

PAAS reveals interpretable local structural microenvironments across proteome-scale AlphaFold models

Andre Falcao

aofalcao@fc.ul.pt

Faculdade de Ciências, Universidade de Lisboa <https://orcid.org/0000-0002-3588-8746>

Rita Guedes

Faculdade de Farmácia, Universidade de Lisboa

Article

Keywords: protein structure, structural bioinformatics, local structural microenvironments, AlphaFold, protein annotation, structural fingerprints

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10578007/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

PAAS reveals interpretable local structural microenvironments across proteome-scale AlphaFold models

Andre O. Falcao^{1,2*} and Rita C. Guedes^{3,4}

^{1*}Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisboa, 1749-06, Portugal.

²BioISI, Biosystems and Integrative Sciences Institute Centre, Lisbon, Portugal.

^{3*}Faculdade de Farmácia da Universidade de Lisboa, Av. Prof Gama Pinto, Lisboa, 1649-003, Portugal.

⁴Research Institute for Medicines (iMed.Ulisboa), Av. Prof Gama Pinto, Lisboa, 1649-003, Portugal.

*Corresponding author(s). E-mail(s): aofalcao@ciencias.ulisboa.pt;

Contributing authors: rguedes@ff.ulisboa.pt;

Abstract

Protein structures are increasingly available at proteome scale, but general-purpose representations that are compact, residue-traceable and suitable for downstream analysis remain limited. We introduce Proximal Amino Acid Sets (PAAS), an unsupervised framework that converts sequence-order-independent three-dimensional residue neighbourhoods into a discrete, substitution-aware vocabulary of structural annotations. A vocabulary of 20,736 PAAS labels was constructed from approximately 5.4×10^5 AlphaFold models of reviewed UniProtKB/Swiss-Prot proteins. Each protein is represented by a sparse structural fingerprint, while every label remains linked to the residues that generated it. Although PAAS was not trained to identify functional sites, label information scores recovered catalytic residues, ligand-contact regions and structural motifs with median AUROCs of 0.766, 0.675 and 0.857, respectively. PAAS outperformed sequence-derived controls across all benchmarks and improved over exact structural hashing for catalytic residues and structural motifs, while ligand-binding results were more heterogeneous. PAAS therefore provides a scalable and interpretable structural annotation layer for protein comparison, residue prioritization and downstream machine-learning applications.

Keywords: protein structure, structural bioinformatics, local structural microenvironments, AlphaFold, protein annotation, structural fingerprints

1 Introduction

Protein function is often determined by local three-dimensional arrangements of residues rather than by global fold alone. Catalytic centers, ligand-binding cavities, transport pathways and interaction interfaces are defined by small constellations of amino acids that may recur across proteins with weak sequence similarity or even distinct overall architectures. Detecting and comparing such local structural microenvironments is therefore central to functional annotation, drug discovery and the interpretation of convergent biochemical behavior. Classical sequence-search methods remain indispensable for recognizing homologous relationships, and structural alignment methods extend comparative analysis to proteins whose folds are conserved despite substantial sequence divergence [1–4]. However, these approaches are usually optimized for detecting sequence homology, assessing global fold similarity or performing pairwise structural alignment, rather than for representing proteins as a collection of reusable, residue-grounded local environments.

The scale of this problem has changed substantially with the arrival of highly accurate protein-structure prediction. The Protein Data Bank provided the experimental foundation for structure-based biology [5], but AlphaFold and the AlphaFold Protein Structure Database have expanded structural coverage to proteome-scale collections spanning a large fraction of known sequence space [6, 7]. As a result, the limiting step is increasingly not whether a structure is available, but whether millions of predicted structures can be organized into representations that are compact, comparable and biologically interpretable. In particular, methods are needed to identify local structural patterns directly from AlphaFold-scale datasets and map those patterns back to the residues that generate them.

Proximal Amino Acid Sets (PAAS) address this need by representing proteins as collections of local residue neighborhoods. A PAAS is an order-independent set of spatially proximal residues extracted from a protein structure. Rather than requiring exact amino-acid identity, PAAS represents residues in a substitution-informed embedding space derived from BLOSUM62 [8], such that conservative substitutions remain close in the resulting representation. Recurrent local environments are then grouped into a discrete vocabulary of PAAS labels that can be used to annotate proteins, construct sparse structural fingerprints and assign residue-level information scores. Because each PAAS remains linked to its source residues, this representation connects proteome-scale structural encoding with residue-level interpretation.

This work builds on an earlier PAAS-inspired representation in which three-dimensional residue neighborhoods were converted into hashed structural fingerprints for downstream modeling [9]. That initial formulation showed that proximal residue sets could support scalable structure-derived representations, but exact or near-exact hashing is brittle to conservative substitutions and does not naturally organize related local

environments into shared motif classes. Here, PAAS is reformulated as a substitution-aware vocabulary of local chemical-structural microenvironments. The aim is not only to compare proteins, but also to identify the local residue environments that carry the strongest structural signal within each protein.

Related work and positioning of PAAS

Several families of methods address different aspects of the local-structure annotation problem. Binding-site detectors identify candidate pockets, clefts, or hot spots on individual protein structures. Geometry-based methods such as Fpocket efficiently detect ligandable cavities from protein shape, whereas probe- or energy-based approaches such as SiteHound and FTMap identify regions that are energetically favorable for molecular interactions [10–12]. Template-based methods, including 3DLigandSite and COFACTOR, infer functional sites by transferring information from related structures with known ligands or annotations [13, 14]. These methods are powerful for site prediction, but their primary goal is to locate or transfer specific functional sites, rather than to construct a general, reusable vocabulary of residue-defined microenvironments for proteome-scale annotation.

A second line of work focuses on conserved local functional environments. Methods such as Evolutionary Trace and ConSurf highlight residues under evolutionary constraint, often revealing catalytic surfaces or interaction regions when suitable sequence families are available [15, 16]. Structural site-comparison methods, including SiteEngine, PocketMatch, and ProBiS, compare local binding-site geometry and chemistry, while sequence order-independent profile comparisons and studies of pocket conservation have shown that local functional relationships may persist even when global similarity is weak [17–21]. These approaches support the central motivation for PAAS: biologically meaningful structural features can be conserved in localized residue-defined environments. PAAS differs by making such environments the primitive units of a proteome-scale annotation vocabulary, rather than treating them primarily as pairwise sites to align or compare.

A third relevant tradition is structural tokenization. Structural alphabets such as Protein Blocks and 3D-BLAST encode proteins as sequences of local backbone states, and FragBag represents proteins as bags of short contiguous fragments for rapid alignment-free retrieval [22–24]. More recent methods, such as Geometricus and Foldseek, demonstrate that structural information can be discretized into reusable symbols at a very large scale: Geometricus uses rotation-invariant local shape descriptors, whereas Foldseek uses a tertiary-interaction alphabet to enable fast structure search [4, 25]. These methods are conceptually close in spirit to PAAS in that they transform 3D structure into compact, scalable vocabularies. Their basic units, however, are typically contiguous fragments, local shape descriptors or per-residue interaction states, whereas PAAS uses multi-residue, sequence-order-independent neighborhoods that explicitly encode local amino acid composition.

Machine-learning approaches have also advanced binding-site and interaction prediction. DeepSite uses 3D convolutional neural networks over voxelized protein environments, P2Rank learns from local surface neighborhoods, and MaSIF applies geometric deep learning to molecular surfaces [26–28]. Although these models can

118 achieve strong predictive performance, their learned internal representations are gener-
119 ally optimized for prediction rather than for producing a discrete, directly interpretable
120 vocabulary of residue constellations that can be counted, compared and mapped back
121 onto proteins.

122 PAAS occupies a distinct position at the intersection of these approaches. Like pocket
123 detectors, it can highlight putative functional regions; like site-comparison methods, it
124 targets recurring microenvironments; and like structural alphabets, it converts protein
125 3D structure into reusable discrete descriptors. Its distinguishing features are (i) the use
126 of sequence-order-independent sets of spatially proximal residues as primitive units, (ii) a
127 substitution-aware amino acid embedding, (iii) the unsupervised construction of a motif
128 vocabulary from large structural collections, and (iv) direct residue-level traceability.
129 Together, these properties make PAAS a suitable intermediate representation between
130 global structure comparison and explicit functional-site annotation, particularly in the
131 AlphaFold era, in which the central challenge is to convert proteome-scale structural
132 availability into interpretable biological insight.

133 2 Results

134 2.1 Structural mapping of PAAS information scores

135 To illustrate the biological relevance of PAAS-derived information scores, two represen-
136 tative proteins with well-characterized ligand-binding regions were examined in detail:
137 the human serotonin 2A receptor (5-HT_{2A}R; UniProtKB P28223) and the immune-
138 checkpoint protein programmed cell death 1 ligand 1 (PD-L1; UniProtKB Q9NZQ7).
139 For each protein, residue-level PAAS information scores were mapped onto the three-
140 dimensional AlphaFold model and visualized using a spectrum ranging from blue,
141 indicating low information content, to red, indicating high information content. This
142 analysis was designed to determine whether residues associated with high-information
143 PAAS labels localize to structurally and functionally relevant regions. Importantly,
144 ligand-binding annotations and functional-site information were not used during PAAS
145 construction.

146 Figure 1 shows the distribution of PAAS information scores across the structures of
147 5-HT_{2A}R (Fig. 1A) and PD-L1 (Fig. 1B). In both proteins, the highest-scoring residues
148 are not distributed uniformly across the structure, but instead form spatially coherent
149 clusters. These clusters correspond to distinctive local structural environments that
150 occur less frequently in the global structural corpus and therefore contribute more
151 strongly to the information content assigned by PAAS.

152 To assess whether the highest-scoring PAAS regions coincide with biologically
153 relevant binding sites, the PAAS information maps were compared with experimentally
154 determined ligand poses and protein-protein interaction surfaces. Figure 2 shows that
155 the experimental ligands are positioned close to regions enriched in high PAAS scores,
156 indicating that PAAS captures local structural environments associated with molecular
157 recognition despite being derived solely from structural-neighborhood patterns.

158 For 5-HT_{2A}R, residues with high PAAS scores were concentrated within the ligand-
159 binding cavity formed by the transmembrane helical bundle, consistent with the
160 location of residues involved in serotonergic-ligand recognition. The experimental poses

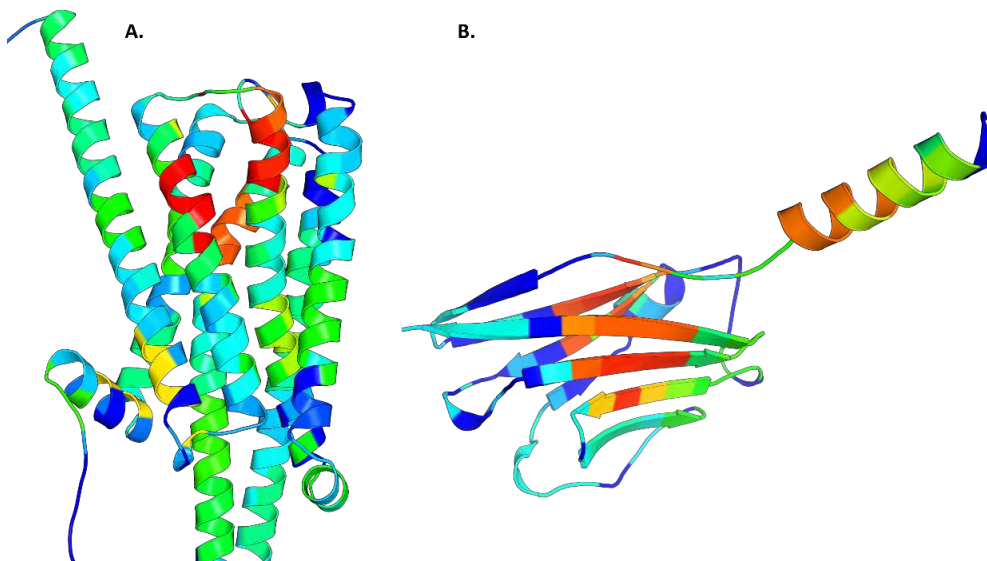


Fig. 1 Structural mapping of PAAS information scores onto the serotonin 5-HT_{2A} receptor and the immune-checkpoint protein PD-L1. (A) AlphaFold model of human 5-HT_{2A}R. (B) AlphaFold model of human PD-L1. Proteins are shown in cartoon representation, with residues colored according to their PAAS information score, from low information content in blue to high information content in red. The figures were prepared using PyMOL.

of LSD, lisuride, lumateperone, and IHCH-7086 (PDB entries 7WC6, 7WC7, 7WC8, and 7WC9, respectively) occupy a region enriched in high PAAS scores [29]. Several residues within this pocket showed elevated PAAS information scores, including Met97 (0.24), Trp151 (0.38), and Phe339 (0.32), the latter corresponding to position 6.51 in the Ballesteros–Weinstein numbering scheme. These residues lie near the conserved aminergic ligand anchor Asp155 and the transmembrane-helix-6 region associated with receptor activation.

For PD-L1 (Fig. 2B), high-scoring PAAS residues were observed near both the PD-1–PD-L1 interaction surface and the small-molecule binding region. The structure of the human PD-1–PD-L1 complex defines the protein–protein interaction interface [31], whereas the BMS-202 complex shows that this inhibitor binds within a hydrophobic cleft formed at a PD-L1 dimer interface and occludes the PD-1-binding surface [30]. Residues associated with this region, including Tyr56, Met115, and Tyr123, showed PAAS scores of 0.16, 0.22, and 0.20, respectively. This spatial correspondence suggests that PAAS identifies structurally informative local environments that overlap with ligand-recognition and protein-interaction regions in both 5-HT_{2A}R and PD-L1.

2.2 Benchmark recovery of functional and structural regions

We evaluated whether PAAS-derived residue scores recover independently annotated structural regions across three benchmark sets: catalytic residues (Benchmark A),

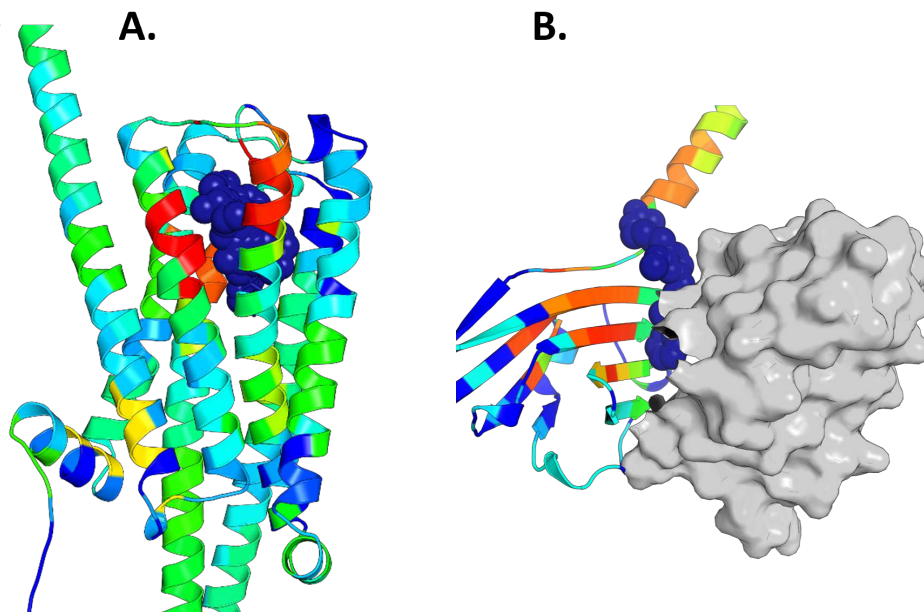


Fig. 2 Relationship between PAAS information scores and experimentally determined binding regions. (A) Human serotonin 5-HT_{2A}R shown in cartoon representation, with ligands from PDB entries 7WC6, 7WC7, 7WC8, and 7WC9 shown as deep-blue spheres. These structures contain LSD, lisuride, lumateperone, and the non-hallucinogenic psychedelic analogue IHCH-7086, respectively [29]. (B) Human PD-L1 shown in cartoon representation, with the small-molecule inhibitor BMS-202 from PDB entry 5J89 shown as deep-blue spheres [30] and PD-1 from the human PD-1-PD-L1 complex (PDB entry 4ZQK) shown as a grey molecular surface [31]. Residues of 5-HT_{2A}R and PD-L1 are colored according to their PAAS information score, from low information content in blue to high information content in red. The figures were prepared using PyMOL.

180 ligand-contact and drug-binding residues (Benchmark B), and annotated structural
 181 motif regions (Benchmark C). Statistical significance was assessed using within-protein
 182 permutation tests and paired comparisons with false-discovery-rate correction, as
 183 detailed in Methods. Each entry was treated as a within-protein residue-ranking task,
 184 using AUROC to measure global recovery, top-10% enrichment to measure hotspot
 185 concentration, and an empirical permutation statistic adjusted by the Benjamini-
 186 Hochberg procedure to assess whether annotated residues occupy unusually high-ranked
 187 positions. Because benchmark annotations were not used to construct PAAS, PHASH,
 188 SAAS or SHASH, these comparisons test whether biologically meaningful regions are
 189 recovered from unsupervised motif vocabularies rather than from supervised site labels.

190 PAAS showed the strongest overall performance across the three benchmarks
 191 (Table 1). In Benchmark A, PAAS achieved a median AUROC of **0.766**, compared
 192 with 0.645 for PHASH, 0.623 for SAAS and 0.524 for SHASH. The corresponding
 193 significant-recovery rate was 24/51 entries for PAAS, whereas PHASH had no significant
 194 entries and the sequence baselines had only two or three. Paired tests confirmed
 195 that PAAS significantly outperformed all baselines for AUROC and for the empirical

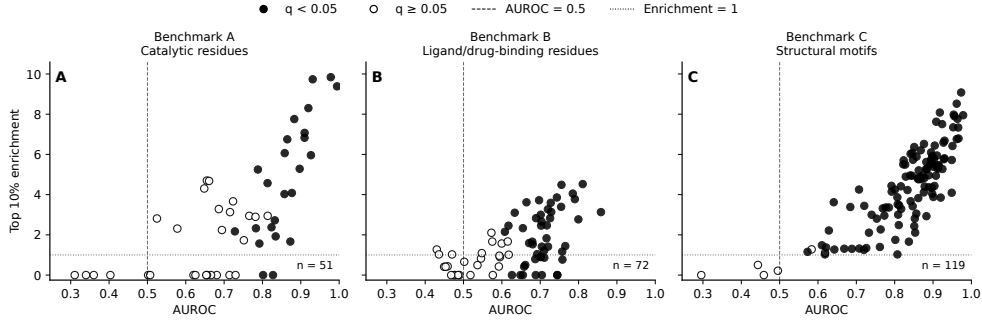


Fig. 3 PAAS-derived residue scores recover annotated functional and structural regions across three independent benchmarks. Each point represents one protein or annotation instance. The horizontal axis shows AUROC for ranking annotated residues or motif positions above background residues from the same protein, while the vertical axis shows enrichment of annotations among the top 10% highest-scoring residues. Dashed reference lines denote random global ranking (AUROC = 0.5) and no top-10% enrichment (enrichment = 1). Point fill indicates whether the Benjamini–Hochberg-adjusted empirical permutation value is below 0.05. Benchmark A evaluates catalytic residues, Benchmark B evaluates ligand-contact and drug-binding residues, and Benchmark C evaluates annotated structural motif regions. Baseline scatter plots are provided in the Supplementary Information; the corresponding quantitative comparisons are summarized in Tables 1–2.

statistic, including the comparison with the previous structural hashing representation (Table 2). This indicates that catalytic residues are not recovered merely because of local sequence words or exact structural hashing; the substitution-aware clustering of three-dimensional residue neighborhoods adds measurable signal.

Benchmark B showed a positive but more heterogeneous ligand-binding signal. PAAS had the highest median AUROC among the four approaches (0.675), and 47/72 ligand-binding annotation instances were significant at $q < 0.05$. PAAS clearly outperformed the sequence-derived baselines SAAS and SHASH in AUROC and in the empirical statistic, but did not significantly outperform PHASH under the conservative two-sided paired tests. PHASH also showed higher median top-10% enrichment in this benchmark. This result is consistent with the heterogeneous nature of ligand-contact annotations: some ligand contacts correspond to conserved structural environments, whereas others reflect ligand-specific, state-specific, shallow or crystallographic contacts. Thus, PAAS improves over sequence-only baselines for ligand-contact recovery, but the ligand benchmark does not support a claim that PAAS uniformly exceeds exact structural hashing.

Benchmark C provided the strongest support for PAAS as a local structural motif representation. PAAS reached a median AUROC of **0.857** and a median top-10% enrichment of **4.43**, exceeding PHASH, SAAS and SHASH on both metrics. Paired differences were significant for all three baselines, including a median AUROC gain of 0.065 and a median top-10% enrichment gain of 1.93 relative to PHASH. Although both PAAS and PHASH detected many structural motif regions at $q < 0.05$, producing a ceiling effect in the binary significance comparison, PAAS consistently improved the ranking and concentration of motif residues among the highest-scoring positions.

Table 1 Aggregate benchmark performance for PAAS and baseline representations. Rows denote benchmark entries and parentheses indicate the number of unique UniProt accessions. Benchmark B is reported at the annotation-instance level because some proteins have more than one ligand-binding annotation.

Benchmark	Method	Rows (acc.)	Median AUROC	Median top-10% enrichment	Significant $q < 0.05$
A: Catalytic residues	PAAS	51 (51)	0.766	2.72	24/51 (47.1%)
	PHASH	51 (51)	0.645	1.21	0/51 (0.0%)
	SAAS	54 (54)	0.623	1.76	3/54 (5.6%)
	SHASH	54 (54)	0.524	0.00	2/54 (3.7%)
B: Ligand binding	PAAS	72 (50)	0.675	1.19	47/72 (65.3%)
	PHASH	72 (50)	0.641	1.67	39/72 (54.2%)
	SAAS	72 (50)	0.558	1.75	14/72 (19.4%)
	SHASH	72 (50)	0.513	0.88	0/72 (0.0%)
C: Structural motifs	PAAS	119 (119)	0.857	4.43	113/119 (95.0%)
	PHASH	119 (119)	0.776	2.32	109/119 (91.6%)
	SAAS	120 (120)	0.777	3.09	95/120 (79.2%)
	SHASH	120 (120)	0.532	1.09	20/120 (16.7%)

Table 2 Abridged paired comparison of PAAS against each baseline. Effects are medians of paired PAAS–baseline differences on the shared benchmark instances; values in parentheses are Benjamini-Hochberg-adjusted two-sided Wilcoxon q -values. For the empirical permutation statistic, positive values indicate lower adjusted empirical values for PAAS and consequent stronger statistical support. The final column gives PAAS-only/baseline-only significant instances at $q < 0.05$, with the adjusted paired binomial q -value in parentheses.

BnchBaseline	n	Δ AUROC (q)	Δ top-10% (q)	Δ empirical (q)	Significant gain (q)
A PHASH	51	+0.099 (4.4e-5)	+0.20 (0.001)	+0.179 (4.8e-6)	24/0 (2.7e-7)
	SAAS	+0.156 (4.5e-6)	+0.00 (0.027)	+0.214 (8.2e-8)	22/0 (8.6e-7)
	SHASH	+0.208 (3.6e-6)	+1.63 (0.025)	+0.637 (2.1e-8)	23/1 (4.5e-6)
B PHASH	72	+0.002 (0.91)	+0.02 (0.95)	-0.000 (0.097)	11/3 (0.065)
	SAAS	+0.091 (8.5e-9)	+0.00 (0.50)	+0.196 (9.7e-9)	34/1 (6.3e-9)
	SHASH	+0.143 (2.1e-9)	+0.01 (0.073)	+0.504 (1.6e-11)	47/0 (6.4e-14)
C PHASH	119	+0.065 (7.8e-10)	+1.93 (3.9e-15)	+0.000 (0.022)	5/1 (0.22)
	SAAS	+0.086 (5.4e-16)	+1.03 (9.5e-15)	+0.000 (1.6e-6)	18/0 (9.8e-6)
	SHASH	+0.318 (2.5e-19)	+3.46 (7.9e-19)	+0.381 (7.9e-19)	95/1 (2.2e-26)

220 Taken together, the baseline analysis supports two conclusions. First, local primary
221 sequence context is insufficient to explain the observed recovery of catalytic, ligand-
222 binding and motif regions: SHASH performed close to chance in several settings, and
223 SAAS remained below PAAS despite using the same BLOSUM62-derived representation
224 and clustering scheme. Second, three-dimensional proximity is necessary but does not
225 fully explain the results: PHASH performed well, particularly for ligand-binding and
226 structural motif annotations, but PAAS improved on PHASH for catalytic residues

and structural motifs. These results support the central design of PAAS: unsupervised clustering of substitution-aware three-dimensional amino-acid neighborhoods captures biologically meaningful structural information that is not reducible to local sequence motifs or to exact hashed residue sets. Together, these results establish PAAS as an interpretable intermediate representation that links proteome-scale structural encoding to residue-level functional and structural interpretation.

3 Discussion

PAAS provides an unsupervised route to structural protein annotation by converting local three-dimensional residue environments into a compact vocabulary of recurrent structural microenvironments. The method is learned from the structural organization of a large AlphaFold/Swiss-Prot corpus itself, rather than from labelled catalytic residues, ligand-binding residues or curated motifs. The benchmark results therefore address a central question of this work: whether biologically meaningful functional regions can emerge from unsupervised structural abstraction. Across the three benchmarks, PAAS recovered catalytic, ligand-binding and motif regions in many proteins despite never being explicitly trained to recognize these sites.

The baseline analyses strengthen this conclusion. Sequence-only controls showed that the observed signal is not reducible to local primary-sequence words. A sequence representation using the same substitution-informed encoding and clustering strategy improved over raw sequence hashing in some settings, but remained below PAAS, indicating that three-dimensional proximity contributes essential information. Conversely, the comparison with exact structural hashing showed that structure alone is not the entire explanation: organizing proximal residue environments through substitution-aware clustering improved the recovery of catalytic residues and structural motifs. Ligand-binding annotations were more heterogeneous, and PAAS did not uniformly exceed the structural hashing baseline for all ligand metrics. Nevertheless, PAAS clearly outperformed sequence-derived controls and remained competitive with the previous structure-hashing approach, consistent with the broader diversity of ligand-contact regions.

These results position PAAS as an interpretable intermediate representation between residue-level structure and protein-level annotation. Each protein is summarized by a small set of local structural descriptors, yet these descriptors remain linked to the residues that generated them. This makes PAAS useful both for protein search and for residue-level interpretation: similar proteins, or proteins sharing related local functional environments, can be compared through compact structural fingerprints, while high-information PAAS labels can be mapped back onto candidate catalytic, binding or motif regions. The representation is therefore not only a retrieval tool, but also a mechanism for highlighting structurally meaningful regions within individual proteins.

The computational efficiency of the approach is also central to its practical value. PAAS descriptors can be generated from AlphaFold models in about 3.3 seconds per protein on average, and downstream scoring or search operations are performed at millisecond scale. This makes the method suitable for proteome-scale structural annotation and for iterative search settings where conventional local structural comparison

270 would be too expensive. The release of the complete PAAS pipeline as open-source soft-
271 ware under an MIT-style license further supports reproducibility, reuse and extension
272 by the community.

273 PAAS also offers a promising source of protein descriptors for machine-learning
274 applications, particularly proteochemometric modelling of ligand-binding activity. Such
275 models require informative representations of both ligands and proteins, but curated
276 binding-site annotations and experimentally resolved structures remain incomplete.
277 PAAS can provide compact, structure-derived descriptors from AlphaFold models with-
278 out supervised site labels, potentially complementing sequence, domain, evolutionary
279 and global-structure features. The results presented here do not imply that PAAS
280 captures all determinants of protein function, nor that it replaces detailed biochemical,
281 evolutionary or ligand-specific modelling. They do, however, show that unsupervised,
282 substitution-aware abstraction of local three-dimensional amino-acid environments
283 captures functional structural signal at scale. This makes PAAS a practical and exten-
284 sible framework for protein annotation, structural retrieval and the generation of
285 interpretable descriptors for future predictive models.

286 4 Methods

287 PAAS is designed to convert a protein structure into a set of interpretable local
288 structural annotations. The central idea is to treat a protein not only as a global fold,
289 but also as a collection of recurring residue microenvironments. Each microenvironment
290 is defined by amino acids that are close in three-dimensional space, regardless of their
291 order in the primary sequence. These local residue sets are embedded in a substitution-
292 aware vector space, grouped into a discrete vocabulary of recurring PAAS labels, and
293 then mapped back to the protein structures from which they were extracted. The result
294 is a representation that provides both a protein-level fingerprint and a residue-level
295 annotation that remains directly traceable to local structural regions.

296 The method is intentionally unsupervised. Functional annotations, catalytic residues,
297 binding-site labels, structural motifs and benchmark classes are not used to construct
298 the PAAS vocabulary. They are used only later as external reference data for evaluation.
299 PAAS therefore asks whether biologically meaningful regions emerge from the statistical
300 organization of local three-dimensional residue environments, rather than being imposed
301 by supervised training on known functional sites.

302 4.1 Overview of PAAS construction and protein labeling

303 Figure 4 summarizes the PAAS construction workflow. The first step defines a numerical
304 representation of amino acid identity. In this work, amino acids are represented
305 in a continuous space derived from their substitution relationships in BLOSUM62,
306 so that chemically or evolutionarily conservative substitutions are placed closer to
307 one another than more dissimilar substitutions [8]. Given a protein structure, local
308 spatial neighborhoods are then extracted by identifying residues that lie within a local
309 three-dimensional region around selected atoms or residues in the structure. Each
310 neighborhood is converted into a Proximal Amino Acid Set, or PAAS.

A PAAS is then represented as a vector by combining the embeddings of its constituent amino acids. Across a large collection of structures, identical or similar PAAS vectors are grouped into clusters. Each cluster defines a PAAS label, corresponding to a recurrent local chemical-structural microenvironment. Once this vocabulary has been defined, every PAAS extracted from a protein can be assigned to a label. A protein is therefore represented by the set of PAAS labels present in its structure, while individual residues can be annotated by the PAAS labels of the local environments in which they participate.

This workflow separates the conceptual representation from specific implementation choices. The same general PAAS framework can be instantiated with different structure collections, neighborhood definitions, embedding dimensionalities or clustering algorithms. The essential requirements are that local residue environments remain spatially defined, amino-acid identity is represented in a way that tolerates conservative substitutions, and the resulting labels remain traceable to the residues that generated them.

4.2 Details of the PAAS representation

4.2.1 Substitution-aware amino-acid representation

PAAS requires a numerical representation of amino acids before residue sets can be compared. Rather than treating amino acid identities as unrelated categorical symbols, PAAS embeds amino acids in a continuous space derived from their substitution behavior. The purpose of this step is not to model sequence evolution directly, but to provide a chemically and evolutionarily informed coordinate system in which conservative substitutions are naturally positioned closer together than nonconservative substitutions.

Let $e(a)$ denote the embedding vector assigned to amino acid a . In the present formulation, these vectors are derived from BLOSUM62, which summarizes observed substitution tendencies among amino acids in conserved protein blocks [8]. This embedding gives PAAS a controlled degree of tolerance: two local environments need not have identical amino acid compositions to be considered related if their constituent residues occupy nearby positions in the substitution-informed space.

4.2.2 Extraction of local residue neighborhoods

Given a protein structure, PAAS instances are extracted from local regions of three-dimensional space. For a local center c , all residues with at least one atom in its spatial neighborhood are collected. If x_i denotes an atomic coordinate and R_j denotes a residue instance, a generic neighborhood can be written as

$$\mathcal{N}(c) = \{R_j : \exists x_i \in R_j \text{ such that } d(x_i, c) \leq r\}, \quad (1)$$

where $d(\cdot, \cdot)$ is a spatial distance and r defines the neighborhood scale. The particular choice of center and radius is an implementation parameter and is described separately with the data preparation and benchmark settings.

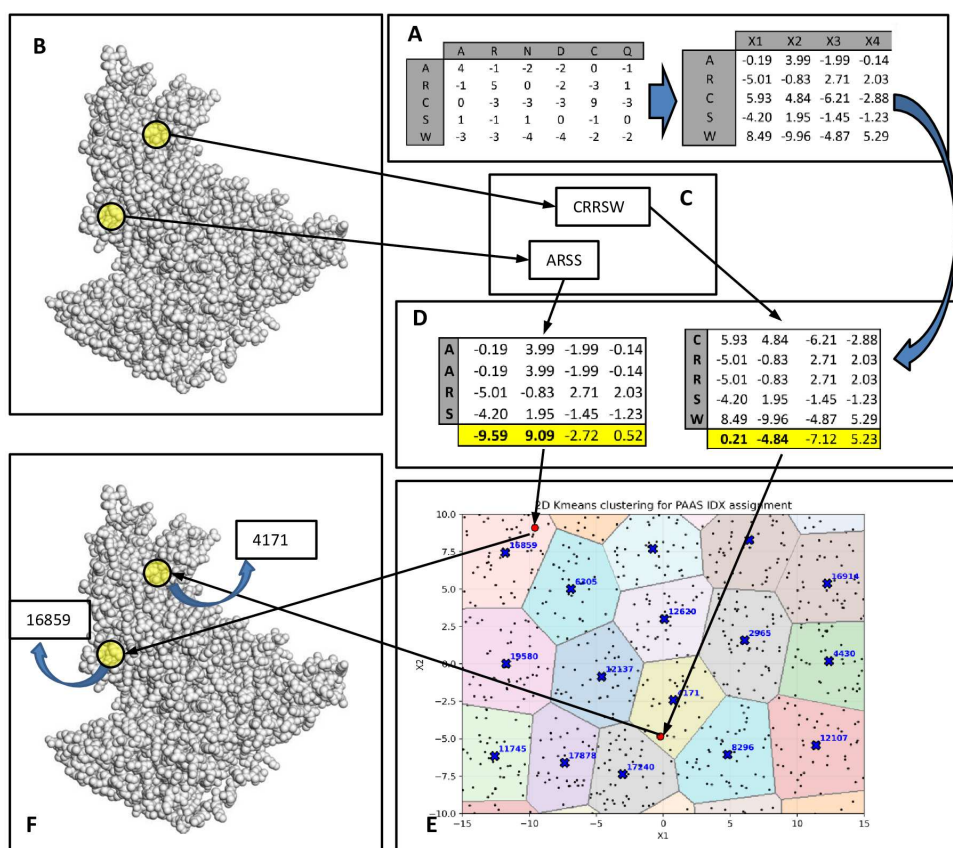


Fig. 4 Schematic overview of PAAS construction and protein labeling. (A) Amino acid identities are embedded in a substitution-informed continuous space derived from BLOSUM62, so that residues with similar substitution profiles occupy nearby positions. (B) For each protein structure, local three-dimensional residue neighborhoods are retrieved by collecting amino acids that fall within a defined spatial radius of a local center, independently of their order in the primary sequence. (C) Each local neighborhood is encoded as a Proximal Amino Acid Set (PAAS), an order-independent multiset of amino acid identities that preserves residue multiplicity while discarding sequence order. (D) The PAAS is converted into a numerical vector by aggregating the BLOSUM62-derived embeddings of its constituent amino acids. (E) PAAS vectors are assigned to discrete PAAS label identifiers by clustering similar vectors in the substitution-aware PAAS space, thereby defining a vocabulary of recurrent local chemical-structural microenvironments. (F) The resulting PAAS labels are mapped back onto the source protein, producing both a protein-level structural fingerprint and residue-level annotations that remain traceable to the local neighborhoods from which they were derived.

350 The neighborhood is converted from residue instances to amino-acid identities while
 351 preserving multiplicity. Thus, if two distinct alanine residues occur in the same local
 352 environment, both contribute to the PAAS. At the same time, the order of residues in
 353 the primary sequence is discarded. A PAAS is therefore best understood as an order-
 354 independent multiset of amino-acid identities associated with a specific local region

355 of a protein structure. This is the key representational choice: PAAS captures spatial
 356 proximity and local composition without requiring contiguous sequence fragments or
 357 global structural alignment.

358 4.2.3 Vectorization of PAAS

359 After extraction, each PAAS is mapped to the same continuous space used to repre-
 360 sent amino acids. If a PAAS P contains amino-acid identities $a_1, \dots, a_m$, its vector
 361 representation is defined as

$$362 \quad V(P) = \sum_{t=1}^m e(a_t). \quad (2)$$

363 This aggregation is order-independent, consistent with the definition of PAAS
 364 as spatial residue sets rather than sequence fragments. It also preserves multiplicity
 365 because each residue instance contributes separately to the sum. As a result, local
 366 environments containing conservatively substituted residues tend to occupy nearby
 367 regions of PAAS space, whereas environments with different amino acid compositions
 368 or different residue multiplicities remain distinct.

369 4.2.4 Construction of a PAAS vocabulary

370 The set of PAAS vectors extracted from a structural collection defines a high-
 371 dimensional catalog of local residue environments. To convert this catalog into a
 372 reusable annotation system, PAAS vectors are grouped into clusters. Each cluster
 373 represents a recurrent class of local chemical-structural environments and is assigned a
 374 discrete PAAS label.

375 This clustering step is conceptually independent of the specific algorithm used
 376 to perform it. Any method that groups nearby PAAS vectors into stable regions of
 377 the embedding space can define a PAAS vocabulary. Once the vocabulary is fixed, a
 378 newly observed PAAS can be assigned to the closest or most appropriate label in that
 379 vocabulary. The important outcome is that similar local residue environments receive
 380 the same or related labels, allowing proteins to be compared through recurring local
 381 structural features rather than through exact neighborhood identity.

382 4.2.5 Protein fingerprints and residue-level annotation

383 After PAAS labels have been assigned, each protein can be represented as a sparse
 384 fingerprint of the labels present in its structure. Let $\mathcal{P}_n$ be the set of PAAS extracted
 385 from protein n , and let $L(P)$ be the label assigned to PAAS P . The protein-level
 386 fingerprint is

$$387 \quad \mathcal{F}_n = \{L(P) : P \in \mathcal{P}_n\}. \quad (3)$$

388 This representation records which local structural microenvironments are present
 389 in a protein. Repeated occurrences can be retained as counts or collapsed to presence-
 390 absence labels, depending on the downstream analysis. In the present manuscript, the
 391 emphasis is on the presence of distinct PAAS labels as interpretable structural features.

392 Because each PAAS originates from a defined local region of a protein structure,
 393 PAAS labels can also be projected back onto individual residues. A residue is annotated

by the labels of all PAAS in which it participates. This residue-level traceability is essential for interpretation: a protein-level fingerprint can be decomposed into the specific local regions that produced it, and the procedure is described next.

4.2.6 From PAAS labels to residue-level information scores

Not all PAAS labels carry the same interpretive value. Some local environments are common across many proteins and may reflect generic structural packing or other broadly recurrent architectural features. Others are rare or selectively enriched in particular functional or structural contexts. To distinguish these cases, each PAAS label can be assigned an information-content score based on its prevalence across the structural collection.

If p_ℓ denotes the empirical prevalence of label ℓ among proteins, a generic information-content score is

$$I_\ell = -\log p_\ell. \quad (4)$$

Labels found in many proteins therefore receive lower scores, whereas labels found in fewer proteins receive higher scores. These scores can be mapped back to residues through the PAAS labels in which each residue participates. A simple residue-level score is the maximum information content among the labels associated with that residue,

$$S(r) = \max_{\ell \in \Lambda(r)} I_\ell, \quad (5)$$

where $\Lambda(r)$ is the set of PAAS labels involving residue r . This rule assigns high scores to residues that participate in at least one high-information local environment. The statistical evaluation of these residue-level scores against external annotations is described separately, together with benchmark construction and performance metrics.

Overall, PAAS provides a bridge between local structural chemistry and proteome-scale annotation. It converts protein structures into a discrete vocabulary of spatial residue environments, provides sparse protein fingerprints for comparison, and preserves the mapping from PAAS labels back to the residues they derive from. Implementation parameters, data sources and benchmark-specific settings are reported separately to keep the methodological definition general.

4.3 Structural corpus and PAAS construction settings

This section defines the structural corpus used to construct PAAS and the independent residue-level benchmarks used to evaluate whether PAAS-derived scores recover structurally meaningful protein regions. The corpus was used to learn and assign PAAS labels, whereas the benchmark annotations were used only after PAAS construction. Thus, catalytic residues, ligand-contact residues, and structural motif regions did not contribute to the construction of the PAAS vocabulary or to the definition of PAAS information scores.

4.3.1 Corpus and construction parameters

The PAAS vocabulary was constructed from AlphaFold-predicted structures corresponding to the reviewed UniProtKB/Swiss-Prot protein set, comprising approximately

5.4 $\times 10^5$ protein models at the time of retrieval. AlphaFold DB provides predicted structures indexed to UniProt accessions, allowing PAAS labels and residue-level scores to be projected directly onto the same canonical residue numbering used by UniProtKB annotations [6, 7, 32].

For all analyses reported here, local residue environments were extracted using a neighbourhood radius of 4.0 Å. This value was selected after some preliminary tests. A radius of 3.0 Å produces mostly sequence triplets, and few structural contact points. On the other hand, a value of 5.0 Å produced 22.3×10^6 unique PAAS, that did not seem to improve substantially the quality of the protein annotations. Amino-acid identities were represented in a substitution-informed space derived from BLOSUM62 [8], and the retained embedding dimension was 16, capturing 98% of the variance of the data. The resulting 2.02×10^6 PAAS vectors were assigned to a discrete vocabulary of 20,736 PAAS labels, obtained from a two-level K-Means clustering, by initially clustering all PAAS vectors into 144 centroids and, on a second phase, the vectors assigned to each first-level cluster were subsequently divided into 144 subclusters, thus ending with 144×144 final labels. This stacked approach was deemed necessary in order to have a sufficiently detailed representation of the different possible structural protein elements. With such an amount of data, we selected the most efficient clustering approach, thus KMeans. Yet, clustering with the required number of centroids was unfeasible with the available computational resources, so a stacked approach was selected.

It is important to state that the radius of 4.0 Å and the clustering of 20,736 clusters are just our experimental conditions, not necessarily the ones that can produce the best results. These parameters were selected before benchmark evaluation on the basis of neighborhood composition, vocabulary size and computational feasibility; they were not optimized against benchmark performance. Changing the radius, the actual number of clusters and the clustering method is deemed to have an impact on the results. Preliminary exploration did not reveal large differences under the tested settings, although systematic optimization of the radius, vocabulary size and clustering procedure was outside the scope of this study.

The full pipeline can be seen in Fig.5. All tools are made available open source on the PAAS-Toolkit repository on GitHub.

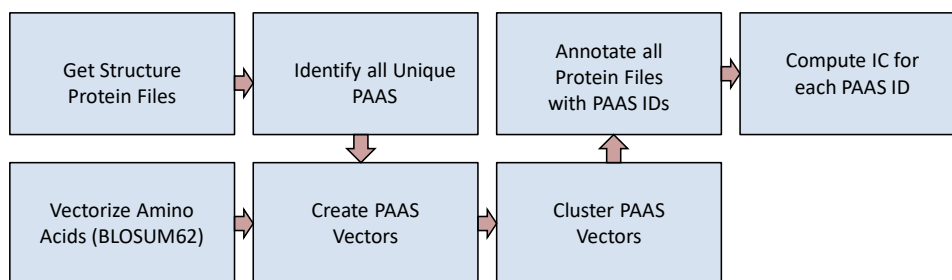


Fig. 5 Overview of the data processing pipeline. This follows the PAAS-Toolkit tools, made available on GitHub

Each protein was subsequently labelled by the PAAS vocabulary present in its structure. For residue-level evaluation, PAAS information scores were projected back to residues by assigning each residue the maximum information content among the PAAS labels in which that residue participated. The benchmark task was therefore formulated as a within-protein residue-ranking problem: externally annotated residues were treated as positives, and all other valid residues in the same protein formed the background ranking set.

The computational effort was mostly in the first phase, where all the UniProt/Swiss-Prot proteins were processed to identify their PAAS instances. This step required an average of 3.3 seconds per protein and, on a 9-year-old Intel Xeon workstation, took approximately 13 hours using 38 cores. The computational cost of the remaining procedures was negligible, including the stacked K-means clustering, protein annotation and information-content computation. On average, each protein was described by 990 unique PAAS IDs.

4.4 Benchmark construction

Three independent benchmark datasets were assembled to test complementary aspects of PAAS behaviour. Benchmark A evaluates recovery of sparse catalytic residues, Benchmark B evaluates recovery of experimentally observed ligand-contact regions, and Benchmark C evaluates recovery of bounded structural motif regions. All benchmark annotations were represented in canonical UniProt numbering wherever possible, so that they could be compared directly with residue-level PAAS scores from the corresponding AlphaFold DB models.

Benchmark A contained 51 enzymes with curated catalytic-residue annotations from the Mechanism and Catalytic Site Atlas (M-CSA), a manually curated resource of enzyme mechanisms and catalytic sites [33]. Proteins were retained only when explicit catalytic residues could be mapped to the canonical UniProt sequence and compared with PAAS residue scores. The positive class consisted of the annotated catalytic residue positions. This benchmark is intentionally stringent because catalytic residues are usually few in number and correspond to residues directly implicated in the catalytic mechanism, rather than to the complete surrounding active-site environment.

Benchmark B contained 72 ligand-binding annotation instances corresponding to 50 UniProt accessions. The benchmark was derived from experimentally determined protein–ligand structures in the Protein Data Bank. Protein residues from the experimental structures were mapped to canonical UniProt numbering using structure-to-sequence correspondence, including SIFTS mappings where available [? ?]. For each retained annotation instance, the positive class consisted of standard amino-acid residues having at least one heavy atom within 5.0 Å of the selected ligand heavy atoms. Only instances passing residue-mapping and annotation quality control were retained. Supplementary Table 1 reports the retained UniProt accession and annotation-instance identifier, together with the corresponding residue-level positive labels in canonical UniProt numbering. Benchmark B is more heterogeneous than Benchmark A because the experimentally observed contact shells may represent inhibitors, substrates, cofactors, allosteric ligands, shallow surface contacts, or conformational states specific to a particular experimental complex.

Table 3 Summary of benchmark datasets used for residue-level PAAS evaluation. The reported counts correspond to the final benchmark entries retained after mapping and quality control.

Benchmark	Annotation target	Entries	Positive residues used for evaluation
A	Catalytic residues	51	Curated catalytic residue positions from M-CSA
B	Ligand/drug-binding regions	72	Residues within 5.0 Å of selected ligand heavy atoms
C	Structural motifs/local modules	119	Residues inside curated UniProtKB or PROSITE/ProRule motif ranges

Benchmark C contained 119 proteins with curated residue-bounded structural motifs or local modules. Candidate annotations were collected from UniProtKB feature records and PROSITE/ProRule-associated annotations when they specified explicit residue ranges suitable for mapping to the canonical sequence [32, 34, 35]. The benchmark included compact or recurrent structural classes such as zinc fingers, EF-hand-like modules, leucine zippers, coiled-coils, WD repeats, leucine-rich repeats, and related local modules. The positive class consisted of all valid residues lying inside the annotated motif ranges. In contrast to the catalytic benchmark, Benchmark C evaluates recovery of extended structural regions rather than isolated functional residues.

Together, the three benchmarks provide complementary tests of whether PAAS-derived residue scores identify structurally meaningful protein regions. Benchmark A asks whether sparse catalytic residues are shifted toward high-information PAAS environments. Benchmark B asks whether experimentally observed ligand-contact shells are enriched among high-scoring PAAS regions. Benchmark C asks whether extended structural motifs and local modules are recovered as high-information residue regions. Because the annotations differ in size, biological origin, and expected spatial extent, the three datasets are interpreted jointly rather than as interchangeable replicates of the same task.

4.5 Baseline representations and ablation controls

To determine which components of PAAS were responsible for benchmark recovery, we evaluated three matched baseline representations using the same benchmark labels and the same residue-level scoring functions. The baselines were designed to address two questions: whether local sequence context alone is sufficient to recover annotated regions, and whether the substitution-aware BLOSUM62 representation with clustering improves on exact hashing of local residue sets.

SHASH is a sequence-only hashing baseline. For each protein, overlapping sequence windows of five residues were extracted, consisting of the central residue and two residues on each side; terminal positions lacking a complete window were not counted. Exact five-residue sequence words were then treated as discrete motifs. The theoretical sequence vocabulary has $20^5 = 3.2 \times 10^6$ possible five-plets, which is of the same order as the approximately two million exact PAAS-like local patterns observed before

539 vocabulary compression. This makes SHASH the closest sequence-only analogue to
540 exact local residue-set enumeration.

541 SAAS is a sequence-only substitution-aware baseline. It uses the same five-residue
542 moving windows as SHASH, but represents amino acids with the same BLOSUM62-
543 derived embedding used by PAAS and assigns windows to the same two-level 144×144
544 clustering scheme. SAAS therefore tests whether the BLOSUM62/clustering strategy
545 is sufficient in sequence space, independently of three-dimensional residue proximity.

546 PHASH is a structural hashing baseline corresponding to the exact local structural
547 residue-set representation used in our previous work [9]. Local structural neighbour-
548 hoods were extracted as in PAAS, but were discretized by hashing rather than by
549 BLOSUM62 vectorization and clustering. To keep the vocabulary scale comparable
550 with PAAS, PHASH labels were assigned using a modulo 144×144 hash space. Thus,
551 PHASH tests whether the present substitution-aware clustering scheme improves on
552 exact structural hashing while preserving the use of three-dimensional proximity.

553 For PAAS and all baselines, motif-level information scores were computed from
554 corpus frequencies and projected back to residues using the same residue-level rule:
555 each residue received the maximum score among the motifs in which it participated.
556 AUROC, top-10% enrichment, and the within-protein empirical permutation statistic
557 were then computed identically for all methods. Sequence-based baselines can be
558 evaluated whenever a protein sequence is available, whereas PAAS and PHASH require
559 an existing AlphaFold structural model. Consequently, sequence baselines have slightly
560 larger coverage in Benchmarks A and C. Paired statistical tests were performed only
561 on benchmark instances shared by PAAS and the corresponding baseline.

562 4.6 Benchmark evaluation of residue-level PAAS information 563 scores

564 PAAS assigns an information score to each residue before any benchmark annotation
565 is considered. This residue score is derived in two steps. First, each PAAS label
566 is assigned an information-content value from its empirical prevalence across the
567 structural corpus, so that common local microenvironments receive lower scores and
568 rarer local microenvironments receive higher scores. Second, this label-level information
569 is projected back onto protein residues: each residue receives the maximum information
570 content among the PAAS labels corresponding to local environments in which that
571 residue participates. Thus, the score assigned to a residue reflects the most informative
572 PAAS-defined microenvironment associated with that residue.

573 The benchmark evaluation asks whether these unsupervised residue-level PAAS
574 information scores preferentially highlight structurally meaningful regions. Each protein
575 is treated as an independent residue-ranking problem. Residues annotated by an
576 external benchmark are treated as positives, and all other valid residues in the same
577 protein are used as the within-protein background. These background residues are not
578 assumed to be biologically irrelevant; rather, they provide the reference distribution
579 against which the annotated residues are ranked. The evaluation therefore tests three
580 complementary properties of the PAAS residue scores: whether annotated residues are
581 globally ranked above the remaining residues, whether they are concentrated among the

highest-scoring PAAS hotspots, and whether their observed placement in the ranking is unlikely under random annotation placement within the same protein.

4.6.1 Global recovery of annotated residues: AUROC

The first question is whether annotated residues tend to receive higher PAAS scores than non-annotated residues across the whole protein. We measure this using the area under the receiver operating characteristic curve (AUROC), a standard rank-based summary for binary ranking problems [36, 37]. In this setting, AUROC can be written as

$$\text{AUROC} = \frac{1}{m(n-m)} \sum_{a \in A} \sum_{b \in B} \left[\mathbf{1}(s_a > s_b) + \frac{1}{2} \mathbf{1}(s_a = s_b) \right]. \quad (6)$$

Thus, AUROC is the probability that a randomly selected annotated residue is ranked above a randomly selected background residue, with ties counted as one half. A value of 0.5 corresponds to random ordering, values above 0.5 indicate preferential ranking of annotated residues, and values below 0.5 indicate the opposite. AUROC is included because it evaluates the full residue ranking and is insensitive to the absolute scale of PAAS scores. This is important for sparse annotations, such as catalytic residues, where a correct signal may shift annotated residues upward without necessarily placing every annotated residue in the extreme top-scoring region.

4.6.2 Hotspot localization: top- α enrichment

The second question is whether annotated residues are over-represented among the most prominent PAAS hotspots. We measure this using enrichment in the top fraction of the PAAS ranking. This follows the same logic as early-recognition and ranked-list enrichment metrics, for which performance depends not only on global ordering but also on whether positives are concentrated near the front of an ordered list [38, 39]. In the main benchmark summaries we use $\alpha = 0.10$, corresponding to the top 10% of residues.

Let T_α be the set of residues in the top α fraction of the PAAS score ranking, with $|T_\alpha| = \lceil \alpha n \rceil$ up to possible expansion when tied scores occur at the threshold. The top- α enrichment is

$$\text{Enrichment}_\alpha = \frac{|A \cap T_\alpha| / |T_\alpha|}{m/n}. \quad (7)$$

The numerator is the fraction of selected top-ranked residues that are annotated, and the denominator is the background annotation rate in the protein. An enrichment of 1 indicates the rate expected under random selection of residues, whereas values greater than 1 indicate that annotations are concentrated in the high-scoring PAAS region. This metric is included because PAAS is intended to support structural inspection and residue prioritization: in practical use, the question is often not only whether annotated residues are ranked above background overall, but whether the strongest PAAS-derived regions correspond to meaningful structural sites. Because a fixed threshold is a convention, top-10% enrichment is interpreted as an effect-size summary rather than as a significance test.

620 4.6.3 Within-protein statistical support: permutation q-value

621 The third question is whether the observed PAAS ranking of annotated residues is
 622 unlikely to arise by chance, given the same protein length and the same number of
 623 annotated residues. We test this using a one-sided empirical permutation procedure
 624 based on the mean PAAS rank percentile of the annotated residues. Let r_i be the rank
 625 of residue i after sorting residues by PAAS score, with the lowest score assigned rank 1
 626 and the highest score assigned rank n ; tied residues are assigned average ranks. The
 627 corresponding rank percentile is

$$628 \quad p_i = \frac{r_i}{n}, \quad (8)$$

629 so larger values indicate higher residues ranked higher by PAAS. The observed statistic
 630 is the mean annotated percentile,

$$631 \quad \bar{p}_A = \frac{1}{m} \sum_{a \in A} p_a. \quad (9)$$

632 For each permutation $b = 1, \dots, B_{\text{perm}}$, a random residue set R_b of size m is sampled
 633 without replacement from the same protein, and its mean percentile $\bar{p}_{R_b}$ is computed.
 634 The empirical one-sided permutation value is

$$635 \quad P_{\text{perm}} = \frac{1 + \sum_{b=1}^{B_{\text{perm}}} \mathbf{1}(\bar{p}_{R_b} \geq \bar{p}_A)}{B_{\text{perm}} + 1}. \quad (10)$$

636 The add-one correction reflects the finite resolution of random permutations and pre-
 637 vents zero empirical P values [40]. Because this test is repeated across benchmark
 638 entries, empirical P values are converted to false-discovery-rate-adjusted q values using
 639 the Benjamini–Hochberg procedure [41]. The permutation q value is interpreted as sta-
 640 tistical support for non-random placement of annotated residues in high PAAS-ranked
 641 positions, whereas AUROC and top- α enrichment are interpreted as complementary
 642 effect-size summaries.

643 4.7 Statistical analysis

644 Each benchmark entry was analysed independently as a within-protein residue-ranking
 645 problem. Residues were ranked by their information scores in ascending order, with
 646 average ranks assigned to tied values. AUROC was calculated using the rank-based
 647 Mann-Whitney formulation, with ties contributing one half. Top-10% enrichment was
 648 calculated as the proportion of annotated residues among the highest-scoring residues,
 649 divided by the overall proportion of annotated residues in the protein. The nominal
 650 top-10% set contained $\lceil 0.10n \rceil$ residues, where n is the number of valid residues; all
 651 residues tied with the score at the selection threshold were included.

652 Statistical support within each benchmark entry was assessed using a one-sided
 653 permutation test based on the mean rank percentile of the annotated residues. For
 654 each entry, 10,000 random residue sets containing the same number of residues as
 655 the annotation were sampled without replacement from the corresponding protein.
 656 The empirical P value was calculated as $(1 + N_{\text{perm}})/(10,000 + 1)$, where N_{perm} is

the number of random sets with a mean rank percentile greater than or equal to that observed for the annotated residues. A fixed random seed of 42 was used to make the analysis reproducible. Empirical P values were adjusted separately within each benchmark using the Benjamini-Hochberg false-discovery-rate procedure, and adjusted values below 0.05 were considered significant.

Comparisons between PAAS and each baseline were restricted to benchmark entries available for both representations. Paired differences in AUROC, top-10% enrichment and permutation-derived scores were evaluated using two-sided Wilcoxon signed-rank tests. Differences in the number of entries significant for only PAAS or only the corresponding baseline were evaluated using paired exact binomial tests. For the metric-level comparisons, P values were adjusted across the 27 benchmark-baseline-metric tests using the Benjamini-Hochberg procedure. P values from the nine paired significant-recovery comparisons were adjusted separately. Effect sizes are reported as the median paired PAAS-minus-baseline difference.

Data availability

The processed datasets generated in this study are available from Zenodo at <https://doi.org/10.5281/zenodo.20082135>. They were generated on 30 April 2026 using the PAAS-Toolkit and AlphaFold DB v2 models corresponding to the reviewed UniProtKB/Swiss-Prot protein set..

Code availability

The source code for the PAAS pipeline is available under an MIT license at <https://github.com/aofalcao/PAAS-Toolkit>. The repository includes the scripts required for PAAS construction, protein annotation and PAAS scoring. The version of the PAAS toolkit used for this work was stored as a snapshot at <https://doi.org/10.5281/zenodo.21724410>.

Supplementary information

Supplementary information accompanies this paper and contains additional methodological details, benchmark definitions, baseline analyses and extended results. The supplementary material includes:

Supplementary Table 1 Benchmark A, B and C definitions, including protein accessions, annotation sources and residue-level labels used for evaluation.

Supplementary Table 2 Detailed performance of PAAS, PHASH, SAAS and SHASH across the three benchmarks, including AUROC, top-10% enrichment and the fraction of benchmark entries with significant permutation support.

Supplementary Table 3 Paired statistical comparisons between PAAS and each baseline representation, reporting Wilcoxon signed-rank tests for AUROC, top-10% enrichment and permutation-derived scores, with Benjamini-Hochberg correction.

696 *Supplementary Figure 1* Benchmark scatter plots for PHASH across the catalytic,
 697 ligand-binding and structural-motif benchmarks.
 698 *Supplementary Figure 2* Benchmark scatter plots for SAAS across the catalytic, ligand-
 699 binding and structural-motif benchmarks.
 700 *Supplementary Figure 3* Benchmark scatter plots for SHASH across the catalytic,
 701 ligand-binding and structural-motif benchmarks.

702 References

- 703 [1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
 704 alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
- 705 [2] Holm, L. & Sander, C. Protein structure comparison by alignment of distance
 706 matrices. *Journal of Molecular Biology* **233**, 123–138 (1993).
- 707 [3] Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based
 708 on the TM-score. *Nucleic Acids Research* **33**, 2302–2309 (2005).
- 709 [4] van Kempen, M., Kim, S. S., Tumescheit, C. *et al.* Fast and accurate protein
 710 structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2024).
- 711 [5] Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Research* **28**, 235–242
 712 (2000).
- 713 [6] Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction
 714 with AlphaFold. *Nature* **596**, 583–589 (2021).
- 715 [7] Varadi, M., Anyango, S., Deshpande, M. *et al.* AlphaFold protein structure
 716 database: massively expanding the structural coverage of protein-sequence space
 717 with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).
- 718 [8] Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein
 719 blocks. *Proceedings of the National Academy of Sciences of the United States of*
 720 *America* **89**, 10915–10919 (1992).
- 721 [9] Oliveira, P. F., Guedes, R. C. & Falcao, A. O. Inferring molecular inhibition
 722 potency with AlphaFold predicted structures. *Scientific Reports* **14**, 8252 (2024).
- 723 [10] Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform
 724 for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
- 725 [11] Ghersi, D. & Sanchez, R. EasyMIFs and SiteHound: a toolkit for the identification
 726 of ligand-binding sites in protein structures. *Bioinformatics* **25**, 3185–3186 (2009).
- 727 [12] Kozakov, D., Grove, L. E., Hall, D. R. *et al.* The FTMap family of web servers
 728 for determining and characterizing ligand-binding hot spots of proteins. *Nature*
 729 *Protocols* **10**, 733–755 (2015).

- 730 [13] Wass, M. N., Kelley, L. A. & Sternberg, M. J. E. 3DLigandSite: predicting ligand-
731 binding sites using similar structures. *Nucleic Acids Research* **38**, W469–W473
732 (2010).
- 733 [14] Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm
734 for structure-based protein function annotation. *Nucleic Acids Research* **40**,
735 W471–W477 (2012).
- 736 [15] Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines
737 binding surfaces common to protein families. *Journal of Molecular Biology* **257**,
738 342–358 (1996).
- 739 [16] Ashkenazy, H., Abadi, S., Martz, E. *et al.* ConSurf 2016: an improved methodology
740 to estimate and visualize evolutionary conservation in macromolecules. *Nucleic*
741 *Acids Research* **44**, W344–W350 (2016).
- 742 [17] Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. Recognition of functional sites
743 in protein structures. *Journal of Molecular Biology* **339**, 607–633 (2004).
- 744 [18] Yeturu, K. & Chandra, N. R. PocketMatch: a new algorithm to compare binding
745 sites in protein structures. *BMC Bioinformatics* **10**, 543 (2009).
- 746 [19] Konc, J. & Janežič, D. ProBiS algorithm for detection of structurally similar
747 protein binding sites by local structural alignment. *Bioinformatics* **26**, 1160–1168
748 (2010).
- 749 [20] Xie, L. & Bourne, P. E. Detecting evolutionary relationships across existing fold
750 space, using sequence order-independent profile–profile alignments. *Proceedings of*
751 *the National Academy of Sciences of the United States of America* **105**, 5441–5446
752 (2008).
- 753 [21] Panjkovich, A. & Daura, X. Assessing the structural conservation of protein
754 pockets to study functional and allosteric sites: implications for drug discovery.
755 *BMC Structural Biology* **10**, 9 (2010).
- 756 [22] Fourier, L., Benros, C. & de Brevern, A. G. Use of a structural alphabet for
757 analysis of short loops connecting repetitive structures. *BMC Bioinformatics* **5**,
758 58 (2004).
- 759 [23] Yang, J.-M. & Tung, C.-H. Protein structure database search and evolutionary
760 classification. *Nucleic Acids Research* **34**, 3646–3659 (2006).
- 761 [24] Budowski-Tal, I., Nov, Y. & Kolodny, R. FragBag, an accurate representation of
762 protein structure, retrieves structural neighbors from the entire PDB quickly and
763 accurately. *Proceedings of the National Academy of Sciences of the United States*
764 *of America* **107**, 3481–3486 (2010).

- 765 [25] Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents
766 protein structures as shape-mers derived from moment invariants. *Bioinformatics*
767 **36**, i718–i725 (2020).
- 768 [26] Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G.
769 DeepSite: protein-binding site predictor using 3d-convolutional neural networks.
770 *Bioinformatics* **33**, 3036–3042 (2017).
- 771 [27] Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and
772 accurate prediction of ligand binding sites from protein structure. *Journal of*
773 *Cheminformatics* **10**, 39 (2018).
- 774 [28] Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular
775 surfaces using geometric deep learning. *Nature Methods* **17**, 184–192 (2020).
- 776 [29] Cao, D. *et al.* Structure-based discovery of nonhallucinogenic psychedelic analogs.
777 *Science* **375**, 403–411 (2022).
- 778 [30] Zak, K. M. *et al.* Structural basis for small molecule targeting of the programmed
779 death ligand 1 (PD-L1). *Oncotarget* **7**, 30323–30335 (2016).
- 780 [31] Zak, K. M. *et al.* Structure of the complex of human programmed death 1, PD-1,
781 and its ligand PD-L1. *Structure* **23**, 2341–2348 (2015).
- 782 [32] UniProt Consortium. UniProt: the universal protein knowledgebase in 2025.
783 *Nucleic Acids Research* **53**, D609–D617 (2025).
- 784 [33] Ribeiro, A. J. M. *et al.* Mechanism and catalytic site atlas (M-CSA): a database
785 of enzyme reaction mechanisms and active sites. *Nucleic Acids Research* **46**,
786 D618–D623 (2018).
- 787 [34] Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic*
788 *Acids Research* **41**, D344–D347 (2013).
- 789 [35] Sigrist, C. J. A. *et al.* ProRule: a new database containing functional and structural
790 information on PROSITE profiles. *Bioinformatics* **21**, 4060–4066 (2005).
- 791 [36] Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver
792 operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
- 793 [37] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**,
794 861–874 (2006).
- 795 [38] Truchon, J.-F. & Bayly, C. I. Evaluating virtual screening methods: good and bad
796 metrics for the “early recognition” problem. *Journal of Chemical Information*
797 *and Modeling* **47**, 488–508 (2007).

- 798 [39] Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach
799 for interpreting genome-wide expression profiles. *Proceedings of the National*
800 *Academy of Sciences of the United States of America* **102**, 15545–15550 (2005).
- 801 [40] Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: calculating
802 exact P-values when permutations are randomly drawn. *Statistical Applications*
803 *in Genetics and Molecular Biology* **9**, Article 39 (2010).
- 804 [41] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and
805 powerful approach to multiple testing. *Journal of the Royal Statistical Society:*
806 *Series B (Methodological)* **57**, 289–300 (1995).

807 Acknowledgements

808 A. O Falcao acknowledges financial support from UID/04046/2025 - Biosystems
809 and Integrative Sciences Institute Centre grant from FCT, Portugal. R. C. Guedes
810 acknowledges financial support from LISBOA-01-0246-FEDER-000017, Lisboa2030-
811 FEDER-03136100, and also acknowledges iMed.Ulisboa for their support from
812 UID/04138/2025 (<https://doi.org/10.54499/UID/04138/2025>), UID/PRR/04138/2025.

813 Author contributions

814 A.O.F. conceived the study, developed the PAAS methodology and software, performed
815 the computational analyses, curated the benchmark datasets, analysed the results and
816 wrote the original manuscript. R.C.G. contributed to study design, interpretation of
817 the structural and ligand-binding analyses, and critical revision of the manuscript.
818 Both authors reviewed and approved the final manuscript.

819 Competing interests

820 The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.zip](#)