Anna Lian, Casper Goverde, Jody Mou, Sergey Ovchinnikov, and Nicholas F. Polizzi. AF2BIND: Predicting small-molecule binding sites using the pair representation of AlphaFold2. *Nature Methods*, 23(3):626–635, 2026. doi: 10.1038/s41592-026-03011-2.

Thomas A. Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2):377–389, 2009. doi: 10.1021/ci800324m.

Hyosoon Jang, Hyunjin Seo, Honghui Kim, Seonghyun Park, Taewon Kim, Yunhui Jang, and Sungsoo Ahn. A systematic evaluation of co-folding model representations for small-molecule learning. arXiv:2602.13249, 2026.

José Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander S. Rose, and Gianni De Fabritiis. DeepSite: Protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017. doi: 10.1093/bioinformatics/btx350.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Jeevan Kandel, Hilal Tayara, and Kil To Chong. PUResNet: Prediction of protein–ligand binding sites using deep residual neural network. *Journal of Cheminformatics*, 13(1):65, 2021. doi: 10.1186/s13321-021-00547-7.

Radoslav Krivák and David Hoksza. P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1):39, 2018. doi: 10.1186/s13321-018-0285-8.

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

Maria Littmann, Michael Heinzinger, Christian Dallago, Konstantin Weissenow, and Burkhard Rost. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, 11:23916, 2021. doi: 10.1038/s41598-021-03431-4.

Alex Morehead, Nabin Giri, Jian Liu, Pawan Neupane, and Jianlin Cheng. Assessing the potential of deep learning for protein–ligand docking. *Nature Machine Intelligence*, 8(1):32–41, 2026. doi: 10.1038/s42256-025-01160-1.

Stelios K. Mylonas, Apostolos Axenopoulos, and Petros Daras. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*, 37(12): 1681–1690, 2021. doi: 10.1093/bioinformatics/btab009.

Lifeng Qiao, He Yan, Gary Liu, Gaoxing Guo, and Siqi Sun. IntelliFold 2: Surpassing AlphaFold 3 via architectural refinement and structural consistency. Release note, IntelliGen-AI, 2026. URL https://raw.githubusercontent.com/IntelliGen-AI/IntelliFold/main/assets/Intellifold_v2_release_note.pdf.

Vít Škrhák, Lukáš Polák, Marian Novotný, and David Hoksza. Seq2Pocket: Augmenting protein language models for spatially consistent binding site prediction, 2026. URL https://doi.org/10.64898/2026.01.28.702257.

Zachary Smith, Michael Strobel, Bodhi P. Vani, and Pratyush Tiwary. Graph attention site prediction (GrASP): Identifying druggable binding sites using graph neural networks with attention. *Journal of Chemical Information and Modeling*, 64(7):2637–2644, 2024. doi: 10.1021/acs.jcim.3c01698.

Marta M. Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Improving detection of protein–ligand binding sites with 3d segmentation. *Scientific Reports*, 10:5035, 2020. doi: 10.1038/s41598-020-61860-z.

Javier S. Utgés and Geoffrey J. Barton. Comparative evaluation of methods for the prediction of protein–ligand binding sites. *Journal of Cheminformatics*, 16(1):126, 2024. doi: 10.1186/s13321-024-00923-z.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Chengxin Zhang, Xi Zhang, Lydia P. Freddolino, and Yang Zhang. BioLiP2: An updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024a. doi: 10.1093/nar/gkad630.

Yang Zhang, Zhewei Wei, Ye Yuan, Chongxuan Li, and Wenbing Huang. EquiPocket: An E(3)-equivariant geometric graph neural network for ligand binding site prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60021–60039. PMLR, 2024b. URL https://proceedings.mlr.press/v235/zhang24bp.html.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_DETRs_Beat_YOLOs_on_Real-time_Object_Detection_CVPR_2024_paper.html.

A Supplementary Related Work

This section expands three distinctions underlying the comparisons in Section 5: how pockets are represented, what information is required at inference, and what constitutes a new test site.

From geometric regions to pocket instances. Structure-based predictors operate on different spatial representations: fpocket groups alpha spheres, SiteMap characterizes candidate regions, and P2Rank classifies surface points before clustering them (Le Guilloux et al., 2009; Halgren, 2009; Krivák & Hoksza, 2018). DeepSite, Kalasanty, PUResNet, and DeepPocket use volumetric representations (Jiménez et al., 2017; Stepniewska-Dziubinska et al., 2020; Kandel et al., 2021; Aggarwal et al., 2022), whereas DeepSurf, GrASP, and EquiPocket encode surface or atomic neighborhoods (Mylonas et al., 2021; Smith et al., 2024; Zhang et al., 2024b). UniSite instead learns pocket masks through one-to-one set matching (Fan et al., 2025). CoPocket adopts this objective and modifies query initialization, feature transfer, and confidence estimation; set prediction itself is not our contribution.

Pretrained information and inference inputs. Sequence-based residue scoring is not necessarily coordinate-free pocket generation. bindEmbed21DL predicts binding residues from sequence embeddings, whereas Seq2Pocket subsequently uses a protein surface to extract pockets (Littmann et al., 2021; Škrhák et al., 2026). IF-SitePred and AF2BIND transfer structure-pretrained features but require a supplied backbone or template (Carbery et al., 2024; Gazizov et al., 2026). Studies of language-model features and co-folding representations further motivate separating the source of pretrained information from downstream inputs (Gamouh et al., 2025; Jang et al., 2026). CoPocket uses sequence and MSA inputs to predict pocket instances; this does not imply that its backbone was pretrained without structural information.

Annotation and generalization. BioLiP2 filters biologically relevant interactions, while LIGYSIS and UniSite consolidate observations across structures (Zhang et al., 2024a; Utgés & Barton, 2024; Fan et al., 2025). ChronoPocket builds on protein-centric annotation but dates consolidated sites rather than individual complexes. This complements chronological complex benchmarks, binding-domain generalization in DockGen, and multi-level similarity splits in PLINDER (Aggarwal et al., 2022; Morehead et al., 2026; Corso et al., 2024; Durairaj et al., 2024). Its discovery subset specifically concerns later-observed sites on proteins with earlier known pockets. Likewise, reporting both residue and pocket metrics links two established evaluation levels rather than replacing them: residue ranking measures binding-residue identification, whereas instance matching also measures site separation and delineation (Utgés & Barton, 2024; Fan et al., 2025).

B Dataset Construction and Splits

B.1 From Structures to Residue Labels

ChronoPocket consolidates ligand contacts in canonical UniProt coordinates. We use PDB biological assembly 1 structures, retain X-ray and electron microscopy entries with reported resolution below 5 Å, and assign protein chains using SIFTS. Chimeric chains are excluded. Processing also excludes files above 100 MiB and assemblies with more than 50 usable protein chains. The retained structural record extends through 2025-12-31.

Ligands come from non-polymer and branched entities, excluding water, peptides, and nucleic acids. We retain ligands with 6–50 non-hydrogen atom records and remove common crystallization additives from X-ray entries. A pocket observation contains the residues of one chain with a heavy atom within 5 Å of the ligand; at least ten contact residues with Cα atoms are required before sequence mapping. Modified amino acids are accepted according to a fixed residue list. Ligand atom counts include alternate conformations and models. The additive and residue lists belong to the extraction configuration rather than additional biological validation; no stratified manual curation was performed.

Observed chain sequences are globally aligned to canonical UniProt sequences (match $+2$, mismatch -1, gap open -10, extension -0.5). We require alignment coverage of at least 0.5 and iden-

tity of at least 0.6. Unmapped contacts are omitted, and the ten-residue minimum is not reapplied after mapping. This step retains 232,891 of 233,128 observations (Table S1).

B.2 Consolidating Repeated Observations

For each protein, we align observed pocket Cα atoms to a protein-only AlphaFold 3 reference (Abramson et al., 2024) using the Kabsch algorithm (Kabsch, 1976). A full-length reference is preferred; otherwise the longest available model window is used. At least three shared residues are required, without an additional fit-quality cutoff. The same transform places ligand centroids in this common frame, where single-linkage clustering at 5 Å groups observations of a site. The 718 observations that cannot be aligned are grouped separately by residue overlap coefficient $|A \cap B| / \min(|A|, |B|) \geq 0.5$.

Each cluster becomes one pocket whose mask is the union of its mapped contact residues. Distinct clusters may overlap in residues. We verify mask lengths, identifier uniqueness, and agreement between the union of pocket masks and the binary binding-residue labels. Labels are projected onto the available model sequence; two proteins in each evaluation split then have no retained pocket. They remain in pocket evaluation, but residue evaluation requires both label classes and therefore uses 2,888 test proteins. Of the 37,269 catalogue clusters, 75 have no contact residues within the selected model windows: 74 are removed during label projection, and one belongs to a protein excluded earlier because its projected training labels are all zero. These comprise 33 training-period, 22 validation-period, and 20 test-period clusters, leaving 37,194 distinct pocket identifiers in the final label files used for all reported metrics.

Table S1: Principal stages of ChronoPocket construction. Final validation and test masks include earlier sites on the same proteins, so split totals are not additive.

Stage	Count
Ligand–chain observations	233,128
Observations retained after sequence mapping	232,891
Observations with nonempty mapped contacts	232,876
Consolidated pocket clusters	37,269
Distinct pockets retained in final labels	37,194
Clusters first observed in train / validation / test periods	25,914 / 5,596 / 5,759
Final training proteins, after cross-period exclusion	15,444
Validation / test proteins	2,842 / 2,890
Final pocket masks: train / validation / test	23,800 / 6,966 / 7,084

B.3 Temporal Splits and Their Scope

Each pocket is dated by its earliest member structure. Dates through 2021-09-30 define training-period sites; dates after 2021-09-30 through 2024-01-01 define validation-period sites; later dates define test-period sites. We use the earliest PDB revision-history date, falling back to the deposition date when unavailable. The fallback can predate public release. A protein receives an evaluation row when it has a new site in that period, and its labels include all sites first observed by that period's end.

Cross-period proteins are excluded entirely from decoder training. Removing the 681 training-eligible proteins with sites first observed in later periods leaves 15,444 training proteins. The Discovery subsets contain 437 validation and 308 test proteins, with 63 shared between them. Neither earlier nor later labels from these proteins enter training. More generally, training and evaluation are accession-disjoint, but validation and test share 152 accessions. Cumulative test labels contain 5,739 test-period sites and 1,345 earlier sites; the discovery analysis isolates later-observed sites.

Temporal assignment applies to site identity, not to each contact in its mask: masks aggregate all member structures, including later observations. For example, 2,132 training-period clusters gain residues from later structures. The evaluation is therefore retrospective, rather than a reconstruction using only annotations available at each cutoff. No sequence-identity threshold defines the splits. An earlier homology audit used the training list before cross-period exclusion; it does not establish homology separation for the final training set.

The cutoff controls decoder supervision, not all sources of pretrained information. The frozen IntelliFold V2-Flash backbone's PDB training cutoff was not documented in the repository and model card available to us, so exposure to test-period structures is unverified. MSAs were obtained with ColabFold/MMseqs2 without a date restriction; the database version was not recorded. Template input is disabled. AlphaFold 3 reference models support annotation and spatial training crops, but are not supplied at inference. These conditions delimit the temporal claim without changing the protein-level exclusion applied to decoder training.

C Model and Training Details

C.1 Architecture

Table S2 collects the configuration. The frozen IntelliFold V2-Flash trunk processes sequence and MSA inputs without its diffusion module. For an input of $N \leq 384$ residues, it supplies residue features with 384 channels and pair features with 128 channels. Two trainable Pairformer blocks refine these features. Padding is excluded from attention, losses, and output masks.

Each residue seeds a proposal mask through a scaled dot product of learned seed and residue projections. Separate heads predict proposal objectness and quality. The 32 residues with the highest product of these scores initialize decoder queries: their projected features are added, in ranked order, to learned query embeddings. Proposal masks provide auxiliary supervision but do not enter the decoder directly.

The decoder uses query self-attention and cross-attention to residue features. Equation 1 adds a partner-averaged pair summary through a separate readout for each layer, head, and query slot. Only final decoder outputs are supervised. For final query q_k and residue feature s_i, the prediction heads are

$$p_{ki} = \sigma\left(\frac{\phi_{\text{mask}}(q_k)^\top \phi_{\text{res}}(s_i)}{\sqrt{384}}\right), \qquad o_k = \sigma(f_o(q_k)), \quad u_k = \sigma(f_u(q_k)), \quad c_k = o_k u_k, \tag{4}$$

where the projections and scalar heads are learned. The backbone remains frozen; the refinement blocks, proposals, decoder, and output heads contain 30.9 million trainable parameters.

C.2 Matching and Supervision

Let $\mathcal{V}$ denote valid residues and $\mathbf{y}_m$ a reference mask restricted to the training crop. Hungarian matching minimizes

$$C_{km} = 2\ \text{BCE}(\hat{\mathbf{s}}_k, \mathbf{y}_m) + 2\, D(\mathbf{p}_k, \mathbf{y}_m) - o_k, \qquad D(\mathbf{p}, \mathbf{y}) = 1 - \frac{2\langle \mathbf{p}, \mathbf{y}\rangle + 1}{\|\mathbf{p}\|_1 + \|\mathbf{y}\|_1 + 1}, \tag{5}$$

where $\hat{\mathbf{s}}_k$ are mask logits, BCE is averaged over $\mathcal{V}$, and all inner products and norms exclude padding. Quality predictions do not affect matching.

The objective in Equation 2 combines objectness BCE over all queries with mask BCE, Dice loss, and quality BCE over matched queries. Matched queries receive objectness target 1 and unmatched queries target 0. The quality target is the detached soft IoU,

$$\tau_{km} = \frac{\langle \mathbf{p}_k, \mathbf{y}_m\rangle}{\max(\|\mathbf{p}_k\|_1 + \|\mathbf{y}_m\|_1 - \langle \mathbf{p}_k, \mathbf{y}_m\rangle, 10^{-6})}. \tag{6}$$

Detaching the target prevents this loss from updating the mask branch through the target itself. It describes crop-local continuous overlap, whereas evaluation uses full-sequence binary IoU. Crops with no visible reference receive only query-objectness supervision from the decoder.

Proposal supervision assigns each binding residue to its largest visible reference pocket when several overlap. Positive proposals receive mask BCE and Dice loss. Proposal objectness is supervised over all valid residues; quality targets are detached soft IoU for positives and zero for negatives. An additional L_1 term penalizes disagreement between proposal objectness and these quality targets. These four principal proposal terms have unit weight, the L_1 term has weight 0.25, and the combined proposal loss has weight 0.5 in the total objective.

C.3 Training and Checkpoint Selection

Training draws 384-residue crops with replacement from pocket-indexed examples, so proteins with more pockets are sampled more often. Spatial crops are used with probability 0.75, preferentially centered on annotated binding residues; other crops are contiguous. A pocket contributes a target if at least one residue remains visible. Spatial crops use predicted structures only during training.

All models trained by us use one seed (42). CoPocket is trained for five epochs with AdamW and a one-cycle cosine learning-rate schedule (Table S2). We evaluate three times per epoch and select the checkpoint maximizing validation pocket AP at IoU 0.3. The selected checkpoint is at step 13,888. Selection used binary-union merging at merge IoU 0.3; final reporting applies probability fusion at merge IoU 0.5 to the same weights. This change is evaluated separately in Appendix D.2, rather than treated as a training improvement.

Table S2: Key architecture and training settings for CoPocket.

Component	Setting
Backbone	Frozen IntelliFold V2-Flash; up to 1,024 MSA rows
Refinement	Two trainable Pairformer blocks
Decoder	32 queries; 10 layers; width 384; 8 attention heads
Training crops	384 residues; spatial/contiguous ratio 0.75/0.25
Batch and budget	Global batch 32; five epochs
Optimizer	AdamW; peak learning rate 10^{-4}; weight decay 10^{-2}
Schedule	One-cycle cosine; 5% warm-up
Precision	bf16 mixed precision

D Inference and Evaluation Details

D.1 Full-Sequence Fusion

Inference covers each sequence with 384-residue windows at stride 192, appending an end-aligned window when needed. Candidate masks are mapped back to sequence coordinates and thresholded at 0.5; empty masks are discarded. We apply no confidence cutoff or per-protein candidate cap.

Candidates are grouped in descending confidence order using the fixed-anchor rule in Section 4.4, with merge IoU 0.5. For group g, the fused probability at residue i is

$$\tilde{p}_{gi} = \frac{\sum_{k \in g:\, i \in W_k} w_k p_{ki}}{\sum_{k \in g:\, i \in W_k} w_k}, \qquad w_k = \max(c_k, 10^{-6}), \tag{7}$$

where W_k is the candidate's window. Only windows covering i enter the average; uncovered positions score zero. Thresholding $\tilde{p}_{gi}$ at 0.5 gives the final mask, and the group's highest member confidence provides its ranking score. Empty fused masks are removed. The same fused probabilities yield residue scores through Equation 3. A protein with no prediction receives zero residue scores and no pocket detections.

D.2 Metrics and Post-Processing Sensitivity

Pocket matching. Native evaluation processes predictions in confidence order and matches each to the highest-IoU unmatched reference on the same protein, provided the evaluation threshold is met. By contrast, the official UniSite rule first chooses the highest-IoU reference among all references; an already matched reference causes a false positive without fallback. We retain the native rule on ChronoPocket and the official rule for both methods in the UniSite comparison. At each IoU threshold, detections are pooled across proteins and AP integrates the monotonically interpolated precision–recall curve. Proteins without references remain in this pool.

Residue metrics. Positive labels are the union of reference masks. Macro metrics average per-protein values; micro metrics pool valid residues. Both exclude proteins without both label classes. ChronoPocket test evaluation therefore covers 2,888 proteins and 1,530,333 residues, including

110,152 positives. AUPRC uses non-interpolated average precision with tied scores grouped at one threshold. No test-selected binarization threshold is required. Method-specific output conversions are in Appendix E.2.

Fusion sensitivity. Table S3 compares post-processing on fixed validation window outputs. Probability fusion gives higher AP at evaluation IoU 0.5 than binary union at each tested merge threshold. Merge IoU 0.3 gives the highest AP at evaluation IoU 0.5 (0.4434); the reported merge IoU 0.5 gives 0.4419. The reporting protocol was adopted after checkpoint selection and is not the optimum of this sweep. Binary union also groups against an evolving union rather than a fixed anchor, so the comparison changes both grouping and aggregation. Separate inference passes produce small numerical variation; only within-pass comparisons isolate post-processing.

Table S3: Validation post-processing sensitivity on fixed window outputs (2,842 proteins, 6,966 reference pockets). All rows use mask threshold 0.5 and objectness–quality ranking. The asterisk marks final reporting. Oracle recall is the fraction of references reachable at IoU 0.5 regardless of ranking.

Aggregation	Merge IoU	AP@0.3	AP@0.5	Oracle recall
Binary union	0.3	0.5425	0.3305	0.4552
Binary union	0.5	0.5499	0.4086	0.5662
Binary union	0.7	0.5353	0.4245	0.6161
Probability fusion	0.3	0.5494	0.4434	0.5594
Probability fusion*	0.5	0.5484	0.4419	0.5897
Probability fusion	0.7	0.5336	0.4320	0.6184

D.3 Discovery Scoring and Uncertainty

ChronoPocket-D tests later-observed sites on proteins withheld entirely from decoder training. Test-period sites are compared with earlier sites from both the training and validation periods. A residue-novel site additionally has IoU below 0.3 with every earlier site and at least half of its residues outside their union. Period-new evaluation includes 527 sites on 308 proteins; its residue-novel subset includes 392 sites on 239 proteins.

Top-k recall matches ranked predictions one-to-one to new sites at IoU 0.5 and divides total hits by the total reference-site count, including incompletely covered sites. Protein-level recovery counts proteins with at least one matched new site. New-site-only AP pools all predictions, treating predictions that match only earlier sites as false positives. Historical-overlap filtering removes predictions with IoU at least 0.3 to any earlier site only after top-k selection, without replacement. The model does not receive earlier-site labels. Discovery uses the same checkpoint as the main benchmark, but a separate export with the MSA-sampling seed reset to 42 for every protein.

For each subset, 10,000 protein-level bootstrap resamples recompute recall as total matched sites divided by total reference sites. The same sampled proteins are used across ranking depths and filtering conditions, allowing paired differences. We report percentile 95% intervals from fixed predictions, not retrained models. At top five, historical-overlap filtering changes recall by -4.9 percentage points (95% CI $[-7.0, -3.1]$) for period-new sites and -0.8 ($[-1.8, 0.0]$) for residue-novel sites. Figure 3b displays intervals for each condition, not for these paired differences. The overlapping subsets are not independent cohorts, and the intervals do not capture training-seed variability, homology dependence, or annotation uncertainty.

E Baselines and Extended Results

E.1 Baseline Configurations

UniSite-3D. We rescore the authors' released prediction files rather than retraining UniSite (Fan et al., 2025). The model uses sequence and receptor structure. Of 2,293 released test proteins, 1,756 share UniProt identifiers with our decoder-training set; excluding them leaves 537 proteins and 822 pockets. This exclusion does not remove homologues. A rerun of the public checkpoint did not

reproduce the released predictions, and its training split was not documented in the model card. We therefore use the released prediction files consistently, with the official matching rule for both methods.

P2Rank. P2Rank 2.5.1 (Krivák & Hoksza, 2018) uses its AlphaFold-specific model and configuration on the AlphaFold 3 predicted receptors for all 2,890 ChronoPocket test proteins. Its native pocket residue lists define instance masks and its pocket scores define their ranking. We do not expand, re-cluster, or truncate these predictions. Residue probabilities are taken directly from P2Rank, with zero assigned to unreported residues.

ESM C comparison. The sequence-only variant uses ESM C 600M (ESM Team, 2024) with the same 32-query, ten-layer set decoder, losses, training labels, five-epoch budget, and final window-fusion protocol as CoPocket. However, it is a system-level comparison, not a single-factor representation ablation. It fine-tunes the last six ESM C layers and final normalization, uses a two-layer residue adapter without pair features, and trains on contiguous rather than spatial crops (75% pocket-centered, 25% random). Backbone and head peak learning rates are 5×10^{-6} and 5×10^{-5}, with 100 warm-up steps. Validation selection uses probability fusion, unlike CoPocket's selection-time binary union. The selected checkpoint is from epoch five; 124.0 million parameters are trainable.

Matched pair-bias ablation. The no-pair-bias model repeats CoPocket's training recipe and selection rule with only the additive cross-attention bias removed. Pair representations still enter the two trainable Pairformer blocks, so this is not a removal of pair information. Its checkpoint is selected at step 14,880, compared with 13,888 for the full model. Each condition has one training run (seed 42), so the small, mixed metric differences in Table 4 establish neither equivalence nor a consistent gain from the bias.

E.2 Converting Outputs for Common Evaluation

For ChronoPocket pocket AP, P2Rank's exported residue identifiers are mapped to reference sequence coordinates and checked against sequence identities. All native pockets are retained and scored with the same reference masks and matching rule as CoPocket. This measures residue-footprint agreement, not geometric DCA or DCC localization.

For ChronoPocket residue metrics, CoPocket and ESM C use the continuous confidence-weighted projection in Equation 3; P2Rank uses its native residue probabilities. The same valid positions and 2,888-protein cohort are used for all methods. The selected ESM C checkpoint was rerun without retraining to export its continuous scores, reproducing the original pocket AP values to three decimal places.

The archived UniSite comparison retains binary pocket masks rather than continuous membership probabilities. Both methods therefore use

$$r_i^{\text{binary}} = \max_k c_k \, \mathbf{1}[i \in M_k], \tag{8}$$

where M_k and c_k are an exported mask and its confidence, with zero outside all masks. All predictions are retained. Reference positives are the union of official pocket masks. All 537 proteins have both label classes, giving 188,057 residues and 12,181 positives. These scores are comparable within the UniSite cohort, but should not be equated with continuous projections on ChronoPocket.

E.3 Extended Benchmark Results

Table S4 supplements the main test results with residue AUROC and per-protein pocket AP. Macro pocket AP averages per-protein AP over proteins with at least one reference pocket. All CoPocket rows use the selected full-model checkpoint unless explicitly marked as the pair-bias ablation. Comparisons are made within cohorts and matching conventions. No multi-seed uncertainty estimates are available for these benchmark comparisons. On validation, CoPocket reaches pocket AP of 0.548/0.442 at IoU 0.3/0.5 under final probability fusion, compared with 0.380/0.245 for ESM C. The archived validation evaluations report pocket AP only; validation residue AUPRC and AUROC were not computed.

Table S4: Extended test results. Panel A uses native matching on ChronoPocket; residue metrics use 2,888 of 2,890 proteins. Panel B uses official UniSite matching on the 537-protein, 822-pocket training-ID-excluded subset. Panel C uses binary-mask residue projection on that same subset.

A. ChronoPocket test

	Pocket AP		Residue AUPRC		Residue AUROC	
Method	IoU 0.3	IoU 0.5	Macro	Micro	Macro	Micro
CoPocket	0.548	0.430	0.707	0.645	0.923	0.922
no pair bias	0.544	0.434	0.703	0.646	0.922	0.920
ESM C variant	0.377	0.230	0.609	0.558	0.887	0.890
P2Rank	0.227	0.074	0.498	0.346	0.838	0.814

B. UniSite subset: pocket evaluation

	Pooled AP		Macro AP	
Method	IoU 0.3	IoU 0.5	IoU 0.3	IoU 0.5
CoPocket	0.492	0.347	0.707	0.550
UniSite-3D (released)	0.459	0.303	0.648	0.485

C. UniSite subset: binary-mask residue evaluation

	Residue AUPRC		Residue AUROC	
Method	Macro	Micro	Macro	Micro
CoPocket	0.564	0.572	0.885	0.900
UniSite-3D (released)	0.517	0.566	0.815	0.833

F Supplementary Analyses and Reproducibility

F.1 Interpreting Confidence and Recovery

The ranking diagnostic in Table 4 reorders the same 87,153 validation masks, so it isolates confidence ranking from mask generation. Oracle recall is unchanged (0.748 at IoU 0.3 and 0.452 at IoU 0.5). The objectness–quality product improves AP over either score alone on these candidates, but this is not an ablation of quality supervision during training. The diagnostic uses binary-union merging at IoU 0.3; its absolute scores should not be compared with final-fusion results.

For residue-novel discovery sites, expanding the ranking from five predictions to the full list increases recovery from 79 to 115 of 392 sites. Thus, 36 recoveries are limited by ranking depth, whereas 277 sites remain unmatched even without a top-five budget. Because annotations are incomplete, unmatched predictions cannot all be interpreted as biologically non-binding regions.

F.2 Illustrative Cases

Table S5 shows selected examples spanning immediate recovery, lower-ranked recovery, and missed sites. A successful prediction must overlap the later-observed reference at IoU at least 0.5 and have IoU below 0.3 to every earlier site. Cases are illustrative, not a prevalence estimate; aggregate discovery results use the formal one-to-one matching in Appendix D.3.

Table S5: Selected residue-novel sites in ChronoPocket-D. Rank is the first successful prediction satisfying the low-historical-overlap condition; parentheses give the best-IoU prediction's rank for a missed site. Dates and PDB entries identify the site's first retained observation.

Protein	UniProt	First observed	PDB	Rank	IoU
NQO1	P15559	2024-01-10	8PQN	1	0.895
PRDX1	Q06830	2025-01-22	9B7A	1	0.917
PPARG	P37231	2024-07-24	8U57	3	0.714
TP53	P04637	2024-11-27	8RBB	5	0.688
WRN	Q14191	2025-10-22	9RTI	7	0.545
WRN	Q14191	2024-04-24	8PFL	(32)	0.167
WRN	Q14191	2025-10-22	9RTI	(73)	0.148

F.3 Computational Scope

A single-GPU probe measured full-length training steps, not the windowed inference pipeline. On an RTX 5880 Ada GPU with bf16, batch size one, and MSA depth 1,024, peak allocated memory was 4.07 GiB at 384 residues and 26.85 GiB at 1,024 residues; second-step times were 0.60 and 7.91 s, respectively. Each length used one protein, and longer-input probes showed sample-dependent numerical instability or memory failures. These measurements motivate bounded input windows but do not establish end-to-end inference throughput. MSA generation and baseline structure prediction were not timed.