zPocket: Better Pocket Finding with Frozen Cofolding Features

Anonymous authors
Paper under double-blind review

Abstract

Identifying protein-ligand binding sites (pockets) is important for understanding protein function and developing drugs. However, no public method has leveraged cofolding models for novel pocket prediction; furthermore, this task lacks a suitable benchmark as existing benchmarks are within the training distribution of cofolding models. We therefore developed **PocketWatch**, a curated dataset of biologically relevant ligand pockets. The PocketWatch test set consists of unseen pockets stratified by sequence similarity, structural similarity, and crypticity to enable evaluation of generalization to novel and challenging pockets. Using this dataset, we develop **zPocket**, a method that trains pocket-prediction heads on top of frozen features from cofolding models. We show that zPocket heads outperform baseline methods across all stratifications. Our best zPocket model achieves overall per-target AUPRC of **0.757** compared to **0.626** for the strongest baseline, VN-EGNN, retrained on the PocketWatch split (**0.544** when using the published weights); the widely used P2Rank reaches **0.503**. We further show that zPocket can be used to outperform baseline methods on top-$N{+}2$ recall of predicted pocket centroids through clustering per-residue predictions or using zPocket to rescore centroids proposed by other models. Finally, we explore two applications of zPocket: crystallographic fragment screens and a proteome-scale scan for uncharacterized pockets on human proteins. Together these results suggest that zPocket heads built on cofolding trunk features can greatly advance ligand binding site prediction.

1 Introduction

Identifying ligand binding sites (pockets) is a foundational problem in structure-based drug design. Classical approaches used geometric cavity detection, evolutionary conservation, and energy-based scoring, and more recent methods apply deep learning with increasing sophistication. However, most methods are still limited by their dependence on input structures. Recent advances in molecular structure prediction provide an opportunity to leverage the rich representations learned by cofolding models to directly predict pockets for a given protein sequence. AF2Bind (Gazizov et al., 2026) showed that the trunk features from AlphaFold2 can be used to predict binding sites, and labs building AF3-era cofolding models also claim impressive results on the blind pocket prediction task (Isomorphic Labs Team, 2026; Anew Labs Team, 2026). However, the field lacks open benchmarks to test these claims, as existing benchmarks predate the training cutoffs of AF3-era cofolding models.

This work addresses this benchmark gap and evaluates two leading open-source cofolding models for this task. First, we build **PocketWatch**, a new temporal-split dataset of ligand binding sites. PocketWatch is a curated, per-UniProt consolidation of biologically relevant pockets whose test set contains unseen pockets with no pre-cutoff record, stratified by sequence similarity and structural similarity to cofolding model training data so we can separate generalization from memorization. We further stratify by crypticity, since cryptic pockets are particularly challenging for structure-based pocket prediction. Second, we develop **zPocket**, a set of per-residue pocket prediction heads over the trunk features of Protenix (Zhang et al., 2026) and ESMFold2 (Candido et al., 2026). We benchmark against P2Rank (Krivák & Hoksza, 2018), a widely used competitive baseline (Utgés & Barton, 2024), alongside two learned detectors evaluated using the published weights as well as retrained on PocketWatch: UniSite (Fan et al., 2025), which reads sequence alone, and VN-EGNN (Sestak et al., 2024), which reads a structure graph. We find that zPocket outperforms these baselines in every

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.

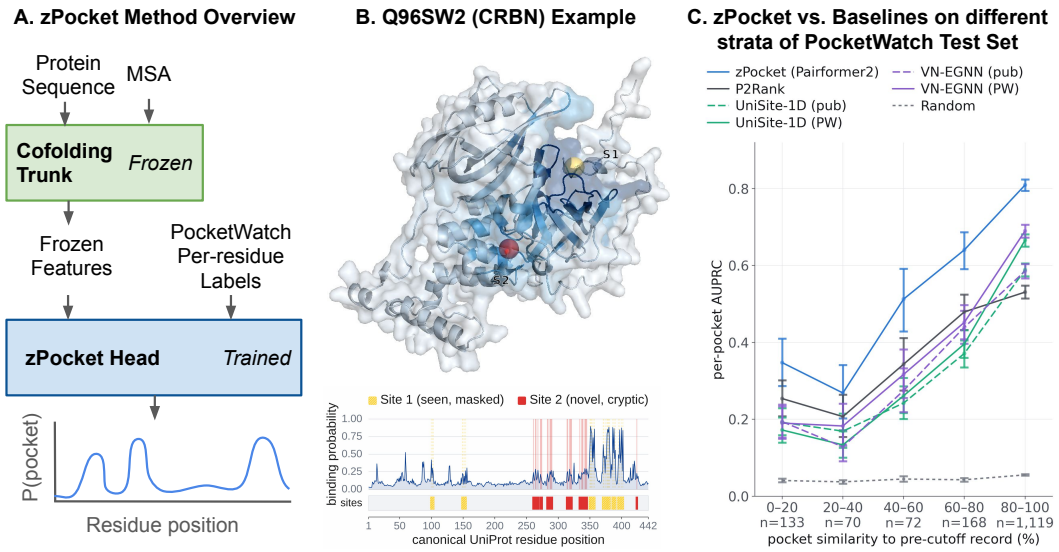


Figure 1: **zPocket predicts pocket residues from frozen cofolding features and beats baselines across different stratifications of the PocketWatch test set. (A)** Method. A sequence is folded by a frozen trunk and its frozen trunk features are cached. A head is trained on these features and per-residue labels from the PocketWatch dataset, emitting a pocket probability for every residue. **(B)** One test target, cereblon (UniProt Q96SW2), with the predicted per-residue score painted on the fold and plotted along the canonical UniProt sequence; spheres mark each annotated site's ligand centroid. Site S_1 is already in the pre-cutoff record, so its residues are masked and not scored; S_2 is novel and cryptic, first appearing after the cutoff, and is scored in full. For more details on this visualization, see Appendix B.5. **(C)** zPocket (Protenix `pairformer2`) against P2Rank, UniSite-1D, and VN-EGNN scored by per-pocket AUPRC against `max_qcov`, the largest fraction of a held-out pocket's residues covered by a pre-cutoff pocket. For UniSite-1D and VN-EGNN, we report results using both the published (pub) weights, as well as the model retrained on PocketWatch (PW). zPocket leads every baseline in every bin. Bin sizes are printed on the axis and bars are bootstrap 95% intervals.

stratum. Figure 1 summarizes the method, provides an illustration of zPocket outputs as well as the per-residue masking the temporal split rests on, and the headline stratified comparison.

2 Related Work

Binding site prediction. Classical pocket finders score surface geometry (Connolly, 1983; Hendlich et al., 1997), evolutionary conservation (Capra et al., 2009), or interaction energetics (Goodford, 1985; Laurie & Jackson, 2005). Since fpocket (Le Guilloux et al., 2009) made cavity enumeration cheap, much of the progress has come from learning to rank cavities rather than to find them: P2Rank (Krivák & Hoksza, 2018) trains a random forest to score solvent-accessible surface points from local physicochemical descriptors and clusters the high-scoring points into ranked pockets, while DeepPocket (Aggarwal et al., 2022) leaves fpocket's proposals untouched and re-ranks them with a 3D CNN over a grid centered on each cavity. Other deep learning methods predict pockets directly from a voxel grid (DeepSite, Jiménez et al., 2017) or a protein graph (VN-EGNN, Sestak et al., 2024). Despite increased sophistication, reliance on predefined features and explicit geometric inputs still limits these methods. A recent method, Lacuna, detects pockets across a generated conformational ensemble rather than on the input structure alone (Moore, 2026).

Pocket-finding benchmarks. UniSite (Fan et al., 2025) and LIGYSIS (Utgés & Barton, 2024) curate pocket datasets along lines similar to ours, but neither provides a temporal split allowing for evaluation of AF3-era cofolding models. PocketWatch extends the record of ligand binding sites found in the RCSB through the end of 2025 to allow for the identification of more novel pockets. Following Runs N' Poses (Škrinjar et al., 2025), which showed that co-folding accuracy collapses on ligand pockets far from the training set, we stratify our test set by structural similarity to better

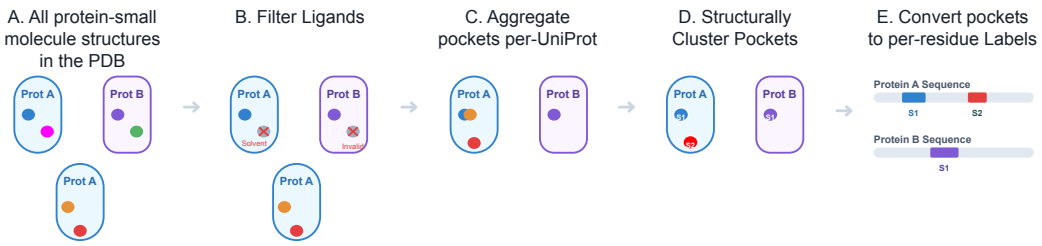


Figure 2: **PocketWatch Construction Overview (A)** We start with every protein–small-molecule structure in the PDB released prior to January 2, 2026. **(B)** We filter to only include ligands that are biologically relevant per BioLiP2 (Zhang et al., 2024) and pass our size filter; more details in Appendix A.1. **(C)** Surviving pocket instances (one receptor chain paired with one ligand occurrence) are aggregated per canonical UniProt sequence, so a protein carries every pocket ever observed on it rather than one per structure. **(D)** Instances are placed in a common frame on a predicted fold and single-linkage clustered by ligand centroid into discrete pocket sites. **(E)** Sites become a per-residue binary label vector indexed by canonical UniProt position, which is what the heads are trained and scored on.

evaluate generalization. Because a binding site open in a ligand-bound structure can be absent from the unbound form, identifying ligand sites from static structures can be challenging. The term "cryptic" is used to describe such pockets, and CryptoBench (Škrhák et al., 2025) formalizes this with an apo–holo pocket RMSD criterion over quality-controlled experimental pairs, and IsoDDE (Isomorphic Labs Team, 2026) additionally conditions crypticity on what its cofolding model saw during training. We borrow from both definitions when stratifying our test set.

Structure predictors as feature extractors. AF2Bind (Gazizov et al., 2026) read binding sites out of AlphaFold2 trunk features, but had to do so indirectly: AF2 takes no ligand input, so AF2Bind appends twenty "bait" amino acids as small-molecule surrogates and classifies the pair representation between them and the target. Using light heads trained on frozen trunk features is a commonly used recipe for other tasks, such as affinity prediction in Boltz-2 (Passaro et al., 2025).

3 Methods

3.1 PocketWatch Dataset

Figure 2 provides an overview of PocketWatch dataset construction. We aggregate ligand-contacting residues ("pocket instances") from the RCSB on a per-UniProt basis and cluster them into pocket sites. For each canonical UniProt sequence in our dataset (a "chain"), we generate a per-residue binary mask based on whether that residue participates in any pocket sites. We then generate a temporal split by assigning each chain to train, validation, or test based on the release date of its earliest member pocket instance. A chain is assigned to test when any of its sites first appeared on or after 2024-01-01, and to validation when any first appeared after 2021-09-30. More details are in Appendix A.1.

Test set stratification. We stratify the test set three ways. Test chains are binned by sequence identity to anything the models have already seen, computed with MMseqs2 (Steinegger & Söding, 2017). Test pockets are binned by structural similarity using PLINDER-style maximum pocket coverage (`max_qcov`) (Durairaj et al., 2024). Finally, we label crypticity following CryptoBench (Škrhák et al., 2025) and IsoDDE (Isomorphic Labs Team, 2026), with additional annotation of nominally cryptic pockets for which the predicted fold and ground truth pocket are actually close in RMSD. Details are in Appendix A.1.1.

3.2 zPocket: Trunk-Feature-Based Pocket Prediction

Training zPocket per-residue heads. Each protein is featurized once by folding its sequence with two (frozen) cofolding models: s and z features are retained for Protenix-v2 (Zhang et al., 2026) and last-layer ESMC-6B embeddings and z are retained for ESMFold2 (Candido et al., 2026).

We train architectures of increasing expressivity on these frozen features to locate where the pocket signal lives. Every z-based head maps a protein's pair representation z to per-residue pocket logits through a common pooling: each residue is represented by the concatenation of its row mean, column mean and self-pair diagonal, which a LayerNorm and a small MLP then reduce to a scalar. The z-based heads differ in how much pair mixing precedes that pooling. **pooled** applies it directly, with no mixing. **axial** projects z to a narrow width and applies a row-then-column self-attention block. **pairformer**K applies K AlphaFold-style triangular pair blocks (Jumper et al., 2021), which let a residue's score depend on geometry it does not directly contact. We also train **single** heads, which apply that readout MLP directly to a per-residue representation (Protenix s or last-layer ESMC-6B embeddings). Heads are trained with per-residue binary cross-entropy loss and we select the checkpoint with the best per-target validation AUPRC. Further details are in Appendix A.2.

Decoding to pocket centroids. Ligand binding site prediction is often evaluated using predicted pocket centroids. We obtain pocket centroids from our per-residue predictions in two ways: **clustering** and **re-ranking** centroids from other pocket predictors. The **clustering** decoder takes residues scoring above a threshold to seed a weighted mean shift over their Cα coordinates, and each resulting density mode is reported as one pocket. The **re-ranking** decoder instead keeps another method's proposals and re-orders them, following the division of labor in DeepPocket (Aggarwal et al., 2022) but with the zPocket head in place of its 3D-CNN: each proposed centroid is scored by aggregating the head's probabilities over its lining residues. We also evaluate "hybrid" decoders, where we combine the zPocket centroid proposals with the base method's centroid proposals before re-ranking. All decoders, and the hyper-parameters chosen on validation, are detailed in Appendix A.2.

3.3 Baselines

We compare against five pocket finders, all run on the same Protenix-predicted folds as zPocket: the geometric detectors **fpocket** (Le Guilloux et al., 2009) and **P2Rank** (Kriv\'ak & Hoksza, 2018) (alphafold configuration), two variants of the ensemble detector **Lacuna** (Moore, 2026), which detects pockets across a generated conformational ensemble rather than on the input fold alone and ranks them with its own fitted model, the graph-based detector **VN-EGNN** (Sestak et al., 2024), and the set predictor **UniSite-1D** (Fan et al., 2025), of which we run the sequence-only 1D model. We run the baseline methods using predicted folds rather than experimental structures; in Appendix B.3 we show that this has no measurable effect on P2Rank's performance.

We additionally retrain VN-EGNN and UniSite-1D from scratch with the PocketWatch train/val split and report the re-trained versions (**PW**) alongside their published weights (**pub**); fpocket, P2Rank and Lacuna appear as published only. Since the published weights are fitted to a different dataset with different ligand policies and contact thresholds, the retrained models allow us to measure the value of the PocketWatch dataset itself. Retraining required no new annotation; both models' supervision signals — per-residue ligandability, and one centroid per pocket for VN-EGNN's position loss — are already carried by our labels. The published and re-trained VN-EGNN models differ slightly in input features, since the authors' current graph builder adds a residue-type one-hot the published checkpoint lacks. Neither VN-EGNN nor UniSite-1D emits per-residue scores natively — VN-EGNN emits ranked points and UniSite-1D a set of scored masks — so both are reduced to a per-residue field as described in Appendix A.3, which also gives retraining protocols, versions and settings for every baseline.

3.4 Evaluation

We evaluate two tasks: per-residue prediction and pocket-centroid prediction. The **per-residue** task ranks a chain's residues by predicted pocket probability. We score it both per target, treating all of a protein's pocket residues as positives, and per pocket, taking one pocket's residues as the positives and masking out every other pocket. We report average precision (AUPRC) as the headline, alongside **AUROC** and **R-precision**. The **pocket-centroid** task scores a ranked list of predicted centroids. Following the top-N+2 recall proposed by Utgés & Barton (2024), we cut the list at the protein's own site count plus two and report **DCC**$_{N+2}$: the fraction of ground-truth sites with a kept centroid within 4 Å of their ligand centroid. A budgeted recall says nothing about what a method emits beyond the budget, so we also report **AP** over the whole list and the **ceiling**, the same DCC rate with no budget. Full definitions are in Appendix A.4.

Table 1: Per-residue results on PocketWatch, for zPocket heads on the **Protenix** trunk against both published and re-trained baselines, scored on the 1,562 test pockets of 1,275 targets in the effective test set. Table A4 compares results from both ESMFold2 and Protenix. All baselines are run on Protenix's predicted folds when structure is needed. Rp = R-precision. Per-target metrics are computed once per target from that target's residue ranking, then averaged over targets with equal weight. Per-pocket metrics are computed once per pocket, with positives from other pockets masked out, and averaged over pockets with equal weight. The last two columns split AUPRC_p by the number of scored pockets for each protein. (pub) is the published weights; (PW) is the same architecture retrained on PocketWatch. UniSite-1D (PW) is the mean of three seeds; every other row is a single model. Bold = best per column.

		per target			per pocket			AUPRC_p by multiplicity	
Method	Variant	AUPRC_t	AUROC_t	Rp_t	AUPRC_p	AUROC_p	Rp_p	1 pocket	>1
Baselines	P2Rank	0.503	0.869	0.473	0.479	0.864	0.447	0.495	0.444
	VN-EGNN (pub)	0.544	0.852	0.500	0.502	0.837	0.460	0.553	0.395
	UniSite-1D (pub)	0.531	0.878	0.507	0.497	0.870	0.474	0.533	0.421
PocketWatch Retrained	VN-EGNN (PW)	0.626	0.888	0.579	0.582	0.869	0.536	0.634	0.472
	UniSite-1D (PW)	0.595	0.898	0.568	0.552	0.883	0.529	0.598	0.458
zPocket	single (s)	0.474	0.874	0.444	0.442	0.863	0.413	0.468	0.390
	`pooled`	0.669	0.938	0.614	0.633	0.927	0.579	0.667	0.560
	`axial`	0.724	0.950	0.658	0.683	0.941	0.620	0.726	0.594
	`pairformer1`	0.746	0.956	0.680	0.702	0.946	0.637	0.750	0.602
	`pairformer2`	**0.757**	**0.958**	**0.692**	**0.714**	**0.948**	**0.650**	**0.761**	**0.617**
random		0.049	0.500	0.049	0.051	0.500	0.051	0.053	0.048

Which pockets are scored is decided once, at the site level, and both tables read the same decision. We mask "seen" pocket sites, those with a member pocket instance predating the test cutoff; "interface-only" pockets, those every one of whose member pockets involves more than one chain, so that P2Rank and the other single-chain methods are not charged for contacts a monomer fold cannot express; and the few novel sites whose residues lie entirely inside an already-public pocket. A pocket observed at an interface in some structures and on a single chain in others is *not* interface-only: it is scored, on its monomer-expressible residues alone. Masked pockets are neither credited nor charged — their residues are excluded from every other pocket's negatives, and a predicted centroid landing on one is dropped from the ranked list rather than counted as a false positive. Of 1,577 test pockets surviving the first two rules, 1,573 are scored over 1,279 chains; the tables report 1,562 over 1,275, restricted to the proteins every method returned a prediction for.

4 Results

4.1 zPocket beats all baselines on per-residue pocket prediction

Table 1 reports zPocket head architectures on Protenix trunk features against P2Rank (Krivák & Hoksza, 2018) run on Protenix folds; Table A4 contains additional results for ESMFold2. Every z-based zPocket head on every trunk beats the published and re-trained baselines by a wide margin. The architecture ordering `pairformer2` > `pairformer1` > `axial` > `pooled` is identical on both trunks and on nearly every metric: pair mixing before the pooled readout buys real signal, but much of the signal can be read directly off z as shown by `pooled`. The best model, Protenix `pairformer2`, reaches 0.757 per-target AUPRC against 0.503–0.626 for the baselines. Table A4 shows that the two trunks are close to interchangeable at matched architectures, with Protenix slightly but consistently beating ESMFold2. Performance mostly saturates after one pairformer block, the second buying $+0.011$. AUROC spans 0.938–0.958 across the four z heads against 0.852–0.898 for the baselines; at a 4.86% positive rate it is dominated by easy negatives and does not really separate the heads.

The pair tensor is where the signal lives. The `single` row in Table 1 applies the same readout to the Protenix s representation; it scores 0.474 against the `pooled` row's 0.669, so most of the

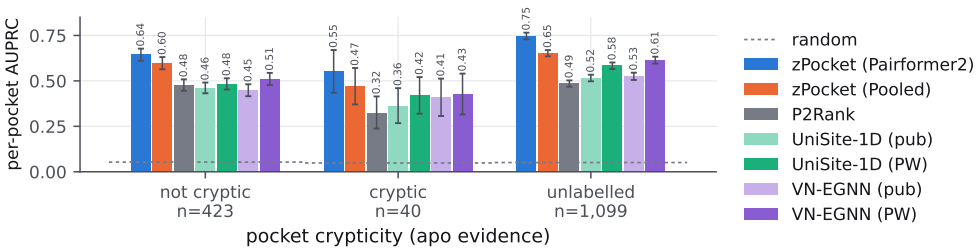


Figure 3: Per-pocket AUPRC by crypticity label. *Cryptic* pockets are those that an experimental apo structure shows closed; *not cryptic* are those an experimental apo structure shows open; *unlabeled* those with no admissible experimental apo structure, which is 70% of the test set and not a neutral remainder; a crypticity label requires a deposited apo structure, which makes the labeled subset a harder, biased sample. Bars are bootstrap 95% intervals; the cryptic column holds 40 pockets and its intervals are wide accordingly. Appendix B.4 crosses this with predicted-fold accuracy.

pocket signal sits in z rather than s, and it is the one zPocket configuration that does not beat P2Rank (0.503). ESMFold2's ESMC embedding is much the stronger per-residue feature (Table A4): it clears Protenix's own single track and all baseline methods with the exception of the VN-EGNN (PW), despite receiving no structural input whatsoever, yet still falls short of its own `pooled` row.

Retraining on PocketWatch improves baselines. VN-EGNN gains $+0.082$ per-target AUPRC over its published weights (0.544 to 0.626) and UniSite-1D $+0.064$ (0.531 to 0.595). This suggests that the PocketWatch benchmark is worth training on rather than only scoring against, and it holds for both a graph network reading structure and a transformer reading sequence. Both retrained baselines still fall short of every z-based zPocket head, so the trunk representation may still buy something the retraining does not. However, we acknowledge that some differences in training procedures (Appendix A.3) may also affect results, so it's not possible to entirely attribute the gains to the new data without further ablations. We also note that the PocketWatch training set is more closely matched to the PocketWatch test set than the training distributions of published methods, which is another mechanism by which retraining can improve performance.

Multi-pocket proteins are harder. The last two columns in Table 1 split per-pocket AUPRC by how many scored pockets a chain holds. Every z head does worse on chains holding more than one, with `pairformer2` falling from 0.761 to 0.617.

Pocket prediction drops for unfamiliar pockets. To test whether zPocket is re-identifying pockets from cofolding training data, we stratify the test set by `max_qcov`, the largest fraction of a held-out pocket's residues covered by a pre-cutoff pocket. Per-pocket AUPRC falls from 0.809 in the most-covered bin to 0.347 below 20% coverage (Fig. 1C), and 1,119 of 1,562 scored pockets — 71.6% — sit in that top bin, so the aggregate in Table 1 is dominated by familiar pockets. Structural similarity is the more informative leakage axis compared to sequence similarity: per-target AUPRC is roughly flat across sequence identity until the $\geq$95% bin, where it falls (Fig. A1). We read that fall as an artifact of seen pockets being masked, leaving only harder pockets. Despite lower results on unfamiliar pockets, `zPocket pairformer2` still leads baselines in every similarity stratum. For structural similarity stratification, the margin has non-overlapping 95% intervals from `max_qcov` 40% upward, and overlapping ones in the two lowest bins, which together hold 203 pockets. In Figure 3, we compare results across the crypticity stratification, once again observing that zPocket beats the baseline methods in every stratum, but we note the small number of cryptic pockets. Appendix B.4 crosses crypticity with predicted-fold accuracy and shows that most of the apparent effect is the fold's, not the pocket's.

4.2 Pocket centroids

In Figure 4, we compare different pocket proposal and pocket ranker combinations in terms of DCC_{N+2}, the fraction of true pocket centroids detected among the top ranked $N+2$ predicted centroids, where N is the number of true pockets for a given protein; we also show the ceiling, which is the total percentage of true pocket centroids detected among all pocket proposals, and therefore the

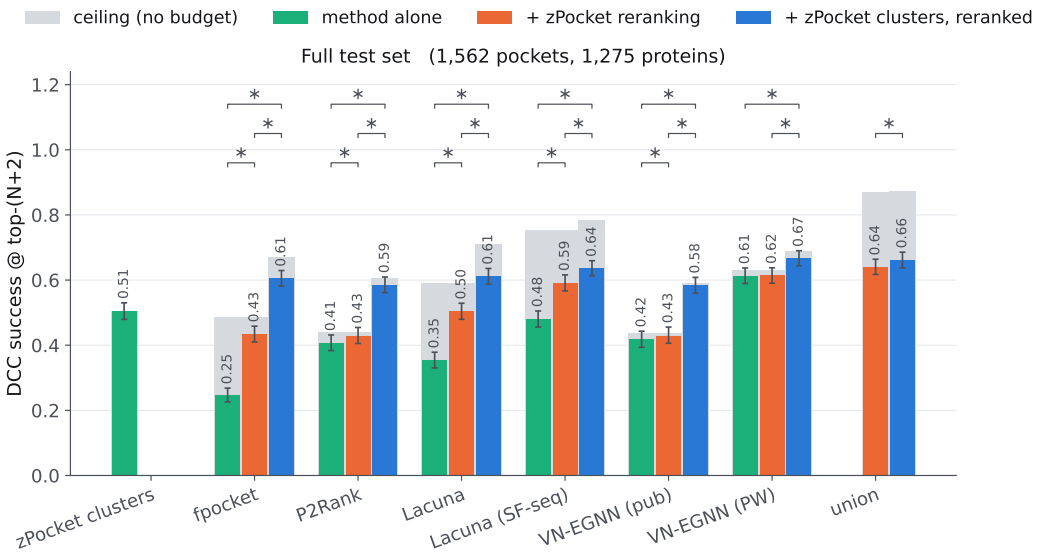


Figure 4: **Pocket centroid prediction on PocketWatch.** DCC_{N+2} over 1,562 pockets of 1,275 targets, at 4 Å. Green is each method's own proposals under its own ranking; orange re-ranks those same proposals with the zPocket head; blue adds the zPocket cluster proposals to them before re-ranking with the zPocket head. Behind *each* bar in gray is that bar's own ceiling — the total number of true pockets detected among all of the proposals. A bar close to its ceiling is limited by what was proposed; a short bar under a tall ceiling is limited by the ranking. *union* pools every method's cavities (other than zPocket clusters), so it has no ranking of its own. An asterisked bracket marks a pair whose *paired* 95% bootstrap over the same pockets excludes zero. Error bars are 95% percentile bootstrap intervals over pockets and are *unpaired*: two bars whose intervals overlap can still differ significantly once the comparison is paired. Re-ranking lifts every method except for the PocketWatch-retrained variant of VN-EGNN (PW). Combining the zPocket clusters with any other method's proposals then re-ranking lifts DCC_{N+2} further.

upper bound on DCC_{N+2}. Table A5 gives the same results numerically along with average precision, precision, and emission rate for each method.

zPocket's clustering is precise but limited. The zPocket clustering decoder emits only 1.1 pockets per protein on average, so cannot reach a protein's second site and its ceiling is 0.505. It is the most precise method by a wide margin — 55% of what it emits lands on a real site, against 18% for retrained VN-EGNN and 6% for P2Rank — but its recall is limited by its low emission rate. Despite this limitation, the zPocket clustering decoder is competitive with the other methods in terms of DCC_{N+2}, beating all published baselines using their published rankers, and only falling short of the PocketWatch-retrained VN-EGNN method.

Re-ranking with zPocket improves ranked pocket recall. Re-ranking a method's pocket proposals with zPocket significantly improved each DCC_{N+2} for each published baseline (comparing orange bars to green bars), and only failed to improve the PocketWatch-retrained VN-EGNN method. Re-ranking increased DCC_{N+2} by an average of 0.081 across these 6 methods, with the biggest gains for fpocket and the two Lacuna variants.

zPocket contributes proposals, not just ranking. We see further improvements when we combine the zPocket clustering proposals with each method's own proposals and then re-rank the union with the zPocket head (blue bars). In fact, every method, even VN-EGNN (PW), sees significant improvement in DCC_{N+2} when combined with zPocket proposals, though we note this is not the case for DCC_N (Table A5). Across these six methods and the union method, the average DCC_{N+2} gain from adding zPocket proposals and re-ranking is 0.101. We note that the highest-scoring method is the re-trained VN-EGNN method combined with zPocket proposals, which reaches 0.666. We note that even though the union of proposals had a higher ceiling, we were not able to improve DCC_{N+2} beyond this through ranking.

Pocket detection deteriorates on less familiar pockets, and ranking gets harder. In Appendix B.2, we split this comparison by `max_qcov`, and observe that the ceiling and DCC_{N+2} both drop for less-familiar pockets; In the $\geq$80% bin the best configurations reach DCC_{N+2} of about 0.79, while below 80% coverage no configuration exceeds 0.37. Furthermore, the gap between the ceiling and the ranked DCC_{N+2} is larger for less-familiar pockets, indicating that ranking is also more difficult.

5 APPLICATIONS

5.1 IN-SILICO FRAGMENT SCREENING

Evaluation dataset. Fragment screening campaigns are used to find preliminary hits against a protein target and characterize possible pockets. The protein is screened against a library of several hundred fragments and a structure is solved for every hit. This process is time consuming and expensive, and ML pocket predictors hold promise in reducing costs through narrower screening of fragments or progressing directly to hit ID. To simulate a fragment screening scenario, we assembled a test set from twelve XChem campaigns released through Fragalysis (details in Appendix C). A fragment is counted once per distinct SMILES, and sites are single-linkage clusters as in PocketWatch. This results in 9 protein targets with 43 pockets unseen by the zPocket heads during training.

Rankings mirror PocketWatch test set. The trends of Figure 4 reproduce on this independent test set (Figure 5a), with the primary exception of deteriorated performance for Lacuna after re-ranking. Re-ranking fpocket's cavities with the zPocket head lifts it from 0.201 to 0.335, and pooling in the head's own cluster proposals lifts it again to 0.357, the best configuration measured. The zPocket clustering decoder alone is once more precise but has limited recall. Overall results demonstrate that a few zPocket-derived protocols are capable of effectively recalling and prioritizing pockets in a screening setting. When considering ranking across methods, however, nine proteins is small and this evidence mostly corroborates earlier results.

Qualitative differences from PocketWatch. Two effects hold the absolute level well below the 0.67 DCC_{N+2} the same pipeline reaches on PocketWatch. The first is site multiplicity: a fragment-screened protein carries 4.8 scored sites on average against PocketWatch's 1.23, and a decoder emitting about one pocket per protein cannot reach a fifth site. Restricting each protein to its dominant site recovers most of the gap without a single prediction changing (0.357 to 0.556). Second, fragment-screening pockets are different from drug-like pockets, as fragments are often smaller and can have much lower-affinity than true binders. The sites that remain genuinely unproposed are small and often sit against a larger neighbor. For example, the eight-fragment site missed in Figure 5b shares five of its six lining residues with the 34 lining the catalytic cleft 12.6 Å away, so the head scores its lining at 0.99 while every decoder folds it into the cleft's centroid.

5.2 NOVEL POCKET HYPOTHESES

Searching for novel pockets. Strong pocket predictors can also be used to generate new ideas in target selection by surfacing high-confidence, but previously unexplored, pockets on proteins with therapeutic potential. As a proof-of-concept, zPocket was run across the ~20,000 reviewed human proteins of UniProt proteome UP000005640, yielding ~22,000 predicted pockets. The resulting pockets were then filtered for basic geometric plausibility (size and buriedness) and novelty. Novelty was determined using both sequence-level UniProt annotations and Q-BioLiP structural annotations. These annotations were used to check the input protein directly (first degree) and analogous proteins (second degree), as judged by shared structural motifs (Folddisco) or shared phylogeny. A pocket that survives these filters is one that would be very unlikely to be found by traditional tools or a database lookup. We describe the full set of novelty filters in Appendix D.

Candidate pocket examples. Of ~22,000 predicted pockets, 359 pockets on 358 proteins survived novelty and plausibility filters if polymer interface sites are not removed, which accommodates, for example, peptide binding pockets. Discarding these leads to 155 on 154 proteins. Two survivors illustrate possible ideas resulting from large-scale screening and filtering (Appendix E): DET1, a substrate-adaptor subunit of a cullin-RING ligase whose pocket is a candidate degrader-recruiter site supported by twelve deposited structures, and COP1, which the filter flags as a known peptide-

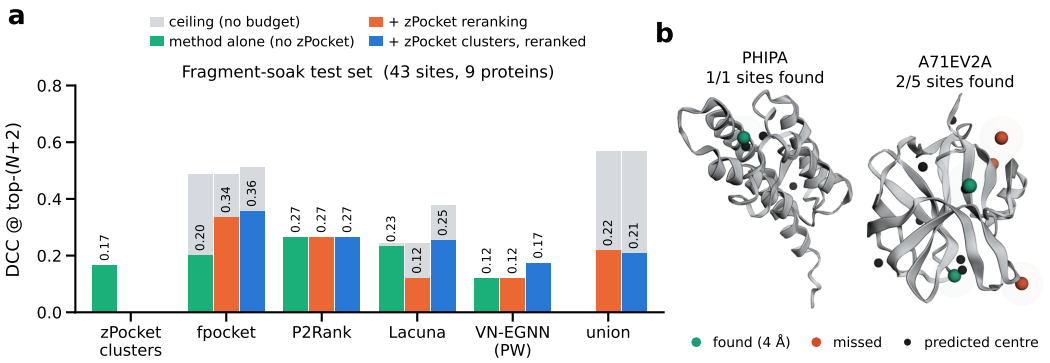


Figure 5: **Pocket prediction on crystallographic fragment screens. (a)** DCC_{N+2} at 4 Å over 43 fragment-soak sites from 9 proteins. As in Figure 4, green is each method's own proposals under its own ranking, orange re-ranks those same proposals with the zPocket head, and blue adds the zPocket cluster proposals before ranking; behind each bar in gray is that bar's own ceiling, the rate with no budget at all. *union* pools every detector's cavities and so has no ranking of its own. The trends of Figure 4 reproduce (with the exception of Lacuna), at roughly a third of the absolute level. zPocket decoder hyperparameters are identical to those in Figure 4. Lacuna is shown with numbers from default ranking, which performs best among the available options. **(b)** Two targets scored by fpocket cavities pooled with zPocket clusters and re-ranked by the head. Spheres mark each scored site's soak ligand centroid, drawn at the 4 Å success radius. PHIPA carries one site and it is found; A71EV2A carries five and two are found, which is the multiplicity effect that separates this test set from PocketWatch.

binding interface, and which no geometric detector proposes at all, yet whose predicted pocket covers fourteen of the fifteen residues contacting the bound degron peptide. In this case, zPocket re-identified a peptide-binding site, despite not being trained with peptide ligands. This pocket is not a completely novel pocket but modality-novel, as there is no known small ligand binder despite zPocket's high-confidence prediction. In discovery, there is an obvious tension between novelty and tractability. Predicted pockets on well-characterized proteins are rarely novel, especially on the reference human proteome. This was a first pass with conservative filters, and future work could refine the filtering criteria used, or focus on particular targets. Future work could also extend to screening variants and proteomes for animal models or microbes.

6 Discussion

Together these results show that modern cofolding models offer a promising foundation for blind binding site prediction. zPocket heads outperform every baseline we ran on per-residue pocket prediction, and improve the rankings of pockets proposed by other methods. PocketWatch fills a benchmark gap, allowing direct comparison of pocket finders on pockets that postdate what cofolding models have seen; its training split is also useful in its own right, lifting both baselines we retrained on it. We hope PocketWatch will be a valuable resource for developers and evaluators of pocket finding methods.

However, challenges remain. As shown by our stratifications, the number of truly novel pockets in the test set is low, and the aggregate performance is dominated by pockets with a close structural precedent; we expect the number of novel pockets available for evaluation to increase as more structures are deposited and PocketWatch is extended. We evaluate monomer pockets only, and extending to interfaces is natural future work, particularly since interface-adjacent sites are where several of our candidate pockets sit.

Finally, we also note that predicted sites may not be functional sites. However, we think this matters less than it once did and may actually be an asset: as degraders, molecular glues, and radioligand conjugates make non-orthosteric sites therapeutically actionable, the useful question shifts from where a protein does its chemistry to where a small molecule can bind at all.

7 AI USE

Tasks with required disclosure. We used generative AI tools, primarily Claude and Claude Code (Anthropic), for several tasks in the required-disclosure category. We used them to *implement methods*, including models, training code, re-training baselines, and analysis scripts. We used them to *clean and reformat data* in constructing PocketWatch, including processing Q-BioLiP data and annotating residues in RCSB structures, clustering pockets, running the SIFTS mapping, quality checking the resulting pockets, constructing the temporal split, and computing the sequence-similarity, structural-similarity and crypticity stratifications. We used them to *design and provide feedback on methodology and experiments* when choosing evaluation metrics, deciding which baselines to run, making curation and stratification choices, and conducting the novel pocket finding exploration. We also used them to *support qualitative and thematic data analysis and interpret results*. We did not use generative AI tools for the remaining required-disclosure tasks.

Tasks with recommended disclosure. Additionally, we used generative AI tools to draft parts of this paper from author notes and outlines and, in the case of methods sections, based on the code itself. We used it to review the manuscript for errors, to write and edit the plotting code behind the figures and tables, for general software engineering beyond the method implementation, and to help us search and understand the existing literature on blind pocket finding.

Review and responsibility. We have reviewed all AI-assisted work. AI-assisted code was read and tested by the authors, and every quantitative result reported in this paper is produced by the committed analysis scripts run on the actual data, never quoted from a model. AI-drafted prose was read, edited and where necessary rewritten by the authors (including this AI-disclosure section!), and every citation was manually checked. All scientific claims and interpretations in this paper are the authors' own. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

8 REPRODUCIBILITY

We plan to make the PocketWatch dataset and its stratifications fully public. We plan to publish weights for our zPocket models along with inference code, to enable running predictions on new proteins. We will also release the MSAs for our test set, allowing results on that set to be fully regenerated from featurization through to metric calculation. We also plan to release all outputs used to generate the figures in the paper. While we can make training code available, releasing all the MSAs and features is not feasible due to space constraints, so re-implementers will be responsible for MSA generation and featurization. Finally, we will release the code and weights of our retrained baselines. To preserve anonymity we do not link these artifacts here.

REFERENCES

Rishal Aggarwal, Akash Gupta, Vineeth 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.

Anew Labs Team. AnewDDE: An agentic drug discovery engine for biomolecular interaction modelling and closed-loop design. Technical report, Anew Labs, September 2026. URL https://anewbt.com/news/anewdde-agentic-drug-discovery-engine. PDF: https://anewlabs-research.tos-cn-beijing.volces.com/anewdde-tech-report.pdf. Accessed 2026-09-24.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

John A. Capra, Roman A. Laskowski, Janet M. Thornton, Mona Singh, and Thomas A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology*, 5(12):e1000585, 2009. doi: 10.1371/journal.pcbi.1000585.

Michael L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983. doi: 10.1107/S0021889883010985.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, 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.

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. *arXiv preprint arXiv:2506.03237*, 2025.

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*, 23:626–635, 2026. doi: 10.1038/s41592-026-03011-2.

Peter J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857, 1985. doi: 10.1021/jm00145a002.

M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, 1997. doi: 10.1016/s1093-3263(98)00002-3.

Isomorphic Labs Team. Accurate predictions of novel biomolecular interactions with IsoDDE. Technical report, Isomorphic Labs, February 2026.

J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose, and G. De Fabritiis. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017. doi: 10.1093/bioinformatics/btx350.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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.

Alasdair T. R. Laurie and Richard M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005. doi: 10.1093/bioinformatics/bti315.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Clayton W. Moore. Lacuna: Cryptic binding pocket discovery via conformational ensemble analysis. *bioRxiv*, 2026. doi: 10.64898/2026.08.14.744956.

Jochem Nelen, Omeir Khan, Ellen Adams, Jasmin C. Aschenbrenner, Warren Thompson, Ali Ebrahim, Eda Capkin, Charlotte Vallee, OpenBind, Elizabeth J. Shotton, Edward J. Griffen, John D. Chodera, Charlotte M. Deane, Frank von Delft, Mohammed AlQuraishi, and Fergus Imrie. The first OpenBind release: an open experimental structure–affinity dataset and benchmark for structure-based AI. *bioRxiv*, 2026. doi: 10.64898/2026.08.27.747600.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Florian Sestak, Lisa Schneckenreiter, Johannes Brandstetter, Sepp Hochreiter, Andreas Mayr, and Günter Klambauer. VN-EGNN: E(3)-equivariant graph neural networks with virtual nodes enhance protein binding site identification. *arXiv preprint arXiv:2404.07194*, 2024. doi: 10.48550/arXiv.2404.07194.

Vít Škrhák, Marian Novotný, and David Hoksza. CryptoBench: cryptic protein–ligand binding sites dataset and benchmark. *Bioinformatics*, 2025.

Peter Škrinjar, Jerome Eberhardt, Gerardo Tauriello, Torsten Schwede, and Janani Durairaj. Have protein-ligand cofolding methods moved beyond memorisation? *bioRxiv*, 2025. doi: 10.1101/2025.02.03.636309.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Javier S. Utgés and Geoffrey J. Barton. Comparative evaluation of methods for the prediction of protein–ligand binding sites. *Journal of Cheminformatics*, 16, 2024.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42:243–246, 2024.

Sameer Velankar, José M. Dana, Julius Jacobsen, Glen van Ginkel, Paul J. Gane, Jie Luo, Thomas J. Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J. Kleywegt. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41(D1): D483–D489, 2013.

Hong Wei, Wenkai Wang, Zhenling Peng, and Jianyi Yang. Q-BioLiP: A comprehensive resource for quaternary structure-based protein–ligand interactions. *Genomics, Proteomics & Bioinformatics*, 22(1), 2024.

Chengxin Zhang, Xi Zhang, Peter L. Freddolino, and Yang Zhang. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, and Wenzhi Xiao. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *ByteDance Seed technical report*, 2026.

A Detailed methods

This section mirrors Section 3: one subsection per methods subsection, in the same order.

A.1 PocketWatch dataset

Initial curation Following LIGYSIS (Utgés & Barton, 2024), we limit our set of ligand interactions to those defined "biologically relevant" per BioLiP2 (Zhang et al., 2024). We started from the Q-BioLiP (Wei et al., 2024) dataset, as it provided the most up-to-date snapshot at the time of writing (annotations through 2026-01-02) and also provides annotations for the number of neighboring chains, making it easy to restrict pockets to those formed by single chains. We match ligand pocket rows

Table A1: PocketWatch statistics. Splits are temporal and assigned per protein. The parenthesized numbers show the number of surviving chains/pockets/residues when both the seen and multi-chain-only pockets are masked. **Effective test set** is the set every number in this paper is computed on; it applies two further exclusions. First we mask 4 pockets that are novel by centroid but whose residues lie entirely inside an already-public pocket, leaving 1,279 chains / 1,573 pockets. Second we drop the 11 pockets on four proteins that VN-EGNN did not support due to unknown residues.

Split	Chains	Pockets	Residues	Positive Residue %
Train	11,016 (9,362)	14,331 (11,850)	4,628,164 (4,037,795)	6.32% (6.18%)
Validation	1,553 (1,266)	2,165 (1,508)	724,783 (585,943)	5.75% (5.12%)
Test	1,566 (1,283)	2,301 (1,577)	766,340 (629,524)	5.73% (4.85%)
Effective test set	1,275	1,562	625,883	4.86%
Total	**14,135 (11,911)**	**18,797 (14,935)**	**6,119,287 (5,253,262)**	**6.18% (5.90%)**

from Q-BioLiP to their corresponding RCSB ligand pockets, which lets us map each contacting residue onto its underlying UniProt sequence through SIFTS and annotate every residue with a heavy atom within 5 Å of the ligand.

Following IsoDDE (Isomorphic Labs Team, 2026), we make further curation decisions to restrict our set of pockets to high-quality "drug-like" pockets. We define a *pocket instance* as one receptor chain paired with one ligand occurrence. Its residues are those with a heavy atom within 5 Å of a ligand heavy atom, computed from the deposited coordinates; we also support taking Q-BioLiP's own published contact set instead, which is the stricter of the two, but every number in this paper uses the 5 Å definition. We retain instances whose ligand has 10–50 resolved heavy atoms (slightly stricter than IsoDDE's 6–50 range), whose chain has $\geq$10 contact residues with $>$80% SIFTS (Velankar et al., 2013) coverage, and which map to exactly one UniProt accession of canonical length 40–1200. We then further remove crystallization additives, metal clusters and metallo-cofactors, tetrapyrroles, lipids and terpenoids, small metabolites, and free amino acids, while retaining organic cofactors, saccharides, and nucleotides. We recognize that curating such a dataset involves a long series of decisions on inclusion/exclusion criteria, contact-residue definition, and pocket clustering protocol; we have parameterized each of these decisions to enable different PocketWatch constructions in the future; here we present results on a first iteration of PocketWatch.

Structural pocket-site clustering. A single binding site can be observed many times with different ligands, so we cluster the resulting set of pocket instances into discrete pocket sites. Following IsoDDE, we place each pocket in a common frame by fitting contact-residue Cα atoms onto the corresponding positions of a single Protenix-predicted structure of the canonical UniProt sequence, then carrying that pocket's observed ligand centroid through the same rigid transform. We then join pocket instances whose centroids lie within 5 Å through single-linkage clustering. Each pocket site is left with a labeled centroid in the frame of the Protenix fold, and a set of contact residues which is the union of its members' contact residues.

Per-residue labels. We emit a per-residue label vector per protein, indexed by canonical UniProt position. A residue is positive if it is part of one of its protein's ligand binding sites, and negative otherwise. Additionally, we provide a per-residue mask; masked residues are neither positive nor negative at training and evaluation time, and are dropped from the loss and from every metric.

Some ligand pockets involve multiple contacting chains. Where every member pocket instance of a site does, we call that site "interface-only". Whether these count as pockets is a policy choice, but we choose to mask them at evaluation time to enable fair comparison to P2Rank and other baselines, which cannot see the neighboring chains when run on single-chain folds. Table A1 summarizes the resulting splits, with and without interface-only pockets.

A.1.1 Test set stratification

Sequence-similarity stratification. We use MMseqs2 (Steinegger & Söding, 2017) to stratify test chains based on similarity to the union of the trunk's pretraining sequences and the PocketWatch

training and validation sequences. Validation sequences are included in the reference set because a split used for model selection is not unseen. A hit counts when it spans at least 50 aligned residues and covers half of *either* sequence, because a pretraining record may be just a single domain; identity is taken as the larger of the alignment-length figure and one normalized by the shorter sequence, since gap columns otherwise penalize a structure with missing unresolved loops. Independently, we flag any protein whose deposited structure appears in the pretraining index outright. Both measures err toward reporting more leakage. Only 16% of scored test proteins fall below 30% identity to anything the model has seen, while 45% sit at or above 70% and 32% at or above 95%, so a single averaged number over this test set would substantially overstate generalization.

Pocket-similarity stratification. A temporal cutoff guarantees that a held-out site is novel *by date*, not *by structure*: a site whose observations are all recent may sit on a fold whose homologues were observed with a bound ligand before the cutoff. Writing P for the residues lining a held-out site's ligand and Q for the residues lining a ligand in a pre-cutoff structure T,

$$\mathrm{qcov}(P, T) \;=\; \frac{\big|\,\{p \in P : p \text{ aligns onto some } q \in Q\}\,\big|}{|P|}, \tag{1}$$

where the alignment is a structural alignment of the query chains onto T's chains, and a site's value is the maximum over all pre-cutoff structures as well as over its own member pocket instances. Equation (1) is the pocket-coverage term of the PLINDER (Durairaj et al., 2024) similarity suite. It's the protein half of the Runs N' Poses (Škrinjar et al., 2025) train–test metric; we deliberately drop the Runs N' Poses ligand-shape term, since we ask whether a site has been seen with a ligand at all rather than whether a specific ligand was seen.

Our setting differs from PLINDER's in one structural respect and three parameters. PLINDER scores a fixed pair of systems whose chains are already in correspondence, whereas we search a database and must therefore choose one. We assign correspondence using a maximum-weight bipartite matching, so that each pocket chain is mapped to at most one target chain and vice versa, so a homomeric target cannot accumulate coverage by matching one query chain against several copies of itself. We take P (the query) as a 5 Å heavy-atom shell, matching the contact definition the PocketWatch labels are built at, and keep PLINDER's 6 Å for Q (the target): the query radius sets the denominator and so belongs to the label definition, whereas the target radius decides what counts as a binding site, so this choice is conservative in that a wider target shell can only surface more precedent, never less. We consider only Q residues contacting a non-water, non-polymer ligand of at least six heavy atoms that is not, following PLINDER, on their ion-and-artifact exclusion list. Finally, we impose no sequence-identity floor on hits, unlike the 0.2 used by Runs N' Poses (Škrinjar et al., 2025), since a remote homologue holding a ligand in the same site is exactly the precedent we mean to count.

The reference set is a Foldseek (van Kempen et al., 2024) database over all RCSB (Berman et al., 2000) biological-assembly mmCIFs, restricted at build time to entries whose initial release date falls on or before 2021-09-30 (182,401 of 246,905 entries), so every hit is admissible by construction. Foldseek retains at most 300 hits per query chain, and scoring reads at most 250 candidate targets per query. We re-scored every read-limited member with that cap effectively removed and re-searched at 5,000 hits per chain every member of a site below the 0.8 threshold whose query chain was truncated; a site already above the threshold can only rise within its own bin. Afterwards 48 sites still retain a chain truncated even at 5,000, so at most 48 further sites could belong at $\mathrm{qcov} \geq 0.8$. We acknowledge that despite conservative choices `max_qcov` remains a lower bound on the precedent that exists; the novel tail of our test set should therefore be read as an upper bound on how novel it is, and the metric could be tightened in future iterations of PocketWatch.

Crypticity. A site is cryptic when no ligand-free experimental structure presents the pocket in a conformation resembling its ligand-bound form. For each site we exhaustively enumerate PDB chains mapping to its UniProt accession, keep those where the site itself is unoccupied, map residues through SIFTS, superpose each apo chain on each holo chain over pocket Cα atoms, and take the minimum all-atom and backbone pocket RMSD. Following CryptoBench (Škrhák et al., 2025) we require a pair to observe the same pocket residues and both structures to be $\leq$2.5 Å, with a 2 Å threshold; following IsoDDE (Isomorphic Labs Team, 2026) we combine a resolution-independent criterion with one restricting apo evidence to structures predating the predictor's training cutoff.

Table A2: zPocket head architectures. All heads share a common readout and differ only in the mixing applied to z beforehand. Parameter counts are for pair width $C = 256$ and exclude the frozen trunk.

Head	Mixing before readout	Width	Params
`pooled`	none	—	264,449
`axial`	1 row+column self-attention block	32	29,537
`pairformer1`	1 triangular pair block	64	231,233
`pairformer2`	2 triangular pair blocks	64	330,177

Because any filter that drops apo–holo pairs can only raise a minimum, filtering can only push a site *toward* cryptic. We therefore make missing evidence explicit: each criterion is cryptic, refuted, or *unknown*, and sites without sufficient evidence are reported as unknown rather than defaulted into either class. Of the 1,562 scored test sites, 463 carry a label — 40 cryptic, 423 not cryptic.

The cryptic label describes experimental structures and says nothing about the predicted structure our models consume. Interestingly, we noticed that the predicted apo fold often reproduces the *holo* pocket to within 2 Å all-atom even for supposedly "cryptic" pockets; furthermore, some sites labeled "not cryptic" had predicted folds with greater than 2 Å all-atom RMSD. This is why we cross the cryptic label with whether the predicted pocket lies within 2 Å all-atom of the experimental holo pocket (Fig. A3, which carries the per-stratum counts). We observe fold accuracy better explains model performance than the cryptic label itself (Appendix B.4). Only **20** scored sites are both cryptic and not already open in the model's predicted fold, too few for a stable AUPRC, so we treat them as a case-study. We note that the previously reported results on cryptic pockets by Isomorphic Labs Team (2026) and Anew Labs Team (2026) do not report the size of the cryptic subset, and followed similar procedures to ours, so it is possible that their reported results are also based on a small number of cryptic pockets.

A.2 zPocket: trunk-feature-based pocket prediction

Featurization. Each protein is featurized once by folding its sequence with two structure predictors, Protenix-v2 (Zhang et al., 2026) and ESMFold2 (Candido et al., 2026), each with MSAs, and persisting the frozen trunk's internal representations rather than only the predicted coordinates. Per sequence we store, in half precision, the pairwise representation $z \in \mathbb{R}^{L\times L\times 256}$ and the raw distogram logits $\in \mathbb{R}^{L\times L\times 64}$, together with a per-residue single representation s (the pairformer single track, $\mathbb{R}^{L\times 384}$, for Protenix; the input single embedding, $\mathbb{R}^{L\times 451}$, for ESMFold2) and, for ESMFold2, the final-layer ESMC-6B language-model embedding $\in \mathbb{R}^{L\times 2560}$. These precomputed features are keyed by sequence hash and consumed by the pocket-prediction heads with the trunks held frozen, so the structure models are run only once and never fine-tuned.

Head architectures. Table A2 contains the details of parameter counts for the heads described in Section 3.2.

Training. Heads are trained with a masked per-residue binary cross-entropy loss,

$$\mathcal{L} \;=\; \frac{\sum_i m_i\,\ell_i}{\sum_i m_i}, \qquad \ell_i \;=\; \mathrm{BCE}_{w,\gamma}\big(\hat{y}_i, y_i\big),$$
$$\mathrm{BCE}_{w,\gamma}\big(\hat{y}, y\big) \;=\; -\Big[\, w\,(1-\hat{y})^{\gamma}\, y \log \hat{y} \;+\; \hat{y}^{\gamma}\,(1-y)\log(1-\hat{y}) \,\Big]. \tag{2}$$

Here y_i is the residue's label, $\hat{y}_i$ the head's predicted probability, $m_i \in \{0,1\}$ the mask of Section 3.1, w the positive-class weight and γ the focal exponent (Lin et al., 2017); $\mathrm{BCE}_{1,0}$ is plain unweighted binary cross-entropy. For the initial version, we did not mask any pockets during training, but in the future we could use it to mask low-confidence pockets. Our reported models use $w = 1$ and $\gamma = 0$, as selected on the validation set.

The trunk is frozen throughout, so only head parameters are updated and each protein is featurized exactly once. Because lengths vary there is nothing to stack into a batch: we take one protein per

Table A3: Re-ranking configurations selected on validation. The grid spans seven aggregates (mean, max, sum, and the mean of the k highest for $k \in \{4, 8, 16, 32\}$) and four linings (the cavity's stored residue list, or a 5, 8 or 10 Å heavy-atom shell around its centroid), giving 28 configurations per detector. *stored list* is the detector's own lining where it publishes one — fpocket's alpha-sphere contacts, P2Rank's surface-point residues, Lacuna's reported residues — and, for VN-EGNN, which emits points rather than cavities, an 8 Å shell computed around each virtual-node centroid at detection time.

Proposals	Aggregate	Lining
fpocket	mean of top 32	8 Å shell
P2Rank	mean	8 Å shell
Lacuna	mean	8 Å shell
Lacuna (SF-seq)	mean	stored list
VN-EGNN (PW)	mean of top 16	10 Å shell
union	mean	stored list

forward pass and accumulate gradients over 8 proteins per optimizer step. We train for 20 epochs with Adam (Kingma & Ba, 2015) at a learning rate of 2×10^{-3} (40 epochs for the `single` heads, which converge later), evaluate every 2 epochs, and select the checkpoint with the best per-target validation AUPRC, which is then scored once on the held-out test split. We mask "seen" pockets in the validation set when computing metrics.

Clustering decoder. To obtain pocket centroids we group the residues the head is confident about by position and take each group as one pocket. Residues scoring at or above a threshold τ seed a weighted mean shift over their Cα coordinates, each residue weighted by its predicted probability; modes closer than half the bandwidth are merged, groups smaller than m residues are discarded as noise, and each surviving group is reported at the probability-weighted centroid of its residues' heavy atoms, ranked by summed residue score. The three free parameters — τ, the bandwidth, and m — are chosen by a grid search on validation scored by DCC top-N at 4 Å and frozen before test is read; the selected setting is $\tau = 0.3$, bandwidth 14 Å, $m = 3$. Two further constants are set rather than tuned: convergence is declared at 0.05 Å or 100 iterations, and a protein whose residues all fall below τ falls back to its ten highest-scoring residues so that every target receives a prediction. Clustering per-residue predictions into discrete pockets is challenging (Fan et al., 2025), and we plan to explore different clustering approaches in future iterations of zPocket work.

Re-ranking decoder. Some cavity detectors have high recall and poor ranking. DeepPocket (Aggarwal et al., 2022) exploits this by keeping fpocket's proposals and re-ranking them with a learned 3D-CNN scored on a grid around each cavity centroid. We follow the same division of labor with the per-residue head in place of the CNN: each cavity is scored by aggregating the head's probabilities over the residues lining it, and the cavities are re-ordered by that aggregate. The lining is either the detector's own residue list or every residue with a heavy atom within a radius of the cavity, and the aggregate is a mean, a max, a sum, or the mean of the k highest — all chosen on validation. We apply this to all five detectors — fpocket, P2Rank (Kriválk & Hoksza, 2018), Lacuna, and VN-EGNN (Sestak et al., 2024) as published and retrained. We also apply this to *union*, the pool of every detector's cavities (other than zPocket), deduplicated at 2 Å so that two centroids on one cavity cannot spend two ranks. Since *union* has no ranking of its own, it only appears re-ranked by zPocket. In the "hybrid" decoder, the zPocket cluster centroids are added to the union list and re-ranked using zPocket. The re-ranking hyper-parameters selected on the validation set are reported in Table A3.

A.3 Baselines

VN-EGNN. We compare against both the released VN-EGNN checkpoint and a retrained version; however these two versions are run on different code paths and so are not directly comparable. The released checkpoint is the authors' own, but it runs on an older version of the code that only supported one site per protein by construction. The newly released training code supports multiple sites per protein, so we use that version to retrain as the PocketWatch dataset carries multiple sites per protein.

The new code path further adds a residue-type one-hot to the graph representation in addition to the ESM-2 embeddings, so the model features are also different from the released version (1301 input dimension against 1280), which may also contribute to the performance difference. We trained for 300 epochs over the training proteins which survive featurization (32 out of 11,016 were dropped), keeping the checkpoint with the best validation DCC (checkpoint 274). While the repo default is 1500 epochs, we found that the validation DCC had plateaued by 100 epochs and so we stopped early. We trained a single model using seed 42.

One forward pass produces both of VN-EGNN's outputs: the ranked centroid proposals used for the pocket-centroid tables and the per-residue logits used for the per-residue comparison. The pocket centroid proposals are where VN-EGNN's virtual nodes settle, and they are ranked by the confidence head. To get per-residue scores, we read the logits from the segmentation head which produces a per-residue logit that is trained with a Dice loss. To enable consistent comparison against other methods, we dropped the 4 proteins from our test set that contained `UNK` residues as these were not supported by either the released or re-trained version of VN-EGNN; however it may be possible to patch this issue in future iterations.

UniSite-1D. The released repository ships inference and evaluation only, so we supply the optimization loop, the loss weighting and a dataset that serves our labels as set-prediction targets, adopting the Mask2Former conventions (classification 2, mask 5, Dice 5). Our retrained model is therefore UniSite's architecture and loss under hyperparameters we chose, not a reproduction of the authors' recipe, and should be read as a lower bound on what retraining this architecture on PocketWatch can reach. We train from scratch using batch size 8, for 30 epochs over the entire training set, selecting on validation per-residue AUPRC; we report the mean of three seeds, whose spread is 0.0025 SD on per-target AUPRC.

UniSite-1D emits a set of scored masks, so for the per-residue tables we take, for each residue, the maximum over queries of the query's objectness times its mask probability there. Ranking by mask probability produces worse results (0.567 to 0.296), because many queries place confident mass on the wrong region and are separated only by their objectness scores. We use the model's continuous mask probabilities rather than the binary masks its post-processing returns. We use this same method to get per-residue scores for both the published and re-trained variants of the model. We note that the frozen model is partly leaked: 216 of our test proteins (13.8%) appear in UniSite-DS, but we report it on the full test set regardless.

Lacuna. We used `lacuna-pockets` 1.2.0, invoked per fold as `lacuna discover` with `--conformers 20`, reading the same Protenix-predicted fold that the other baselines read. We report two variants: the default detector and ranker (`--detector alpha`, `--rank-by learned`), and surface fusion (`--detector surface-fusion`), which pools the geometric and learned-surface detectors and selects its own matching ranker (`learned-fused`); its sequence pass requires the optional `plm` extra and ESM-2 650M. Lacuna truncates its report to the ten best pockets by default, which would cap recall and every ceiling we report, so we removed the cap by raising it to 1,000. The most any test protein produced was 171 cavities under surface fusion and 162 under the default detector.

A.4 Evaluation

Residue-level metrics. Pocket finding is a heavily imbalanced per-residue retrieval problem, so we report threshold-free ranking metrics and their exact random expectations. Our headline metric is **per-target AUPRC**: average precision computed within each protein and then averaged over proteins, which prevents large proteins from dominating. We report **AUROC** for comparability with prior work and **R-precision** (precision among the top-k residues, where k is that target's number of pocket residues). Because a protein's pockets are pooled into a single ranking by any per-target metric, we additionally evaluate each pocket **in isolation**: its residues are the positives, the chain's non-pocket residues the negatives, and every other pocket is masked out. We then average over pockets, assigning equal weight. Every metric is computed only over unmasked residues.

Pocket-centroid metrics. A method emits a ranked list of pocket centroids, scored by distance to real ligands. **DCC** is the distance from a predicted centroid to a site's ground-truth ligand centroid,

and a prediction succeeds if it is within 4 Å, a common criterion in this literature. Following Utgés & Barton (2024), we report the top-$N{+}2$ recall metric: we cut the predicted centroid list based on the protein's own site count plus two (DCC$_{N+2}$) and report the fraction of ground-truth sites found, testing each site against every kept prediction independently. A budgeted recall says nothing about what a method emits outside the budget, so we also report **average precision** (AP) over the whole ranked list, which charges for every false positive ranked above a hit and requires one-to-one matching (a centroid between two sites is one true positive, not two), together with pooled precision and the mean list length $\bar{k}$. Finally we report the **ceiling**: the same DCC rate with no budget at all. The ceiling helps separate a detection failure from a ranking failure.

Table A4: **Protenix against ESMFold2 as a frozen trunk.** Per-residue results on PocketWatch for every zPocket head trained on each of the two trunks, scored on the 1,562 pockets of 1,275 targets in the effective test set. Protenix leads on every matched head, but the margin is small compared to the choice of head architecture. The ranking across zPocket head architectures is the same for both trunks. The `single` rows are not a matched pair: Protenix uses the trunk single representation s, while ESMFold2 uses the final-layer ESMC-6B embedding. P2Rank, run on each cofolding model's own predicted folds, is included as a structure-dependent reference point on the same folds the heads are compared over. Columns: Rp = R-precision. Per-target metrics are computed once per target from that target's residue ranking, then averaged over targets with equal weight. Per-pocket metrics are computed once per pocket, with positives from other pockets masked out, and averaged over pockets with equal weight. The last two columns split AUPRC_p by the number of scored pockets for each protein. Bold = best per column.

		per target			per pocket			AUPRC_p by multiplicity	
Trunk	Head	AUPRC_t	AUROC_t	Rp_t	AUPRC_p	AUROC_p	Rp_p	1 pocket	>1
Protenix	single (s)	0.474	0.874	0.444	0.442	0.863	0.413	0.468	0.390
	`pooled`	0.669	0.938	0.614	0.633	0.927	0.579	0.667	0.560
	`axial`	0.724	0.950	0.658	0.683	0.941	0.620	0.726	0.594
	`pairformer1`	0.746	0.956	0.680	0.702	0.946	0.637	0.750	0.602
	`pairformer2`	**0.757**	**0.958**	**0.692**	**0.714**	**0.948**	**0.650**	**0.761**	**0.617**
ESMFold2	single (ESMC)	0.611	0.918	0.564	0.574	0.906	0.528	0.610	0.499
	`pooled`	0.639	0.924	0.589	0.600	0.913	0.551	0.640	0.516
	`axial`	0.716	0.945	0.652	0.675	0.934	0.613	0.719	0.584
	`pairformer1`	0.737	0.950	0.677	0.695	0.941	0.637	0.741	0.600
	`pairformer2`	0.745	0.951	0.682	0.702	0.941	0.639	0.751	0.599
P2Rank (Protenix folds)		0.503	0.869	0.473	0.479	0.864	0.447	0.495	0.444
P2Rank (ESMFold2 folds)		0.477	0.859	0.449	0.455	0.855	0.425	0.466	0.430
random		0.049	0.500	0.049	0.051	0.500	0.051	0.053	0.048

B Supplementary results

B.1 Additional per-residue results

Table A4 extends the full per-residue comparison in Table 1 with a trunk comparison: every zPocket head on the ESMFold2 trunk alongside the same head on the Protenix trunk, with P2Rank on each trunk's own predicted folds as a reference. The learned baselines are left to Table 1. Figure A1 breaks per-target AUPRC down by sequence identity to anything the model has already seen.

B.2 Additional pocket centroid results

Table A5 gives the numbers behind Figure 4 — DCC at both budgets, the ceiling, average precision, precision and the emission rate — for every proposal and ranker combination. Figure A2 repeats the same comparison split by `max_qcov` (Section A.1.1), the largest share of a test pocket's residues covered by a pocket that was already public before the cutoff.

B.3 How do predicted structures affect P2Rank?

Every number in this paper is computed on a predicted structure. It is possible that pocket prediction tools like P2Rank, which can work on either predicted or experimental structures, might perform differently depending on the input. To check this, we ran P2Rank on a set of experimental apo structures and compared the results to those obtained from the predicted folds.

For every test target with a deposited apo chain we selected one experimental structure and ran the same P2Rank configuration on it. A chain was admitted only if it was ligand-free at every scored pocket residue (no non-polymer component within 5 Å, ignoring crystallization additives and anything under five heavy atoms), contained all annotated pockets, and came from a structure at 2.5 Å or better; among admissible chains we took the best resolution. We then ran P2Rank on this

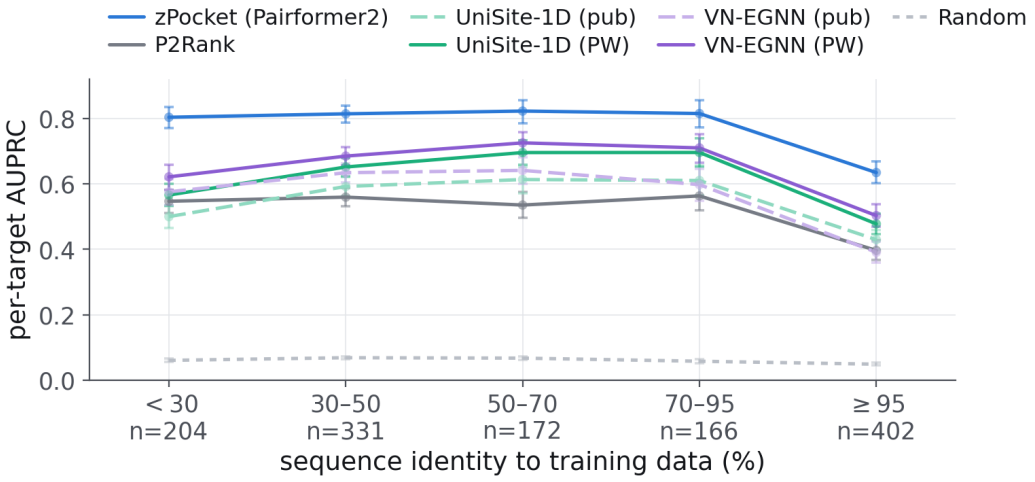


Figure A1: Per-target AUPRC against sequence identity to anything the model has already seen — the trunk's pretraining sequences plus the PocketWatch training and validation sequences. Performance is roughly flat from *novel* through *close* and falls only in the *near-identical* ($\geq$95%) bin. We suspect that this fall is an artifact of the seen-pocket mask rather than a generalization effect: a protein with pockets deposited before the cutoff has most of its familiar pockets masked, leaving the harder remainder to be scored. Bin sizes are printed on the axis and bands are bootstrap 95% intervals. zPocket outperforms all baselines.

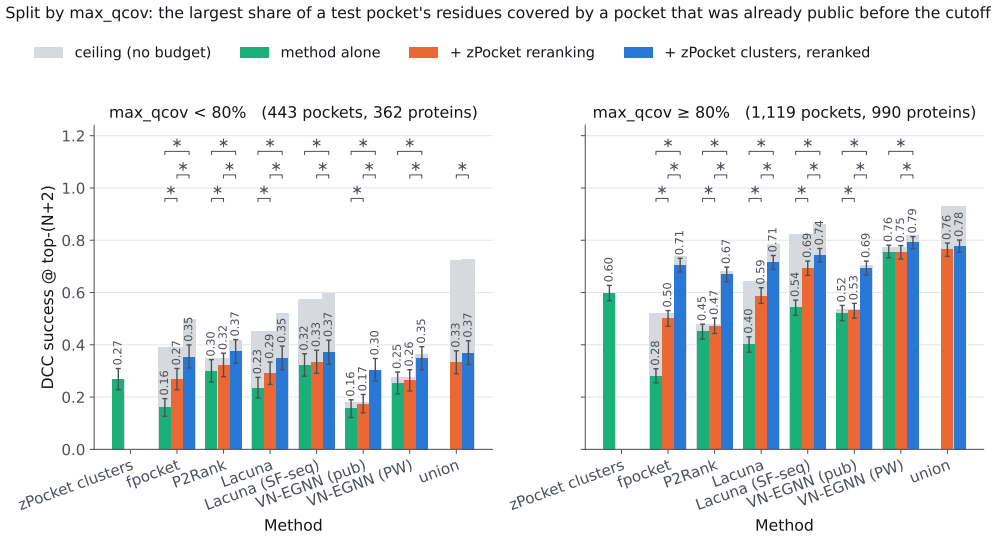


Figure A2: **Pocket centroid prediction split by pocket familiarity.** The comparison of Figure 4, DCC_{N+2} at 4 Å, computed separately on unfamiliar pockets (`max_qcov` below 80%: 443 pockets of 362 proteins) and familiar ones (`max_qcov` at or above 80%: 1,119 pockets of 990 proteins). Colors, gray ceilings, asterisked brackets and error bars are as in Figure 4: green is each method's own proposals under its own ranking, orange re-ranks those proposals with the zPocket head, blue adds the zPocket cluster proposals before re-ranking, an asterisk marks a paired 95% bootstrap excluding zero, and the intervals on the bars themselves are unpaired. Every method degrades sharply on unfamiliar pockets, and the gap between each bar and its own ceiling widens, but adding the zPocket proposals and re-ranking still lifts DCC_{N+2} significantly for most methods in both similarity bins.

experimental apo structure and compared its results to those obtained from the predicted fold of the same target, only considering the canonical positions the experimental chain observes (sometimes the

Table A5: Discrete pocket prediction on PocketWatch: how close a ranked list of predicted centroids lands to the real ligand centroids, over the same 1,562 pockets of 1,275 targets. A prediction succeeds within 4 Å of a ground-truth centroid. DCC N keeps each protein's own pocket count as the budget and $N{+}2$ allows two more; *ceiling* is the rate with no budget at all, which the proposals fix and no ranking can exceed; *prec.* is the share of everything emitted that claimed a pocket, and $\bar{k}$ how many centroids were emitted per protein. Every decoder is the winner of a validation sweep restricted to ranking by a single scorer, so no row mixes the head's score with a detector's own. *union* pools every detector's cavities, folded most-selective first and deduplicated, so it has no ranking of its own to report – only a pool for the head to rank. Bold = best per column.

			DCC success					
Section	Proposals	Ranker	N	$N{+}2$	ceil.	AP	prec.	$\bar{k}$
Base methods	fpocket	fpocket	0.168	0.247	0.485	0.228	0.017	34.1
	P2Rank	P2Rank	0.346	0.408	0.442	0.394	0.064	8.4
	Lacuna	Lacuna	0.191	0.354	0.590	0.296	0.017	42.3
	Lacuna (SF-seq)	Lacuna (SF-seq)	0.275	0.480	0.752	0.410	0.019	49.7
	VN-EGNN (pub)	VN-EGNN (pub)	0.371	0.418	0.436	0.435	0.113	4.7
	VN-EGNN (PW)	VN-EGNN (PW)	**0.534**	0.613	0.631	0.617	0.180	4.3
	zPocket clusters	zPocket	0.495	0.505	0.505	0.549	**0.554**	1.1
Rescoring with zPocket	fpocket	zPocket	0.375	0.435	0.485	0.427	0.017	34.1
	P2Rank	zPocket	0.407	0.430	0.442	0.439	0.064	8.4
	Lacuna	zPocket	0.386	0.504	0.590	0.476	0.017	42.3
	Lacuna (SF-seq)	zPocket	0.398	0.592	0.752	0.527	0.018	49.7
	VN-EGNN (pub)	zPocket	0.392	0.430	0.436	0.452	0.113	4.7
	VN-EGNN (PW)	zPocket	**0.534**	0.615	0.631	0.620	0.180	4.3
	union	zPocket	0.496	0.641	0.871	0.627	0.012	89.8
Hybrid	zPocket + fpocket	zPocket	0.519	0.606	0.671	0.603	0.023	35.1
	zPocket + P2Rank	zPocket	0.521	0.586	0.607	0.591	0.080	9.3
	zPocket + Lacuna	zPocket	0.517	0.611	0.711	0.612	0.020	43.2
	zPocket + Lacuna (SF-seq)	zPocket	0.517	0.637	0.786	0.631	0.019	50.5
	zPocket + VN-EGNN (pub)	zPocket	0.512	0.584	0.590	0.591	0.188	3.8
	zPocket + VN-EGNN (PW)	zPocket	0.528	**0.666**	0.690	0.635	0.210	4.0
	zPocket + union	zPocket	0.530	0.662	**0.874**	**0.650**	0.012	89.1

experimental chain is a subset of the full UniProt sequence). This procedure gave us an evaluation set with 433 targets and 537 pockets (398 X-ray, 35 cryo-EM; median resolution 1.85 Å).

We found that P2Rank performs similarly on both structure sources. Per-target AUPRC is 0.518 on the experimental apo structure against 0.514 on the Protenix fold of the same target, a paired difference of $+0.004$ (95% percentile bootstrap CI over targets, $[-0.013, +0.022]$); per pocket, 0.489 against 0.483 ($+0.006$, $[-0.010, +0.022]$). The interval excludes an advantage for either structure source larger than about 0.02 AUPRC, an order of magnitude smaller than zPocket's margin over P2Rank in Table A4. We therefore read the fold-based numbers throughout as a fair proxy for what these methods would reach on experimental input.

B.4 Cryptocity versus predicted-fold accuracy

Since cryptic sites might be difficult to predict from a single conformation of an unbound protein, we were eager to evaluate whether a method like zPocket, which does not depend on coordinates, would predict these pockets better than methods that do. In the course of our investigation, we found that Protenix often predicted folds close to the ground-truth pocket even for sites labeled cryptic per our definition, and that some sites labeled not cryptic had predicted folds more than 2 Å away, so we decided to stratify our results by fold accuracy as well as by the cryptic label. We think this is a useful finding about the "cryptic" definition used by IsoDDE and CryptoBench: it is a property of the experimental structures, not of the predicted structure, and a predicted apo fold can reproduce a holo pocket even if the model has never seen that holo conformation for this particular protein, perhaps by learning it from homologues.

Fold accuracy is more predictive of binding prediction performance than crypticity: when the predicted fold comes within 2 Å of the ground-truth pocket, both zPocket and P2Rank score higher regardless of crypticity label, and when it does not, every method degrades (Fig. A3). We do see a small dip in AUPRC for cryptic pockets within the "fold within 2 Å of ground-truth pocket" bucket for both zPocket and P2Rank, but three of the five buckets hold only 24, 20 and 20 pockets, so it is hard to draw conclusions from any of them. Keeping in mind these small numbers, we note that zPocket still outperforms baselines across every stratification.

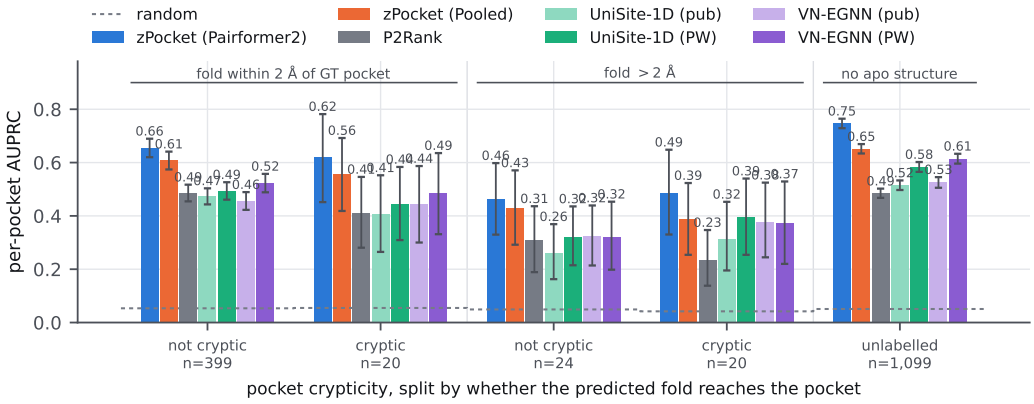


Figure A3: Per-pocket AUPRC by crypticity label, crossed with whether the Protenix predicted structure comes within 2 Å of the ground-truth pocket. zPocket heads against P2Rank and both baselines, each as published (*pub*) and retrained on PocketWatch (*PW*). Bucket sizes are printed beneath each group; bars are bootstrap 95% intervals and the dashed line is the random expectation, computed per bucket from its own positive rate.

B.5 Calibration and the main-text case-study figure

Panel (B) of Figure 1 is rendered from a head trained with a weighted focal loss ($w = 5$, $\gamma = 2$) rather than the plain unweighted BCE ($w = 1$, $\gamma = 0$) of Eq. 2 that produces every quantitative result in this paper. The two heads are otherwise identical. We show this variant because its higher positive weight and flatter calibration make S_2 easier to visualize, and we report the plain-BCE head because it scores higher on the validation set (0.774 against 0.766 AUPRC). In fact, even though the weighted-focal head makes S_2 easier to see, the plain-BCE head actually ranks the residues of S_2 higher, with a per-pocket AUPRC of 0.312 compared to 0.247 for the weighted-focal head. So the main text does not flatter the case study by showing the weighted-focal variant. We show the same target scored by the reported model in Figure A4.

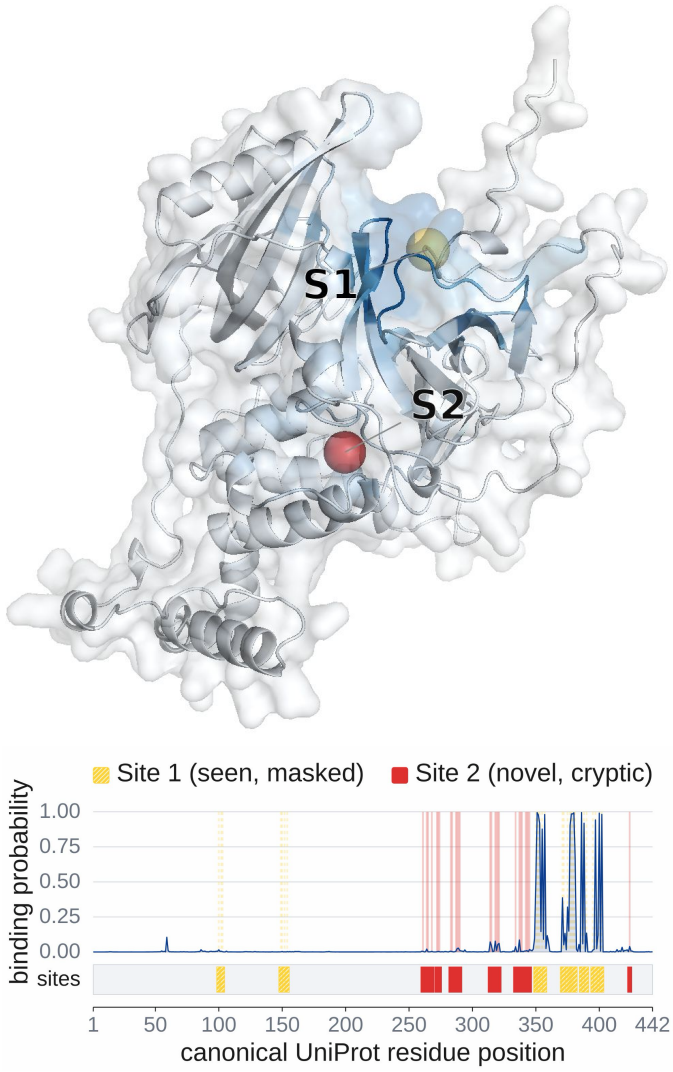


Figure A4: **The main-text case study rendered from the reported model.** The same target and annotated sites as Figure 1(B), scored by the plain unweighted BCE head ($w = 1$, $\gamma = 0$) used throughout the paper. Site coloring and seen-pocket masking follow Figure 1(B).

C SOURCING AND CURATION OF THE FRAGMENT-SOAK TEST SET

Fragment screens were taken from Diamond Light Source's XChem programme through Fragalysis, downloaded on 2026-09-20/21 from the legacy host (per-target archives) and from the production and staging hosts of the current stack (an unauthenticated REST interface), giving 61 distinct campaigns; one of them, the EV-A71 2A protease campaign, is the first OpenBind release (Nelen et al., 2026). Campaigns were selected before any prediction was run, on two independent leakage axes computed for all 61. On the *head* axis a campaign is clean when its accession is absent from the zPocket head's label table and nothing in that table exceeds 50% identity to the construct, and it is discarded when the accession sits in the head's training split or something at $\geq 90\%$ identity does. On the *trunk* axis we count PDB entities at $\geq 70\%$ identity released before the cofolding trunk's 2021-09-30 cutoff, and discard a campaign carrying more than fifteen. Raw counts mislead in both directions, so the selection also recorded the two quantities behind them. The fraction of a protein's entities that predate the cutoff separates campaigns the trunk merely glimpsed from campaigns it knows as well as it ever will: A71EV2A and SOS3ATA both look sparse at 6 and 3 pre-cutoff entities, but A71EV2A has 480 entities in total while all three of SOS3ATA's predate the cutoff, so the second was dropped and the first kept. The count of pre-cutoff relatives at 30% identity separates a sparsely sampled protein from a densely sampled fold: CAMK1DA has ten pre-cutoff relatives at 70% identity but 1,079 at 30%, because it is a kinase, and a per-residue head reading a trunk representation is likelier to transfer at fold level than an identity tier implies, so it too was dropped. The screens excluded this way include the SARS-CoV-2 main protease and nsp13 campaigns, which are trunk-saturated, and the human NUDT, BRD and PTP1B campaigns, whose accessions are in the head's training split. Twelve campaigns survived.

Within a campaign, a hit is the soaked compound's own coordinate file rather than a `HETATM` record, because the deposited complexes also carry acetate, DMSO and ethylene glycol that were never in the library; a fragment is counted once per distinct canonical SMILES however many crystals it appears in. Every bound structure was checked to sit already in its campaign's apo reference frame (Cα deviation below 1.5 Å) and ligand coordinates were used as deposited. Where a second protein chain is present it is either a crystallographic copy of the modeled chain or a genuine partner, and the two demand opposite treatment: ligands on a copy are carried onto the modeled chain by superposition, while ligands at a partner interface (nsp16/nsp10, NXF1/NXT1) are dropped and flagged. The two cases are told apart by sequence identity to the modeled chain and never by chain label, since a campaign's template chain letter frequently disagrees with its bound files. One campaign needs a further distinction: the 925 site observations released for A71EV2A are 113 fragment-screen events followed by 812 lead-like compounds from the optimization phase that followed, and treating all 925 as fragments saturates the active site and inflates the positive base rate to most of the accessible surface, so only the fragment-screen subset is used. Sites are then single-linkage clusters of ligand centroids at 5 Å in the apo frame, the rule PocketWatch already uses. Of the 123 sites this produces, 71 are masked — crystal-packing artifacts, taken as a median symmetry-mate contact fraction ≥ 0.25 over the site's fragments (for A71EV2A, the release's own artifact flag), and singletons carrying one fragment — leaving 52 scored sites over 11 campaigns, one campaign retaining none. Dropping the sites the head has seen in its own label table gives the 43 sites over 9 proteins that are scored in the main text.

D POCKET NOVELTY FILTERS

D.1 WHY KNOWN SITES ARE RESOLVED BY HOMOLOGY

A filter that asks whether a pocket coincides with a known binding site on *its own* protein cannot distinguish two outcomes: the site was tested and is clear, or nothing was ever known about that protein. Keyed on the query accession, our annotation reached only 2,444 of the 20,092 scanned proteins, so 87% of pockets cleared that gate without any evidence being brought to bear. An earlier revision of this pipeline produced a shortlist of four leads on that basis; independent review established that every one was a known functional site invisible to the gate, in each case because no experimental structure of the human protein exists. The filters described below resolve known sites through homologues, and report per gate whether evidence was available at all (§D.7).

Table A6: Pocket filtering funnel over the reviewed human proteome, grouped by how evidence of a known site is propagated to the query: not at all, by sequence homology to Pfam family and clan relatives, or by structural-motif similarity irrespective of homology. Geometric requirements are independent of all three. Counts are pockets remaining after each successive filter; proteins are the distinct accessions those pockets lie on. Filters are conjunctive, so the survivor count is independent of their order; per-row removals are not.

Filter	Pockets	Removed	Proteins
All predicted pockets	21,646	—	17,527
Direct: evidence on the query accession			
Known ligand site on this protein	18,908	2,738	15,254
UniProt ACT_SITE, BINDING or SITE	15,557	3,351	12,468
Polymer interface in its own assemblies	12,480	3,077	9,995
Homology-propagated: transferred from Pfam family and clan relatives			
Transferred ligand site	9,038	3,442	7,268
Transferred active site	7,044	1,994	5,709
Transferred interface (family only)	4,489	2,555	3,590
Catalytic site localized, if an enzyme	4,456	33	3,564
Geometry			
9–30 lining residues	1,024	3,432	1,017
Burial $\geq$ 39.9 heavy atoms	765	259	758
Mean pLDDT $\geq$ 80	612	153	606
Mean head confidence $\geq$ 0.70	306	306	303
Motif-propagated: structural similarity, no homology required			
Motif holds no ligand elsewhere in the PDB	**155**	151	**154**

Table A7: The funnel with all interface evidence removed, for comparison with Table A6. Geometric and motif requirements are unchanged. The surviving set is a strict superset of the 155 reported there.

Filter	Pockets	Removed	Proteins
All predicted pockets	21,646	—	17,527
Direct			
Known ligand site on this protein	18,908	2,738	15,254
UniProt functional site, interactions excluded	15,602	3,306	12,501
Homology-propagated			
Transferred ligand site	11,297	4,305	9,110
Transferred active site	8,956	2,341	7,268
Catalytic site localised, if an enzyme	8,914	42	7,235
Geometry			
9–30 lining residues	2,381	6,533	2,366
Burial $\geq$ 39.9 heavy atoms	1,828	553	1,816
Mean pLDDT $\geq$ 80	1,579	249	1,568
Mean head confidence $\geq$ 0.70	754	825	750
Motif-propagated			
Motif holds no ligand elsewhere in the PDB	**359**	395	**358**

D.2 The ligand-site annotation

Ligand-contact evidence throughout — both for the direct gate and as the donor pool for homology transfer — comes from a curated structural annotation derived from Q-BioLiP, in which every ligand-binding residue of every deposited structure is resolved to a canonical UniProt position through per-residue SIFTS. Contacts are admitted under a fixed policy: a literal 5 Å heavy-atom contact

definition; at least 10 contacting residues per chain; ligands of 5–50 heavy atoms; at least 80% SIFTS coverage of the chain with at most a 0.2 mismatch fraction; a single primary accession per chain; and a UniProt length of 40–1,200 residues.

Two properties of this policy propagate into the results. First, the annotation contains *nonpolymer ligands only* — no peptide, nucleic-acid or protein partners — which is why polymer interfaces are computed separately from PDB biological assemblies rather than read from the same table. Second, the 10-residue minimum censors the pocket-size distribution from below. The length cap is also the reason the direct gate reaches only 2,444 of the 20,092 scanned proteins: our scan admits chains up to 2,560 residues, so large proteins are annotated in the reference set only where a shorter construct exists.

D.3 Donor construction

Three classes of site were assembled for all organisms, keyed by UniProt accession. *Ligand contacts* were taken from a cross-species structural annotation of nonpolymer ligands: 19,203 accessions, 322,466 ligand pockets and 6,122,832 contact residues, at a literal 5 Å contact definition. *Annotated sites* were UniProt ACT_SITE and BINDING features, excluding range features spanning more than 30 residues, which are deletions rather than point annotations. *Interface residues* were inter-chain contacts computed from PDB biological assemblies and mapped to canonical numbering through per-residue SIFTS, covering 11,999 accessions. Donors carrying at least one site of any class number 275,471, spanning 11,231 Pfam families.

D.4 Transfer procedure

Each scanned protein was resolved to its Pfam families and, through `Pfam-A.clans`, to its clans. Donors sharing a family were ranked first, then donors sharing only a clan, with ties broken by agreement of the three-level EC number and then by site count; the top 60 candidates were aligned. Alignment used global Needleman–Wunsch (parasail, BLOSUM62, gap open 11, extend 1), discarding alignments below 25% identity over aligned columns, and the 12 highest-identity donors were retained. Donor site positions were mapped to query numbering through the alignment, and a pocket was discarded if any lining residue lay within 5 Å of a transferred position in the predicted fold.

Two restrictions apply to clan-level transfer. First, a clan is not always a functional family: CL0015 (MFS) shares a substrate cavity across its members, whereas CL0196 (DSRM) shares only a small structural motif between otherwise unrelated proteins. Clan-level transfers therefore require corroboration by at least two independent donors, which suppresses single spurious alignments without weakening transfer within genuine fold families. Second, interface transfers are restricted to family level throughout: a binding cavity is conserved across a clan, but a protein's interaction partners are not. With these settings the gate reaches 16,754 of 20,092 proteins.

D.5 Geometric thresholds

Thresholds were calibrated against 78,370 experimentally determined pocket instances measured on the same predicted folds. Burial (39.9 protein heavy atoms within 8 Å of a lining-residue centroid, excluding the pocket's own atoms) and the upper size bound of 30 residues are the 10th and 95th percentiles of that reference set. The lower size bound is not a percentile. The reference distribution is censored near ten residues by the `minimum_contact_residues_per_chain` policy of the annotation that produced it, so it carries no information about smaller pockets and cannot justify a floor in that range. Lowering the bound without compensation admits low-confidence output — pockets of 3–5 residues have a median head confidence of 0.29 against 0.82 for those of 16–30 — which shows that the size cut had been acting as an implicit confidence filter. We therefore set the size floor at 9 and impose an explicit confidence floor of 0.70, which together retain a set of comparable size to the previous thresholds but at higher median confidence.

Per-residue pLDDT is the mean B-factor over all atoms of a residue in the predicted structure, and a pocket's pLDDT is the mean over its lining residues. Residues without a value are excluded from that mean; pockets on proteins for which no per-residue values had been extracted were

previously discarded by the pLDDT gate as missing rather than poor, affecting 3,834 pockets, and were backfilled.

D.6 Motif-propagated evidence

Neither direct nor homology-propagated evidence can reach a pocket that is uncharacterized on its own protein and across its family, yet recurs as a ligand-binding structural motif in proteins with no detectable evolutionary relationship to it. Each survivor's lining residues were submitted as a motif to Folddisco against 233,054 PDB structures. Matched target residues were resolved to chain and residue identifiers and intersected with the ligand-contact annotation of §D.2; a motif with any matched residue that is a ligand contact in a deposited structure was discarded. A second rule, flagging sub-ångström matches to a target chain of the same family or EC class, fired on a single pocket and is not load-bearing: most matched target chains carry no annotation to compare against, so this route is under-powered and the ligand-contact rule carries the filter.

D.7 Evidence coverage

Each gate reports one of three states — the site is hit, the site is clear, or no evidence was available — and each pocket carries the number of gates that had evidence. Of the 155 survivors, 1 cleared three gates on evidence, 11 cleared two, 54 cleared one and 89 cleared none. The majority of the output is therefore untested rather than tested and clear, and should be read as such; homology resolution reduces this fraction (from 87% to 57%) but does not eliminate it, because a protein whose Pfam family contains no annotated member remains unreachable by any transfer.

D.8 Alternative pocket proposals

To test whether the head's own decoder limits recall, pocket proposals from fpocket, P2Rank and Lacuna were generated for all scanned proteins (1,797,462 candidates) and scored by the same head. Of those clearing the confidence and size thresholds, 735 on 346 proteins passed the direct and homology-propagated filters. Every one overlapped a pocket the head had already proposed; none was a site at a location the head had not itself identified, and none of the proteins for which the head predicted no pocket was rescued. The proposal stage is therefore not the binding constraint on this pipeline.

E Candidate Pockets

Two pockets are described in the main text. They are not the two highest-scoring pockets; they are the ones for which a mechanistic hypothesis exists, and they were selected after the filters rather than by them. Each is reported with the evidence against it, and with the extent to which conventional cavity detectors agree that the site is a pocket at all.

E.1 Agreement with geometric cavity detectors

fpocket, P2Rank and Lacuna were run natively on the same predicted structures, and each tool's own score order was used — not a re-ranking by the head. Table A8 reports the rank at which each tool first proposes a pocket overlapping ours.

Table A8: Agreement of geometric cavity detectors with the predicted site. Each tool was run natively and its own score order used. Entries give the rank of the tool's best-overlapping proposal, out of the number it made on that protein, with the Jaccard overlap in parentheses. A dash means no proposal reached a Jaccard of 0.25, so the tool does not propose this site at any rank.

	zPocket	P2Rank	fpocket	Lacuna
DET1	1	4 / 17 (0.71)	39 / 40 (0.70)	1 / 50 (0.42)
COP1	1	— (0.06)	— (0.16)	24 / 52 (0.42)

The two cases differ. For **DET1** the classical detectors disagree with each other: P2Rank ranks the site fourth of seventeen and fpocket thirty-ninth of forty, while Lacuna, itself a learned method, ranks it first. The site is thus found by all three, but a user relying on fpocket's ordering would never reach it. For **COP1** no geometric detector proposes the site at all. P2Rank returns only four proposals on the entire 731-residue protein, of which the highest-scoring is a 40-residue region sharing three residues with the predicted pocket; the remaining three share none. Only Lacuna proposes a comparable site, at rank 24 of 52. This is consistent with the site being a shallow peptide groove, which cavity-geometry methods are least able to score, and it is the case where the head's proposal is not reproduced by an existing tool.

E.2 DET1

DET1 (UniProt Q7L5Y6, 550 residues) is a substrate-adaptor subunit of the DCX(DET1–COP1) cullin–RING ligase, with RBX1, DDB1, CUL4A and COP1. The predicted pocket comprises 17 residues drawn from five sequence-distant segments spanning the chain (57–63, 142–143, 218, 478–479, 517–539), with a mean head confidence of 0.831, burial of 43.8 heavy atoms and mean pLDDT of 90.1. Because DET1 carries no RING domain and is not the catalytic subunit, a ligand at this site cannot inhibit the ligase; it can only alter substrate recognition, which is the architecture exploited by molecular glues. Twelve experimental structures of DET1 are deposited, and across the 16 DET1 chains they contain, no partner atom approaches within 6 Å of any lining residue; the nearest partner contact is 14.9 Å. The pocket also avoids the E2-binding region and the DDB1-anchoring segment.

Limitations. The site is not a pre-formed cavity. Across all 16 experimental copies the lining is solvent-exposed, with a mean relative SASA of 0.295 and no fully buried lining residue, and fpocket ranks the overlapping groove in the bottom half of its own list. This measurement requires care: computed on holo structures with the ligand deleted, the same statistic rejects apo cereblon, which scores worse than DET1 — every validated site appears enclosed once a ligand has opened it. Measured against genuinely apo references the result is that DET1's site is not pre-formed, not that it is unliganddable, and the lining is mildly more mobile than the protein average (1.31 Å side-chain spread against 1.09 Å).

E.3 COP1

COP1 (Q8NHY2, 731 residues) is the substrate receptor of the same ligase complex as DET1, degrading JUN, p53 and the ETV1/4/5 transcription factors. The predicted pocket comprises 15 residues, all 15 inside annotated WD40 repeats and spanning five of the seven blades — the top face of the β-propeller. Superposition on the COP1–TRIB1 complex shows that 14 of the 15 are within 4.5 Å of the bound degron peptide, and that the three-residue Val–Pro anchor of that peptide engages ten COP1 residues, all ten of which are in the predicted pocket. The site is therefore not merely the degron surface but specifically its hot-spot subsite, which is the configuration that makes a protein–protein interface tractable to small molecules.

Limitations. Occupancy blocks substrate recognition, so a ligand is first an inhibitor, and a cereblon-style glue is not available at this pocket. Direction of effect is substrate-dependent and cannot be separated chemically, since p53, C/EBPα and ETV4/5 all dock through the same slot. The site has been attacked: an affinity-selection campaign reported six COP1-binding peptides at potencies comparable to the native motif, and no small molecule is known; ChEMBL contains no COP1 target record at all. The experimental literature describes the VP pocket as shallow.