Artificial Intelligence in the Life Sciences 9 (2026) 100166

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research article

Protein family membership governs exosite predictability across the structural proteome

Folorunsho Bright Omage [a,b,*], Ivan Mazoni [a], Inácio Henrique Yano [a], Goran Neshich [a,*]

[a] *Computational Biology Research Group, Embrapa Digital Agriculture, Av. André Tosello 209, Campinas, SP 13083-886, Brazil*
[b] *Biological Chemistry Laboratory, Department of Organic Chemistry, Institute of Chemistry, University of Campinas (UNICAMP), R. Monteiro Lobato 270, Campinas, SP 13083-970, Brazil*

ARTICLE INFO

Dataset link: https://www.exosite.cbi.cnptia.embrapa.br/, https://github.com/omagebright/STINGExoFind

Keywords:
Exosite prediction
Machine learning
Protein structure
Gradient boosting
Nanoenvironment descriptors
Drug discovery

ABSTRACT

Exosites, defined as protein surface regions that mediate macromolecular recognition at sites distinct from catalytic centers, represent emerging targets for selective drug design, yet their structural diversity has precluded systematic computational identification. Here we demonstrate that exosite prediction performance varies substantially across protein families, ranging from Matthews correlation coefficient (MCC) of 0.47 for coagulation factors to 0.14 for kinases. Using ExositeDB, we developed STINGExoFind, a gradient boosting framework leveraging 87 structural descriptors from the STINGRDB2 database, and evaluated 180 proteins under leave-one-protein-out cross-validation (LOPO-CV). Coagulation proteases achieved 50% success rates at the MCC $\geq$ 0.5 threshold, whereas kinases and caspases remained largely unpredictable. Ten structures spanning six families exceeded MCC $\geq$ 0.7, including MAPK/ERK2 (MCC = 0.86) within the otherwise challenging kinase family, indicating that high-confidence predictions remain achievable for specific proteins even in poorly-performing families. These results establish exosite prediction as a family-specific rather than universal challenge: computational approaches can meaningfully guide experimental validation for coagulation factors and similarly consistent protein families, while structurally diverse families require experimental characterization. STINGExoFind is provided as a community resource to support future method development and exosite-targeting drug discovery.

1. Introduction

The dominant paradigm in structure-based drug design targets orthosteric binding sites, which are the substrate pockets and catalytic centers where enzymes perform their biochemical functions. This approach, while responsible for numerous therapeutic successes, faces inherent limitations: orthosteric sites are often highly conserved across protein family members, complicating the design of selective inhibitors, and natural substrates typically bind with high affinity, necessitating potent competitors [1]. An alternative strategy targets exosites, which are secondary binding regions located distant from the catalytic center that mediate substrate recognition, tethering, and specificity through long-range interactions. Unlike allosteric sites that modulate catalytic activity through conformational changes, exosites function primarily by orienting substrates or recruiting binding partners, substrate selectivity without altering intrinsic catalytic efficiency [2]. Since exosites mediate substrate recognition through direct binding interactions distinct from the catalytic center, they provide opportunities for modulating enzyme specificity. In thrombin, exosite I serves as a recognition surface for multiple ligands including fibrinogen and the cofactor thrombomodulin; when thrombomodulin occupies exosite I, it excludes procoagulant substrates such as fibrinogen while positioning protein C for activation, demonstrating how exosite occupancy can redirect enzyme function [3]. While allosteric modulators achieve selectivity through conformational effects that alter active site properties [4], exosites operate as recognition surfaces that recruit and orient substrates through direct binding interactions. The functional consequence differs accordingly: allosteric modulation changes how efficiently an enzyme processes its substrates, whereas exosite occupancy determines which substrates are presented to the active site.

Beyond its role in substrate recognition, thrombin's exosite architecture has enabled therapeutic innovation. In addition to exosite I, thrombin possesses exosite II, which binds heparin, GPIbα, and prothrombin fragment 2 [3]. The bivalent inhibitor bivalirudin exploits this architecture by simultaneously occupying the active site and exosite I. Because exosite I presents structural features distinct from other coagulation proteases, this dual engagement achieves selective inhibition that would be difficult to attain through active-site targeting alone [5].

[*] Corresponding authors.
E-mail addresses: omagefolorunsho@gmail.com (F.B. Omage), goran.neshich@embrapa.br (G. Neshich).

https://doi.org/10.1016/j.ailsci.2026.100166
Received 8 January 2026; Received in revised form 26 February 2026; Accepted 10 March 2026
Available online 19 March 2026

Subsequent research identified exosite-like secondary binding regions across diverse enzyme families. Protein kinases employ docking grooves (including the D-recruitment site, DEF pocket, and αG-helix region) to recognize linear motifs on substrates and scaffold proteins, enabling specificity within the 500-member human kinome [6]. Protein phosphatases achieve selectivity through regulatory subunit interfaces that direct the catalytic core to specific substrates [7]. Matrix metalloproteinases use hemopexin domains and exosite loops for collagen recognition and processing [1]. This functional convergence, where secondary surfaces govern specificity across unrelated enzyme families, suggests that exosites represent a general organizational principle of enzyme regulation [8].

Despite their biological importance and therapeutic potential, no computational method has been specifically developed for predicting protein exosites. While tools exist for identifying other functional sites, such as surface cavities [9], cryptic binding sites [10], and allosteric sites [11–13], these address fundamentally different phenomena and are not applicable to exosite prediction. This gap persists because exosites present unique computational challenges: they lack the geometric enclosure of traditional binding pockets, they do not require conformational change for accessibility like cryptic sites, and unlike allosteric sites, they do not couple ligand binding to functional conformational changes.

A fundamental question precedes dedicated method development: do exosites possess sufficient structural regularity to enable computational prediction? The prospects appear uncertain a priori. Active sites benefit from catalytic constraints, as the chemical requirements of enzymatic mechanisms impose convergent structural features that machine learning can exploit. Exosites, serving diverse recognition functions across unrelated protein families, may lack such global conservation. However, an alternative hypothesis draws from the protein nanoenvironment concept: even structurally diverse functional sites may share local physicochemical signatures detectable through per-residue descriptors characterizing the immediate atomic neighborhood [14,15]. This approach has proved successful across multiple prediction challenges, including protein–protein interface residues [16], secondary structure element environments [17,18], allosteric site-forming residues [12,13], and CRISPR-Cas9 off-target activity [19]. Whether similar local signatures characterize exosites remains untested. The extreme class imbalance inherent to residue-level prediction (typically 1 to 2% of surface residues constitute exosites) compounds this challenge, as does the heterogeneous provenance of exosite annotations in the primary literature.

We addressed this question through systematic analysis of exosite predictability across the protein structure landscape. Using ExositeDB [20], to the best of our knowledge the only available database providing residue-level exosite annotations, we developed STINGExoFind, a machine learning framework integrating structural descriptors from the STINGRDB2 database [15,21,22] with gradient boosting classification. We evaluated performance through LOPO-CV, which represents the most stringent assessment of generalization to novel targets. Our results establish that exosite predictability is not a universal property of the prediction task but varies dramatically with protein family.

2. Methods

2.1. Data source

Exosite annotations were obtained from ExositeDB [20] (https://exosite.cbi.cnptia.embrapa.br), a curated database containing over 600 entries with 456 unique PDB structures and 350 unique proteins, with residue-level exosite designations supported by experimental evidence from peer-reviewed literature. For the present study, we selected a subset of 193 PDB structures representing 178 unique proteins based on the following criteria: (1) availability of complete residue-level exosite annotations; (2) PDB coordinate files with resolution ≤3.0 Å; (3) structures with ≤10% missing residues to ensure descriptor completeness; and (4) availability of corresponding structural descriptors in the STINGRDB2 database. The dataset comprises 145,397 total residues, of which 1914 (1.32%) are annotated as exosite residues.

The 10% missing residue threshold was chosen to balance dataset size with descriptor reliability. Of the 193 selected structures, 20 (10.3%) had complete atomic coordinates with no missing residues, 63 (32.5%) had 1%–5% missing residues, and 42 (21.6%) had 5%–10% missing residues, yielding a median missing residue fraction of 6.66% (mean 8.47%). Because STINGRDB2 computes per-residue descriptors using local atomic neighborhoods within 3–7 Å radii, missing residues beyond the immediate neighborhood have minimal impact on descriptor values for neighboring residues; descriptor incompleteness is localized to the missing region rather than propagating globally. Residual descriptor incompleteness arising from missing coordinates was encoded as NaN and handled by XGBoost's native missing value protocol, which learns optimal split directions for missing features during training. Furthermore, the LOPO-CV evaluation protocol independently validates each protein, ensuring that any residual incompleteness does not systematically bias predictions across the dataset.

Structural quality was assessed through multiple criteria. The resolution distribution of the 193 structures showed a median of 2.30 Å (mean 2.36 Å, range 1.09–4.20 Å), with 63.2% of structures resolved at ≤2.5 Å and 88.4% at ≤3.0 Å. The dataset comprises predominantly X-ray crystallographic structures (179, 92.7%), with 10 cryo-EM structures, 2 NMR structures, and 2 structures determined by other methods (Supplementary Table S3). While explicit filtering by R-work, R-free, or Ramachandran outlier percentages was not applied, several factors ensure structural quality: (a) all structures derive from ExositeDB, which curates entries from peer-reviewed literature describing experimentally validated exosites; (b) PDB deposition requires validation through the wwPDB validation pipeline, which checks geometric quality, including bond lengths, bond angles, and Ramachandran statistics; (c) STINGRDB2 performs coordinate validation during descriptor computation; and (d) the LOPO-CV protocol inherently tests robustness to structural quality variation, as each structure is independently evaluated against models trained on remaining data.

2.2. Structural descriptors and feature selection

Per-residue structural descriptors were retrieved from STINGRDB2 [15,21,22], a database providing pre-computed standardized structural analysis of PDB coordinate files. STINGRDB2 contains approximately 1280 per-residue protein structural descriptors [14,23,24]; these become internal protein nanoenvironment descriptors when applied to characterize specific functional regions. From these, we performed domain-knowledge-based feature selection to retain only descriptor categories relevant to surface binding site characterization. Excluded descriptor categories included: *Residue_Properties* (amino acid-specific physicochemical constants providing no position-specific information), *Internal_Contacts* and *Core_Residue* descriptors (designed to characterize buried residues rather than surface sites), *Catalytic_Site_Metrics* (focused on active site geometry, whereas exosites are by definition distinct from catalytic centers), *Cavity_Descriptors* (designed for pocket detection, whereas exosites are typically flat or shallow surfaces), *B_Factor_Derived* features (crystallographic quality metrics rather than functional descriptors), and various redundant secondary structure annotations. This domain-knowledge filtering reduced the feature space to selected 151 descriptors across eight categories specifically designed to capture surface geometry, local packing, spatial organization, and network topology, properties hypothesized to distinguish exosite regions from general protein surface.

For this study, these 151 descriptors span eight categories: *Density_Sponge* (40 features): local atomic density and packing patterns measured at radii 3 to 7 Å for Cα and last heavy atom (LHA) representations, with interface forming residue (IFR) normalized variants capturing packing density relative to protein–protein interface propensity [15]. *DSSP* (36 features): secondary structure assignments, hydrogen bonding patterns including donor and acceptor energies and residue numbers, backbone dihedral angles (ϕ, ψ, κ), accessibility, and structural motif indicators (helices, sheets, turns, bends, bridge pairs). *Cross_Presence_Order* (27 features): residue co-occurrence patterns quantifying neighbor distributions at sequence separation thresholds of 5, 10, and 15 residues for Cα, Cβ, and LHA atom types [14]. *Graph_Descriptor* (16 features): protein structure network properties including degree centrality, betweenness centrality, closeness centrality, clustering coefficient, eccentricity, PageRank, and local topological measures computed on the residue contact network. *Side_Chain_Orientation* (15 features): sidechain angular distributions at radii 3 to 7 Å, including average angles and neighbor sidechain orientations. *Weighted_Contact_Number* (9 features): distance-weighted contact counts and catalytic profile combination scores at multiple neighborhood sizes (k = 2 to 5). *Stride* (5 features): STRIDE-computed secondary structure assignments, solvent accessibility, and backbone dihedral angles. *Distances* (3 features): normalized distance to N-terminus, C-terminus, and center of geometry.

Further feature selection was performed exclusively on training data to prevent information leakage. Three sequential filters were applied: (1) *Variance filtering*: Features with variance <0.01 across training residues were removed, eliminating near-constant descriptors that provide no discriminative information (12 features removed). (2) *Univariate statistical testing*: ANOVA F-test assessed each remaining feature's ability to discriminate between exosite and non-exosite residues; features with $P \geq 0.05$ after Benjamini–Hochberg false discovery rate correction were removed (38 features removed). (3) *Information-theoretic filtering*: Mutual information between each feature and class labels was computed; features with MI = 0 (no dependence on class) were removed (14 features removed).

The final feature set comprises 87 descriptors (57.6% retention from the initial 151), with composition: Density_Sponge (28 features, 70% retention), Cross_Presence_Order (15 features, 56% retention), Graph_Descriptor (15 features, 94% retention), DSSP (9 features, 25% retention), Weighted_Contact_Number (5 features, 56% retention), Side_Chain_Orientation (6 features, 40% retention), Distances (2 features, 67% retention), and Stride (1 feature, 20% retention). Missing values arising from incomplete coordinate data were encoded as NaN and handled during model training through XGBoost's native missing value protocol.

2.3. Model architecture and training

Data were partitioned at the PDB structure level rather than residue level to prevent information leakage, ensuring all residues from the same PDB structure remain in the same partition. Protein family was used as stratification variable to ensure balanced representation across partitions. Initial experiments explored approaches that ignored family membership: a unified model trained on all proteins regardless of family achieved mean MCC = 0.225 under LOPO-CV, with only 11.1% of structures reaching MCC ≥ 0.5. Additional non-stratified approaches on held-out test data showed similar limitations: baseline XGBoost yielded MCC = 0.21, SMOTE oversampling degraded performance to MCC = 0.19, and graph neural networks failed entirely with MCC = −0.01 (worse than random). In a paired comparison on the same set of 190 structures, family-specific model routing increased the success rate from 11.1% to 15.3% of structures exceeding MCC ≥ 0.5, a 38% relative improvement (paired t-test $P = 0.52$ for mean MCC, indicating that the benefit manifests in the distribution tail rather than the central tendency). Under fully optimized per-family threshold tuning (evaluated on 180 eligible structures; see Section 2.4), the overall success rate reached 21.7%. These systematic comparisons demonstrate that family-aware model architecture is essential for meaningful exosite prediction, as a single unified model cannot effectively learn the diverse structural signatures across unrelated protein families.

The final partitions comprised: training set with 113 PDB structures (58.5%), 81,613 residues, 1054 exosite residues (1.29%); validation set with 36 PDB structures (18.7%), 25,560 residues, 434 exosite residues (1.70%); and test set with 44 PDB structures (22.8%), 38,224 residues, 426 exosite residues (1.11%). Family proportions were preserved within 5% relative deviation across all partitions. For LOPO-CV, 180 of the 193 PDB structures met the minimum requirements of ≥20 residues and ≥1 annotated exosite; the remaining 13 structures were excluded due to unstable MCC estimates on very small evaluation sets.

We employed XGBoost gradient boosting classifiers [25], selected for their robustness to class imbalance and native handling of missing values. Hyperparameters were determined through grid search on a held-out validation set, exploring 216 parameter combinations: maximum tree depth of 7 (searched: 3, 5, 7), learning rate of 0.1 (searched: 0.01, 0.05, 0.1), 200 estimators (searched: 100, 200, 300), minimum child weight of 5 (searched: 1, 3, 5), subsample fraction of 0.8 (searched: 0.8, 1.0), column subsample of 1.0 (searched: 0.8, 1.0), and scale_pos_weight of 76.43 computed from the training class ratio.

Class imbalance was addressed through the scale_pos_weight parameter, which weights positive class samples by the inverse class frequency (76.43 = 80,559 negative/1054 positive in training data). This weighting ensures that the loss function penalizes false negatives proportionally to class rarity.

Family-specific models were trained for families with ≥5 structures in training data: thrombin (18 structures), kinase (14), metalloproteinase (12), factor X (6), phosphatase (4), caspase (4), and other (55). At prediction time, proteins were routed to family-specific models based on UniProt family annotation; proteins without family annotation or from unrepresented families were evaluated with the general (other) model.

Classification thresholds were optimized per family through grid search on validation data, maximizing MCC. Optimal thresholds ranged from 0.05 (caspase, favoring recall) to 0.92 (kinase, favoring precision), reflecting family-specific class distributions and exosite characteristics.

2.4. LOPO-CV and performance metrics

LOPO-CV was implemented by iterating over all 180 eligible structures. For each held-out protein: (1) remove all residues of the held-out protein from training data; (2) retrain family-specific models on remaining data; (3) reoptimize classification thresholds on remaining validation structures; (4) predict exosite probabilities for held-out protein residues; (5) apply family-appropriate threshold to generate binary predictions; and (6) compute performance metrics. This protocol ensures complete independence of evaluation from training, simulating application to genuinely novel proteins.

The primary metric was MCC, computed as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

MCC ranges from −1 (perfect anti-correlation) through 0 (random) to +1 (perfect prediction), providing a balanced measure appropriate for imbalanced datasets. Structures with TP = FP = 0 or TN = FN = 0, yielding undefined MCC (0/0), were assigned MCC = 0. Secondary metrics included precision = TP/(TP+FP), recall = TP/(TP+FN), and F1 = 2 × precision × recall / (precision + recall). Performance thresholds were defined as: MCC ≥ 0.7 (high confidence, suitable for guiding experiments), MCC ≥ 0.5 (acceptable, useful for hypothesis generation), and MCC < 0.5 (limited utility).

2.5. Statistical analysis and baseline classifiers

Family-level comparisons used the Kruskal–Wallis H-test, a non-parametric alternative to one-way ANOVA appropriate for non-normal distributions. Post-hoc pairwise comparisons used Dunn's test with Bonferroni correction for multiple testing. Significance threshold was α

= 0.05 throughout. Confidence intervals for mean MCC were computed via bootstrap resampling: 1000 bootstrap samples were drawn with replacement from each family's structure-level MCC values, and 95% CIs were estimated as the 2.5th and 97.5th percentiles of the bootstrap distribution. Standard deviations reported for family-level metrics characterize within-family variation across structures, not estimation uncertainty.

Two baseline classifiers were implemented for comparison. The random baseline predicts exosite status for each residue with probability $p = 0.0129$ (training class prevalence), yielding expected MCC = 0 by construction and empirical MCC = 0.00 ± 0.03 across 1000 random trials. The accessibility baseline computes median STRIDE accessibility for each structure and predicts residues with accessibility above the median as exosite, testing whether exosite prediction reduces to simple accessibility thresholding.

3. Results

3.1. Dataset characteristics and structural feature representation

We utilized ExositeDB [20], a curated database containing experimentally validated residue-level exosite annotations derived from peer-reviewed literature. From this resource, we extracted 193 PDB structures representing 178 unique proteins, comprising 145,397 residues with 1914 (1.32%) annotated as exosite residues (Fig. 1a,b).

The dataset spans seven protein families, with coagulation factors (thrombin, factor X) most represented, reflecting historical research emphasis on these well-characterized exosite systems (Fig. 1a). The severe class imbalance (1.32% exosite residues, 76:1 ratio) necessitated class-weighted loss functions during model training (Fig. 1b).

Feature selection on training data (see Methods) retained 87 of 151 descriptors (57.6%). Solvent accessibility emerged as most discriminative ($F = 200.2$, $P = 2.2 \times 10^{-45}$), consistent with exosite surface exposure. Local density descriptors ranked second, with Density_Sponge features achieving F-statistics exceeding 150 ($F = 153.2$, $P = 3.7 \times 10^{-35}$). This feature profile reflects established exosite biology: accessible surface patches with characteristic local geometry.

3.2. LOPO-CV reveals family-dependent predictability

We evaluated STINGExoFind under LOPO-CV, the most stringent assessment of generalization to novel proteins. In each of 180 iterations, a single structure was withheld entirely from training (including any family-specific model components) and the classifier was evaluated on the held-out protein. This protocol simulates real-world application to proteins absent from the training database.

Aggregate LOPO-CV performance was modest: mean MCC = 0.25 ± 0.25, with 5.6% of structures (10/180) achieving MCC ≥ 0.7 and 21.7% (39/180) exceeding 0.5 (Fig. 2a–c). However, this aggregate obscures striking heterogeneity across protein families that constitutes our central finding (Table 1).

Coagulation proteases demonstrated high predictability. Factor X structures (n = 10) achieved mean MCC = 0.47 ± 0.24, with 50% exceeding the 0.5 acceptance threshold. Thrombin (n = 31) performed comparably at MCC = 0.47 ± 0.20, with 52% above threshold. These values exceed the accessibility baseline (MCC = 0.08) by factors of 5 to 6, confirming that structural descriptors capture meaningful exosite signatures in coagulation factors. The consistency across two independently evolved serine protease families suggests that exosite architecture in this enzyme class is amenable to computational characterization.

Intermediate families showed partial conservation. Metalloproteinases achieved MCC = 0.26 ± 0.22, with 10.5% exceeding threshold, representing performance intermediate between coagulation factors and unpredictable families. Phosphatases reached MCC = 0.27 ± 0.41, with notably high variance reflecting heterogeneous exosite architecture across the phosphatase superfamily. For these families, a subset of structures yielded acceptable predictions, suggesting that exosite features are partially conserved but more variable than in coagulation proteases.

Table 1
LOPO-CV performance by protein family. Analysis restricted to families with $n \geq 5$ structures for statistical validity. MCC, Matthews correlation coefficient. Values reported as mean ± standard deviation characterizing within-family variation. The Kruskal–Wallis test confirms significant differences across families ($H = 41.6$, $P < 0.001$); post-hoc Dunn's tests identify significant differences between coagulation factors and kinases ($P < 0.01$, Bonferroni-corrected).

Family	*n*	Mean MCC	Mean F1	% MCC ≥ 0.5
Factor X	10	0.47 ± 0.24	0.47 ± 0.25	50.0
Thrombin	31	0.47 ± 0.20	0.45 ± 0.19	51.6
Phosphatase	7	0.27 ± 0.41	0.26 ± 0.38	28.6
Metalloproteinase	19	0.26 ± 0.22	0.23 ± 0.21	10.5
Other	85	0.19 ± 0.22	0.17 ± 0.20	11.8
Kinase	22	0.14 ± 0.27	0.13 ± 0.25	18.2
Caspase	6	0.11 ± 0.13	0.10 ± 0.12	0.0
Overall	180	0.25 ± 0.25	0.23 ± 0.24	21.7

Non-stratified unified model (paired comparison on n = 190 structures): MCC = 0.23 ± 0.18, F1 = 0.20, 11.1% MCC ≥ 0.5 (vs 15.3% with family stratification under identical conditions). The overall 21.7% success rate in the table above reflects fully optimized per-family threshold tuning on 180 eligible structures (see Supplementary Table S2 for per-structure comparison).

Kinases and caspases proved largely unpredictable. Kinases achieved mean MCC = 0.14 ± 0.27, barely exceeding the accessibility baseline, with only 18.2% above threshold. Caspases performed worst at MCC = 0.11 ± 0.13, with no structure achieving MCC ≥ 0.5. These families appear to lack conserved exosite structural signatures amenable to our descriptor set, consistent with the documented diversification of kinase docking mechanisms [6].

Statistical analysis confirmed that family differences are significant (Kruskal–Wallis $H = 41.6$, $P < 0.001$), rejecting the null hypothesis that all families share a common predictability distribution. Post-hoc Dunn's tests with Bonferroni correction identified significant differences between coagulation factors (thrombin, factor X) and kinases ($P < 0.01$). Critically, these results indicate that exosites do not share a common nanoenvironment that can be described as a canonical exosite signature. Rather, our analysis reveals diverse nanoenvironments across protein families, each characterized by distinct residue-level structural descriptors, such that machine learning cannot capture them within a single predictive rule.

To assess whether the improvement from family-specific modeling generalizes across protein families or is driven by a small subset of well-performing targets, we performed a paired comparison of unified versus family-specific models on the same 190 structures (Supplementary Table S2). The overall mean MCC improvement was modest (0.225 to 0.232, paired t-test $P = 0.52$), but the proportion of structures achieving MCC ≥ 0.5 increased from 11.1% (21/190) to 15.3% (29/190). Importantly, the improvement was not driven by a single over-represented target. Factor X showed the largest benefit: its success rate increased from 10.0% (1/10) to 50.0% (5/10) under family-specific routing. Thrombin improved moderately from 35.5% (11/31) to 45.2% (14/31), distributed across multiple structures rather than concentrated in one or two outliers. In contrast, the "Other" category (n = 91), encompassing structurally heterogeneous families, showed no improvement (4.4% to 2.2%), confirming that the benefit of family-specific routing is biologically specific to families with conserved exosite architecture rather than an artifact of model aggregation.

3.3. High-confidence predictions span multiple families

While family-level statistics reveal broad patterns, individual high-confidence predictions provide insight into protein-specific determinants of predictability. Ten structures achieved MCC ≥ 0.7 under

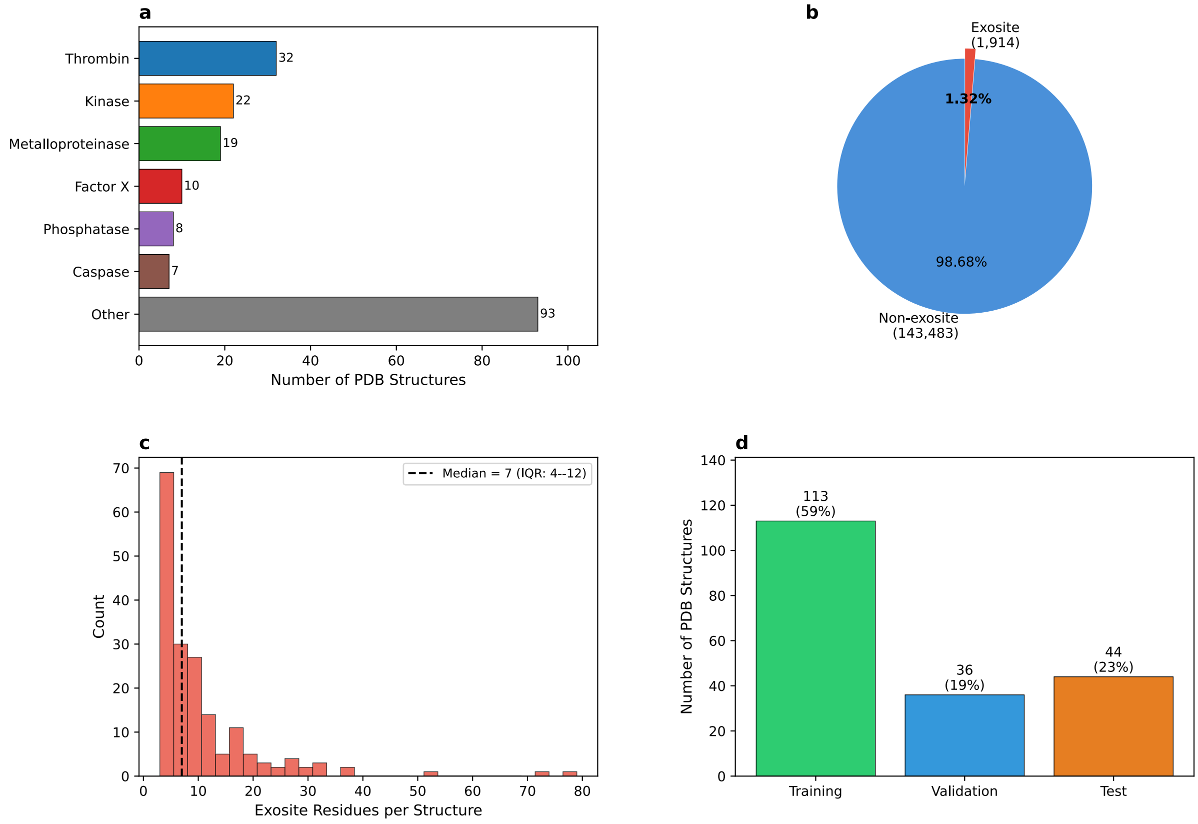

Fig. 1. ExositeDB composition and class distribution. **a**, Distribution of PDB structures across protein families. Coagulation factors (thrombin, factor X) are most represented, reflecting historical research emphasis. The “Other” category comprises 93 structures from families with fewer than 5 representatives, including ADAMTS ($n = 2$) and plasminogen ($n = 1$). **b**, Class composition across the complete dataset. Exosite residues (red, 1.32%, n = 1914) constitute a small minority of total residues (n = 145,397), establishing severe class imbalance as a fundamental challenge. **c**, Distribution of annotated exosite residues per structure. Median = 7 residues (IQR: 4–12). Structures with extensive exosite annotations derive primarily from thrombin (exosites I and II combined). **d**, Data partitioning into training (113 structures, 59%), validation (36 structures, 19%), and test (44 structures, 23%) sets. Family proportions are preserved across partitions within 5% relative deviation.

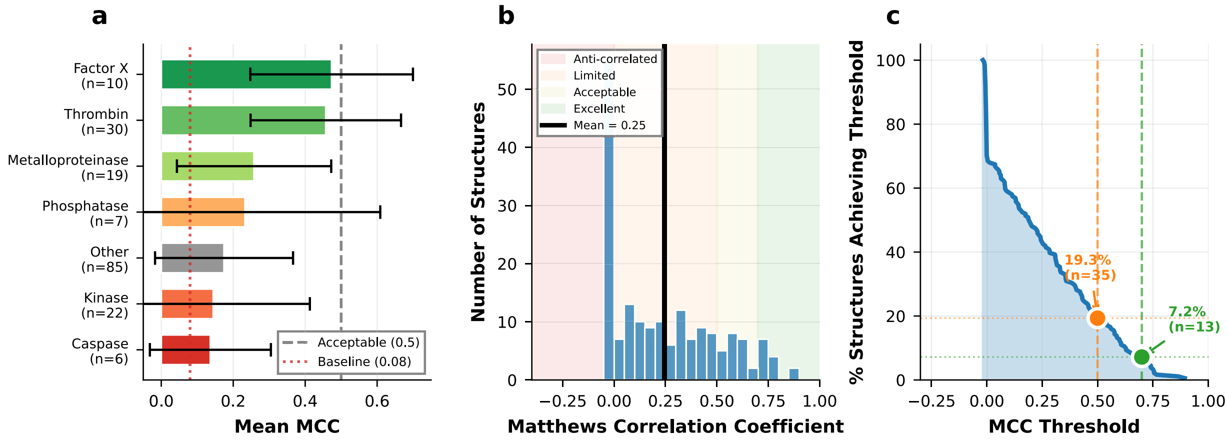

Fig. 2. LOPO-CV reveals family-dependent performance. **a**, Per-family mean MCC with 95% bootstrap confidence intervals. Horizontal dashed lines indicate performance thresholds: MCC = 0.5 (acceptable prediction, gray) and MCC = 0.7 (high confidence, black). Random baseline (MCC = 0) and accessibility baseline (MCC = 0.08) shown as reference. Coagulation factors significantly outperform kinases and caspases (Kruskal–Wallis $P < 0.001$). **b**, Distribution of structure-level MCC values across 180 evaluated proteins. Colors indicate performance categories: red (MCC < 0.5, limited utility), yellow (0.5 ≤ MCC < 0.7, acceptable), green (MCC ≥ 0.7, high confidence). Distribution is right-skewed with mode near zero. **c**, Cumulative distribution function showing percentage of structures exceeding each MCC threshold. Vertical lines indicate key thresholds: 21.7% exceed MCC = 0.5; 5.6% exceed MCC = 0.7.

LOPO-CV, spanning six protein families; these high-confidence predictions showed balanced precision (>0.7) and recall (>0.6) without systematic bias.

The highest performance was achieved by protein phosphatase 1β (PDB: 3EGH, MCC = 0.89), with perfect precision (1.0) and 80% recall. PP1's regulatory subunit binding groove, which engages diverse targeting proteins through a conserved RVxF motif, represents a well-defined exosite with consistent structural features across PP1 isoforms. The model correctly identified 4 of 5 annotated exosite residues while making no false positive predictions.

MAPK/ERK2 (PDB: 2ERK, MCC = 0.86) achieved high performance despite the kinase family's poor aggregate statistics. This apparent contradiction resolves upon structural analysis: ERK2's D-recruitment site, a hydrophobic groove engaging the DEJL motif of substrates and phosphatases, is structurally conserved across the MAPK subfamily and extensively characterized crystallographically. The model's success reflects this conservation; kinase structures with less characterized or more variable docking mechanisms failed prediction.

Coagulation factors achieved consistently high performance, with factor X (1IOE, MCC = 0.77), factor XIa (6I58, MCC = 0.71), and thrombin (8TQS, MCC = 0.75) all exceeding the 0.7 threshold. Additional high-confidence predictions included prolyl oligopeptidase (1QFS, MCC = 0.82), insulin-degrading enzyme (2WBY, MCC = 0.71), and metalloproteinase ADAMTS13 (3GHN, MCC = 0.74).

These results establish an important principle: aggregate family statistics obscure protein-level variation. Accurate prediction remains achievable for specific proteins within families showing poor mean performance, suggesting that individual structural features, rather than merely taxonomic assignment, determine predictability.

3.4. Error analysis identifies structural determinants of prediction failure

To understand why certain proteins resist accurate prediction under LOPO-CV, we systematically analyzed structures with MCC $\leq$ 0.0, where predictions were anti-correlated with ground truth.

Three failure modes emerged. First, proteins with *distributed exosites*, characterized by multiple small patches rather than contiguous surfaces, consistently failed. Our local descriptors, computed within 7 Å radii, cannot capture binding interfaces distributed across the protein surface. The histone deacetylase HDAC8, whose exosite comprises residues spanning 40 Å, exemplifies this pattern.

Second, *conformationally dynamic exosites*, which are surface regions that are buried or disordered in the crystallographic structure but functional in alternative conformations, showed poor performance. Our descriptors characterize static structures; exosites requiring conformational change for accessibility escape detection. Several kinase structures fell into this category, with exosites annotated from solution studies but occluded in crystal structures.

Third, *annotation heterogeneity* affected some failures. Literature exosite definitions vary: some studies designate any protein–protein contact surface as an exosite, while others restrict the term to regulatory sites with demonstrated functional consequences. Structures annotated with the broader definition may include non-regulatory surfaces that lack distinctive structural signatures.

These failure modes suggest specific improvements: integration of sequence conservation to detect distributed functional sites, ensemble analysis of multiple conformations from molecular dynamics or alternative crystal forms, and refinement of exosite definitions through functional annotation standards.

4. Discussion

This study demonstrates that exosite prediction performance varies substantially across protein families, with coagulation factors achieving mean MCC of 0.47 compared to 0.14 for kinases under LOPO-CV evaluation. This order-of-magnitude difference in predictability represents the central empirical finding of our work and has immediate implications for how computational exosite identification should be approached.

The observed family-dependent performance likely reflects differences in structural consistency of exosite architecture within each family. Coagulation proteases share a common serine protease fold with well-characterized exosite regions, including thrombin's electropositive exosite I and heparin-binding exosite II [2]. The structural descriptors we employed, which capture local density, graph topology, and spatial organization within 7 Å of each residue, appear sufficient to distinguish these conserved exosite features from surrounding protein surface. Literature suggests that coagulation proteases diverged early in vertebrate evolution and have remained under strong functional constraint [26], which may explain the structural regularity we observe. In contrast, kinases exhibit diverse docking mechanisms across subfamilies [6], and this structural heterogeneity likely underlies their poor aggregate predictability.

The success of MAPK/ERK2 (MCC = 0.86) within the otherwise unpredictable kinase family illustrates that aggregate family statistics can obscure protein-specific predictability. ERK2's D-recruitment site is well-characterized and structurally conserved across the MAPK subfamily. This suggests that subfamily-specific or hierarchical modeling approaches might extend predictability to protein classes that perform poorly under family-level analysis.

The paired comparison of unified versus family-specific models (Supplementary Table S2) provides nuanced insight into where family stratification adds value. Although the overall mean MCC difference is not statistically significant ($P = 0.52$), the distribution of improvements reveals a biologically coherent pattern. Factor X structures benefit most dramatically, with success rate (MCC $\geq$ 0.5) increasing from 10% to 50%, consistent with the highly conserved exosite architecture shared by factor X family members. Thrombin shows moderate improvement (35.5% to 45.2%), broadly distributed across 14 of 31 structures rather than concentrated in one or two outliers. In contrast, the “Other” category, a heterogeneous collection of families with fewer than 5 representatives, slightly worsens under family-specific routing (4.4% to 2.2%). This pattern confirms that the improvement from family stratification is not an artifact of over-representation or outlier-driven inflation, but reflects genuine structural conservation within specific protein families that family-specific models can exploit.

A deliberate design choice in STINGExoFind was the exclusive use of structure-derived Most Relevant Nanoenvironment Descriptors (MRNDs) from STINGRDB2, without incorporating sequence conservation features. This decision reflects the observation that many exosites, particularly in the “Other” category, represent orphan functional sites with limited homologous examples. Sequence conservation features would bias the model toward well-studied protein families while providing no information for poorly characterized targets. By relying solely on local structural descriptors computed from individual coordinate files, our approach can be applied to any protein with a solved structure, regardless of the availability of homologs or multiple sequence alignments. Future work may explore whether integrating conservation features improves prediction for well-characterized families without compromising performance on orphan targets.

Our findings should be interpreted in context of existing binding site prediction tools. CryptoSite identifies sites requiring conformational change for ligand accessibility (AUROC 0.83) [10], while fpocket detects geometric cavities (70%–85% success) [9]. Exosites differ from both: they are typically constitutively accessible surfaces rather than cryptic sites, and often flat rather than cavity-like. STINGExoFind addresses this distinct prediction problem, building upon prior work using STINGRDB2 descriptors for allosteric site prediction [12].

The practical implications are straightforward. For coagulation factors, where 50% of structures achieve MCC $\geq$ 0.5, computational prediction can meaningfully prioritize candidate exosite residues for experimental validation. For kinases, computational guidance is unreliable and experimental approaches such as hydrogen–deuterium exchange mass spectrometry or systematic mutagenesis remain necessary.

5. Conclusion

This study provides the first systematic evaluation of computational exosite prediction across diverse protein families. Through LOPO-CV assessment of 180 protein structures, we demonstrate that prediction performance varies substantially by protein family: coagulation factors (thrombin, factor X) achieve mean MCC of 0.47 with 50% of structures exceeding MCC ≥ 0.5, while kinases achieve only 0.14. Ten structures across six families exceeded MCC ≥ 0.7, including MAPK/ERK2 (MCC = 0.86), demonstrating that high-confidence predictions remain achievable for specific proteins even within poorly-performing families.

These results establish that exosite prediction should be approached as a family-specific rather than universal problem. For well-characterized families such as coagulation proteases, STINGExoFind provides actionable predictions that can guide experimental validation. For structurally diverse families such as kinases, experimental characterization remains essential. STINGExoFind is provided as a community resource to support future method development and drug discovery applications targeting protein exosites.

CRediT authorship contribution statement

Folorunsho Bright Omage: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ivan Mazoni:** Writing – review & editing, Resources, Data curation. **Inácio Henrique Yano:** Writing – review & editing, Resources, Data curation. **Goran Neshich:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Funding

This research was funded by the São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP) (grant nos. 2023/02691-2, 2025/23708-6).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Folorunsho Bright Omage reports financial support was provided by São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the institutional support provided by EMBRAPA, which granted access to their high-performance computing cluster. This facility was essential for our extensive data processing and machine learning analyses. We also thank our collaborators and reviewers for their valuable feedback, which greatly improved this work.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ailsci.2026.100166.

Data availability

ExositeDB is publicly available at https://www.exosite.cbi.cnptia.embrapa.br/ with RESTful API access for programmatic queries. The database is provided under CC-BY 4.0 license.

STINGExoFind source code, trained models, and analysis scripts are available at https://github.com/omagebright/STINGExoFind under MIT license. The repository includes all scripts for reproducing the complete analysis pipeline, from data preparation through model training and evaluation.

References

[1] Bock PE, Panizzi P, Verhamme IM. Exosites in the substrate specificity of blood coagulation reactions. J Thromb Haemost 2007;5:81–94, Review article on exosites in coagulation.

[2] Huntington JA. Thrombin plasticity. Biochim et Biophys Acta (BBA)-Proteins Proteom 2012;1824:246–52.

[3] Lane DA, Philippou H, Huntington JA. Directing thrombin. Blood 2005;106:2605–12.

[4] Lin H. Substrate-selective small-molecule modulators of enzymes: Mechanisms and opportunities. Curr Opin Chem Biol 2023;72:102231. http://dx.doi.org/10.1016/j.cbpa.2022.102231, review on substrate-selective enzyme modulation including allosteric and exosite mechanisms.

[5] Weitz JI, Eikelboom JW, Samama MM. New antithrombotic drugs: Antithrombotic therapy and prevention of thrombosis, 9th ed: American college of chest physicians evidence-based clinical practice guidelines. Chest 2012;141:e120S–51S. http://dx.doi.org/10.1378/chest.11-2294.

[6] Reményi A, Good MC, Lim WA. Docking interactions in protein kinase and phosphatase networks. Curr Opin Struct Biol 2006;16:676–85.

[7] Krishnan N, Koveal D, Miller DH, Xue B, Akshinthala SD, Kragelj J, Jensen MR, Gaber CM, Bhatta R, Gaber A, et al. Targeting the disordered c terminus of ptp1b with an allosteric inhibitor. Nat Chem Biol. 2014;10:558–66.

[8] Omage FB, Mazoni I, Yano IH, Neshich G. Protein exosites: Structural determinants of specificity and emerging targets for drug discovery. 2026a, Manuscript submitted for publication.

[9] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 2009;10:1–11.

[10] Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Wolber RA, Keiser MJ, Shoichet BK, Sali A. Cryptosite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. J Mol Biol 2016;428:709–19.

[11] Greener JG, Sternberg MJ. Allopred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC Bioinformatics 2015;16:1–7.

[12] Omage FB, Salim JA, Mazoni I, Yano IH, Borro L, Gonzalez JEH, de Moraes FR, Giachetto PF, Tasic L, Arni RK, et al. Protein allosteric site identification using machine learning and per amino acid residue reported internal protein nanoenvironment descriptors. Comput Struct Biotechnol J 2024;23:3907–19.

[13] Omage FB, Salim JA, Mazoni I, Yano IH, Hernández González JE, Giachetto PF, Tasic L, Arni RK, Neshich G. Stingallo: a web server for high-throughput prediction of allosteric site-forming residues using internal protein nanoenvironment descriptors. Brief Bioinform 2025;26:bbaf424.

[14] Neshich G, Mancini AL, Yamagishi ME, Kuser PR, Fileto R, Pinto IP, Palandrani JF, Krauchenco JN, Baudet C, Montagner AJ, Higa RH. Sting millennium: A new suite of programs for a comprehensive analysis of protein structure and sequence. Prot Sci 2003;12:103.

[15] Neshich G, Rocchia W, Mancini AL, Romani-Ber M, Santoro MM, Honig B. Sting millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. Nucleic Acids Res 2006b;34:W209–13.

[16] de Moraes FR, Neshich IA, Mazoni I, Yano IH, Pereira JG, Salim JA, Jardine JG, Neshich G. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. PLoS One 2014;9:e87107.

[17] Mazoni I, Borro LC, Jardine JG, Yano IH, Salim JA, Neshich G. Study of specific nanoenvironments containing α-helices in all-α and $(\alpha+\beta)+(\alpha/\beta)$ proteins. PLoS One 2018;13:e0200018.

[18] Mazoni I, Salim JA, de Moraes FR, Borro L, Neshich G. A comparison between internal protein nanoenvironments of α-helices and β-sheets. PLoS One 2020;15:e0244315.

[19] Mak JK, Bendandi A, Salim JA, Mazoni I, de Moraes FR, Borro L, Störtz F, Rocchia W, Neshich G, Minary P. Learning to utilize internal protein 3d nanoenvironment descriptors in predicting crispr-cas9 off-target activity. NAR Genom Bioinform 2025;7:lqaf054.

[20] Omage FB, Mazoni I, Yano IH, Neshich G. Sting-exositedb: An ai-assisted curated database of protein exosites for drug discovery. 2026b, Manuscript submitted for publication.

[21] Neshich G, Togawa RC, Mancini AL, Kuser PR, Yamagishi ME, Pappas Jr G, Torres WV, Fonseca e Campos T, Ferreira LL, Luna FM, et al. The diamond sting server. Nucleic Acids Res 2005b;33:W29–35.

[22] Neshich G, Borro LC, Higa RH, Kuser PR, Yamagishi ME, Franco EH, Fileto R, Ribeiro AA, Bezerra GB, Velludo TM, et al. Sting_rdb: A relational database of structural parameters for protein analysis with support for data warehousing and data mining. Genet Mol Res 2013;12:4120–46.

[23] Neshich G, Borro LC, Higa RH, Kuser PR, Yamagishi ME, Franco EH, Krauchenco JN, Fileto R, Ribeiro AA, Bezerra GB, et al. Sting report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the sting database. Nucleic Acids Res 2005a;33:D269–74.

[24] Neshich G, Mancini AL, Togawa RC, Kuser PR, Yamagishi ME, Fileto R, Pappas Jr G, Higa RH, Ribeiro AA, Baudet C, et al. The sms structure-function database and comparative structure analysis tool (blue star sting). In Silico Biol 2006a;6:197–207.

[25] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 785–94.

[26] Doolittle RF. The evolution of vertebrate blood coagulation as viewed from a comparison of puffer fish and sea squirt genomes. Phil Trans R Soc B 2009;364:2149–59.