Auditable Algebraic Counting Field for Cryptic-Pocket Detection from Apo Structures

Shan Yu[1] and Xuening Wu[2,†]
[1]*Independent Researcher, Hangzhou, China*
[2]*Pfizer, Shanghai, China*
[†]*E-mail: xuening.wu@pfizer.com*

Cryptic ligand-binding pockets are not apparent in experimentally determined apo structures, making them difficult to identify from unbound receptor geometry. A complementary challenge is to make the structural measurements and learned evidence behind each prediction directly inspectable. We introduce a supervised algebraic counting field (ACF) for predicting cryptic-pocket residues from apo structures. ACF compiles explicit geometric, physicochemical, and topological features into compact, integer-weighted lookup tables. Each prediction score can be reconstructed from feature values, training counts, table weights, and spatial aggregation, without sequence search, structural-template transfer, or a protein language model at inference. We evaluate ACF on CryptoBench and two locked external collections, separating ranking performance from the effects of residue-calling budgets. On an external set of 57 post-CryptoBench apo–holo units, ACF exceeded P2Rank by +0.044 in mean paired ROC-AUC (multiplicity-adjusted 95% CI [+0.010, +0.079]). The advantage was dataset-dependent: official-fold ROC-AUC and matched-budget F1 differences against P2Rank remained unresolved, and a second external evaluation did not confirm gains from added structural features. ACF thus provides a compact predictor with externally validated signal and an inspectable path from structural measurements and training counts to residue scores.

Keywords: cryptic binding sites; protein function; structure-based prediction; apo proteins; interpretable machine learning; algebraic counting field

1. Introduction

Cryptic ligand-binding sites may be absent or closed in an observed apo structure, although local geometry and chemistry may indicate where opening is possible.[1–4] Identifying their lining residues therefore tests structure-based functional inference beyond annotation transfer from homologous sequences or structural templates.

Structure-based predictors span explicit geometry, energetic probes, simulation, and learned representations. The geometry-based method Fpocket detects cavities from the spatial arrangement of protein atoms.[5] Other established approaches include geometric and physicochemical site scoring in SiteMap and DoGSiteScorer and energetic probe mapping in FTMap.[6–8] DeepSite, Kalasanty, and DeepPocket instead formulate pocket detection or segmentation with three-dimensional convolutional networks.[9–11] P2Rank learns ligandability scores for local chemical

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

neighborhoods on the protein surface.[12] PocketMiner uses a graph neural network to identify residues likely to participate in cryptic-pocket opening.[13] CryptoBench adds a supervised protein-language-model baseline (pLM-NN), drawing on the sequence representations learned by large protein language models, and sequence-clustered splits for cryptic-site evaluation.[14,15] These methods show that single structures and learned representations contain useful site information. They also motivate a complementary question: can a predictor remain competitive while allowing each score to be mechanically reconstructed from explicit local measurements and training counts? This focus on traceability distinguishes explanations constructed after prediction from models whose predictive computations are directly inspectable.[16]

We answer this question with an *algebraic counting field* (ACF)—a discrete counting architecture over local structural neighborhoods in which every score decomposes into explicit table lookups and integer-weighted additions. Within each protein chain, structural features are rank-discretized into four levels and paired to address compact lookup tables compiled from training-fold residue counts. A regularized compilation step assigns bounded integer multiplicities to the tables. The deployed score requires no neural-network forward pass or external database search, although the method is *not* training-free: its table counts, residue propensity, and integer multiplicities are derived from the training fold.

The contribution is an explicit structural representation with a compact, exactly decomposable readout, evaluated against strong baselines with ranking and operating-point effects separated. Two external reads support transferable signal but delimit accuracy claims; controlled training-fold probes identify additional structural information whose benefit was not confirmed externally. Figure 1 summarizes the training and prediction workflow.

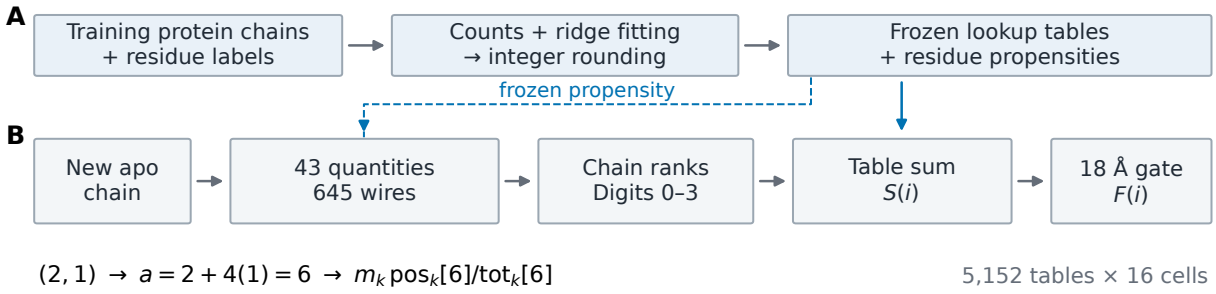


Fig. 1. ACF workflow. (A) Training uses apo protein chains and residue labels to compile lookup-table counts, integer weights, and residue-type propensities, which are then frozen. (B) Prediction uses a new apo chain's structural features and within-chain ranks with the frozen parameters to obtain residue scores through table lookup and spatial aggregation; no residue labels are required.

2. Materials and Methods

2.1. *Benchmark, evaluation units, and leakage controls*

We use the official CryptoBench apo split, in which the training and test partitions are disjoint under MMseqs2 clustering at 10% sequence identity.[15,17] Receptor structures are retrieved from

the RCSB Protein Data Bank (RCSB PDB),[18] and every coordinate and label file is verified against a recorded SHA-256 hash. Data loading halts if the manifest is absent, a file hash has changed, the declared clustering metadata differ from the recorded values, or a cluster identifier occurs in both partitions.

Each evaluation unit is identified by a PDB entry–chain pair. The official analysis includes 192 single-chain units drawn from 190 PDB entries. Thirty-eight cases requiring multi-chain residue mapping, plus one chain not retrievable in legacy PDB format, are excluded. The multi-chain cases are excluded because the benchmark labels do not unambiguously assign residues to individual chains. This exclusion is an explicit scope restriction. All structure-based predictors receive only apo-receptor coordinates. A runtime guard rejects inputs containing non-solvent heteroatoms or other non-polymer atoms, preventing ligand coordinates from entering receptor-only predictions. Table 1 summarizes the frozen evaluation partitions.

Table 1. Frozen CryptoBench partitions. Each unit is a PDB entry–chain pair.

Partition	Units	Residues	Role
Training fold	770	234,838	ACF compilation
Fit half	386	118,716	Candidate-readout fitting
Selection half	384	116,122	Configuration selection
Official test subset	192	—	Final paired evaluation

2.2. *Local quantities, wires, and ACF compilation*

Each residue has 43 explicit base quantities: 35 geometric and topological descriptors, seven fixed physicochemical constants, and one Laplace-smoothed residue-type propensity estimated only from the training fold. The structural descriptors cover surface exposure, local spectral geometry, density-field calculus, ultrametric organization, curvature, anisotropy, and global position. The seven fixed physicochemical quantities are Kyte–Doolittle hydropathy,[19] Zamyatnin side-chain volume,[20] aromatic-ring count, formal charge, hydrogen-bond donor and acceptor counts, and the number of rotatable side-chain bonds. The public implementation documents all definitions, sources, neighborhoods, boundary cases, and expansion rules. Geometric and physicochemical extensions require recompilation and separate evaluation. Propensity is fixed during prediction and re-estimated for a different training fold.

For residue i, define $N_r(i) = \{j : \|c_i - c_j\| \le r\}$. Here, c_i is the centroid of its observed heavy atoms, and the intra-chain neighborhood includes i. Each quantity x is used directly and through its neighborhood mean, standard deviation, centered difference, and local rank:

$$\mathrm{sd}_r[x]_i = \sqrt{\max\left(\overline{x^2}_{N_r(i)} - \left(\overline{x}_{N_r(i)}\right)^2, 0\right)}, \quad D_r[x]_i = x_i - \overline{x}_{N_r(i)}, \quad K_r[x]_i = \frac{|\{j \in N_r(i) : x_j < x_i\}|}{|N_r(i)|}.$$

Each resulting residue-level scalar feature is called a wire. Neighborhood means are evaluated at 6, 10, 14, 20, and 26 Å, whereas dispersion, centered difference, and local rank are evaluated at 6, 14, and 20 Å. These predefined transformations retain all 15 versions of each base quantity, yielding $43(1 + 5 + 3 + 3 + 3) = 645$ wires without a further feature-selection step.

Each wire is discretized separately within the scored chain. For a chain of n residues, let $r_j(i)$ be the zero-based rank of wire j, with ties assigned their mean rank. Its digit is $d_j(i) = \min\{3, \lfloor 4r_j(i)/\max(n-1,1)\rfloor\}$. This discretization is invariant to strictly increasing transformations of that wire; ties need not produce equally populated bins. No training-set bin edges are transferred.

A table's *width* is the number of input wires. For a width-2 table with ordered wire indices (u_k, v_k), the address is $a_k(i) = d_{u_k}(i) + 4d_{v_k}(i) \in \{0, \ldots, 15\}$. Thus a table has $4^2 = 16$ cells: for example, digits $(2, 1)$ select cell 6. The bank is generated by 16 successive permutations of all 645 wire indices using NumPy `default_rng(20260725)`. Adjacent indices are paired and the final singleton is discarded in each round; pairs are not selected by labels or biological criteria. Each round contributes $\lfloor 645/2 \rfloor = 322$ tables, hence $16 \times 322 = 5{,}152$ tables and $5{,}152 \times 16 = 82{,}432$ cells. A wire is used at most once per round; partners can recur across rounds. With the frozen seed, 629 wires occur in 16 tables and 16 wires in 15 tables, so none is absent from the complete bank.

For table k, cell a stores two training-fold counts: $\mathrm{tot}_k[a]$, the number of residues reaching it, and $\mathrm{pos}_k[a]$, the number labeled cryptic. Its value is the unsmoothed fraction $\mathrm{pos}_k[a]/\mathrm{tot}_k[a]$, or the training-fold base rate for an unseen cell. There are 825 unseen cells (1.00%) in the full-fold compilation.

The unaggregated residue score is

$$S(i) = \sum_{k=1}^{5{,}152} m_k \frac{\mathrm{pos}_k[a_k(i)]}{\mathrm{tot}_k[a_k(i)]}, \qquad m_k \in \mathbb{Z} \cap [-32, 32],$$

with the training-fold base rate substituted for an unseen cell. The deployed score is

$$F(i) = S(i) + \overline{S}_{18}(i) \frac{\mathrm{sd}(S)}{\mathrm{sd}(\overline{S}_{18})}.$$

Here, $\overline{S}_{18}(i)$ is the mean score among residues within 18 Å of residue i, and the standard deviations are computed across residues in the scored chain. This spatial gate favors coherent clusters of high-scoring residues over isolated high scores. The standard-deviation ratio places the neighborhood term on the same dispersion scale as S. Radius 18 Å and weight 1 were selected on a UniProt accession-disjoint training half and frozen; this internal split does not guarantee sequence-cluster separation. The gate is skipped if the neighborhood mean has zero standard deviation.

The integer multiplicities m_k are obtained by solving a ridge-regularized closed-form system over training-fold table outputs and rounding the resulting coefficients. Let Σ be the within-class scatter of table outputs divided by $\max(N-2, 1)$, where N is the number of training residues, and let μ_1, μ_0 be the class means. The solution is

$$w = \left[\Sigma + \left(0.03\,\frac{\mathrm{tr}\,\Sigma}{K} + 10^{-12}\right) I\right]^{-1} (\mu_1 - \mu_0), \qquad K = 5{,}152.$$

Weights are $m_k = \mathrm{round}(32\, w_k / \max_j |w_j|)$, or all zero when the solution is zero. The ridge penalty scales with the mean diagonal, not the unnormalized trace. Compilation uses a real-valued solve; inference uses floating-point cell fractions, integer coefficients, and spatial statistics. Of the 5,152 compiled tables, 4,797 have nonzero multiplicity in the full-training-fold ACF.

2.3. *Baselines and operating points*

Rigid geometry (`geometric_foundation`) is a training-free static-coordinate baseline that ranks free-space points by ray enclosure and a local cavity-volume proxy. It uses neither ACF's 43 quantities nor its 645-wire neighborhood expansion; it is not an ACF ablation.

P2Rank 2.5.1 is executed locally with a pinned binary and Java environment. Ranking metrics use its native residue probability. Its shipped binary output is the native pocket assignment. PocketMiner is rebuilt and checked on its published self-test, where the reconstructed score is 0.8677 versus the reported 0.87. The pLM-NN baseline uses the benchmark's ESM2-3B embedding and trained head; reconstruction fidelity to the released example is 0.9987 mean cosine similarity. A logistic regression on the same 645 continuous wires, standardized on the fit half, tests the readout on the training selection half. For post-hoc external evaluation, standardization and logistic coefficients were refitted on all 770 training chains, retaining the training-derived propensity and the logistic model's training-selected gate (18 Å, weight 0.5); the deployed ACF retains its 18 Å gate with weight 1. Both models scored the same residues. Set A and Sets B+C were evaluated separately using 4,000 paired chain-bootstrap resamples (seed 20260929) and unadjusted 95% intervals. This supplementary comparison was not preregistered.

The counting field's deployed binary call is the top 0.09 fraction of residues per chain, selected on the training fold. To remove an operating-point asymmetry, we additionally call the top 0.09 fraction of residues in each chain for both the field and P2Rank. Applying the same grid search to P2Rank's training-fold predictions also selects 0.09, so the common-budget and separately training-selected rules coincide. Threshold curves from 2% to 40% are descriptive robustness analyses; they do not change the deployed operating point. P2Rank's test-fold-best threshold is shown only as a descriptive comparison in Fig. 3.

2.4. *External validation*

The first external set (Set A) contains 57 apo–holo units and 905 positive residues from depositions released after 2024-05-08, the newest structure date in CryptoBench. It retains one unit per UniRef50 cluster and shares no cluster with either CryptoBench partition. Coordinates, labels, inclusion decisions, and hashes were frozen before scoring. A committed analysis plan names three co-primary paired ROC-AUC comparisons—the deployed field against P2Rank, pLM-NN, and PocketMiner—and applies a Bonferroni correction. Binary comparisons use the same matched top-9% budget as the official-fold robustness analysis.

A second external read (Sets B+C) tests whether the training-fold lift from four additional feature families transfers. These sets contain 45 scorable single-particle cryo-EM apo–holo units, are disjoint at UniRef50 from CryptoBench, Set A, and each other, and were frozen before the additional families were developed. One prespecified unit was excluded before scoring because its three-character chain identifier could not be represented by the legacy PDB writer. After the expanded 1269-column field was fixed, a committed plan pinned four co-primary mean paired ROC-AUC comparisons: expanded field minus the deployed 645-wire field, pLM-NN, P2Rank, and PocketMiner. All five methods were evaluated on the same residue universe; no unit or residue was lost to the intersection. Bonferroni-adjusted intervals control the four

comparisons. The repository commit graph is part of the chronology evidence for both external reads and must accompany the archival submission.

2.5. *Statistical analysis*

Metrics are computed per unit and averaged across units. We report both ROC-AUC and PR-AUC because precision–recall summaries are informative under class imbalance.[21] For the official-fold and external comparisons and the internal logistic comparison, intervals for paired differences use a paired nonparametric bootstrap over units.[22] The main evaluations use 10000 resamples; the internal and post-hoc external logistic comparisons use 4,000. Undefined per-unit metrics are excluded from the corresponding paired comparison; they are never imputed as zero. "Resolved" denotes a paired interval excluding zero, with multiplicity adjustment where comparisons are co-primary. The mean paired ROC-AUC remains the primary estimand. A 20% trimmed mean that was selected on the training partition is reported as exploratory because 4 indexed reads of the official fold, 3 from the present architecture lineage, preceded its registration.

Structural-family intervals are instead $\bar{d} \pm 1.96\, s_d/\sqrt{S}$, where $\bar{d}$ and s_d are the mean and sample standard deviation of paired ROC-AUC gains across $S = 12$ half-splits. These descriptive intervals ignore dependence from reused data and support no confirmatory significance or population-coverage claims.

2.6. *Training-only structural-family probes*

To test what information the deployed representation omits, we appended predefined feature families while holding the deployed pair bank, gate, ridge, fan-out cap, and quantization fixed. Mean per-unit ROC-AUC lift over the frozen deployed detector was measured on 12 UniProt accession-disjoint halvings of the CryptoBench training fold (seeds 20260725–20260736). Each family was compared with a size-matched control that added the same number of tables over already deployed wires. For families that read additional atom-level data, a second control permuted rows within each chain, preserving column marginals while destroying residue assignment. These probes are exploratory analyses of the representation; they do not use the official or external folds for candidate selection or change the deployed detector.

3. Results

3.1. *Official-fold ranking and strong baselines*

Table 2 and Fig. 2 summarize the official-fold results. The counting field reached ROC-AUC 0.799 [0.766, 0.830], PR-AUC 0.376 [0.339, 0.411], MCC 0.304 [0.270, 0.338], and F1 0.333 [0.301, 0.365] on the official fold. P2Rank reached 0.793, 0.358, 0.285, and 0.302, respectively. The paired mean ROC-AUC difference was +0.0058 [-0.0168, +0.0276] ($p = 0.598$); PR-AUC was likewise unresolved (+0.0177 [-0.0094, +0.0450]). The field was ahead on 115 chains and behind on 77, with a median difference of +0.0412. The exploratory 20% trimmed mean was +0.0281 [+0.0117, +0.0427], but it cannot repair the prior exposure of this fold and is not the primary claim.

Table 2. Official-fold results. The first three rows report all four metrics; the last two report ROC-AUC for the reconstructed baselines.

Method	ROC-AUC	PR-AUC	MCC	F1
Counting field	0.799	0.376	0.304	0.333
P2Rank 2.5.1	0.793	0.358	0.285	0.302
Rigid geometry	0.665	0.276	0.239	0.287
PocketMiner	0.7523	—	—	—
pLM-NN	0.8235	—	—	—

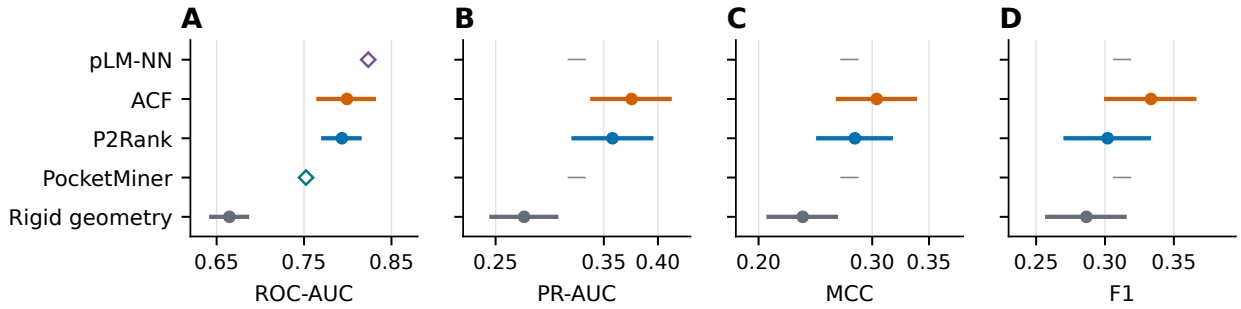


Fig. 2. Official-fold means (192 units). Filled points: means with 95% bootstrap intervals; open diamonds: reported means without marginal intervals; dashes: metrics not shown. MCC/F1 use method-specific calls (ACF: top 9%; P2Rank: native pockets). Axes differ. The paired ACF–P2Rank ROC-AUC difference remains unresolved.

The stronger baselines bound the accuracy claim. The field-minus-pLM-NN ROC-AUC difference was -0.0243 [-0.0465, -0.0033]; the field scored lower on 111 of 192 chains. At a common calling budget, its F1 and MCC differences were -0.0522 [-0.0786, -0.0274] and -0.0561 [-0.0863, -0.0269]. pLM-NN remained ahead at all 39 calling fractions. The field exceeded PocketMiner by +0.0468 [+0.0251, +0.0682], but P2Rank also exceeded PocketMiner by +0.0410 [+0.0211, +0.0614]. Differences in label definitions may contribute to the PocketMiner gap; these comparisons do not isolate its cause or establish general methodological superiority.

On the training selection half, the same-wire logistic model reached ROC-AUC 0.8008; the field-minus-logistic difference was +0.0038 [-0.0057, +0.0125]. This training-fold comparison does not establish an accuracy advantage for the table architecture.

3.2. *The F1 conclusion depends on the operating convention*

Fig. 3 compares the shipped and matched operating conventions. Under shipped outputs, counting-field F1 exceeded P2Rank by +0.0315. This compares the field's top-fraction rule (0.09 of residues) with P2Rank's native pocket assignment, which called 1.65 times as many residues. When both methods called the top 0.09 fraction of residues in each chain, their mean F1 values were 0.3334 and 0.3194, and the paired difference was +0.0140 [-0.0085, +0.0364] ($p = 0.224$). MCC was also unresolved under the common budget (+0.0169 [-0.0092, +0.0424]). These results do not support a convention-independent F1 advantage.

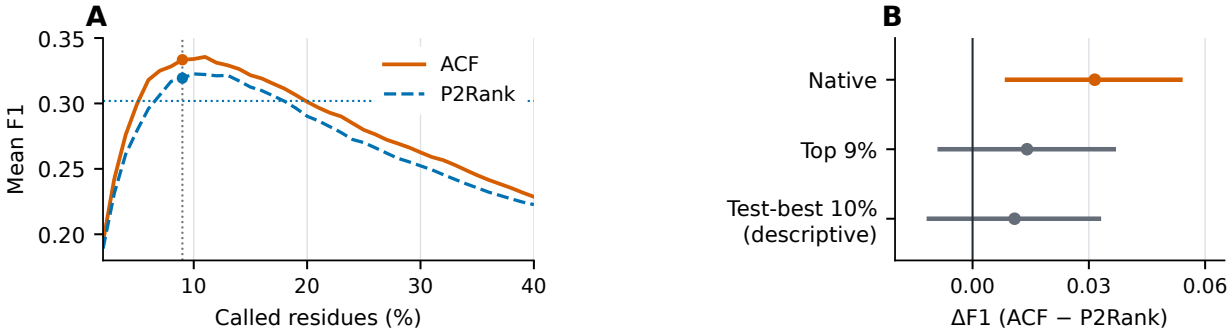


Fig. 3. Calling-budget sensitivity. (A) F1 curves; dotted lines mark training-selected 9% calls (vertical) and P2Rank native-pocket F1 (horizontal). (B) Paired differences with 95% intervals: native rules, matched top-9%, and ACF at 9% versus P2Rank's test-fold-best 10% (descriptive). Matched-budget F1 is unresolved.

Table 3. Locked external validation on 57 apo–holo units. Differences are counting field minus comparator in mean paired ROC-AUC; intervals are unadjusted. Bonferroni-adjusted intervals appear in Fig. 4A.

Comparator	Comparator AUC	Difference	95% paired CI
P2Rank	0.7968	+0.0443	[+0.0162, +0.0724]
pLM-NN	0.8751	-0.0340	[-0.0701, -0.0006]
PocketMiner	0.7734	+0.0677	[+0.0328, +0.1013]

Across all 39 thresholds from 2% to 40%, the field's F1 remained directionally above P2Rank's, although pointwise intervals excluded zero at only 10 thresholds. pLM-NN remained above the field throughout, with intervals excluding zero at all 39 thresholds. This robustness curve supports a directional comparison but does not create 39 independent confirmatory tests.

3.3. *First locked external validation*

Table 3 and Fig. 4A report the predeclared external analysis. The field reached mean ROC-AUC 0.8411. Its paired advantage over P2Rank was +0.0443 [+0.0162, +0.0724], with a Bonferroni-adjusted interval of [+0.0102, +0.0792]; it was ahead on 42 of 57 units. This is the paper's principal confirmatory accuracy result.

The point-estimate deficit to pLM-NN persisted externally (-0.0340 [-0.0701, -0.0006]), although its multiplicity-adjusted interval [-0.0784, +0.0071] included zero. The advantage over PocketMiner was also replicated (+0.0677 [+0.0328, +0.1013]). At the matched top-9% budget, external F1 against P2Rank was unresolved (+0.0246 [-0.0224, +0.0708]), agreeing with the official-fold matched-threshold analysis. P2Rank abstained on 1 external unit; ranking analysis retained its residue scores, whereas undefined threshold metrics followed the declared missing-value rule.

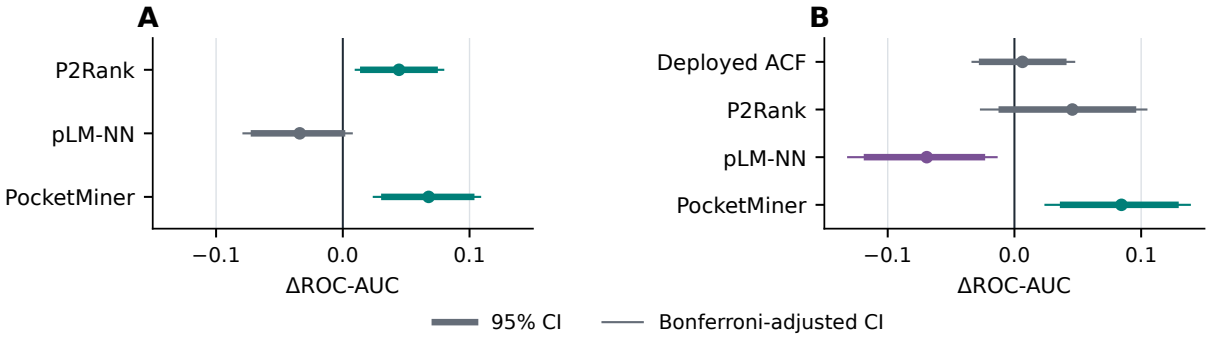


Fig. 4. Locked external evaluations. (A) Set A, 57 units: deployed ACF minus comparator. (B) Sets B+C, 45 cryo-EM units: expanded ACF minus comparator. Points are paired mean ROC-AUC differences; thick/thin lines denote 95%/Bonferroni-adjusted intervals (three comparisons in A, four in B). Positive favors the named ACF variant. Expanded-versus-deployed improvement is unresolved; evaluations are not pooled.

3.4. *Second external read of the expanded representation*

The expanded field reached mean per-unit ROC-AUC 0.7980 on the 45 cryo-EM units, compared with 0.7917 for the deployed field, 0.7523 for P2Rank, 0.7136 for PocketMiner, and 0.8672 for pLM-NN (Fig. 4B). The primary expanded-minus-deployed difference was +0.0063 [-0.0255, +0.0385], ahead on 27 units and behind on 18. Its interval contains zero and the training-fold estimate +0.0121; the read therefore does not confirm that the four added families improve the deployed representation.

Against P2Rank, the expanded field differed by +0.0457 [-0.0100, +0.0935] and remained unresolved after multiplicity correction ([-0.0264, +0.1041]). It exceeded PocketMiner by +0.0844 [+0.0384, +0.1271], with the adjusted interval [+0.0245, +0.1384] excluding zero. Conversely, pLM-NN remained ahead: the expanded-minus-pLM-NN difference was -0.0692 [-0.1163, -0.0257] (adjusted [-0.1312, -0.0141]). Thus, the expanded representation transfers signal relative to PocketMiner, without a resolved gain over the deployed field or P2Rank.

In the post-hoc comparison, the deployed 645-wire ACF and logistic regression achieved mean ROC-AUCs of 0.8411 and 0.8257 on Set A (57 units; ACF minus logistic: +0.0154, 95% CI [−0.0012, +0.0347]) and 0.7917 and 0.7968 on Sets B+C (45 units; −0.0051, 95% CI [−0.0415, +0.0333]), respectively. Neither interval excludes zero; these results establish neither ACF superiority nor statistical equivalence.

3.5. *Discarded atom-level structure carries training-fold signal*

The controlled family ladder distinguishes new structural information from re-encoding. A chemistry family that mainly resolves residue-identity collisions in the existing physicochemical wires changed mean ROC-AUC by only +0.0002 [-0.0007, +0.0010]. In contrast, atom-level backbone descriptors improved ROC-AUC by +0.0044 [+0.0033, +0.0055], side-chain conformational descriptors by +0.0048 [+0.0032, +0.0063], and void-connectivity descriptors by +0.0026 [+0.0012, +0.0039]. The combined backbone–side-chain family yielded +0.0072 [+0.0055, +0.0089], and adding void geometry produced +0.0095 [+0.0081, +0.0109]. All three

corresponding size-matched controls remained near zero; within-chain row permutation yielded negative gains for the backbone, side-chain, and void families.

The same screen identified a fourth source outside the deployed coordinate summary. Within-chain ranks and ratios derived from deposited temperature factors, occupancies, and alternate locations produced a +0.0070 [+0.0050, +0.0090] lift, driven primarily by temperature-factor quantities (+0.0063 [+0.0049, +0.0077]) rather than alternate-conformer quantities (+0.0005 [-0.0007, +0.0017]). These training-fold results support a specific representation-level interpretation: useful information remains in atom positions, empty-space connectivity, and structure-relative displacement records that are lost when each residue is reduced to a centroid and a contact neighborhood. They do not establish an accuracy gain for a revised detector on held-out data.

The repository contains the full family ladder and controls. At the manuscript snapshot, the expanded field's only held-out evaluation was Sets B+C; neither the official fold nor Set A had been scored with it. Later probes on the spent official fold remain exploratory.

3.6. *Architecture sensitivity, auditability, and cost*

Sensitivity models were fitted on the fit half and compared on the training selection half; each row reports its best gate from the stated training-only search, so the sweep is not uniformly a fixed-gate ablation. Three, four, and five quantization levels yielded ROC-AUC 0.7995, 0.8045, and 0.7996. Fan-out caps 16, 32, and 64 yielded 0.8018, 0.8045, and 0.8010. Ridge values 0.03, 0.1, and 0.3 yielded 0.8045, 0.8042, and 0.8023. At four levels, within-chain ranking reached 0.8045 versus 0.7905 for pooled ranking. The selected architecture ranked first on all 4 CryptoBench training folds and all 25 accession-disjoint half-splits. These results concern predictive performance, not stability of individual integer weights. On the selection half, the frozen field scored 0.7873 without the spatial gate and 0.8045 with it. This is not an independently held-out estimate of the gate benefit.

Exact recomputation of four residue scores from addresses, cell fractions, multiplicities, and the spatial gate agreed with stored scores to within 5×10^{-16}. Fig. 5 shows two rule-selected structural disagreement cases. In these audited cases, the coarse geometry-family margin between labeled hits and misses was +15.2, and the gate added +4.4. No coarse family separated true positives from false positives by more than 1.3, so the decomposition identifies where the score came from but is not a causal biological explanation.

On the matched Apple M4 environment, warm end-to-end median time was 0.298 s per chain for the field and 0.162 s for P2Rank: ACF took 1.84 times as long and used 2.95 times as much CPU time as P2Rank. Feature extraction consumed 0.212 s, or 71% of the field runtime, while table scoring consumed 0.0873 s. In a cold one-chain invocation, P2Rank took 5.98 times as long as ACF because of JVM startup overhead. Thus, compact model size, rather than faster served inference, is the supported efficiency advantage (1.79 MiB, 0.39 MiB compressed, versus 22.01 MiB for the P2Rank model).

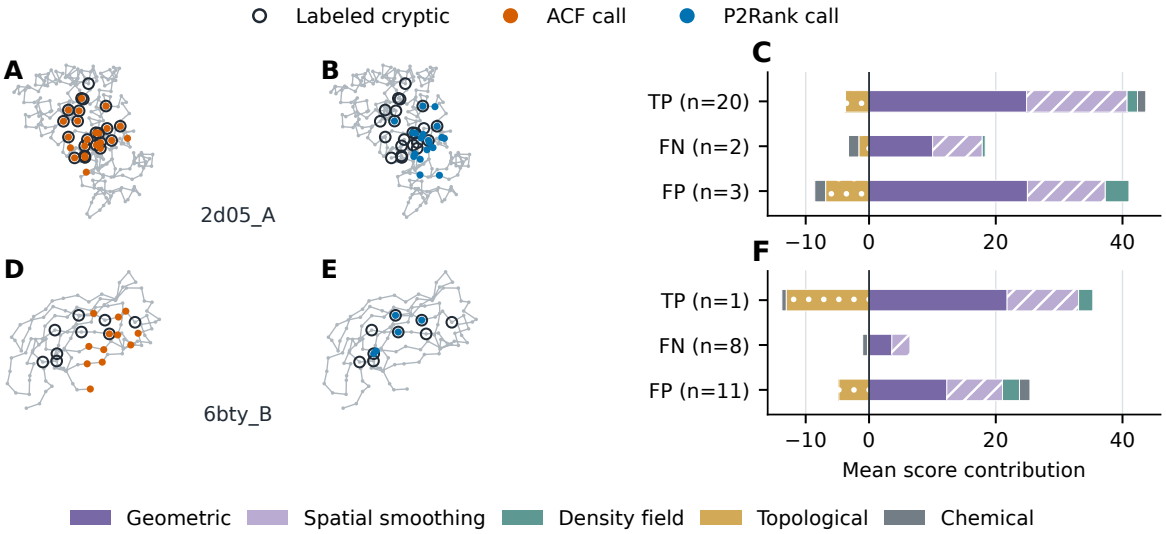


Fig. 5. Rule-selected disagreement cases: 2d05_A (A–C) and 6bty_B (D–F). A/D show ACF calls; B/E show P2Rank calls under deployed rules. Rings mark cryptic labels. Cα projections preserve the coordinate-defined viewing direction and equal spatial scale within each case. C/F show mean ACF contributions for true positives (TP), false negatives (FN), and false positives (FP); n counts residues. Cases illustrate numerical traceability, not confirmatory evidence or causal mechanisms.

4. Discussion

4.1. *Positioning the contribution*

ACF's accuracy advantage over P2Rank is specific to Set A; the cryo-EM comparison did not confirm gains from the expanded representation. Same-wire logistic comparisons likewise do not establish a benefit of the table readout. With pLM-NN stronger on the official fold and directionally stronger externally, ACF's supported contributions are numerical traceability and compact storage relative to P2Rank, not leading accuracy or faster warm inference.

4.2. *Limitations*

The official test fold was inspected repeatedly: the manuscript snapshot records three principal development rescoring rounds, 13 indexed reads (12 from this architecture lineage and one from another), and 12 scored architectures. The subsequently updated repository reports 20 scored architectures; the original counts describe the snapshot, not the current development history. Small official-fold margins are therefore descriptive rather than confirmatory. Set A mitigates but does not eliminate this concern: it contains only 57 units, excludes ambiguous intermediate-displacement cases through its labeling guard band, and may be easier than CryptoBench. The second external read contains only 45 scorable units and consists entirely of single-particle cryo-EM structures, so its wider intervals and modality shift limit interpretation. The decline of all coordinate-reading methods relative to the official fold is consistent with, but does not prove, reduced usable local geometric signal in these models. Both external collections are now spent and cannot confirm a future revision. Chronology must remain verifiable from the full repository commit graph.

The existing sensitivity and family-addition analyses do not establish that all 645 wires are essential, quantify their redundancy, or replace systematic deletion ablations. Nonzero table weights indicate use, not independent biological importance. Coordinate-perturbation robustness and stability of individual weights across training splits and pairing seeds remain untested here; bootstrap-seed stability addresses a different question. B-factor-derived features may reflect structure quality or experimental refinement as well as motion. Numerical auditability can support score inspection, but an actionable experimental benefit has not been demonstrated.

The analysis excludes 38 multi-chain units and does not establish performance at interfaces or in assemblies. Exact numerical decomposition is not the same as mechanistic or causal explanation, particularly because hits and false positives have similar coarse-family contributions. The target is a benchmark cryptic-residue label, not ligand identity, binding affinity, druggability, free-energy change, or biological efficacy. Broader locked datasets, a frozen chain-aware mapping, and prospective experimental or simulation follow-up are needed.

5. Reproducibility

Code, analysis protocols, and result artifacts are available at `https://github.com/shaneraphel/geoaudit-cryptobench`. The original results, macros, and deployed configuration are pinned to commit `5d1726a`. The artifact `data/cryptobench_apo/TABLE_FIELD.json` records wire names, pairs, counts, weights, propensity, operating point, and code/manifest hashes. The manuscript uses `paper/frozen_numbers.tex` from this snapshot. Defining modules are `wide_descriptors.py`, `table_bank.py`, and `table_field.py` under `src/pocket_bench/methods/`. The post-hoc logistic protocol and results are in the companion repository (commit `6c5a640`). Pinned versions distinguish manuscript artifacts from later development.

`make recompute` regenerates metrics from archived predictions; `make verify` checks artifacts. Inference and compilation require the frozen inputs, caches, baseline binaries/weights, and environment. `make extorder` checks lock/read ancestry in the retained Git history; this supports chronology but cannot exclude unrecorded reads.

Runtime measurements used Apple M4, macOS 26.2 (arm64), Python 3.12.8, NumPy 2.5.1/Accelerate, OpenJDK 17, and P2Rank 2.5.1. The record separates warm/cold wall time, CPU time, threading, extraction, scoring, and peak RSS (1180 MiB). These results are environment-specific; artifact checks are not experimental reruns. This work received no funding and used only public data (sources in Methods). X.W. is employed by Pfizer; the authors declare no competing interests related to this work.

6. Conclusion

ACF provides compact, numerically traceable structure-based prediction without homology or template searches at inference. On Set A, its mean paired ROC-AUC advantage over P2Rank was 0.0443 (adjusted 95% CI [0.0102, 0.0792]). Official-fold ROC-AUC and matched-budget F1 differences against P2Rank remained unresolved; cryo-EM evaluation did not confirm gains from added features. These findings support an auditable supervised predictor, without establishing lookup-table superiority or a causal explanation of pocket opening.

References

1. S. Vajda, D. Beglov, A. E. Wakefield, M. Egbert and A. Whitty, Cryptic binding sites on proteins: definition, detection, and druggability, *Current Opinion in Chemical Biology* **44**, 1 (2018).
2. P. Cimermancic, P. Weinkam, T. J. Rettenmaier, L. Bichmann, D. A. Keedy, R. A. Woldeyes, D. Schneidman-Duhovny, O. N. Demerdash, J. C. Mitchell, J. A. Wells, J. S. Fraser and A. Šali, CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites, *Journal of Molecular Biology* **428**, 709 (2016).
3. D. Beglov, D. R. Hall, A. E. Wakefield, L. Luo, K. N. Allen, D. Kozakov, A. Whitty and S. Vajda, Exploring the structural origins of cryptic sites on proteins, *Proceedings of the National Academy of Sciences of the United States of America* **115**, E3416 (2018).
4. G. R. Bowman and P. L. Geissler, Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites, *Proceedings of the National Academy of Sciences of the United States of America* **109**, 11681 (2012).
5. V. Le Guilloux, P. Schmidtke and P. Tufféry, Fpocket: an open source platform for ligand pocket detection, *BMC Bioinformatics* **10**, p. 168 (2009).
6. T. A. Halgren, Identifying and characterizing binding sites and assessing druggability, *Journal of Chemical Information and Modeling* **49**, 377 (2009).
7. A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann and M. Rarey, Combining global and local measures for structure-based druggability predictions, *Journal of Chemical Information and Modeling* **52**, 360 (2012).
8. D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, L. Luo, B. Xia, D. Beglov and S. Vajda, The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins, *Nature Protocols* **10**, 733 (2015).
9. 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**, 3036 (2017).
10. M. M. Stepniewska-Dziubińska, P. Zielenkiewicz and P. Siedlecki, Improving detection of protein–ligand binding sites with 3D segmentation, *Scientific Reports* **10**, p. 5035 (2020).
11. R. Aggarwal, A. Gupta, V. Chelur, C. V. Jawahar and U. D. Priyakumar, DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks, *Journal of Chemical Information and Modeling* **62**, 5069 (2022).
12. R. Krivák and D. Hoksza, P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure, *Journal of Cheminformatics* **10**, p. 39 (2018).
13. A. Meller, M. Ward, J. Borowsky, M. Kshirsagar, J. M. Lotthammer, F. Oviedo, J. Lavista Ferres and G. R. Bowman, Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network, *Nature Communications* **14**, p. 1177 (2023).
14. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma and R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, *Proceedings of the National Academy of Sciences of the United States of America* **118**, p. e2016239118 (2021).
15. V. Škrhák, M. Novotný, C. P. Feidakis, R. Krivák and D. Hoksza, CryptoBench: cryptic protein–ligand binding sites dataset and benchmark, *Bioinformatics* **41**, p. btae745 (2025).
16. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* **1**, 206 (2019).
17. M. Steinegger and J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, *Nature Biotechnology* **35**, 1026 (2017).
18. S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta *et al.*, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology

and energy, *Nucleic Acids Research* **47**, D464 (2019).
19. J. Kyte and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology* **157**, 105 (1982).
20. A. A. Zamyatnin, Protein volume in solution, *Progress in Biophysics and Molecular Biology* **24**, 107 (1972).
21. T. Saito and M. Rehmsmeier, The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE* **10**, p. e0118432 (2015).
22. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC, New York, 1993).