Supporting Information

Systematic Profiling of Encoder Architectures and Multimodal Embeddings in Protein Foundation Models Reveals Dataset-Driven Separability Regimes in Prediction of Allosteric and Orthosteric Binding Sites

Alina Bazarova,[1,2] *Gennady M. Verkhivker*[3,4*]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, Wilhelm-Johnen-Strasse, 52428 Jülich, Germany

[2]Helmholtz AI Munich, Ingolstädter Landstraße 1 85764 Neuherberg, Germany

[3]Keck Center for Science and Engineering, Schmid College of Science and Technology, Chapman University, Orange, CA 92866, United States of America

[4]Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine, CA 92618, United States of America

*Correspondence: verkhivk@chapman.edu; Tel.: +1-714-516-4586 (G.V)

Text S1. Pre-training Architectures

Text S1.1 Structural Encoders

Structural representations were generated using two complementary structural encoders integrated within OneProt: a graph-based structural encoder and a token-based structural encoder. These encoders capture complementary aspects of protein three-dimensional organization across multiple spatial scales.

Graph-based structural representations were obtained using the ProNet architecture, a hierarchical graph neural network that learns protein representations at three complementary levels of structural granularity: amino-acid, backbone, and all-atom. Given a protein structure containing Lamino-acid residues, the protein is represented as a graph $G = (V,E,P)$, where Vdenotes the set of residue nodes, Erepresents the edges describing covalent and spatial relationships between residues, and P contains the structural information associated with the protein. Through hierarchical modeling at the amino-acid, backbone, and all-atom levels, ProNet captures complementary structural features spanning local geometric environments and higher-order protein organization.

At each level, node representations are iteratively refined using message passing according to the general update rule:

$$m_v^{(t)} = \text{AGGREGATE}_{u\in\mathcal{N}(v)}\left(\text{MESSAGE}^{(t)}(h_v^{(t)},h_u^{(t)},e_{uv})\right)$$

$$h_v^{(t+1)} = \text{UPDATE}^{(t)}(h_v^{(t)},m_v^{(t)})$$

where $\mathcal{N}(v)$denotes the neighborhood of node v, e_{uv}represents geometric edge features derived from the structural information contained in P, and MESSAGE, AGGREGATE, and UPDATE are learnable functions at layer t.

Through hierarchical message passing across the amino-acid, backbone, and all-atom representations, ProNet learns structural embeddings that capture fine-grained local interactions together with the higher-order three-dimensional organization of the complete protein fold. Within OneProt, these graph embeddings are projected into the shared multimodal latent space and aligned with sequence, pocket, text, and MD representations during contrastive pretraining.

Token-based structural representations were generated using Foldseek structural tokens processed with an ESM2-35M transformer encoder. Foldseek converts protein structures into sequence-like representations by discretizing local three-dimensional residue environments into a learned structural alphabet. Each structural token encodes the geometric relationship between a residue and its nearest spatial neighbor, capturing local tertiary structural information in a discrete representation.

The resulting structural token sequence:

$$S_{struct} = (t_1, t_2, \ldots, t_L)$$

where each token $t_i \in \mathcal{T}$ belongs to a learned structural alphabet of size $|\mathcal{T}| = 20$, is processed by an ESM2 transformer encoder to produce contextual structural embeddings:

$$H_{struct} = \mathrm{ESM2}(S_{struct}) \in \mathbb{R}^{L \times 1280}$$

In contrast to graph-based structural representations, structural-token embeddings model proteins as sequences of discretized local structural environments, enabling transformer self-attention to capture long-range dependencies and global fold organization while preserving information about local tertiary geometry.

Text S1.2 Pocket Encoder

Pocket embeddings were obtained using the binding-site encoder integrated within the OneProt framework. Binding pockets are represented as three-dimensional residue graphs and processed using a graph neural network based on the ProNet architecture. Graph-based representations are particularly suitable for binding-site modeling because they preserve spatial residue relationships and local geometric organization while remaining invariant to global coordinate transformations.

For each binding site, a binding-pocket center $\mathbf{c} \in \mathbb{R}^3$was defined. For each residue i, the residue centroid was computed as the average Cartesian coordinate of all atoms belonging to that residue:

$$\mathbf{r}_i = \frac{1}{|A_i|} \sum_{a \in A_i} \mathbf{x}_a$$

where $A_i$denotes the set of atoms comprising residue i, and $\mathbf{x}_a \in \mathbb{R}^3$is the Cartesian coordinate of atom a.

Residues were ranked according to their Euclidean distance from the binding-site center:

$$d_i = \| \mathbf{r}_i - \mathbf{c} \|_2$$

and the $K = 100$nearest residues were selected to define the binding pocket.

The resulting pocket was represented as a residue graph:

$$G_{pocket} = (V_{pocket}, E_{pocket}, P_{pocket})$$

where V_{pocket}denotes the set of selected residue nodes, E_{pocket}represents the graph connectivity defined by the ProNet encoder, and P_{pocket}contains the associated three-dimensional structural information.

For each residue, graph node features included the residue identity, backbone atom coordinates (N, Cα, and C), backbone torsion-angle embeddings, and side-chain torsion-angle embeddings. Backbone torsion angles were encoded using sine and cosine transformations:

$$\mathbf{b}_i = [\cos \phi_i, \cos \psi_i, \cos \omega_i, \sin \phi_i, \sin \psi_i, \sin \omega_i]$$

whereas side-chain torsion angles were represented as:

$$\mathbf{s}_i = [\sin \chi_{i,1}, \sin \chi_{i,2}, \sin \chi_{i,3}, \sin \chi_{i,4}, \cos \chi_{i,1}, \cos \chi_{i,2}, \cos \chi_{i,3}, \cos \chi_{i,4}]$$

The pocket graph was then processed by the ProNet graph encoder to obtain a latent structural representation:

$$\mathbf{h}_{pocket} = f_{ProNet}(G_{pocket})$$

where f_{ProNet} denotes the hierarchical graph neural network operating on the residue graph.

Within OneProt, this latent representation was further transformed by a learnable projection head and L2 normalization:

$$\mathbf{z}_{pocket} = \text{Normalize}(\text{Proj}(\mathbf{h}_{pocket}))$$

where Proj denotes the learnable projection module and Normalize represents L2 normalization. The resulting embedding $\mathbf{z}_{pocket}$ constitutes the final pocket representation used for multimodal contrastive alignment.

Text S1.3 Text Encoder

Textual representations were derived from UniProt functional annotations and processed using the BiomedBERT encoder integrated within OneProt model. BiomedBERT is a domain-adapted BERT model pretrained on biomedical literature, achieving improved performance on biomedical natural language processing tasks.

UniProt annotations consist of concatenated text fields including protein names and families, organism information, function descriptions, subcellular localization, and pathway and similarity information. Given a textual annotation $T = (w_1, w_2, \ldots, w_M)$ of length M tokens, the BiomedBERT encoder produces contextualized representations:

$$H_{\text{text}} = \text{BiomedBERT}(T) \in \mathbb{R}^{M \times 768}$$

The final text embedding is obtained through mean pooling across token positions:

$$\mathbf{h}_{\text{text}} = \frac{1}{M} \sum_{i=1}^{M} H_{\text{text}}[i] \in \mathbb{R}^{768}$$

The pooled representation is subsequently projected into the shared OneProt latent space through a trainable projection head:

$$\mathbf{z}_{\text{text}} = P_{\text{text}}(\mathbf{h}_{\text{text}}) \in \mathbb{R}^{1024}$$

where P_{text} denotes the projection network. This modality captures semantic and functional biological information associated with protein activity, molecular function, and biochemical context. Previous OneProt studies demonstrated that inclusion of textual modalities substantially improves multimodal latent-space organization and downstream transfer performance.

Text S1.4 MD Encoder

We used the extended OneProt architecture, which incorporates MD trajectory representations. Because allosteric regulation is fundamentally associated with conformational fluctuations, ensemble redistribution, and long-range dynamical coupling, static experimental structures may incompletely capture the conformational landscapes underlying regulatory behavior. MD trajectories therefore provide complementary information describing protein flexibility and time-dependent structural variability beyond single static conformations.

The MD modality was implemented using the latent MD encoder derived from MDGen, a transformer-based generative framework for MD trajectory modeling. Unlike conventional structure encoders operating on individual conformations, the MDGen encoder processes full MD trajectories as temporally ordered structural ensembles.

Given an input trajectory $X_{MD} = \{X_1, X_2, \ldots, X_T\}$ comprising T simulation frames and L residues, each frame is represented using SE(3)-invariant roto-translational offsets together with backbone and side-chain torsion-angle representations. The encoder produces latent trajectory representations:

$$H_{MD} = f_{MDGen}(X_{MD}) \in \mathbb{R}^{B \times T \times L \times D}$$

where B is the batch size, T is the number of trajectory frames, L is the number of residues, and D is the latent embedding dimension (1024).

Architecturally, the latent MD encoder is based on a Scalable Interpolant Transformer (SIT) framework augmented with long-context Hyena operators replacing conventional temporal attention layers, enabling efficient modeling of long-timescale temporal dependencies across molecular trajectories. Spatial features are processed through transformer-based structural layers,

while temporal evolution is captured through Hyena-based sequence modeling across trajectory frames.

To obtain fixed-length embeddings compatible with the OneProt multimodal framework, temporal pooling across trajectory frames was first applied:

$$\mathbf{h}_{t,l} = \frac{1}{T}\sum_{t=1}^{T} H_{MD}\,[t,l] \in \mathbb{R}^{D}$$

followed by spatial pooling across residues:

$$\mathbf{h}_{MD} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{h}_{t,l} \in \mathbb{R}^{1024}$$

The resulting pooled trajectory representations were subsequently projected into the shared OneProt latent space using modality-specific MLP projection heads, producing 1024-dimensional MD embeddings aligned with the remaining OneProt modalities.

MD trajectories used during pretraining were curated from mdCATH, GPCRmd, and ATLAS databases through the MDRepo repository. These datasets provide complementary coverage of protein families, conformational regimes, and simulation timescales, including membrane proteins, globular proteins, and structurally diverse protein folds. To reduce heterogeneity between simulation sources, trajectories were temporally standardized to a common sampling interval of $\Delta t = 0.1$ns and converted into torsion-angle and rigid-frame representations compatible with the latent MD encoder.

Text S1.5 Cross-Modal Representation Alignment and Transfer

Modality-specific encoders within OneProt are jointly aligned using multimodal contrastive learning based on the InfoNCE objective, conceptually analogous to multimodal representation-learning frameworks such as CLIP and ImageBind. Embeddings derived from different biological modalities are projected into a shared latent space, where alignment is performed pairwise with the sequence modality acting as the central anchor representation.

For a batch of Nprotein samples, let $\mathbf{z}_{\text{seq}}^{(i)}$, $\mathbf{z}^{(i)}_{\text{pocket}}$, $\mathbf{z}^{(i)}_{\text{text}}$, $\mathbf{z}^{(i)}_{\text{struct}}$, and $\mathbf{z}^{(i)}_{\text{MD}}$denote the normalized embeddings for sample $i$from the sequence, pocket, text, structural, and MD modalities, respectively.

The contrastive loss for aligning sequence and pocket modalities is defined as:

$$\mathcal{L}_{\text{seq-pocket}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{z}_{\text{seq}}^{(i)} \cdot \mathbf{z}^{(i)}_{\text{pocket}}/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{z}_{\text{seq}}^{(i)} \cdot \mathbf{z}^{(j)}_{\text{pocket}}/\tau\right)}$$

where $\tau > 0$is a temperature parameter controlling the sharpness of the softmax distribution, and the dot product $\mathbf{z}_{\text{seq}}^{(i)} \cdot \mathbf{z}^{(i)}_{\text{pocket}}$measures the cosine similarity between normalized embeddings.

The total multimodal alignment loss is the sum over all modality pairs anchored to the sequence modality:

$$\mathcal{L}_{\text{total}} = \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{seq-}m}$$

where $\mathcal{M} = \{\text{pocket,text,struct,MD}\}$.

This framework enables transfer of structural, functional, and dynamical information across heterogeneous biological modalities. During pretraining, all modality-specific encoders are jointly

optimized to minimize $\mathcal{L}_{\text{total}}$, resulting in a shared latent space where embeddings from different modalities corresponding to the same protein are pulled together while those from different proteins are pushed apart.

Previous OneProt studies demonstrated that multimodally aligned representations substantially improve downstream transfer performance compared to their unimodal counterparts, indicating that biological information is propagated between modalities during pretraining. Consequently, modality-specific embeddings within OneProt represent enriched multimodal representations rather than isolated unimodal encodings.

The extent and nature of this cross-modal knowledge transfer depend on the specific encoder configuration participating in multimodal alignment. Previous OneProt studies and subsequent extensions incorporating MD representations demonstrated that different modality combinations produce distinct downstream transfer characteristics and latent-space organization properties. In the present study, these encoder-level ablations therefore provide a framework for probing how structural, functional, and dynamical modalities influence representations associated with allosteric regulation.

Although downstream classifiers operate only on extracted sequence, pocket, and text embeddings, additional modalities such as graph-based structural representations, structural tokens, and MD trajectories may still influence downstream prediction indirectly through their contribution to latent-space organization during multimodal pretraining.

Text S2. Pocket Construction and Structural Processing

Binding pocket representations were constructed following the preprocessing and multimodal representation framework used in the OneProt protein foundation model. OneProt aligns heterogeneous biological modalities—including protein sequence, three-dimensional structure, binding pockets, textual annotations, and MD trajectories—within a shared latent embedding space through modality-specific encoders and multimodal contrastive learning.

Text S2.1 Pocket Construction for PPI-Site Dataset

For the PPI-Site dataset, binding pockets were reconstructed from the cavity annotations provided in the original resource. Annotated ligands were first identified within the corresponding Protein Data Bank (PDB) structures, and the geometric centroid of the ligand atoms was used to define the pocket center:

$$\mathbf{c}_{\text{pocket}} = \frac{1}{N_{\text{lig}}} \sum_{i=1}^{N_{\text{lig}}} \mathbf{x}_i^{\text{lig}}$$

where $\{\mathbf{x}_i^{\text{lig}}\}_{i=1}^{N_{\text{lig}}} \subset \mathbb{R}^3$ are the coordinates of ligand atoms.

For the associated protein chain, residue centroids were computed as the average coordinates of all atoms belonging to each standard amino-acid residue:

$$\mathbf{c}_r = \frac{1}{N_r} \sum_{a \in r} \mathbf{x}_a$$

where N_r is the number of atoms in residue r.

Residues were subsequently ranked according to the Euclidean distance between their centroids and the ligand center:

$$d(r,\mathbf{c}_{\text{pocket}}) = \| \mathbf{c}_r - \mathbf{c}_{\text{pocket}} \|_2$$

and the $K = 100$nearest residues were retained to construct fixed-size pocket representations:

$$\mathcal{R}_{\text{pocket}} = \{r_{(1)}, r_{(2)}, \ldots, r_{(K)}\}$$

where $r_{(1)}, r_{(2)}, \ldots$denote residues sorted by increasing distance to the pocket center. All atoms belonging to the selected residues were preserved, yielding atom-level pocket structures together with their associated residue sequences. The complete protein-chain sequence was stored separately for sequence-based representation extraction.

Text S2.2 Pocket Overlap and Conformational Variability Analysis

Because the PPI-Site dataset contains structurally resolved pocket annotations spanning both allosteric and orthosteric competitive binding regimes within shared protein contexts, it enables direct comparison of local binding environments across regulatory classes. Equivalent analyses were not performed on the AlloDiverse dataset because ASD pockets are reconstructed from residue-level annotations rather than experimentally resolved cavity definitions, preventing direct structural comparison of matched pocket pairs. Similarly, the KinSite dataset predominantly reflects highly conserved ATP-binding environments under strong class imbalance, limiting the interpretability of residue-overlap analyses across regulatory states.

To assess the extent of structural overlap between allosteric and orthosteric competitive binding environments in PPI-Site, we compared pockets across both classes for proteins sharing the same protein and chain identifiers. Comparisons were performed at both the residue-composition level and the atom-coordinate level.

Text S2.3 Residue-Level Overlap Analysis

Residue-level overlap was quantified using the Jaccard similarity coefficient. Given residue sets A and Bcorresponding to two pockets, the Jaccard similarity was defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$denotes the number of shared residues and $|A \cup B|$denotes the total number of unique residues across both pockets.

In addition to Jaccard similarity, we computed the overlap coefficient:

$$O(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$$

which normalizes the shared residue count by the size of the smaller pocket and is therefore more sensitive to partial containment relationships between pockets of different sizes.

Text S2.4 Coordinate-Level Geometric Similarity

To assess coordinate-level geometric similarity, atom-level nearest-neighbor statistics were computed using KD-tree nearest-neighbor search. Given atomic coordinate sets $X = \{\mathbf{x}_i\}_{i=1}^{N}$and $Y = \{\mathbf{y}_i\}_{i=1}^{M}$corresponding to two pockets, the fraction of atoms in pocket Xlocated within a distance threshold $\tau$of any atom in pocket $Y$was defined as:

$$f_{X \to Y} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\min_j \| \mathbf{x}_i - \mathbf{y}_j \|_2 < \tau\right)$$

where $\mathbf{1}(\cdot)$denotes the indicator function and $\tau = 4$Å was used as the spatial proximity threshold.

Continuous geometric similarity was further quantified using the mean nearest-neighbor atom distance:

$$d_{X \to Y} = \frac{1}{N} \sum_{i=1}^{N} \min_j \| \mathbf{x}_i - \mathbf{y}_j \|_2$$

Using this procedure, we identified 100 pocket pairs (<5% of the dataset) exhibiting identical residue composition ($J = 1.0$), indicating that the same 100 residues were selected for both allosteric and orthosteric competitive pockets within the same protein structure. These cases arise when proteins are crystallized in distinct ligand-bound regulatory states, such that nearest-residue pocket extraction yields identical residue membership despite differing ligand identities and binding regimes.

Despite complete residue-level overlap, the corresponding pockets frequently exhibited substantial coordinate-level geometric divergence. Mean nearest-neighbor atom distances ranged from <0.1 Å to >18 Å, indicating that structurally similar residue environments can adopt markedly different conformations depending on ligand-binding state. These observations suggest that allosteric and orthosteric regulation may, in some cases, emerge from distinct conformational organizations within partially shared local structural environments rather than from entirely disjoint residue compositions. Given the relatively small number of such cases and their substantial coordinate-level variability, these instances were retained in the dataset.

Text S2.5 Pocket Construction for ASD Dataset

For the positive instances from ASD dataset, representative allosteric regions were reconstructed from residue-level annotations using spatial clustering and geometric pocket reconstruction. Residues annotated as allosteric were first mapped onto experimentally resolved protein structures using their residue identifiers. Each residue was represented by the three-dimensional coordinates of its Cα atom.

Spatially contiguous allosteric regions were identified using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which groups residues according to local spatial density without requiring a predefined number of clusters. DBSCAN was selected because it naturally accommodates disconnected and sparsely annotated allosteric regions frequently encountered in residue-level annotations.

Given a set of residue coordinates $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^3$, two residues $\mathbf{x}_i$and $\mathbf{x}_j$are considered neighbors if their Euclidean distance satisfies:

$$\| \mathbf{x}_i - \mathbf{x}_j \|_2 < \varepsilon$$

where εdefines the neighborhood radius. In this study we used $\varepsilon = 10$Å and minimum number of samples minPts = 1. The selected radius allows grouping of residues forming spatially contiguous allosteric regions while remaining permissive to structurally sparse annotations frequently observed in experimentally characterized allosteric sites. Setting minPts = 1 ensures that isolated annotated residues are retained as valid singleton clusters rather than discarded as noise, accommodating sparse residue-level annotations frequently encountered in experimentally characterized allosteric regions.

When multiple disconnected clusters were identified within the same protein structure, the largest cluster was retained as the representative allosteric region. The pocket center $\mathbf{c}$was subsequently defined as the centroid of the cluster coordinates:

$$\mathbf{c} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

where Ndenotes the number of residues in the selected cluster. Residues were then ranked according to the Euclidean distance between their centroids and the reconstructed pocket center,

and the $K = 100$nearest residues were selected to define the final fixed-size pocket representation following the same procedure described for the PPI-Site dataset.

S2.6 Harmonized Ligand-Centred Pocket Construction for the ASD Dataset

To assess whether the distinct pocket-construction procedure used for ASD contributed to the high classification performance observed for AlloDiverse, we generated an alternative set of ASD allosteric pockets using a ligand-centred construction analogous to that applied to the PPI-Site and KinSite datasets. This analysis was designed as a sensitivity analysis of the pocket definition rather than as a replacement for the primary ASD dataset. The primary ASD construction retains the curated residue-level annotations as the basis for defining the allosteric region, whereas the harmonized construction uses experimentally resolved allosteric ligands to define the pocket center.

For each ASD entry from section S2.2, the experimentally resolved allosteric ligand corresponding to the curated allosteric-site annotation was identified in the associated PDB structure. Candidate ligand-centered pockets were generated for the annotated protein chain using the geometric centroid of the ligand atoms as the pocket center,

$$\mathbf{c}_{lig} = \frac{1}{N_{lig}} \sum_{i=1}^{N} \mathbf{x}_i$$

where N_{lig} is the number of atoms belonging to the ligand and denotes the coordinate vector of ligand atom $\mathbf{c}_{lig}$. Residue centroids were calculated from the atomic coordinates of each standard amino-acid residue, and residues were ranked according to their Euclidean distance from . The 100 nearest residues were retained, together with all atoms belonging to these residues, yielding fixed-

size pocket representations consistent with the ligand-centred construction used for the other datasets.

Because a PDB structure may contain multiple ligands or ligand-associated pockets, an additional matching step was performed to ensure that the reconstructed ligand-centered pocket corresponded to the same allosteric region represented by the original ASD residue annotation. Candidate pockets were restricted to the same protein and chain as the original ASD-derived pocket. For each candidate, the Euclidean distance between the centroid of the ligand-centered pocket $\boldsymbol{c}_{lig}$ and the centroid of the corresponding annotation-derived ASD pocket $\mathbf{c}_{ASD}$ was calculated,

$$d = ||\boldsymbol{c}_{lig} - \mathbf{c}_{ASD}||_2$$

When multiple candidate ligand-centered pockets were available for the same protein and chain, the candidate with the minimum centroid distance was retained. Matches more than 5 Å apart were excluded to reduce the possibility of assigning a spatially distinct ligand-binding site to the curated ASD annotation. Entries were also excluded when the corresponding experimentally resolved allosteric ligand could not be identified or when an appropriate ligand-centered pocket could not be reconstructed.

Application of this procedure reduced the number of ASD-derived allosteric pocket instances from 14043 to 11279. The resulting harmonized ASD set was subsequently combined with the PPI-Site and KinSite instances using the same dataset-construction and splitting framework as in the primary AlloDiverse analysis. The resulting harmonized AlloDiverse dataset contained 15116 train 1672 validation and 1555 test binding-site instances, Table S1.

The harmonized dataset is smaller and compositionally different from the primary AlloDiverse dataset because only ASD annotations for which a corresponding experimentally resolved ligand-

centered pocket could be confidently recovered were retained. Consequently, the harmonized analysis was treated as a sensitivity analysis of pocket-construction methodology rather than as a directly matched replication of the primary AlloDiverse benchmark.

Text S3. Downstream Allostery Classification Framework

Text S3.1 Hyperparameter Optimization of Downstream Classifiers

For downstream allostery classification, embeddings were first extracted from the frozen OneProt backbone using the selected combinations of pocket, sequence, and text embeddings. These embeddings were subsequently used to train Multi-Layer Perceptron (MLP) classifiers for discrimination between allosteric and orthosteric competitive binding sites.

To optimize classifier performance, we performed a systematic hyperparameter sweep over both architectural and training parameters. The evaluated search space comprised learning rates (0.001, 0.01), batch sizes (32, 64), maximum epochs (50), hidden-layer configurations ([256], [512, 256]), dropout rates (0.10, 0.25), layer normalization (enabled, disabled), activation functions (ReLU, GELU), and residual connections (enabled, disabled).

For each dataset, encoder architecture, embedding combination, and random seed, all hyperparameter combinations were evaluated. Model selection was performed using validation-set ROC-AUC, and the best-performing configuration was subsequently used for test-set evaluation. Early stopping was employed to reduce overfitting, with training terminated when validation performance ceased to improve.

This strategy enabled efficient exploration of a diverse range of classifier architectures while maintaining the OneProt backbone frozen. Consequently, downstream performance primarily reflects information encoded within the pretrained multimodal representations rather than

extensive task-specific adaptation of the foundation model. The use of precomputed embeddings substantially reduced computational cost relative to end-to-end fine-tuning, allowing systematic evaluation across multiple datasets, encoder configurations, embedding combinations, and random seeds.

Text S3.2 Performance Metrics

Performance was evaluated using the following metrics.

ROC-AUC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to rank positive instances higher than negative instances, integrated over all possible classification thresholds:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t)$$

where $\text{TPR}(t) = \text{TP}(t)/(\text{TP}(t) + \text{FN}(t))$is the true positive rate at threshold t, and $\text{FPR}(t) = \text{FP}(t)/(\text{FP}(t) + \text{TN}(t))$is the false positive rate.

PR-AUC (Area Under the Precision-Recall Curve) is particularly informative for imbalanced datasets, measuring the trade-off between precision $\text{Precision}(t) = \text{TP}(t)/(\text{TP}(t) + \text{FP}(t))$and recall $\text{Recall}(t) = \text{TP}(t)/(\text{TP}(t) + \text{FN}(t))$:

$$\text{PR-AUC} = \int_0^1 \text{Precision}(r)\, dr$$

TPR (True Positive Rate) and TNR (True Negative Rate) are class-specific accuracy measures:

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$$

which is the proportion of allosteric sites correctly identified, and

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$$

which is the proportion of orthosteric sites correctly identified.

Precision quantifies the fraction of predicted allosteric sites that are truly allosteric / (TP + FP)

Accuracy is the fraction of all correctly classified examples

$$(\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$$

F1 score is the harmonic mean of precision and recall and provides a threshold-dependent summary of positive-class classification performance

$$\mathrm{F1} = 2 \times (\mathrm{Precision} \times \mathrm{Recall}) / (\mathrm{Precision} + \mathrm{Recall})$$

Matthews Correlation Coefficient (MCC) summarizes binary classification performance while accounting for all four entries of the confusion matrix and is particularly useful under class imbalance:

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / \sqrt{[(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})]}$$

MCC ranges from −1 to 1, where 1 indicates perfect classification, 0 indicates no better than chance-level association between predictions and labels, and −1 indicates complete disagreement.

The probabilistic prediction quality as the mean squared difference between predicted positive-class probabilities and the observed binary outcomes:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

where p_i is the predicted probability for sample i and $y_i \in \{0,1\}$ is the corresponding class label. Lower Brier scores indicate more accurate probabilistic predictions. Because the Brier score

reflects both discrimination and calibration, it is interpreted here as a measure of overall probabilistic prediction quality rather than as a standalone formal calibration analysis.

Prevalence-only Brier baseline was calculated by assigning every test example the observed positive-class

$$p_{i,0} = \pi = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$Brier_0 = \frac{1}{N}\sum_{i=1}^{N} (\pi - y_i)^2$$

A model Brier score below this prevalence-only value indicates probabilistic predictions that are more informative than predicting the test-set prevalence for every example.

For metrics evaluated across 10 repeated random seeds or independently shuffled cross-validation partitions, 95% confidence intervals of the mean were calculated using the Student *t*-distribution:

$$95\%\ \mathrm{CI} = \bar{x} \pm t_{0.975,\,n-1} \times s / \sqrt{n}$$

where $\bar{x}$ is the mean metric value across runs, s is the sample standard deviation, and n is the number of runs. The predefined held-out test set remained fixed across runs; consequently, these confidence intervals quantify variability associated with repeated model training or model-selection seeds rather than sampling uncertainty over test observations.

Together, ROC-AUC and AUPR characterize threshold-independent ranking performance, whereas precision, recall, specificity, accuracy, F1, and MCC characterize threshold-dependent classification behavior. Brier scores provide a complementary assessment of probabilistic

prediction quality. This combination of metrics was used to distinguish ranking performance from operational positive-class recovery, particularly in the strongly imbalanced KinSite dataset.

Text S3.3 Class Imbalance

For the AlloDiverse dataset, positive samples substantially outnumbered negative samples ($N_{\text{allo}} = 15{,}904$, $N_{\text{ortho}} = 5{,}203$). To mitigate training bias introduced by class imbalance, positive instances were randomly subsampled during each training iteration to maintain balanced mini-batches while leaving validation and test distributions unchanged. Specifically, for each mini-batch, we sampled $B/2$positive and $B/2$ negative instances uniformly at random from the training set, where $B$is the batch size. This balanced sampling strategy ensured that the classifier received equal exposure to both classes during training, preventing the model from simply learning the prior class distribution.

S3.4 Within-Protein Pocket-Level Control Analysis

A total of 10 proteins in the held-out test set contained at least one allosteric and at least one orthosteric pocket. For each such protein p, every allosteric pocket was paired with every orthosteric pocket belonging to the same protein. This resulted in 47 allosteric–orthosteric within-protein pocket pairs. No cross-protein comparisons were included in this analysis.

For each trained classifier, we obtained the predicted probability of the allosteric class for every pocket. For an allosteric pocket and orthosteric pocket belonging to the same protein, a pairwise comparison was considered correctly ranked when

$$P(y = 1|x_i) > P(y = 1|x_j)$$

Where $P(y = 1|x)$ denotes the classifier-predicted probability of the allosteric class. A comparison was assigned a score of 1 when the allosteric pocket received the higher probability, 0 when the orthosteric pocket received the higher probability, and 0.5 in the case of an exact tie.

For each protein , the within-protein ranking accuracy was calculated as

$$A_p = \frac{1}{n_p^+ n_p^-} \sum_{i \in P_{p,}^+, j \in P_p^-} s(i,j)$$

Where P_p^+ and P_p^- denote the sets of allosteric and orthosteric pockets, respectively, n_p^+ and n_p^- are their corresponding numbers, and *s(i,j)* is the pairwise score defined above. An accuracy of 0.5 therefore corresponds to chance-level ranking, whereas a value of 1 indicates that every allosteric pocket is assigned a higher allosteric probability than every orthosteric pocket from the same protein.

The analysis was conducted independently for each random seed and encoder–embedding configuration using predictions from the corresponding held-out test set. Summary statistics were subsequently calculated across seeds. Importantly, the analysis does not retrain or fine-tune the classifiers on the subset of proteins used for the within-protein comparison; it evaluates the predictions of the classifiers trained using the same procedure as in the primary classification experiments.

S3.5 Handcrafted Feature Baselines

We evaluated simple handcrafted baselines to determine whether classification could be achieved from coarse sequence or pocket characteristics alone. For the amino-acid-composition baseline, each protein was represented by the relative frequencies of the 20 standard amino acids in the full

protein sequence. For the pocket-geometry baseline, each pocket was represented by pairwise distances between pocket residues, using the mean, median, standard deviation, and maximum pairwise distance as global measures of pocket compactness and spatial extent. Because these representations are low-dimensional, we used L1-regularized logistic regression rather than the MLP architecture utilized for the high-dimensional OneProt embeddings. Prior to model fitting, all input features were standardized to zero mean and unit variance using parameters estimated from the training data. The regularization strength was controlled through the inverse penalty parameter (C). Candidate values of (C) were sampled on a logarithmic grid spanning (10^{-4} to 10^{4}) using 20 values. The optimal (C) was selected exclusively from the training data by 5-fold cross-validation, using mean ROC-AUC across the cross-validation folds as the selection criterion. After selection, the corresponding L1-regularized logistic regression model was refitted on the full training set and evaluated once on the predefined test set. No test or original validation samples were used during hyperparameter selection or fitting the model, in order to avoid data leakage.

For each residue *i*, a residue centroid $c_i \in R^3$ was calculated as the arithmetic mean of the Cartesian coordinates of all atoms assigned to that residue,

$$\mathbf{c}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \mathbf{x}_{ik}$$

where N_i is the number of atoms in residue *i* and $\mathbf{x}_{ik}$ denotes the coordinate vector of atom *k*.

For a pocket containing *N* residues, let d_{ij} denote the pairwise distance between residues *i* and *j*,

$$d_{ij} = ||\boldsymbol{c}_i - \mathbf{c}_j||_2, i < j$$

with $i<j$, giving $M = N(N-1)/2$ pairwise distances. The four pocket-geometry descriptors were calculated as

$$\bar{d} = \frac{1}{M}\sum_{i<j} d_{ij}$$

$$d_{med} = \text{median}_{i<j}(d_{ij})$$

$$\sigma_d = \sqrt{\frac{1}{M}\sum_{i<j}\left(d_{ij} - \bar{d}\right)^2}$$

$$d_{max} = \max_{i<j} d_{ij}$$

Amino-acid composition was represented by the fractional abundance of each of the 20 standard amino acids. For amino-acid type , its fraction was calculated as

$$f_a = \frac{n_a}{N}$$

where n_a is the number of residues of type *a* and *N* is the total number of residues in the pocket or protein for the pocket-level and whole-protein composition baselines, respectively. Thus, the pocket baseline comprised 24 features (four geometric and 20 compositional descriptors), whereas the whole-protein composition baseline comprised 20 features.

Before classification, features were standardized to zero mean and unit variance. Standardization was included within the model-fitting pipeline such that scaling parameters were estimated exclusively from the respective training data. Classification was performed using logistic regression with an L1 penalty and the liblinear solver (max_iter = 5000).

The inverse regularization strength was optimized over 20 logarithmically spaced values between 10^{-4} and 10^{4}. Model selection was performed exclusively on the training data using 5-fold cross-validation, with mean ROC-AUC across folds as the optimization criterion. The value of yielding the highest mean cross-validation ROC-AUC was selected, and the model was subsequently refit on the complete training set before evaluation on the predefined held-out test set.

For analyses involving merged datasets, the predefined training partitions from the individual source datasets were concatenated prior to model fitting, and their corresponding test partitions were concatenated independently to construct the held-out test set.

Performance was strongly dataset dependent: AUROC was 0.59 ± 0.009 (95% CI 0.583–0.596) on PPI-Site, 0.763 ± 0.004 (0.760–0.766) on KinSite, 0.720 ± 0.002 (0.719–0.721) on Dual-Site, and 0.782 on AlloDiverse. The particularly strong KinSite result demonstrates that coarse amino-acid composition and/or pocket geometry already encode substantial class-associated information in this restricted kinase setting. Consistent with its severe class imbalance, the corresponding prevalence-sensitive metrics provided a more moderate assessment (AUPR 0.252 [0.250–0.254], precision 0.131, recall 0.384, MCC 0.171). In contrast, the baseline remained substantially below sequence-only and the strongest pocket-containing representations in Dual-Site and below both the source-only baseline and strongest learned representations in AlloDiverse, indicating that coarse composition and geometry alone do not reproduce their discrimination.

Text S4. Implementation Details

All experiments were implemented using Python 3.12 with PyTorch 2.7.0 as the deep learning framework. The OneProt multimodal backbone was obtained from the official repository and used with frozen pretrained weights. All MLP classifiers were implemented using PyTorch interface.

Computational resources consisted of NVIDIA A100 GPUs with 40GB memory for model inference and training. The total computational cost for all experiments, including the ten independent runs per configuration, was approximately 2,000 GPU-hours.

Code, dataset splits, and pretrained model weights will be made available upon publication to ensure reproducibility of all results.

Supporting Figures

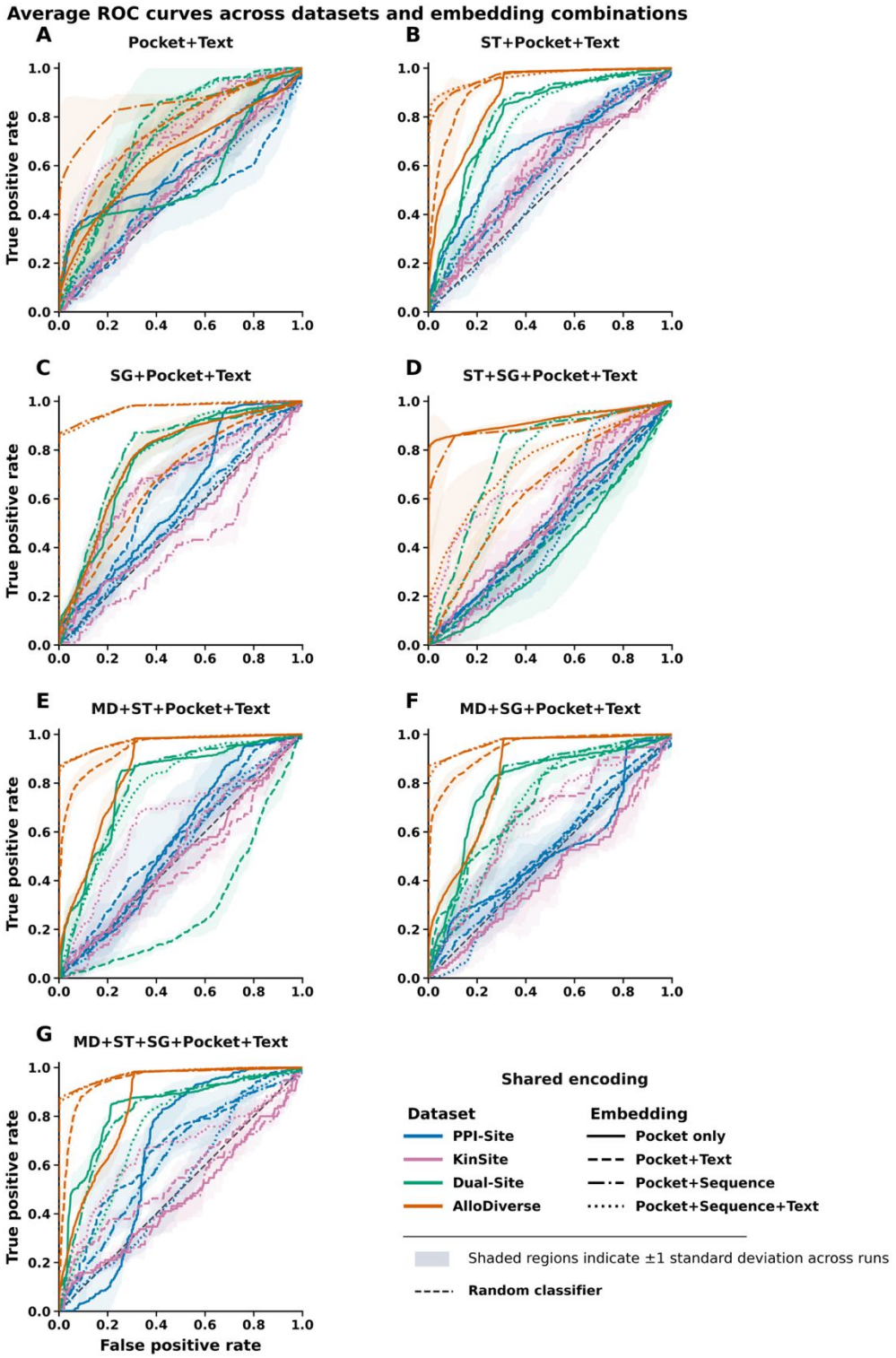


Figure S1. Dataset-dependent ROC-AUC performance across OneProt encoder architectures and embedding combinations. Receiver operating characteristic (ROC) curves illustrating the performance of seven OneProt encoder architectures (individual panels) evaluated across four allosteric discrimination datasets representing distinct separability regimes. Colors indicate datasets: PPI-Site (blue), KinSite (purple), Dual-Site (green), and AlloDiverse (red). Line styles denote the embedding combinations used for downstream classification: Pocket+Text (solid), Pocket+Sequence (dashed), and Pocket+Sequence+Text (dotted). Curves represent averages across independent runs, and shaded regions indicate ±1 standard deviation. Across all architectures, a consistent progression from low- to high-separability regimes is observed. PPI-Site remains close to random performance regardless of architecture or embedding composition. KinSite exhibits moderate separability and substantial variability. Dual-Site shows markedly improved discrimination, particularly for architectures incorporating structural and molecular-dynamics information. AlloDiverse consistently achieves near-ceiling ROC-AUC values across architectures. These results confirm that dataset composition exerts a substantially greater influence on allosteric discrimination performance than architectural variation alone.

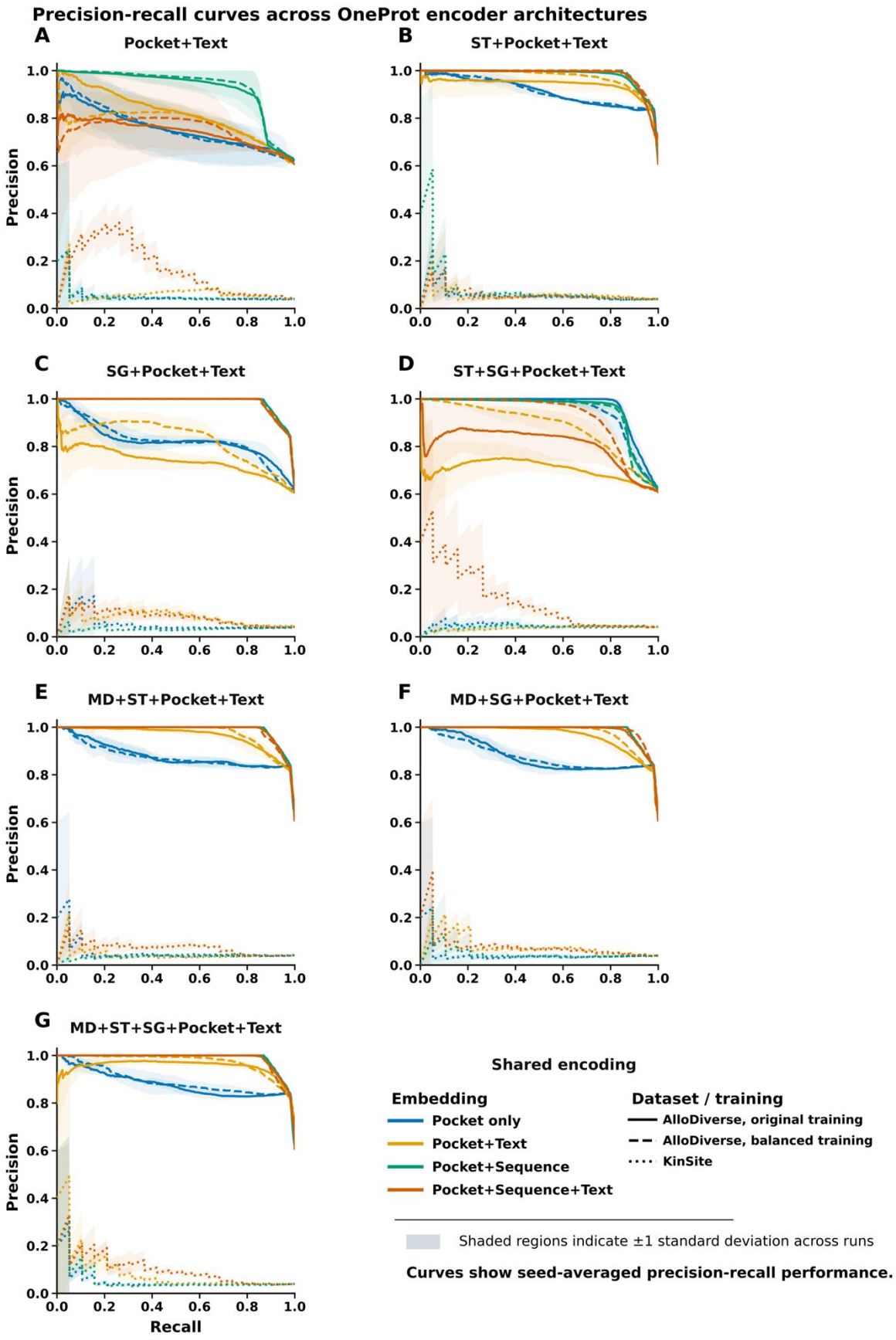

Precision-recall curves across OneProt encoder architectures
A
Pocket+Text
B
ST+Pocket+Text
C
SG+Pocket+Text
D
ST+SG+Pocket+Text
E
MD+ST+Pocket+Text
F
MD+SG+Pocket+Text
G
MD+ST+SG+Pocket+Text
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Shared encoding
Embedding
Pocket only
Pocket+Text
Pocket+Sequence
Pocket+Sequence+Text
Dataset / training
AlloDiverse, original training
AlloDiverse, balanced training
KinSite
Shaded regions indicate ±1 standard deviation across runs
Curves show seed-averaged precision-recall performance.

Figure S2. Precision–recall performance across OneProt encoder architectures under different class-imbalance regimes. Precision–recall (PR) curves for seven OneProt encoder architectures (individual panels) evaluated using three embedding combinations: Pocket+Text (blue), Pocket+Sequence (green), and Pocket+Sequence+Text (red). The x-axis within each panel corresponds to the seven OneProt encoder architectures: Pocket+Text (baseline), ST+Pocket+Text (structure-token), SG+Pocket+Text (structure-graph), MD+ST+Pocket+Text (MD + structure-token), MD+SG+Pocket+Text (MD + structure-graph), ST+SG+Pocket+Text (both structural encoders), and MD+ST+SG+Pocket+Text (full multimodal). Line style denotes dataset or training condition: AlloDiverse original training (solid lines), AlloDiverse balanced training (dashed lines), and KinSite (dotted lines). Curves represent averages across independent runs, and shaded regions indicate ±1 standard deviation. The figure highlights how model behavior changes under different class-imbalance and separability conditions. In the highly separable AlloDiverse regime, most architectures achieve near-perfect precision across a broad range of recall values, particularly when sequence-containing embeddings are included. In contrast, the strongly imbalanced KinSite dataset exhibits substantially lower precision–recall performance, with pronounced sensitivity to embedding composition and architecture. Sequence-containing embeddings (green and red) generally maintain higher precision at increasing recall levels, whereas Pocket+Text (blue) shows more rapid performance degradation. Architectures incorporating structural and molecular-dynamics modalities frequently preserve higher precision over a wider recall range, suggesting that structural and dynamical information contributes complementary signal for recovering rare allosteric sites. These results underscore the importance of evaluating allostery prediction using metrics sensitive to positive-class recovery and further support the existence of distinct dataset-dependent learnability regimes.

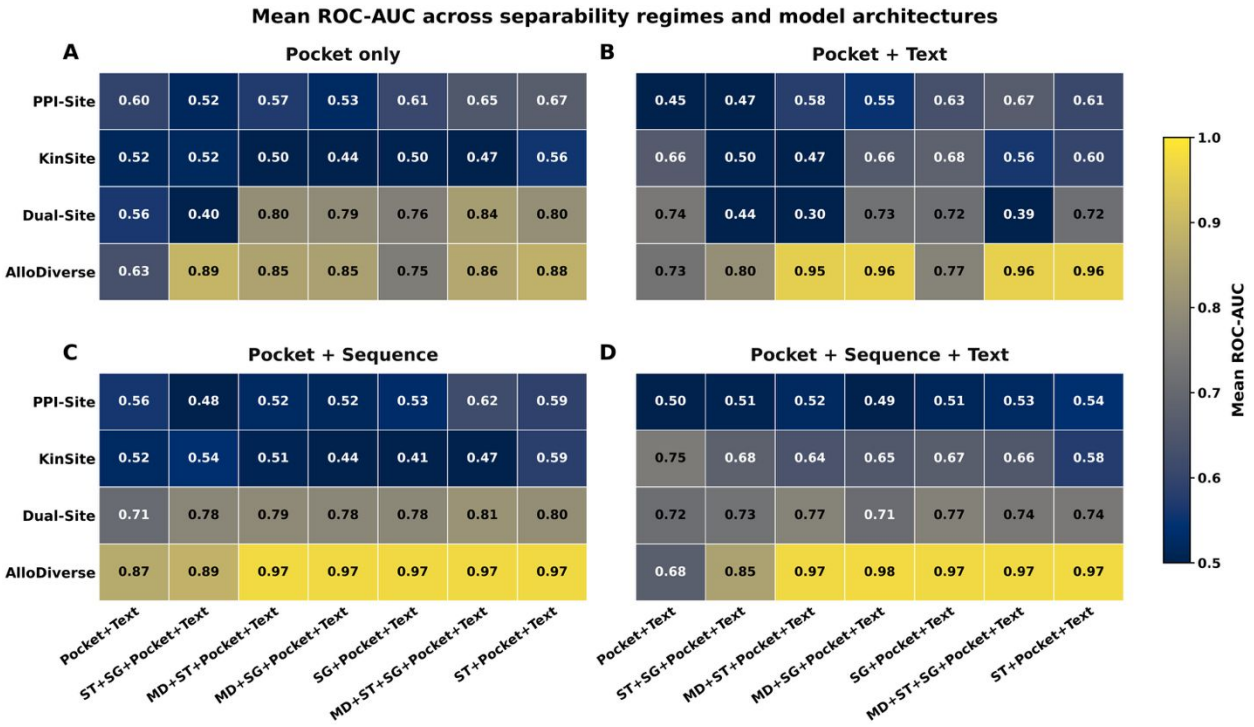


Figure S3. Mean ROC-AUC across separability regimes, embedding combinations, and OneProt encoder architectures. Heatmaps summarize the mean ROC-AUC values averaged across independent runs for all combinations of dataset regime (rows) and encoder architecture (columns). Panels correspond to the four downstream embedding combinations: (A) Pocket only, (B) Pocket+Text, (C) Pocket+Sequence, and (D) Pocket+Sequence+Text. Rows represent the four datasets arranged according to increasing separability: PPI-Site, KinSite, Dual-Site, and AlloDiverse. Columns correspond to the seven evaluated OneProt encoder architectures. Cell colors indicate mean ROC-AUC, with numerical values shown in each cell. Performance is primarily determined by dataset regime rather than architecture, with a clear progression from low performance in PPI-Site to near-ceiling performance in AlloDiverse.

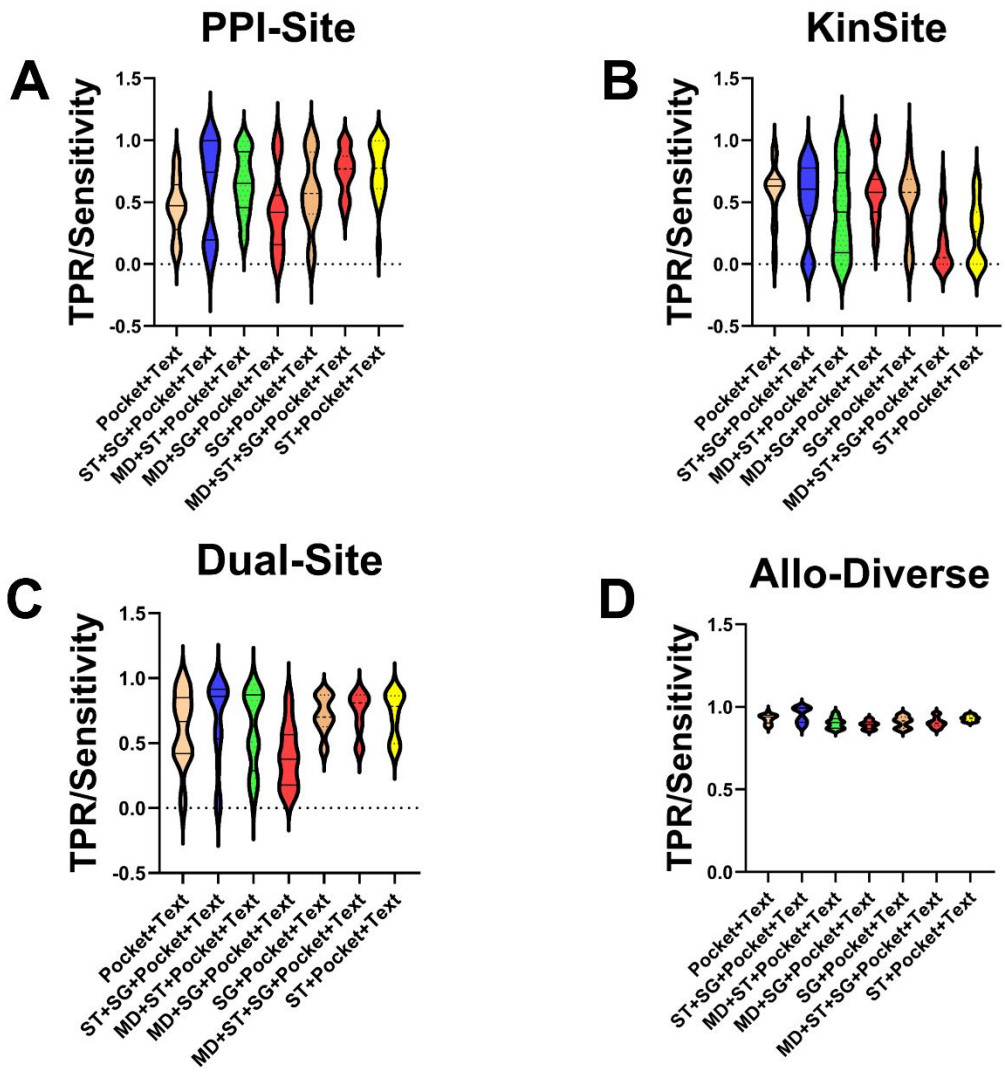


Figure S4. TPR/Sensitivity (true positive rate) across allosteric binding site discrimination regimes. Violin plots show the distributions of sensitivity (true positive rate for allosteric sites) across the four datasets representing distinct separability regimes: PPI-Site (A), KinSite (B), Dual-Site (C) , and AlloDiverse (D). The x-axis within each panel corresponds to the seven OneProt encoder architectures evaluated in this study: Pocket+Text (baseline), ST+Pocket+Text (structure-token), SG+Pocket+Text (structure-graph), MD+ST+Pocket+Text (MD + structure-token), MD+SG+Pocket+Text (MD + structure-graph), ST+SG+Pocket+Text (both structural encoders), and MD+ST+SG+Pocket+Text (full multimodal). Colored points within each violin indicate the

embedding combination used for downstream classification. Boxes indicate the median and interquartile range across random seeds, while individual points represent single runs. The dashed horizontal line denotes random-expectation level (0.5). The figure reveals a progressive transition from low-separability (PPI-Site) to high-separability (AlloDiverse) regimes. In PPI-Site, sensitivity exhibits substantial variability and remains close to random performance across all architectures and embedding combinations, indicating limited discriminative signal. KinSite displays uniformly low sensitivity across configurations, reflecting the extreme difficulty of recovering allosteric sites under severe class imbalance ($\rho \approx 51.1$). Dual-Site shows improved sensitivity, with sequence-containing embeddings (green and red) consistently outperforming pocket-only and pocket+text combinations across most architectures. In AlloDiverse, sensitivity approaches ceiling performance across all architectures and embedding combinations, indicating that most models readily identify allosteric sites.

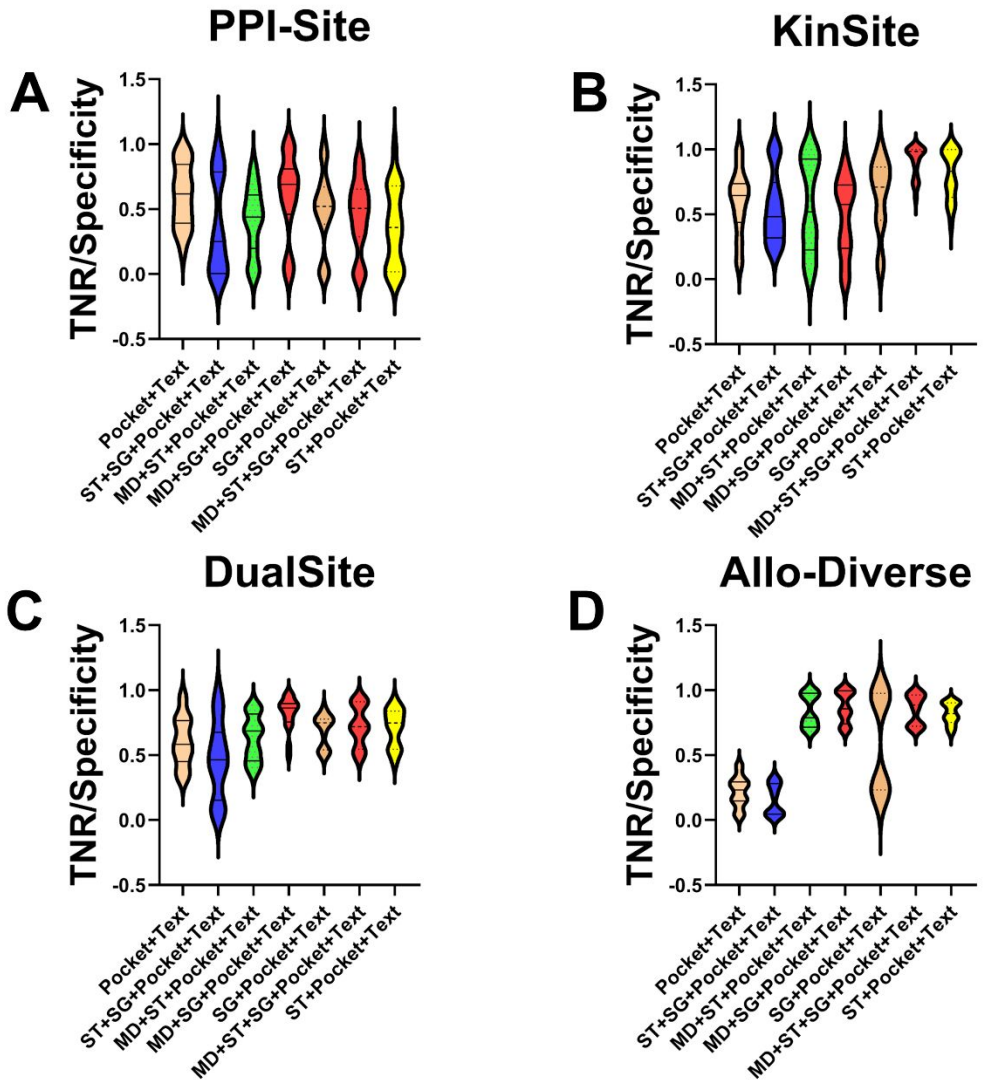


Figure S5. TNR/Specificity (true negative rate) across allosteric binding site discrimination regimes. Violin plots show the distributions of specificity (true negative rate for orthosteric competitive sites) across the four datasets representing distinct separability regimes: PPI-Site (A), KinSite (B), Dual-Site (C), and AlloDiverse (D). The x-axis within each panel corresponds to the seven OneProt encoder architectures evaluated in this study: Pocket+Text (baseline), ST+Pocket+Text (structure-token), SG+Pocket+Text (structure-graph), MD+ST+Pocket+Text (MD + structure-token), MD+SG+Pocket+Text (MD + structure-graph), ST+SG+Pocket+Text (both structural encoders), and MD+ST+SG+Pocket+Text (full multimodal). Colored points

within each violin indicate the embedding combination used for downstream classification. Boxes indicate the median and interquartile range across random seeds, while individual points represent single runs. The dashed horizontal line denotes random-expectation level (0.5). The figure reveals a progressive transition from low-separability (PPI-Site) to high-separability (AlloDiverse) regimes. In PPI-Site, specificity exhibits substantial variability and remains close to random performance across all architectures and embedding combinations, indicating limited discriminative signal. KinSite displays higher specificity compared to sensitivity (Figure S2), reflecting the relative ease of identifying orthosteric ATP-binding sites under extreme class imbalance ($\rho \approx 51.1$). Dual-Site shows improved specificity, with sequence-containing embeddings (green and red) consistently outperforming Pocket+Text across most architectures. In AlloDiverse, specificity remains more variable than sensitivity, indicating that models readily identify allosteric sites but differ in their ability to reject orthosteric sites.

Supporting Information Tables

Table S1. Number of binding-site instances in the training, validation, and test partitions for each dataset used in this study. The PPI-Site, KinSite, Dual-Site, and AlloDiverse datasets correspond to the benchmark datasets described in the main text. AlloDiverse Text denotes the subset of the AlloDiverse dataset for which text annotations were available and was therefore used in experiments incorporating text embeddings. ASD positive lists the positive (allosteric) binding sites extracted from the Allosteric Database (ASD) before pairing with orthosteric sites to construct the balanced AlloDiverse benchmark for the ASD pocket construction with DBSCAN and without it (harmonized).

Dataset	**train**	**valid**	**test**
PPI-Site	2767	421	341
KinSite	2502	540	493
Dual-Site	5269	961	834
AlloDiverse	17346	1942	1819
AlloDiverse Text	16640	1727	3913
ASD positive	12077	981	985
AlloDiverse harmonized	15116	1672	1555

AlloDiverse Text harmonized	14534	1625	1504
ASD positive harmonized	9847	711	721

Table S2. Performance of the handcrafted pocket geometry and amino-acid composition baseline across allosteric-site classification datasets. L1-regularized logistic regression was trained using four pairwise residue-centroid distance descriptors and 20 pocket amino-acid composition features. The inverse regularization strength was selected by 5-fold cross-validation on the training set using ROC-AUC as the optimization criterion. Results are reported as mean ± standard deviation across 10 independently shuffled cross-validation partitions, while the predefined held-out test set was kept fixed across runs. AUROC, area under the precision–recall curve (AUPR), Matthews correlation coefficient (MCC), accuracy, F1 score, precision, and recall are reported. For AlloDiverse dataset the results were invariant across seeds, and therefore the standard deviation and condidence intervals are not reported. No test-set information was used for hyperparameter selection.

Dataset	AUROC	AUPR	MCC	Accuracy	F1	Precision	Recall
PPI-Site	0.589 ± 0.009 (0.583–0.596)	0.454 ± 0.013 (0.445–0.463)	0.132 ± 0.012 (0.124–0.141)	0.545 ± 0.005 (0.542–0.549)	0.519 ± 0.007 (0.515–0.524)	0.430 ± 0.005 (0.426–0.433)	0.656 ± 0.010 (0.649–0.663)
KinSite	0.763 ± 0.004 (0.760–0.766)	0.252 ± 0.003 (0.250–0.254)	0.171 ± 0.012 (0.162–0.180)	0.878 ± 0.003 (0.876–0.880)	0.195 ± 0.009 (0.188–0.202)	0.131 ± 0.006 (0.127–0.135)	0.384 ± 0.025 (0.366–0.402)
Dual-Site	0.720 ± 0.002	0.367 ± 0.013	0.209 ± 0.004	0.775 ± 0.004	0.345 ± 0.001	0.353 ± 0.006	0.337 ± 0.004

	(0.719–0.721)	(0.357–0.376)	(0.206–0.212)	(0.772–0.777)	(0.344–0.346)	(0.349–0.358)	(0.335–0.340)
AlloDiverse	0.782	0.868	0.191	0.653	0.774	0.651	0.956

Table S3. Within-protein discrimination of allosteric and orthosteric pockets in the AlloDiverse test set. Within-protein pairwise ranking accuracy was evaluated for proteins containing both allosteric and orthosteric pockets. For each protein, all allosteric–orthosteric pocket pairs were compared, and a pair was considered correctly ranked when the allosteric pocket received a higher predicted allosteric probability than the orthosteric pocket. Pairwise ranking accuracy was first calculated separately for each protein and then averaged across proteins, ensuring that each protein contributed equally irrespective of the number of pockets. The analysis comprised 10 proteins and 47 within-protein pocket pairs. Values represent mean ± SD across ten random seeds. A value of 0.5 corresponds to random pairwise ranking. Bold values indicate the highest mean accuracy within each encoder configuration.

Encoder	Embedding combination	Within-protein pairwise ranking accuracy
ST+SG+Pocket+Text	Pocket	0.637 ± 0.041
	Pocket + Sequence	0.648 ± 0.055
	Pocket + Sequence + Text	**0.775 ± 0.127**
	Pocket + Text	0.480 ± 0.111
MD+ST+SG+Pocket+Text	Pocket	0.693 ± 0.137
	Pocket + Sequence	0.683 ± 0.049
	Pocket + Sequence + Text	**0.817 ± 0.071**
	Pocket + Text	0.689 ± 0.028
MD+ST+Pocket+Text	Pocket	0.300 ± 0.000
	Pocket + Sequence	0.558 ± 0.169
	Pocket + Sequence + Text	**0.693 ± 0.033**
	Pocket + Text	0.617 ± 0.177
MD+SG+Pocket+Text	Pocket	0.340 ± 0.065
	Pocket + Sequence	**0.732 ± 0.090**
	Pocket + Sequence + Text	0.716 ± 0.061
	Pocket + Text	0.681 ± 0.103

Pocket + Text	Pocket	0.480 ± 0.291
	Pocket + Sequence	0.552 ± 0.072
	Pocket + Sequence + Text	0.578 ± 0.084
	Pocket + Text	**0.818 ± 0.081**
SG + Pocket + Text	Pocket	0.633 ± 0.225
	Pocket + Sequence	**0.809 ± 0.086**
	Pocket + Sequence + Text	0.733 ± 0.044
	Pocket + Text	0.805 ± 0.127
ST + Pocket + Text	Pocket	0.400 ± 0.056
	Pocket + Sequence	0.706 ± 0.116
	Pocket + Sequence + Text	**0.856 ± 0.026**
	Pocket + Text	0.766 ± 0.020