JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access
This article is licensed under CC-BY-NC-ND 4.0

pubs.acs.org/jcim
Article

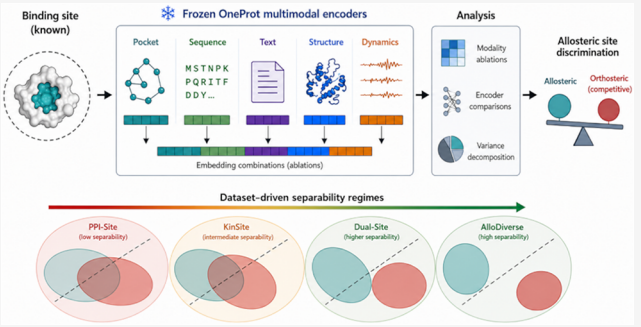
Systematic Profiling of Encoder Architectures and Multimodal Embeddings in Protein Foundation Models Reveals Data Set-Driven Separability Regimes in Prediction of Allosteric and Orthosteric Binding Sites

Alina Bazarova and Gennady M. Verkhivker*

Cite This: https://doi.org/10.1021/acs.jcim.6c03114
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Allosteric regulation represents a fundamental mechanism of protein function, yet distinguishing allosteric from orthosteric protein-binding sites remains a persistent computational challenge. While multimodal protein foundation models offer the potential to integrate complementary biological signals, including sequence, structure, functional annotations, and conformational dynamics, their performance determinants in allosteric binding site detection remain poorly understood. We introduce a unified computational framework for profiling multimodal protein foundation models across distinct binding site separability regimes. Rather than evaluating models solely by predictive accuracy, the framework combines systematic modality embedding ablations, encoder architecture comparisons, and variance decomposition to characterize how evolutionary, structural, functional, and dynamical information contributes to allosteric site discrimination. Using the OneProt multimodal model, we evaluate two complementary levels of multimodal integration: (a) encoder architectures that differ in the modalities incorporated during pretraining and (b) downstream combinations of pocket, sequence, and text embeddings used for classification. To systematically probe the determinants of model performance, we benchmark these configurations across four assembled data sets of protein complexes representing a spectrum of biological complexity and a range of structural, dynamic, and evolutionary contexts for orthosteric and allosteric binding sites. Through comprehensive embedding ablations, encoder architecture comparisons, and variance decomposition, we demonstrate that model performance is governed primarily by intrinsic data set properties rather than architectural complexity, with data set identity accounting for 63.7% of explainable variance. Across the data sets examined, we identify substantial differences not only in overall separability but also in the information accessible from different pretrained representations. Single-modality controls show that high predictive performance does not necessarily imply reliance on the explicitly provided pocket representation: depending on data set composition, substantial discrimination can already be obtained from sequence- or text-based multimodal representations. Within-protein comparisons nevertheless demonstrate that pocket-containing representations retain binding site-specific discriminatory information when protein identity and the associated sequence and text representations are held constant. At the encoder level, structural and molecular dynamics encoders provide the greatest benefit in intermediate and high separability settings. Example-level analysis further shows that combining data sets can reshape which binding sites are preferentially recognized. These findings establish that the success of multimodal foundation models depends critically on alignment between available modalities and the biological signatures that distinguish regulatory classes in each data set.



INTRODUCTION

The discovery of druggable binding sites in proteins—particularly those that are cryptic, transient, or allosteric—remains one of the most persistent bottlenecks in structure-based drug design and chemical biology.[1−4] Allosteric sites, in particular, offer tremendous therapeutic potential due to their capacity for exquisite selectivity, tunable activity control, and resistance evasion, yet they remain notoriously difficult to identify computationally. Unlike orthosteric ligands, which directly compete with endogenous substrates at catalytic or active sites, allosteric modulators influence protein behavior indirectly through long-range conformational coupling and regulatory communication within the protein structure.[5−8]

Received: September 5, 2026
Revised: October 8, 2026
Accepted: October 12, 2026

ACS Publications
© XXXX The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

Because allosteric modulation can provide increased selectivity, tunable activity control, and reduced toxicity, allosteric drug discovery has emerged as an important strategy in modern pharmacology and chemical biology.[9] Orthosteric pockets are often evolutionarily conserved and structurally well-defined, whereas allosteric sites tend to exhibit greater heterogeneity in their geometry, physicochemical composition, and spatial relationship to functional regions.[10−13] In many proteins, allosteric regulation is strongly coupled to conformational dynamics and transient structural states that are not fully captured by static structural descriptions alone.[14] Consequently, distinguishing allosteric binding sites from orthosteric binding sites requires integrating information across multiple biological scales, including sequence conservation, local pocket geometry, functional annotations, and conformational variability. Computational strategies for binding site prediction have evolved from geometric and energetic approaches such as Fpocket,[15] SiteHound,[16] and MetaPocket 2.0[17] to machine learning (ML) methods such as P2Rank,[18,19] which predicts binding-site residues by scoring local chemical neighborhoods using a random forest classifier trained on known protein−ligand complexes. Deep learning (DL) has further advanced the field, with voxel-based approaches such as DeepSite,[20] object detection architectures exemplified by FRSite,[21] segmentation networks,[22] and dynamics-aware models BiteNet[23] achieving high accuracy on benchmarks. However, these structure-based methods exhibit performance drops when predicting binding sites in the ligand-bound holo complexes using apo structures and identifying cryptic sites that emerge during dynamic changes, revealing their limitations in allosteric site detection, where conformational heterogeneity and transient pocket formation are important features underpinning allosteric regulatory events.[24−26]

Recent advances in protein foundation models provide a promising framework for integrating complementary signals across the biological scales. Protein language models (PLMs) have demonstrated that evolutionary-scale representations encode substantial information about protein structure and function.[27,28] Trained on hundreds of millions of sequences, PLMs learn evolutionary constraints and structural propensities directly from primary sequence, enabling scalable proteome-wide predictions of binding sites and ligand binding without explicit structural input.[29−31] PLM fine-tuning strategies can adapt model parameters during task-specific training, which can be achieved through full fine-tuning, in which all transformer layers are updated,[32] or via parameter-efficient fine-tuning (PEFT) methods that introduce a small set of trainable parameters while freezing the original PLM weights.[33,34] A range of fine-tuning protocols for the prediction of cryptic binding sites was evaluated using benchmarking against frozen-embedding baselines, revealing that sequence-based models could be enhanced by combining domain-informed fine-tuning with multitask learning and high-quality curated data sets.[35] However, the reliance on evolutionary conservation can often render PLMs less robust in detection of allosteric and cryptic sites, which are often evolutionarily labile and lack deep sequence conservation.[31,35] Recent approaches have fused PLM embeddings with structural graphs (EquiBind,[36] DiffSBDD[37]), dynamics simulations (PocketMiner[38]), and physics-informed strategies (AlloPred,[39] Allosite,[40] and DeepAllo[41,42]).

More recently, multimodal protein foundation models have extended this paradigm by jointly incorporating sequence, structural, biochemical, textual, and dynamic information into shared latent embedding spaces learned through self-supervised or contrastive pretraining.[43−47] By aligning multiple biological modalities, these models aim to capture complementary aspects of protein organization not accessible from sequence or structure alone. A prominent example is the OneProt model,[48] which integrates (a) sequence by using the ESM2 model, (b) structural information through Foldseek[49] tokens and ProNet graphs,[50] (c) binding pockets via ProNet, and (d) text annotations with BiomedBERT[51] through contrastive alignment to a sequence anchor. An extended version further incorporates molecular dynamics (MD) trajectories via a transformer-based MDGen encoder,[52] capturing protein flexibility and showing that MD inclusion alters latent-space organization and produces distinct downstream transfer characteristics.

Despite this progress, robust detection of functional allosteric and cryptic binding sites remains elusive, as most methods are optimized for stable, evolutionarily conserved orthosteric pockets, whereas allosteric sites are sparsely populated in structural databases.[53] A recent investigation integrated local binding geometry, coevolution, and dynamic allostery into a three-parameter model to identify potentially hidden allosteric sites in ensembles of protein structures with orthosteric ligands.[54] The performance dichotomy raises a fundamental question: what physical principles govern the evolutionary design of allosteric sites, and how can we leverage these principles to build more interpretable predictive models? To systematically interrogate the physical determinants of binding site predictability, we have recently developed a dual analytical framework that integrates a fine-tuned PLM with a physics-based energy landscape-centric interpretability layer derived from energy landscape theory and a frustration analysis approach.[55−62] Applying this framework to human kinases, we demonstrated that orthosteric and allosteric sites occupy fundamentally different energetic regimes, where orthosteric ATP-binding sites can be reliably identified with high precision, whereas allosteric sites are detected with substantially lower confidence across kinase families and conformational states.[31] This analysis revealed that predictive success is governed by the local energetic embedding of binding sites: orthosteric pockets are located in minimally frustrated basins that generate strong evolutionary and structural signatures, whereas allosteric pockets occupy predominantly neutrally frustrated zones associated with conformational plasticity and reduced evolutionary constraint.[31] Most existing studies focus on maximizing predictive performance on individual benchmarks, while comparatively little attention has been given to how intrinsic data set properties influence the learnability of allosteric regulation. In particular, it remains unclear how structural diversity, conformational heterogeneity, functional annotation coverage, and class imbalance affect the relative contribution of different biological modalities. To address this gap, we develop a unified analytical framework for systematically profiling multimodal protein foundation models across distinct binding-site separability regimes. The framework enables systematic dissection of how different sources of biological information contribute to allosteric-site discrimination across diverse biological contexts. Using the OneProt multimodal framework, we systematically evaluate both encoder architectures, which differ in the biological modalities incorporated during multimodal pretraining, and downstream combinations of pocket, sequence, and text embeddings

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

extracted from the corresponding pretrained OneProt encoders across multiple data sets containing experimentally annotated allosteric and orthosteric binding sites. Rather than treating allostery prediction as a single benchmark problem, we interpret model behavior across four designed data sets (PPI-Site, KinSite, Dual-Site, and AlloDiverse) reflecting distinct biological regimes of separability. The PPI-Site data set comprises binding sites annotated by their spatial relationship to protein−protein interaction interfaces, where allosteric and orthosteric sites often share the same structural context. The KinSite data set contains kinase inhibitor complexes, where orthosteric ATP-binding sites are evolutionarily conserved while allosteric sites are structurally heterogeneous. The Dual-Site data set combines PPI-Site and KinSite to assess how increased sample and biological heterogeneity influences allosteric site discrimination. AlloDiverse data set further incorporates allosteric sites spanning multiple protein families providing a broader biological setting for evaluating representation behavior. We investigate how modality composition, structural diversity, conformational variability, and class imbalance shape predictive performance and representation quality across different allosteric learning regimes. Our results reveal that model performance is governed primarily by data set characteristics and composition rather than architectural complexity alone. Across data sets, we identify three distinct regimes: (a) low-separability settings, where current representations underperform in reliably distinguishing allosteric and orthosteric pockets; (b) intermediate regimes, where multimodal integration substantially improves ranking and sensitivity; and (c) highly separable regimes, where most architectures converge to near-ceiling performance. We further show that the contributions of pocket, sequence, and text embedding combinations, as well as those of structural and MD encoder architectures, depend on data set composition and structural diversity, highlighting the context-dependent role of multimodal protein representations in modeling allosteric regulation. Importantly, single-modality controls demonstrate that overall separability and the biological scale at which discriminative information is encoded are distinct properties: strong classification performance does not necessarily imply reliance on explicitly provided local binding-pocket information. Example-level analyses further reveal data set-specific patterns in which binding sites are preferentially recognized across the different separability regimes. In PPI-Site, the low observed separability must be interpreted in the context of the task formulation, as the PPI interface used to define the class labels is not explicitly provided to the classifier. Nevertheless, local pocket geometry may provide indirect information relevant to this distinction, consistent with pocket representations showing the strongest, although still limited, discrimination. KinSite presents a different regime: despite the shared kinase context, simple whole-protein amino-acid composition provides substantial discrimination, indicating that class-associated signal is present at the level of coarse protein sequence characteristics. At the same time, the strongly imbalanced nature of the data set and the heterogeneous behavior of individual OneProt representations emphasize that performance in this setting cannot be attributed straightforwardly to a single biological feature or representation modality. In the combined Dual-Site setting, substantially stronger performance is accompanied by improved recognition of examples originating from both constituent data sets. Relative to KinSite-only training, the merged setting recovers additional orthosteric kinase sites as well as rare allosteric examples, while several PPI-Site allosteric pockets are classified more consistently under Dual-Site training than when the classifier is trained on PPI-Site alone. Together with the strong performance of sequence-based OneProt representations on Dual-Site, these prediction-level patterns suggest that combining the two data sets provides complementary training signals that enable the downstream classifier to establish a more effective decision boundary within the pretrained representation space. These findings establish that the success of multimodal foundation models depends critically on alignment between available modalities and the biological signatures that distinguish regulatory classes in each data set. Through this framework, we provide a systematic characterization of when and why multimodal integration benefits allostery prediction. By linking predictive performance to underlying biological and statistical properties and grounding our analysis in physical principles, this study establishes a foundation for developing more robust, context-aware, and physically informed models for allosteric site discovery.

MATERIALS AND METHODS

General Formulation and Multimodal Downstream Prediction Pipeline

In this study, we investigate the discrimination of allosteric and orthosteric competitive binding sites using multimodal protein foundation-model representations. Unlike conventional ligand-binding prediction tasks, our objective is not to identify whether a protein region binds ligands but rather to distinguish between distinct functional and regulatory binding regimes, assuming that the binding site location is known a priori.

Formally, let a protein P be represented by its amino acid sequence $S \in \mathcal{A}^L$, where $\mathcal{A}$denotes the alphabet of 20 standard amino acids and L is the sequence length. Associated with P is a three-dimensional structure $\mathcal{X} \in \mathbb{R}^{N\times 3}$ comprising N atoms, and a set of functional annotations T derived from curated biological databases. Where available, conformational dynamics are represented by an MD trajectory $M \in \mathbb{R}^{N\times 3\times \tau}$, where τ denotes the number of trajectory frames. For a given binding site B, we define its spatial extent as a set of K =100 residues $\mathcal{R}_B = \{r_1, r_2, ..., r_K\}$ proximal to the bound ligand. During multimodal pretraining, the structural modality X may be encoded either through graph-based structural representations X_{SG} or through structure-token representations X_{ST}, depending on the selected OneProt encoder architecture.[48]

The downstream task is formulated as binary classification, where experimentally annotated allosteric pockets are treated as positive instances y = 1 and orthosteric competitive pockets are treated as negative instances y = 0. Let the encoder input space be defined as in (eq 1):

$$\mathcal{I} = \{S, X_{SG}, X_{ST}, R_B, T, M\} \tag{1}$$

For a given OneProt encoder architecture, only a subset $\mathcal{J}' \subseteq \mathcal{J}$ is used during multimodal pretraining. The pretrained model is therefore represented by (eq 2):

$$\Phi: \mathcal{J}' \rightarrow \{\mathbf{z}_k\}_{k=1}^{|\mathcal{J}'|} \tag{2}$$

where each $\mathbf{z}_k \in \mathbb{R}^d$ is a modality-specific embedding vector. For downstream allostery prediction, only the pocket $\mathbf{z}_{pocket}$, sequence $\mathbf{z}_{seq}$, and text $\mathbf{z}_{text}$ embeddings are considered. Depending on the embedding ablation setting, these embeddings are concatenated to form

$\mathbf{z} = \text{concat}(\{\mathbf{z}_i\}_{i=1}^{k}) \in \mathbb{R}^{kd}$ where $k \in \{1,2,3\}$ depending on the selected embeddingss combination. The downstream classifier $f: \mathbb{R}^{kd} \rightarrow [0, 1]$ predicts the probability $\hat{y} = P(y = 1|\mathbf{z})$. Rather than

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

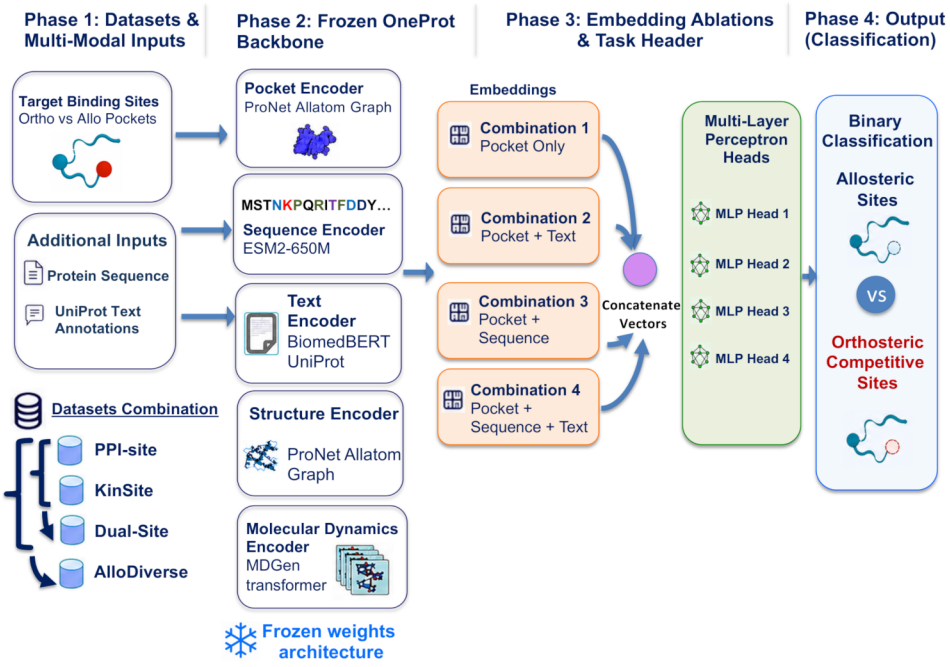


Figure 1. Analytical framework for investigating multimodal representations for binding-site detection. The study is organized into four interconnected stages. Phase 1 defines the allosteric-site prediction task and introduces four benchmark datasets—PPI-Site, KinSite, Dual-Site, and AlloDiverse—that span distinct biological regimes and collectively define a progression from low to high separability. Phase 2 employs a frozen OneProt multimodal backbone comprising pocket, sequence, text, structure, and molecular-dynamics encoders to generate protein representations. Phase 3 systematically evaluates two levels of multimodal integration: encoder architectures that differ in the modalities incorporated during pretraining (ST, SG, MD variants) and downstream embedding combinations (pocket-only, pocket+text, pocket+sequence, pocket+sequence +text). Phase 4 performs binary classification of allosteric versus orthosteric competitive sites and enables quantitative comparison of predictive performance across data sets, architectures, and embedding combinations, providing a systematic assessment of how data set composition, modality selection, and architectural design influence allosteric-site prediction.

using OneProt solely as a predictive classifier, we interpret multimodal representation behavior across data sets as a means of probing which biological modalities contribute to allosteric discrimination under different structural and statistical conditions.

The full multimodal representation pipeline, illustrated in Figure 1, proceeds through four integrated phases. First, we combine three complementary data sets. The PPI-Site data set was derived from a large-scale pocket-centric collection of protein−ligand binding sites and protein−protein interaction (PPI) interfaces[63] derived from experimentally resolved Protein Data Bank[64] structures. The KinSite data set was assembled from multiple structural and activity resources, including PDB,[64] KLIFS database,[65] KinLigDB,[66] which provides structures of 4219 kinase-ligand complex structures covering 297 human kinases, and the Kinase Conformation Resource (KinCoRe).[67,68] The Allosteric Database (ASD) datset[69,] includes allosteric ligand-protein complexes that include diverse protein families, including kinases, phosphatases, GPCRs, ion channels, and transcription factors. For each binding pocket, we assemble the input modalities required by the frozen pretrained OneProt encoders to generate modality-specific embeddings for downstream fine-tuning: (a) the protein sequence S (amino acid string), (b) the three dimensional pocket structure X (atomic coordinates of the K=100 residues surrounding the binding site), and (c) UniProt[70] text annotations T (including protein names, protein family assignments, organism information, functional descriptions, pathway annotations, subcellular localization).

This multimodal input design is critical because allostery is encoded not only in sequence conservation and geometric pocket shape but also in the functional context and evolutionary history of the protein. We use pretrained OneProt encoder architectures[48] to generate modality-specific embeddings for the downstream allostery classification. Throughout this study, the pretrained encoders remain frozen, while only a lightweight multilayer perceptron (MLP), taking the selected embedding combinations as input, is trained on the allosteric versus orthosteric classification task by using a binary cross-entropy loss. The downstream classifier receives only combinations of the pocket, sequence, and text embeddings extracted from pretrained encoders. However, the pretrained OneProt architectures differ in the encoder modalities incorporated during the multimodal pretraining. In particular, some architectures additionally include structure tokens, structure graphs, and MD encoders. Although embeddings from these additional encoders are not used directly for downstream classification, they influence the shared latent representation learned during pretraining and thereby shape the resulting pocket, sequence, and text embeddings. This design enables us to distinguish the contribution of multimodal pretraining from that of the embedding combinations used by the downstream classifier. For each modality, a dedicated encoder extracts the task-relevant features. Pocket Encoder (ProNet Atom Graph) is a graph neural network operating on the atomic graph of the binding pocket, capturing local geometric and chemical features, including atom types, bond distances, and residue contacts. Sequence Encoder (ESM2 Transformer) is a transformer-based encoder that processes the amino acid sequence, capturing evolutionary constraints, coevolutionary signals, and context-dependent residue relationships. Text Encoder (BiomedBERT UniProt) is a biomedical language model fine-tuned on UniProt text annotations that extracts semantic information from functional descriptions, domain annotations, and pathway information. Optional encoders include structural encoders (ProNet Atom Graph + ESM2 Transformer). The ProNet encoder[50] processes the protein structure as an atom graph, whereas the ESM2 transformer operates on FoldSeek-derived structure tokens. Together, these complementary structural encoders provide alternative representations of the same three-dimensional protein structure, whose outputs are integrated during multimodal pretraining. The MD Encoder (MDGen transformer encoder) is an encoder that processes trajectory data, capturing dynamical features such as residue fluctuations and conformational ensembles. The outputs of the modality-specific encoders are

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

projected into a shared latent space during multimodal pretraining. The pretrained OneProt backbone is kept frozen to examine how different encoder architectures shape the learned latent representation without modification during downstream training.

To characterize how different OneProt representations contribute to allosteric site discrimination, we performed a systematic embedding ablation analysis. We train separate classifiers using four distinct embedding combinations: combination 1: pocket only (local structural features alone); combination 2: pocket + text (local structure + functional annotations); combination 3: pocket + sequence (local structure + evolutionary sequence patterns); and combination 4: pocket + sequence + text (trimodal integration); and additionally evaluate sequence-only and text-only representations to determine how much discrimination is accessible without explicitly providing a pocket embedding to the downstream classifier. OneProt multimodal pretraining aligns representations generated by distinct modality-specific encoders within a shared representation space. Consequently, the resulting pocket, sequence, and text embeddings can contain partially overlapping information and are not intended to represent independent biological information channels. Indeed, the original OneProt study showed that these representations exhibit different performance characteristics across protein-binding site discrimination tasks. We therefore use the ablations to examine how information relevant to allosteric-site discrimination is distributed across pretrained OneProt representations and how their relative utility and benefit of combination vary across data sets.

Data Sets and Biological Regimes

We evaluated the proposed framework on three complementary data sets containing experimentally annotated allosteric and orthosteric binding sites. Rather than serving solely as independent benchmarks, the selected data sets were deliberately chosen to represent distinct structural, functional, and statistical regimes, enabling systematic investigation of how data set composition influences allosteric discriminability and modality-dependent representation behavior. Where predefined train/validation/test partitions were not available, proteins were clustered using MMseqs2[59] at 30% sequence identity prior to data set splitting to ensure that homologous proteins did not appear across splits. Where training/validation splits were available (ASD data set), we inherited those while randomly splitting the original validation set into two equally sized validation and test sets.

PPI-Site Data Set

The PPI-Site data set[63] was constructed through a multistage filtering pipeline incorporating experimental quality constraints, including X-ray structures with resolution ≤3.5 Å, ligand filtering, and interface-aware classification of binding pockets. Binding pockets were identified using VolSite.[71] A defining feature of the PPI-Site data set is that binding pockets are classified according to their spatial relationship to PPI interfaces, yielding three distinct types: PLOC (Protein−Ligand Orthosteric Competitive), where the ligand pocket directly overlaps with the PPI epitope; PLONC (Protein−Ligand Orthosteric Non-Competitive), where the ligand pocket is near but nonoverlapping with the PPI interface; and PLA (Protein−Ligand Allosteric), where the ligand pocket is spatially distinct from the PPI interface. In this study, PLOC pockets were treated as negative instances (orthosteric) and PLA pockets as positive instances (allosteric); PLONC pockets were excluded to maintain a clear binary distinction between direct competitive binding and allosteric modulation. Because PPI-Site class labels arise from spatial relationships rather than independent regulatory mechanisms, allosteric and orthosteric pockets often share closely related structural and functional contexts. Importantly, the PPI interface itself is not explicitly provided as an input to the downstream classifier in the present framework. The classifier therefore has to infer any interface-related distinction indirectly from pocket, sequence, and text representations. PPI-Site contains structurally consistent pocket annotations across a broad range of protein families with moderate class imbalance, comprising 1845 PLOC and 1689 PLA sites across 1278 unique protein chains spanning 847 CATH families.

KinSite Data Set

In the KinSite data set, KinLigDB[66] contains 4219 curated kinase-ligand complex structures of 297 human kinases from 106 kinase families, with ∼75% of these kinases having at least two structures with different ligands and 92 kinases having more than 10 complex structures. For this database, there are reported type I inhibitors (3167 entries), type II inhibitors (283 entries), type III inhibitors (23 entries), type IV inhibitors (9 entries), competitive inhibitors (554 entries), covalent inhibitors (31 entries), activators (49 entries), and other allosteric ligands (103 entries).[66] KinLigDB entries are combined with KinCoRe,[67,68] which classifies inhibitors into five mechanistically distinct classes. Type I inhibitors are ATP-competitive and orthosteric, binding to the DFG-in, αC-in conformation. Type I.5 inhibitors are ATP-competitive and orthosteric, binding to the DFG-in, αC-out conformation, extending into the back pocket. Type II inhibitors are ATP-competitive and orthosteric, binding to the DFG-out conformation with back pocket extension. Type III inhibitors are allosteric and proximal, binding >6.0 Å from the hinge region. Type IV inhibitors are allosteric and distal, binding >6.5 Å from both the hinge region. Types I, I.5, and II were treated as negative instances (orthosteric), while Types III and IV were treated as positive instances (allosteric). Kinase ATP-binding sites are highly conserved, generating strong evolutionary signals, whereas allosteric kinase sites are structurally heterogeneous. This creates an intermediate-separability regime where sequence information becomes essential for discrimination. After rigorous filtering (temporal: structures deposited after 2020 only; sequence: ≥30% sequence identity removed; pocket-level: ≥40% residue identity to training pockets removed; quality: resolution ≤3.0 Å; occupancy: ligand atoms with occupancy <0.5 excluded), the KinSite data set comprises 3363 orthosteric and 172 allosteric high-confidence complexes spanning 246 kinases.

ASD Data Set

Unlike PPI-Site and KinSite, ASD primarily provides residue-level annotations rather than predefined pocket structures. To obtain representations compatible with OneProt, residue-level annotations were converted to fixed-size structural pocket representations through spatial clustering and geometric reconstruction. ASD introduces substantially greater allosteric diversity, spanning diverse protein families, regulatory mechanisms (including both orthosteric and allosteric modulation), and evolutionary contexts.

The ASD-derived data set comprises 14,043 allosteric pocket instances across 1842 unique proteins (spanning 247 CATH families and 531 Pfam families), substantially increasing the diversity of the positive class and enabling investigation of multimodal representation behavior under conditions of increased allosteric heterogeneity.

Data Set Combinations and Regime Progression

To systematically investigate how data set composition influences learnability, we constructed four evaluation configurations that systematically vary total sample size, allosteric diversity, and class balance, enabling controlled investigation of how data set composition influences discriminability and modality-dependent behavior. The first configuration uses the PPI-Site data set alone, comprising 1845 PLOC (negative) and 1689 PLA (positive) sites, with a class imbalance ratio ρ = 1.09. This represents the low-separability regime under the current task formulation, consistent with the closely related structural contexts of the two classes and the absence of the class-defining PPI-interface relationship from the classifier inputs.

The second configuration uses the KinSite data set alone, comprising 3363 orthosteric (negative) and 174 allosteric (positive) high-confidence complexes, with ρ = 51.1. This represents an intermediate-separability regime where sequence information becomes essential for discrimination, reflecting the conserved nature of kinase ATP sites contrasted with heterogeneous allosteric pockets. The third configuration, Dual-Site, combines PPI-Site and KinSite, yielding 5203 orthosteric (negative) and 1861 allosteric (positive) sites, with ρ = 2.79. This creates an intermediate-separability regime in which the combination of two biologically distinct data sets

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

provides complementary training signals that can be exploited within the pretrained representation space. The fourth configuration, AlloDiverse, combines ASD with PPI-Site and KinSite, yielding 5203 orthosteric (negative) and 15,904 allosteric (positive) sites, with $\rho = 0.33$. This represents a high-separability regime with extensive allosteric diversity and a positive-class majority, providing multiple potential sources of discriminative signal. This progression from low to high separability enables controlled investigation of how data set composition influences discriminability and modality-dependent behavior, with each configuration strategically positioned to probe a distinct aspect of the allosteric discrimination problem (Table S1).

Multimodal Backbone Architecture

We adopted the OneProt multimodal protein foundation model as the representation-learning framework for all downstream experiments.[48] The conceptual role of OneProt in our downstream pipeline is described in the previous section outlining the multimodality downstream prediction pipeline. Here, we briefly summarize the architectural components relevant to the present study. Complete details of the model architecture, multimodal pretraining objectives, and optimization procedure are available in the original OneProt model.[48] Protein sequence representations are generated using the OneProt sequence encoder based on the ESM2−650 M transformer architecture[43]. The encoder comprises 33 transformer layers with 20 attention heads per layer and a hidden dimension of 1280. Given an input protein sequence $S = (s_1, s_2,\cdots,s_L)$, the encoder produces contextualized residue-level representations as shown in (eq 3):

$$H_{\text{seq}} = \text{ESM2}(S) \in \mathbb{R}^{L\times 1280} \tag{3}$$

A protein-level sequence representation is obtained by mean pooling across residue positions as defined in (eq 4):

$$\bar{\mathbf{h}}_{\text{seq}} = \frac{1}{L}\sum_{i=1}^{L} H_{\text{seq}}[i] \in \mathbb{R}^{1280} \tag{4}$$

and is subsequently mapped into the shared OneProt latent space using a trainable projection head, as described in (eq 5):

$$\mathbf{z}_{\text{seq}} = P_{\text{seq}}(\bar{\mathbf{h}}_{\text{seq}}) \in \mathbb{R}^{1024} \tag{5}$$

where P_{seq} denotes the projection network and d is the dimension of the shared multimodal embedding space.

Structural representations were generated using two complementary encoders integrated within OneProt. The graph-based structural encoder uses ProNet,[50] which processes protein structures as hierarchical graphs through message passing across amino-acid, backbone, and all-atom levels to capture fine-grained structural details and higher-order geometric organization. The token-based structural encoder uses Foldseek structural tokens[49] processed with an ESM2 transformer to discretize local three-dimensional environments into sequence-like tokens, capturing broader structural context and fold-level organization. Pocket embeddings were obtained using the binding-site encoder integrated within OneProt. Binding pockets are represented as three-dimensional residue graphs and processed using a graph neural network based on ProNet.[50] For a pocket comprising $K = 100$ residues with centroids $\{c_i\}_{i=1}^{K}$, edges are established between residues i and j when $\|c_i - c_j\|_2 < \delta$, where $\delta = 8$ Å. Through hierarchical message passing, the encoder captures localized chemical interactions and higher-order structural organization. The final pocket embedding is obtained by pooling across graph nodes.

Textual representations were derived from UniProt functional annotations[49] and processed using the BiomedBERT encoder,[51] a domain-adapted BERT model pretrained on biomedical literature. Given a textual annotation $T = (w_1, w_2,...,w_M)$, the encoder produces contextualized token-level representations, which are mean-pooled to obtain a fixed-dimensional text embedding.

The extended OneProt architecture incorporates MD trajectory representations using the latent MD encoder derived from MDGen,[52] a transformer-based generative framework that processes full MD trajectories as temporally ordered structural ensembles. MD trajectories used during pretraining were curated from mdCATH, GPCRmd, and ATLAS databases through MDRepo.[72] Consistent with the ablation framework, MD embeddings were not directly concatenated into downstream classifier inputs; their contribution arises indirectly through multimodal latent space shaping during pretraining.

During multimodal pretraining, modality-specific representations are projected into a shared latent space of dimension 1024 and aligned using a multimodal contrastive learning objective based on InfoNCE,[73] with the sequence modality serving as the central alignment anchor. The total alignment loss is computed across all available modality pairs, enabling the model to learn a unified representation spanning the sequence, structure, pocket geometry, functional annotations, and conformational dynamics. For complete details of the pretraining loss formulation, architectural hyperparameters, and optimization procedure, we refer the reader to the original OneProt model described in our recent study.[48] More details on encoder architecture are available in Supporting Information Text S1.

Pocket Construction and Structural Processing

For the PPI-Site data set, binding pockets were reconstructed from cavity annotations. The geometric centroid of the annotated ligand atoms defined the pocket center. Residue centroids were computed as the average coordinates of all the atoms in each residue. Residues were ranked by Euclidean distance to the ligand center, and the $K = 100$ nearest residues were retained to construct fixed-size pocket representations.

Unlike PPI-Site and KinSite, which provide ligand-associated structural pockets, ASD primarily provides curated residue-level annotations of allosteric regions. We therefore used these annotations directly to define the allosteric region, preserving the site definition provided by ASD rather than imposing a ligand-centered definition where one was not consistently available. Spatially contiguous annotated residues were identified using DBSCAN clustering.[74]

Given residue coordinates $\{x_i\}_{i=1}^{N}\mathbb{R}^3$, two residues are neighbors if $\|x_i - x_j\|_2 < \varepsilon\|$, where $\varepsilon = 10$ Å and minPts = 1. The largest cluster was retained and its centroid defined the pocket center, followed by selection of the 100 nearest residues.

Because this procedure differs from the ligand-centered construction used for PPI-Site and KinSite, we additionally performed a harmonized sensitivity analysis in which ASD pockets with identifiable experimentally resolved allosteric ligands were reconstructed by using the same ligand-centered procedure. Full reconstruction, matching, and exclusion criteria are provided in Supporting Information 2.7.

To assess structural overlap between allosteric and orthosteric pockets in PPI-Site, we compared pockets sharing the same protein and chain identifiers at both residue-composition and atom-coordinate levels. Residue-level overlap was quantified using the Jaccard similarity coefficient and the overlap coefficient. Coordinate-level similarity was assessed using KD-tree nearest-neighbor search, measuring the fraction of atoms in one pocket within 4 Å of any atom in the other pocket and the mean nearest-neighbor distance. We identified 100 pocket pairs (<5% of the data set) with identical residue composition (Jaccard = 1.0) but substantial coordinate-level divergence (mean distances ranging from <0.1 Å to >18 Å), indicating that allosteric and orthosteric regulation can emerge from distinct conformations within shared local environments.

Remaining details on pocket construction and analysis are available in Supporting Information Texts S2.1−2.6.

Encoder Ablation Studies

To investigate how different biological modalities contribute to allosteric discrimination, we conducted systematic multimodal ablation experiments across all data sets. Rather than treating modality selection solely as an architectural optimization problem, these experiments were designed to probe which biological signals encoded within the pretrained latent space remain informative under different structural and statistical regimes. The ablation strategy operates on two distinct levels. At the encoder level, we control which

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Modalities and Encoders Assessed for Their Contribution to Binding Site Prediction Performance

Abbreviation	Modality	Input Type	Information Captured	References
ST	Structural Token	Foldseek tokens	Global fold-level organization; long-range structural context	Foldseek[49]
SG	Structural Graph	Protein graph (nodes: residues; edges: spatial contacts)	Fine-grained local geometry; contact relationships	ProNet[50]
MD	Molecular Dynamics	MD trajectory frames	Conformational fluctuations; ensemble behavior; temporal dynamics	MDGen[52]

modalities participate in multimodal pretraining and, thus, shape the organization of the shared latent space. At the embedding level, we control which modality-specific embeddings are extracted from the pretrained backbone and provided as input features for the downstream classifier. This two-level design distinguishes between two complementary effects: (a) how the inclusion of different modalities during pretraining shapes latent-space organization and (b) how different combinations of extracted embeddings contribute to downstream classification. The OneProt framework allows for flexible inclusion or exclusion of modality-specific encoders during multimodal pretraining. All configurations retain the sequence encoder as the central anchor for the multimodal alignment. The pocket encoder is required to obtain binding site-specific representations for annotated pockets. The text encoder is retained because previous OneProt experiments demonstrated that textual functional annotations improve the multimodal alignment and downstream transfer performance.

Against this common sequence−pocket−text backbone, we systematically ablate three structural and dynamical encoders (Table 1). The first encoder, structural tokens (ST), employs token-based structural representations derived from Foldseek structural tokens. This modality captures global fold-level organization and long-range structural context by encoding protein structures as discrete tokens analogous to sequence tokens. The second encoder, structural graph (SG), utilizes graph-based structural representations from ProNet. Unlike the global token-based approach, this modality captures fine-grained local geometry and contact relationships by representing proteins as graphs of residues with edges defined by spatial proximity. The MD encoder leverages MD trajectory representations from MDGen.

We evaluated seven encoder configurations representing a progression from the minimal backbone to the full multimodal model. Pocket+Text is a baseline configuration. Only the sequence, pocket, and text encoders participated in pretraining. No structural or MD information was available. This establishes the performance floor against which all structural and dynamical contributions are measured. The ST+Pocket+Text encoder combination adds token-based structural representations (ST) to the baseline. This configuration tests whether global structural context provides useful signal for allosteric discrimination beyond sequence, pocket geometry, and functional annotations. The SG+Pocket+Text encoder combo adds graph-based structural representations (SG) to the baseline. This evaluates whether fine-grained local geometry and contact relationships provide complementary information for token-based representations. The MD+ST+Pocket+Text encoder adds MD trajectories to the token-based structural configuration. This evaluates whether conformational dynamics provide additional signal beyond the static structural context. The MD+SG+Pocket+Text encoder adds MD trajectories to the graph-based structural configuration. This evaluates whether dynamics complement the fine-grained static geometry. The ST+SG+Pocket+Text encoder combines both structural encoders without MD. This evaluates whether token-based and graph-based representations encode overlapping or complementary information. The MD+ST+SG+Pocket+Text encoder represents a full multimodal configuration. All encoders participated in pretraining. This establishes the performance ceiling. For each encoder configuration, we used the corresponding pretrained OneProt model with frozen weights. The configuration determines which modalities contribute to latent space organization during pretraining. Downstream classifiers then receive embeddings extracted from available modalities.

Embedding-Level Ablations: Feature Combinations for Downstream Classification

For each encoder configuration described above, we evaluate four different combinations of extracted embeddings as input features to the downstream MLP classifier (Table 2). These combinations allow us to compare the discriminative information accessible through local pocket-centered, global sequence-anchored, and functional-annotation representations, individually and in combination. We construct the following feature vectors, where $z_{\text{pocket}}, z_{\text{seq}}, z_{\text{text}} \in \mathbb{R}^{1024}$ denote the normalized embeddings extracted from the frozen OneProt backbone for a given protein sample.

Table 2. Feature Combinations for Downstream MLP Classification

Combination	Features	Dimension	Modalities Included
Pocket	z_pocket	1024	Structural (pocket only)
Pocket + Text	z_pocket + z_text	2048	Structural + Text
Pocket + Sequence	z_pocket + z_seq	2048	Structural + Sequence
Pocket + Sequence + Text	z_pocket+z_seq + z_text	3072	Structural + Sequence + Text

The pocket representation is generated by ProNet from the local three-dimensional binding-site graph, the sequence representation by ESM2 from the full protein amino-acid sequence, and the text representation by BiomedBERT from UniProt functional annotations. The corresponding embeddings are concatenated to construct the input feature vectors for the different ablation configurations. For each OneProt encoder configuration, a separate MLP classifier is trained for each available embedding combination. Because these representations are aligned during OneProt multimodal pretraining, the embedding-level ablations compare the discriminative information accessible through different pretrained representations and their combinations rather than decomposing performance into independent contributions from the underlying biological modalities. The experimental design can be summarized as follows. For encoder-level (what modalities shape the latent space), we employ seven configurations (Pocket+Text, ST+Pocket+Text, SG+Pocket+Text, MD+ST+Pocket+Text, MD+SG+Pocket+Text, ST+SG+Pocket +Text, MD+ST+SG+Pocket+Text). For embedding-level (what features are given to the classifier), we used four combinations (Pocket, Pocket+Text, Pocket+Sequence, Pocket+Sequence+Text). In total, the full experimental design involves 28encoder−embedding combinations per data set, with each combination evaluated across 10 random seeds, thus amounting to 1120 independent runs.

Downstream Allostery Classification

For downstream tasks, MLP classifiers were trained on embeddings extracted from the frozen OneProt backbone. The MLP consisted of an input layer matching the embedding dimension, hidden layers, and a single sigmoid output unit for binary classification. Hyperparameters, including the number and size of hidden layers, activation function, dropout rate, learning rate, batch size, and layer normalization, were optimized as described in the Supporting Information. For the selected architecture with two hidden layers, the classifier computes the output shown in (eq 6):

$$\hat{y} = \sigma(W_2\,\mathrm{ReLU}(W_1\,\mathrm{ReLU}(W_0\mathbf{e} + b_0) + b_1) + b_2) \quad (6)$$

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

where **e** denotes the extracted embedding vector and $\sigma(\cdot)$ is the sigmoid activation. Training minimized the binary cross-entropy loss defined in (eq 7):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{I=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{7}$$

The downstream MLP classifier was optimized by using the AdamW optimizer. Models were trained for up to 50 epochs with early stopping based on the validation loss (patience of 20 epochs). For each run, the checkpoint achieving the lowest validation loss was selected and subsequently evaluated on the validation and test sets. Each experimental configuration was evaluated across ten independent random seeds. For each run, the model achieving the highest validation ROC-AUC was selected, and the corresponding test metrics were aggregated across the runs. Details of the downstream hyperparameter search are provided in Supporting Information. For AlloDiverse, positive samples (N_{allo} = 15,904) substantially outnumbered negative samples (N_{ortho} = 5,203). To mitigate training bias, positive instances were randomly subsampled during each training iteration to maintain balanced minibatches ($B/2$ positive and B/2 negative per batch), while validation and test distributions remained unchanged.

As an additional control, we evaluated simple handcrafted representations to determine how much discrimination could be achieved from coarse protein- and pocket-level characteristics without using OneProt embeddings. Whole-protein sequence representations comprised of the relative frequencies of the 20 standard amino acids, while pocket representations combined amino acid composition with simple geometric descriptors of pocket compactness and spatial extent. These low-dimensional features were evaluated using L1-regularized logistic regression, with regularization strength selected by cross-validation on the training data. Full feature definitions, preprocessing, model selection, and evaluation procedures are provided in Supporting Information Text S3.5.

Within-Protein Pocket Comparison

To determine whether pocket-dependent discrimination persists when global protein identity is held constant, we performed an additional within-protein analysis on the held-out AlloDiverse test set. We identified proteins represented by both allosteric and orthosteric pockets based on identical full-chain protein sequences. 10 proteins met this criterion, yielding 47 within-protein allosteric−orthosteric pocket pairs. No additional training or model fitting was performed for this analysis. Instead, we used the predicted allosteric probabilities already generated by the corresponding classifiers for these samples in the standard AlloDiverse test-set evaluation. For each trained classifier, we compared the predicted allosteric probability of each allosteric pocket with that of each orthosteric pocket belonging to the same protein. Pairwise ranking accuracy was defined as the fraction of comparisons in which the allosteric pocket received a higher predicted probability than the orthosteric pocket, with ties assigned a value of 0.5. We calculated the metric, obtained by first calculating pairwise ranking accuracy separately for each protein and then averaging equally across proteins. The analysis was performed separately for each random seed and encoder−embedding configuration. Because pockets belonging to the same protein share identical sequence and UniProt text representations, the pocket representation is the only input component that varies within each allosteric−orthosteric comparison. Any difference in predicted probability within a pair must therefore depend on pocket-varying information, although the way this information is used by the classifier may depend on the broader representation context learned during training. Above-chance within-protein ranking therefore demonstrates that discrimination depends on pocket-varying information and cannot be explained by global protein identity, sequence, or protein-level annotation alone. Full information on the comparison procedure is available in the Supporting Information Text S3.4.

Variance Decomposition Analysis

To quantify the relative contributions of data set composition, embedding selection, encoder architecture, and stochastic variation, we performed factorial ANOVA. Each run was represented by ROC-AUC with four categorical factors: Data set (4 levels), Embedding (4 levels), Architecture (7 levels), and Random Seed (10 levels). A global ANOVA model was first fitted across all experimental runs to estimate the overall contributions of data set, embedding combination, and encoder architecture together with their pairwise interactions as specified in (eq 8):

$$\mathrm{ROC-AUC} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl} \tag{8}$$

where α_i, β_j, γ_k, and δ_l denote data set, embedding combination, encoder architecture, and random seed effects respectively, while $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2)$is the residual error.

To characterize modality contributions within each biological regime, separate ANOVA models were subsequently fitted for each data set. Since the data set factor is fixed in these analyses, the model includes only embedding combination, encoder architecture, and their interaction, as given in (eq 9):

$$\mathrm{ROC-AUC} = \mu + \beta_j + \gamma_k + \delta_l + (\beta\gamma)_{jk} + \varepsilon_{jkl} \tag{9}$$

For both the global and data set-specific analyses, sums of squares were normalized by the total sum of squares to obtain the percentage variance explained by each factor using (eq 10):

$$\text{Variance Contribution}(\theta) = \frac{\mathrm{SS}_\theta}{\mathrm{SS}_{\mathrm{total}}} \times 100\% \tag{10}$$

All statistical computations were performed in Python 3.10 using SciPy (v1.10) for hypothesis testing and statsmodels (v0.14)[75] for multiple comparison correction. Visualization of distributions was generated with Matplotlib (v3.7)[76] and GraphPad Prism version 11.0.

Performance Evaluation

Classification performance on the predefined test sets was evaluated using complementary ranking-, threshold-, and probability-based metrics. Receiver operating characteristic area under the curve (ROC-AUC) was used as the primary threshold-independent ranking metric. Because class prevalence varied substantially across data sets, particularly in the highly imbalanced KinSite data set, we additionally report precision-recall area under the curve (PR-AUC). Threshold-dependent performance was characterized using accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC), with MCC providing a balanced measure of classification performance under class imbalance. Probabilistic prediction quality was assessed using the Brier score, where lower values indicate better probabilistic predictions, and these were compared with a prevalence-only predictor that assigns the training-set positive-class prevalence to all test samples. Metrics were computed independently for each random seed on the corresponding predefined test set and were reported as the mean across seeds, with variability expressed as the standard deviation across random seeds. Where applicable, 95% confidence intervals are additionally reported. Complete metric values for all data sets, architectures, and embedding configurations are provided in the Supporting Text S3.2.

RESULTS

A Multi-Stage Analytical Framework for Investigating Multimodality in Differentiating Orthosteric and Allosteric Binding Sites

To systematically investigate how multimodal protein representations capture allosteric regulation across diverse biological contexts, we designed a staged analytical framework that progresses from data set characterization through modality

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897149/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

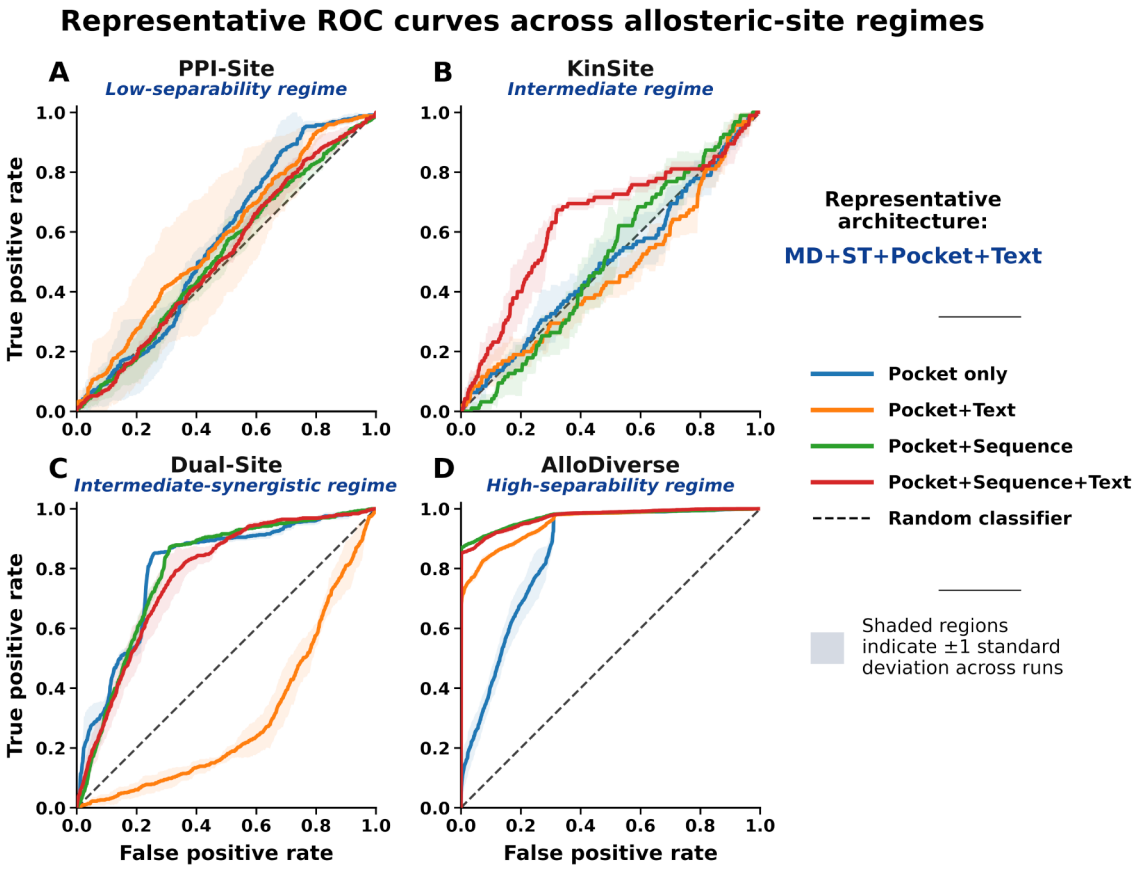


Figure 2. ROC curves for OneProt performance across data sets and embedding combinations. Panels show results for the four data sets: (A) PPI-Site, (B) KinSite, (C) Dual-Site, and (D) AlloDiverse. Receiver operating characteristic (ROC) curves illustrate the predictive performance of a representative OneProt architecture including sequence, text, structure token, and molecular dynamics encoders across four data sets. Different colors correspond to the different embedding combinations used for predictions: Pocket only (blue), Pocket+Text (orange), Pocket+Sequence (green), and Pocket+Sequence+Text (red). The dashed diagonal line represents random performance (AUC = 0.5). Each curve represents the average across ten independent random seeds; shaded regions indicate ±1 standard deviation. PPI-Site shows the weakest discrimination (curves near the diagonal), while AlloDiverse achieves near-ceiling performance across all embedding combinations. KinSite and Dual-Site occupy intermediate regimes with progressive improvement. The vertical dashed lines in each panel indicate the operating points corresponding to the best-performing embedding combination for each data set, highlighting how the optimal sensitivity-specificity trade-off varies across regimes.

ablation to interpretative synthesis (Figure 1). This framework integrates three complementary pillars: (a) a curated set of protein-binding site data sets spanning distinct structural and statistical regimes, (b) the OneProt multimodal foundation model with systematic encoder and embedding ablations, and (c) a variance decomposition framework to quantify the relative contributions of data set composition, modality selection, and architectural choice. Our analytical strategy proceeds through four interconnected stages.

We first characterize the intrinsic separability of allosteric and orthosteric binding environments across data sets with varying structural diversity, conformational heterogeneity, and class balance. We then systematically ablate pocket-containing embedding combinations—testing pocket-only, pocket+text, pocket+sequence, and full multimodal configurations—to determine which biological modalities contribute most strongly to allosteric discrimination under different data set conditions. In parallel, sequence-only and text-only representations are evaluated as single-modality controls to determine how much discrimination is accessible from global protein context in the absence of explicit pocket-specific information. Next, we evaluate distinct encoder architectures, including structure-token (ST), structure-graph (SG), and MD encoders, to assess whether the specific implementation of structural and dynamical representations influences predictive performance. Finally, we perform variance decomposition to quantify the relative contributions of data set identity, embedding composition, and architectural choice, establishing a rigorous statistical basis for our conclusions.

This staged design addresses three central questions: First, does the utility of multimodal integration depend on the intrinsic separability of the data set? Second, which modality representations provide complementary information for allosteric discrimination, and does this vary across biological contexts? Third, can we identify distinct learnability regimes that guide the application of multimodal protein foundation models to allostery prediction?

Overall Performance Reveals a Hierarchy of Data Set Separability

Across all evaluated architectures, embedding combinations, and random seeds, a clear hierarchy of classification performance emerged (Figure 2). The PPI-Site data set consistently represented the most challenging prediction regime, with ROC curves remaining close to the diagonal and median ROC-AUC values only modestly above random expectation (AUC ≈0.55−0.60) (Figure 2A). This near-random performance

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

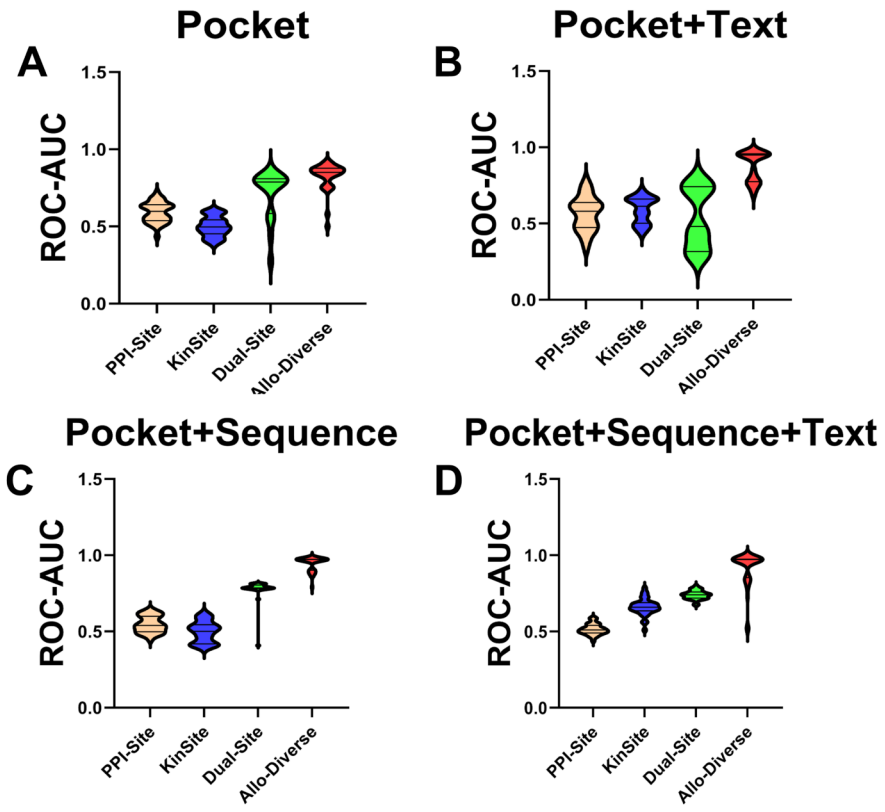


Figure 3. Distribution of ROC-AUC values across all encoder architectures and random seeds. Panels correspond to the four data sets: (A) PPI-Site, (B) KinSite, (C) Dual-Site, and (D) AlloDiverse. Violin plots show the distribution of ROC-AUC values across all encoder architectures and random seeds for each data set and embedding combination. Different colors indicate embedding combinations: Pocket only (blue), Pocket+Text (orange), Pocket+Sequence (green), and Pocket+Sequence+Text (red). Each point represents a single run (one architecture, one random seed). The horizontal dashed line indicates random performance (AUC = 0.5). The figure reveals a systematic progression across data sets: PPI-Site (low-separability), KinSite (intermediate), Dual-Site (intermediate-synergistic), and AlloDiverse (high-separability).

indicates that allosteric and orthosteric sites in PPI-Site (Figure 2A) share extensive structural and functional similarities that current representations cannot reliably resolve. The challenge arises from the data set-defining characteristic: class labels are assigned based on spatial relationships to protein−protein interaction interfaces rather than intrinsic physicochemical differences between binding sites, information not explicitly provided to the classifier. Consequently, the same protein, the same ligand-binding pocket, and often the same set of residues can be classified as either allosteric or orthosteric depending on subtle conformational arrangements relative to a protein−protein interface.

The KinSite and Dual-Site data sets occupied intermediate positions between these extremes (Figure 2B,C). KinSite yielded AUC values ranging from approximately 0.55 to 0.70 (Figure 2B), with stronger configurations reaching ROC-AUC values of up to 0.75 (Supporting Information, Figure S3D). Given the strong class imbalance, prevalence-sensitive metrics provide an important complementary perspective on this performance. Pocket-containing configurations with ROC-AUC values around 0.66−0.68 typically achieved PR-AUC values of approximately 0.07−0.09 and MCC values around 0.08−0.13, whereas stronger configurations reached PR-AUC values of approximately 0.18−0.21 and MCC values of approximately 0.15−0.20 (Supporting Information, Tables S6−S11). Thus, KinSite contains appreciable class-discriminative signal, although its magnitude is more moderate when assessed using metrics sensitive to the rare allosteric class than suggested by ROC-AUC alone. This pattern is consistent with the challenge of discriminating conserved ATP-binding sites from structurally heterogeneous allosteric pockets under extreme class imbalance ($\rho \approx 51.1$).

The moderate performance improvement over PPI-Site arises from the strong evolutionary signals that distinguish orthosteric ATP-binding clefts—which are under intense purifying selection—from allosteric regulatory sites, which are evolutionarily permissive. Dual-Site exhibited progressively improved discrimination performance (AUC ≈0.77−0.80) relative to either constituent data set alone (Figure 2C). Importantly, this improvement was also evident in PR-AUC and MCC for the stronger sequence-containing representations, indicating that the gain is not restricted to ROC ranking. Because the pretrained OneProt encoders remain frozen during downstream training, this improvement should not be interpreted as the merged data set causing OneProt to learn new or more robust representations. Rather, combining PPI-Site and KinSite provides complementary training signals that allow the downstream classifier to establish a more effective decision boundary within the pretrained representation space. The extent to which this gain reflects data set provenance, global protein-level information, or pocket-level information is examined below.

In contrast, the AlloDiverse data set yielded near-ceiling performance across nearly all evaluated configurations (AUC ≈0.90−0.98), indicating the presence of a strong and readily separable signal distinguishing allosteric and orthosteric sites

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(Figure 2D). The expanded metrics confirm that this is not solely an ROC-ranking effect: high-performing configurations also achieved PR-AUC values of approximately 0.98−0.99 and MCC values of approximately 0.81−0.84. Their Brier scores were approximately 0.06−0.08, substantially below the prevalence-only Brier baseline of approximately 0.24, indicating substantially improved probabilistic prediction quality relative to a prevalence-only predictor. Nevertheless, as shown by the source and single-modality controls below, a substantial component of this separability is associated with data set provenance and global protein context. The incorporation of ASD substantially broadens the range of protein families and allosteric-site architectures represented in this setting, creating multiple potential sources of discriminative signal that are accessible across different OneProt representations. As shown by the source and single-representation controls below, however, the high separability of AlloDiverse is not attributable exclusively to binding-site-level information, and its interpretation therefore requires consideration of the composition of the assembled data set. Full details on the ROC-AUC curves are provided in Supporting Information Figure S1 and Table S4. Since the extreme class imbalance in KinSite makes precision−recall analysis particularly relevant, we additionally report PR-AUC in Supporting Information Figure S2 and Table S5. Information on the other metrics is available in the Supporting Text S3.2 and Supporting Information Tables S6−S11.

The data set-dependent differences in separability became even more apparent when performance distributions across all encoder architectures and random seeds were examined collectively (Figure 3).The PPI-Site data set formed a low-learnability regime characterized by broad overlap with random performance (distributions centered near a median ROC-AUC of 0.55) and substantial sensitivity to model configuration (Figure 3A). Critically, embedding composition had a minimal effect within this regime: pocket-only, pocket +text, pocket+sequence, and full multimodal combinations all produced overlapping distributions, indicating that additional modalities could not compensate for the weak underlying discriminative signal. This observation underscores a limitation of the PPI-Site task formulation: when allosteric and orthosteric sites share structural and evolutionary contexts, even sophisticated multimodal representations cannot reliably distinguish them.

The KinSite data set exhibited modest but reproducible improvements over PPI-Site, with AUC distributions centered around 0.70−0.75 and relatively compact interquartile ranges (Figure 3B). Unlike PPI-Site, embedding composition began to differentiate performance here: sequence-containing embeddings (pocket+sequence and pocket+sequence+text) consistently outperformed pocket-only and pocket+text combinations. This suggests that while overall performance is constrained by the extreme class imbalance ($\rho \approx 51.1$), the evolutionary signals captured by sequence context provide a consistent discriminative advantage. The compact interquartile ranges further indicate that this signal is robust across architectures, with structural and MD encoders providing modest but reproducible gains.

The combined Dual-Site data set shifted toward substantially higher ROC-AUC values (AUC ≈0.80−0.85), with broader distributions reflecting the increased complexity and diversity of the merged data set (Figure 3C). The widening of distributions compared to KinSite indicates that combining data sets with complementary biological characteristics—structural diversity from PPI-Site and evolutionary signals from KinSite—creates a more heterogeneous prediction landscape. Within this regime, embedding composition exerted its strongest effect: the gap between pocket-only and sequence-containing embeddings widened considerably, with pocket +sequence+text achieving the highest median performance. Architectural effects also became more apparent, as the addition of structural encoders consistently shifted distributions toward higher AUC values.

Finally, incorporation of ASD-derived allosteric sites produced a highly separable regime in which most evaluated models achieved consistently strong performance (AUC ≈0.95−0.98) with remarkably narrow distributions (Figure 3D). Pocket-only representations showed considerably lower discrimination, whereas multimodal representations, particularly those incorporating sequence or both text and sequence context, reached ROC-AUC values above 0.95 and up to approximately 0.97−0.98 for the strongest configurations. As demonstrated by the controls below, data set provenance and associated protein-family composition provide a strong baseline signal in AlloDiverse. However, the strongest multimodal representations exceed the performance attainable from source identity alone, indicating that provenance contributes substantially to, but does not fully account for, the high-separability regime.

Variance Decomposition Identifies Data Set Regimes as Primary Performance Determinants

The observed performance hierarchy remained stable despite extensive variation in both embedding composition and encoder architecture. To quantify the relative contributions of these factors, we performed variance decomposition across all evaluated data sets, embedding combinations, architectures, and random seeds (Figure 4). This analysis partitions the total variance in ROC-AUC into components attributable to data set identity, embedding composition, encoder architecture, their interactions, and stochastic variation, providing a rigorous statistical basis for understanding what drives performance differences. Data set identity accounted for the dominant fraction of explainable variance, at 63.7%, substantially exceeding the contributions of embedding composition (1.4%) and architecture choice (3.4%). Random seed effects were negligible (<0.5%). This overwhelming dominance of data set identity indicates that the primary determinant of predictive performance is not architectural optimization or modality selection, but rather the separability of allosteric and orthosteric binding environments within each data set. In other words, the biological signal available for discrimination is largely fixed by the data itself, and model design choices can only marginally modulate what is already present. Notably, interaction terms between data set and representation components explained additional variance, with data set−embedding and data set−architecture interactions contributing 8.3% and 5.8%, respectively (Figure 4, left). These interactions are particularly informative: they reveal that neither embedding choice nor architecture exerts a universal effect on performance. Instead, their influence depends strongly on the biological context represented by a particular data set. For example, sequence embeddings that are essential for discriminating kinase binding sites may be irrelevant or even detrimental when applied to PPI-Site data, where allosteric and orthosteric pockets share evolutionary and functional contexts.

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897149/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

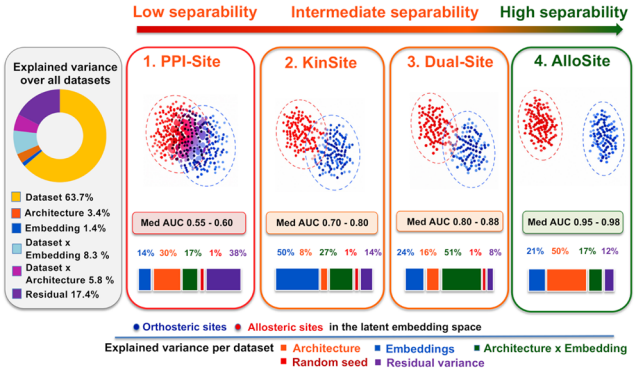


Figure 4. Variance decomposition of ROC-AUC across data sets, embedding combinations, architectures, and random seeds. (Main panel) global variance decomposition across all experimental runs, showing the proportion of total variance attributable to data set identity, embedding composition, encoder architecture, their pairwise interactions, random seed, and residual error. Data set identity accounts for the dominant fraction (63.7%), followed by data set−embedding interaction (8.3%) and data set−architecture interaction (5.8%). (Panels 1−4) Data set-specific variance decompositions for PPI-Site, KinSite, Dual-Site, and AlloDiverse, respectively, showing the proportion of variance attributable to embedding composition, encoder architecture, their interaction, random seed, and residual error within each data set. Each panel is scaled to its own total variance, enabling direct comparison of the relative importance of each factor within each regime. Shaded regions represent the proportion of variance explained by each component.

This context dependence motivates the data set-specific analysis that follows.

To further examine these context-dependent effects, we performed variance decomposition separately for each data set (Figure 4). These analyses reveal how the determinants of performance shift systematically across the separability spectrum, providing insights into the underlying biological signals available in each regime.

In the low-separability PPI-Site regime, architectural variation explained the largest fraction of performance variance (30%), although a substantial residual component (38%) remained unexplained. This pattern reflects the weak discriminative signal and high uncertainty observed across all configurations: when allosteric and orthosteric sites are nearly indistinguishable, architectural refinements become the primary source of performance differences, but they are fundamentally limited by the absence of a strong biological signal. The large residual variance further underscores that performance in this regime is dominated by irreducible uncertainty rather than systematic factors under our control. In KinSite, embedding composition dominated the variance (50%), indicating that modality selection is the primary determinant of performance in this intermediate regime. This finding aligns with the biological dichotomy of kinase binding sites: orthosteric ATP-binding clefts generate strong sequence conservation signatures that protein language models can capture, while allosteric regulatory sites produce weak or absent sequence signals. In this regime, the critical decision is selecting the right biological signal—sequence, text, or structural—rather than choosing a particular encoder to extract it. The dominance of embedding composition over architecture suggests that once the appropriate modality is selected, the specific implementation matters less. The Dual-Site data set exhibited the strongest architecture × embedding interaction (50%), demonstrating that predictive performance depends not only on the chosen representations but also on how effectively a particular architecture exploits them.

This synergy reflects the increased complexity of the merged data set, where combining structural diversity (from PPI-Site) and evolutionary signals (from KinSite) creates complementary information sources. The strong interaction indicates that the optimal representation strategy must be matched to the appropriate architectural implementation—a combination that is not obvious from either data set alone. In the highly separable AlloDiverse data set, architecture emerged as the dominant contributor (50%), while residual variance remained low. This indicates that once strong discriminative information is available, architectural differences become the principal determinant of how effectively the pretrained representations exploit that information. This result is complementary to the source and single-representation controls described below: those analyses address the origins of the high baseline separability of AlloDiverse, whereas the variance decomposition identifies the factors governing relative performance among OneProt configurations within that setting. The convergence of several structurally informed architectures at high performance further suggests that architectural design modulates how effectively the available multimodal information is represented and integrated in this high-separability regime.

Collectively, these data set-specific variance decompositions reveal a systematic progression in the determinants of performance across the separability spectrum (Table 3). The dominant contribution of data set identity suggests that the primary challenge in allosteric site prediction is not necessarily architectural optimization but rather the separability of allosteric and orthosteric binding environments within a given data set. In the PPI-Site data set, all evaluated models remained clustered near-random performance regardless of representation choice, suggesting substantial overlap between allosteric and orthosteric competitive sites within the learned latent space. By contrast, the AlloDiverse data set contained a strong discriminative signal that was readily captured by a broad range of representations. The KinSite and Dual-Site data sets occupied an intermediate regime in which performance differences between architectures and embedding combinations became more apparent (Table 3).

Table 3. Summary of Separability Regimes and Key Determinants of Performance

Data set	Regime	Dominant Modality	Key Architectural Finding
PPI-Site	Low	Pocket geometry	All architectures perform near-random (AUC ∼0.55−0.58); weak signal dominates.
KinSite	Intermediate	Pocket + Text	Structural encoders provide modest gains; modality selection drives performance.
Dual-Site	Intermediate	Sequence	Architecture−embedding interaction is the primary determinant of performance.
AlloDiverse	High	Sequence + Text	Multiple structural architectures achieve near-ceiling performance (∼0.96−0.98 AUC).

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

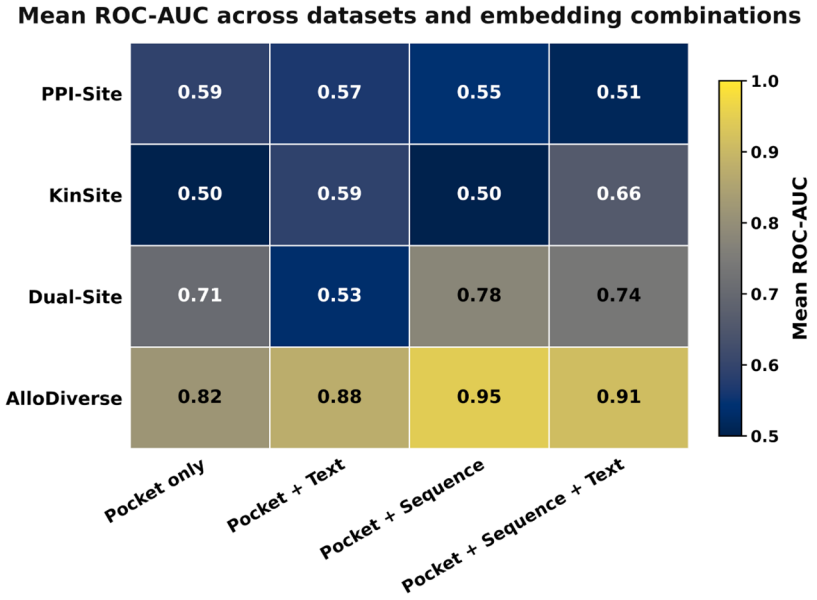


Figure 5. Mean ROC-AUC heatmap across data sets and embedding combinations. Rows correspond to the four data sets—PPI-Site (0.59, 0.57, 0.55, 0.51), KinSite (0.50, 0.59, 0.50, 0.66), Dual-Site (0.71, 0.53, 0.78, 0.74), and AlloDiverse (0.82, 0.88, 0.95, 0.91)—ordered from low to high separability. Columns correspond to the four embedding combinations: Pocket only, Pocket+Text, Pocket+Sequence, and Pocket+Sequence+Text. Color intensity ranges from blue (low AUC) to red (high AUC), with numerical values displayed in each cell. The relative ordering of embedding combinations changes systematically across data sets: pocket-only performs best in PPI-Site, while text and sequence embeddings excel in other regimes. This regime-dependent progression reflects the distinct biological signals available in each data set.

Contribution of Multimodal Embeddings across Separability Regimes

Having established that data set composition is the dominant determinant of performance and that modality selection drives performance in intermediate regimes, we next examined how different embedding combinations contribute within each separability regime.

The average ROC-AUC values across data sets and embedding combinations revealed that no universally optimal representation strategy exists (Figure 5). Instead, the utility of individual modality embeddings depended strongly on data set composition, reflecting the distinct biological signals available in each regime. Importantly, performance did not increase monotonically with the number of concatenated modality embeddings: the highest-dimensional Pocket+Sequence+Text representation was not consistently optimal and, in some data sets, performed worse than lower-dimensional representations. This nonmonotonic behavior argues against representation dimensionality alone as the primary explanation for differences between embedding combinations.

In the low-separability PPI-Site regime, pocket-only embeddings achieved the strongest average performance (AUC = 0.59 ± 0.04 SD), while the addition of text and sequence representations progressively reduced predictive performance: Pocket+Text (AUC = 0.57 ± 0.08 SD), Pocket +Sequence (AUC = 0.55 ± 0.03 SD), and Pocket+Sequence +Text (AUC = 0.51 ± 0.03 SD) (Figure 5). Consistent with this pattern, sequence-only and text-only OneProt representations were not independently discriminative, achieving mean AUROC values of approximately 0.54 ± 0.04 SD and 0.47 ± 0.05 SD, respectively. Thus, little discriminatory information is accessible through these representations in PPI-Site without explicitly providing the local pocket representation to the downstream classifier. The expanded metrics support this interpretation: PR-AUC and MCC remained modest across representations, and several configurations showed Brier scores close to or above the prevalence-only baseline. Overall, no representation provided robust discrimination in this regime, consistent with the absence of explicit PPI-interface information from the model inputs (Supporting Information, Tables S6−S10). This striking result—where adding more information consistently degrades performance, with the full multimodal combination performing worst—reflects the biological characteristics of the PPI-Site data set. As established in our data set characterization, allosteric and orthosteric sites in PPI-Site share the same protein family context, the same structural architecture, and often the same residue composition, differing only in conformational arrangement relative to protein−protein interaction interfaces. Under these conditions, global evolutionary and functional context—which is largely shared between the two pocket classes within a given protein—introduces variance rather than complementary information. Moreover, the pocket embedding retains local three-dimensional structural information through the ProNet graph encoder and may therefore capture geometric features that indirectly correlate with this interface-related distinction. The pocket-overlap analysis supports this interpretation: we identified 100 pocket pairs (<5% of the data set) with identical residue composition (Jaccard = 1.0) but substantial coordinate-level divergence (mean distances ranging from <0.1 Å to >18 Å), indicating that regulatory discrimination depends primarily on subtle local structural differences. This finding aligns with the variance decomposition analysis, where residual variance dominated PPI-Site (38%), reflecting the weak discriminative signal and irreducible uncertainty that limit performance regardless of embedding strategy.

A fundamentally different pattern emerged in the KinSite data set. Here, embedding combinations produced a clear hierarchy: Pocket-only (AUC = 0.50 ± 0.05 SD) < Pocket +Sequence (AUC = 0.50 ± 0.04 SD) < Pocket+Text (AUC =

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

0.59 ± 0.03 SD) < Pocket+Sequence+Text (AUC = 0.66 ± 0.03 SD) (Figure 5). The single-modality controls reinforce the differential contribution of the global modalities: sequence-only representations remained close to chance on average (AUROC ≈0.51 ± 0.08 SD), whereas text-only representations achieved appreciably higher discrimination (AUROC ≈0.63 ± 0.08 SD). Thus, the pretrained text representation provides a measurable protein-level signal in KinSite, whereas sequence-only representations contribute little on average, with neither global representation yielding consistently stable discrimination across seeds. This distinction is broadly reflected in the pocket-containing combinations, where adding text to the pocket representation improved performance (AUC = 0.59 ± 0.03 SD), while adding sequence alone did not (AUC = 0.50 ± 0.04 SD). The improvement from pocket-only to the full multimodal representation (+0.16 AUC) indicates that incorporating global protein-level representations alongside the pocket embedding substantially improves discrimination in this regime, with the single-modality and embedding-combination results pointing to the text representation as the stronger individual contributor. The expanded metrics further characterize this representation-dependent improvement under the severe KinSite class imbalance. Configurations achieving the strongest ROC-AUC also generally showed improved positive-class performance, with the best-performing models reaching precision of approximately 0.12, while recall varied substantially across configurations and reached approximately 0.75−0.80 in some cases. PR-AUC and MCC nevertheless remained modest in absolute terms, reflecting the difficulty of identifying the rare allosteric class in this strongly imbalanced data set. Thus, the additional metrics broadly support the representation-dependent differences observed by ROC-AUC, while showing that improved discrimination does not eliminate the challenge imposed by the low prevalence of allosteric examples. This pattern reflects the biological characteristics of the KinSite data set: while kinase ATP-binding clefts generate strong sequence conservation signatures, the most informative signal for distinguishing allosteric from orthosteric sites comes from functional annotations that capture protein-specific regulatory mechanisms—mechanistically diverse processes such as myristoyl pocket binding in ABL, αC-helix displacement in BRAF, and activation loop phosphorylation across many kinases. This finding is consistent with the variance decomposition analysis, where embedding composition dominates performance in this intermediate regime (50% of explained variance).

In the Dual-Site data set, the embedding hierarchy shifted compared to KinSite: Pocket+Text (AUC = 0.53 ± 0.03 SD) < Pocket-only (AUC = 0.71 ± 0.03 SD) < Pocket+Sequence +Text (AUC = 0.74 ± 0.02 SD) < Pocket+Sequence (AUC = 0.78 ± 0.03 SD) (Figure 5). The single-modality controls further revealed that sequence-only representations achieved a mean AUROC of approximately 0.82 ± 0.01 SD, exceeding the average performance of the pocket-containing combinations, whereas text-only representations reached approximately 0.72 ± 0.09 SD. Thus, global sequence-anchored representations provide a particularly strong and stable ranking signal in Dual-Site, while the lower performance of Pocket+Sequence+Text relative to Pocket+Sequence shows that adding further representations is not uniformly beneficial.

The sequence-only advantage in ROC-AUC, however, did not extend uniformly to threshold-dependent classification. For several matched architectures, Pocket+Sequence achieved higher MCC despite somewhat lower AUROC. Most notably, for SG+Pocket+Text, sequence-only achieved AUROC 0.817 ± 0.035 but MCC 0.092 ± 0.245, whereas Pocket+Sequence achieved an AUROC of 0.785 ± 0.002 and an MCC of 0.37 ± 0.041. Similar, although smaller, differences were observed for other architectures. These results indicate that global sequence-anchored representations provide strong overall ranking in Dual-Site, while the inclusion of pocket information can improve balanced class discrimination for particular encoder architectures.

This pattern aligns with the variance decomposition finding that the architecture-embedding interaction is the primary determinant of performance in Dual-Site (50%), reflecting the increased complexity of the merged data set, where different modalities contribute differently depending on the biological context.

In the AlloDiverse data set, the hierarchy was: Pocket-only (AUC = 0.82 ± 0.03) < Pocket+Text (AUC = 0.88 ± 0.02) < Pocket+Sequence+Text (AUC = 0.91 ± 0.02) < Pocket +Sequence (AUC = 0.95 ± 0.01) (Figure 5). The single-modality controls showed that substantial discrimination was already accessible from global protein-level representations: sequence-only achieved a mean AUROC of approximately 0.92 ± 0.02, while text-only reached approximately 0.87 ± 0.02. The strong performance of both sequence-only and Pocket +Sequence representations identifies the pretrained sequence representation as a major contributor in this regime, while the lower performance of Pocket+Sequence+Text relative to Pocket+Sequence indicates that adding text is not uniformly beneficial.

The expanded metrics confirm that the strong performance of successful AlloDiverse models is not restricted to ROC ranking. The strongest pocket-containing architectures achieved PR-AUC values of approximately 0.98−0.99 and MCC values of approximately 0.81−0.84, together with Brier scores of approximately 0.06−0.08, substantially below the prevalence-only baseline of approximately 0.24. Direct comparison with sequence-only representations nevertheless indicates an additional contribution from pocket information. MD+ST Pocket+Sequence achieved an AUROC of 0.973 (95% CI 0.971−0.976) and MD+ST+SG+Pocket+Text achieved 0.973 (95% CI 0.972−0.973), compared with 0.969 (95% CI 0.966−0.971) for the strongest sequence-only configuration (ST+SG). Pocket+Sequence also produced lower Brier scores than sequence-only in five of seven matched encoder configurations. Thus, pocket information can provide additional ranking and probabilistic information beyond the already strong sequence representation, although this benefit is architecture-dependent.

Taken together, these results indicate that pocket information contributes to discrimination in AlloDiverse beyond the signal captured by protein-level representations, although a substantial component of the observed separability remains accessible without explicit pocket information. The extent to which this high separability reflects data set provenance and associated protein-level differences is examined directly in the following section.

This behavior is consistent with the variance decomposition result showing that architecture becomes the dominant factor (50%) in this high-separability regime, while residual variance remains low—indicating that once sufficient biological diversity is present, architectural refinements rather than

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

modality selection become the primary source of performance differences.

Collectively, these results demonstrate that the contribution of multimodal information is highly context-dependent and follows a systematic progression across the separability spectrum. In low-separability settings where allosteric and orthosteric sites share structural and evolutionary contexts, additional modalities introduce noise rather than complementary information. In intermediate regimes where differential evolutionary signals distinguish regulatory classes, sequence and text embeddings provide essential discriminative power. In high-separability settings, where broad structural and functional diversity creates strong discriminative signals, multimodal integration provides marginal but consistent improvements. This context-dependent behavior, summarized in Table 3, has direct implications for model development: the optimal embedding strategy must be matched to the biological characteristics of the target data set.

Importantly, these modality-dependent differences do not follow a simple relationship with the representation size or the number of concatenated embeddings. Increasing the number of embeddings did not produce monotonic improvements in performance: pocket-only representations performed as well as or better than multimodal combinations in PPI-Site, whereas specific additional modalities improved performance in KinSite, Dual-Site, and AlloDiverse. Moreover, the largest representation, Pocket+Sequence+Text, was not uniformly optimal across data sets. The optimal combination varied across data sets, indicating that the contribution of additional embeddings depends on the information relevant to each classification regime rather than representation complexity alone.

The complete ROC-AUC matrix across architectures and embedding combinations is provided in Supporting Table S4 and Supporting Figure S3, with corresponding PR-AUC and threshold-dependent metrics reported in Supporting Information Tables S6−S10 and Supporting Information Figures S4,S5.

Dissecting the Origin of the Data Set-Dependent Performance Gains

To determine whether the improved Dual-Site performance could be explained by data set provenance, we first constructed a source-only baseline. Because KinSite is strongly dominated by orthosteric examples, KinSite samples were assigned to the orthosteric class, whereas PPI-Site samples were assigned according to its approximately balanced class distribution. This source-only classifier achieved a ROC-AUC of 0.66. Provenance therefore carries appreciable discriminatory information but does not fully explain Dual-Site performance: sequence-only representations achieved a mean ROC-AUC of approximately 0.82 ± 0.01 across encoder variants, text-only representations achieved approximately 0.72 ± 0.09, and individual pocket-only configurations reached approximately 0.84 ± 0.01. Thus, Dual-Site discrimination cannot be attributed solely to data set source and is accessible from both global protein-level and local pocket representations.

The corresponding provenance effect was substantially stronger in AlloDiverse. A source-only classifier assigning ASD-derived samples to the positive class and samples from the constituent negative data sets to the negative class achieved ROC-AUC values of 0.931−0.935. Sequence-only and text-only representations achieved lower mean ROC-AUC values of approximately 0.92 ± 0.02 and 0.87 ± 0.02, respectively. Thus, data set provenance alone accounts for a substantial component of AlloDiverse separability, and neither global representation independently exceeds this source-associated signal on average. At the same time, the strongest pocket-containing OneProt representations reached approximately 0.97−0.98 (±0.003), indicating that source identity alone does not fully account for the achievable discrimination.

The within-protein analysis of the held-out AlloDiverse test set identified 10 proteins containing both allosteric and orthosteric pockets, comprising 47 allosteric−orthosteric pocket pairs (Supporting Information, Table S3). Several pocket-containing representations achieved substantial within-protein ranking accuracy across seeds. For Pocket+Sequence +Text embeddings, ranking accuracy reached 0.817 ± 0.071 for the MD+ST+SG+Pocket+Text encoder, 0.775 ± 0.127 for ST+SG+Pocket+Text, and 0.856 ± 0.026 for ST+Pocket +Text. Because sequence and text representations are identical for pockets belonging to the same protein, successful within-protein ranking necessarily depends on pocket-varying information, although its effective use remains representation-dependent. Thus, although the source-only and sequence-only controls show that substantial overall AlloDiverse discrimination can be achieved from data set provenance and global protein-level representations, the within-protein analysis provides complementary evidence for additional pocket-dependent discrimination. Given the limited number of proteins represented by both pocket classes, we interpret this analysis as a targeted control rather than as an independent estimate of overall predictive performance.

As an additional nonfoundation-model control, we evaluated an L1-regularized logistic-regression baseline using simple pocket geometry and amino-acid composition features (Supporting Text 3.5 and Table S2). Performance was strongly data set-dependent, with the baseline performing particularly well on KinSite (AUROC = 0.763 ± 0.004), indicating that substantial class-associated signal in this data set is accessible from simple handcrafted features. In contrast, the baseline remained substantially below OneProt representations in Dual-Site and AlloDiverse, indicating that these simple features do not account for the discrimination achieved by the pretrained representations in these settings.

Finally, high AlloDiverse performance was retained when ASD pockets were reconstructed using the same ligand-centered procedure as the other data sets, with the strongest configurations reaching ROC-AUC values of approximately 0.96−0.98 (±0.004, Tables S12−S15). The corresponding source-only control reached 0.910−0.915, indicating that high separability persists beyond that attributable to data set provenance alone and is not explained by differences in pocket-construction procedure.

Structural and Dynamical Encoder Architectures Contribute Differently across Separability Regimes

While embedding composition influenced performance in a data set-dependent manner, substantial variability also remained between encoder architectures. We therefore examined whether different structural and dynamical encoders capture complementary aspects of allosteric regulation and how these architectural choices affect classification performance. To isolate architectural effects, we compared OneProt variants incorporating structure-token (ST), structure-graph (SG), and MD encoders while evaluating a fixed downstream

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897149/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

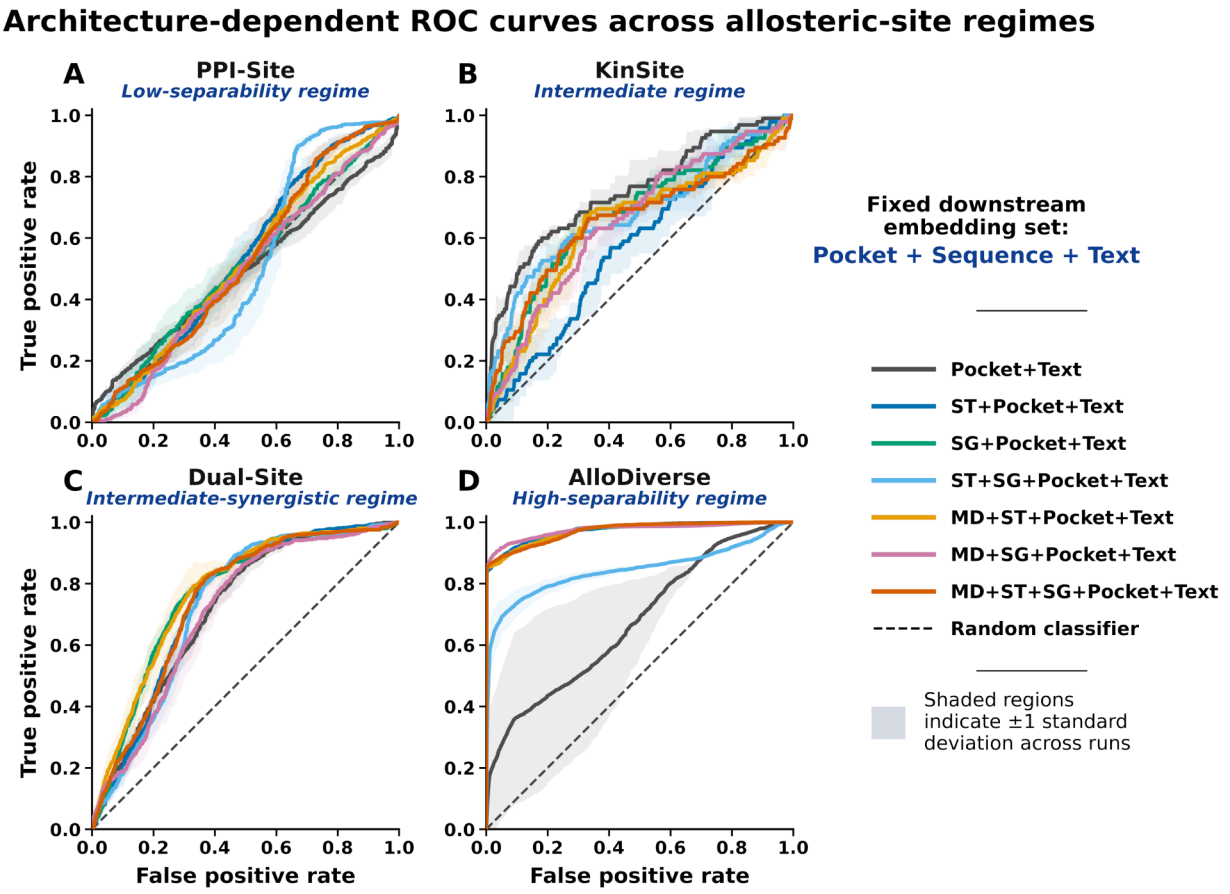


Figure 6. ROC curves for different architectural encoder combinations with fixed embedding combination. Panels show results for the four data sets: (A) PPI-Site, (B) KinSite, (C) Dual-Site, and (D) AlloDiverse. Different colors indicate encoder configurations: Pocket+Text (baseline, no structural encoders; gray), ST+Pocket+Text (structure-token; blue), SG+Pocket+Text (structure-graph; green), MD+ST+Pocket+Text (MD + structure-token; orange), MD+SG+Pocket+Text (MD + structure-graph; purple), ST+SG+Pocket+Text (both structural encoders; teal), and MD +ST+SG+Pocket+Text (full multimodal; red). The dashed diagonal line represents random performance (AUC = 0.5). Architectural effects vary dramatically across data sets: in PPI-Site, all architectures cluster near random performance; in KinSite, structural encoders provide modest improvements; in Dual-Site, structural encoders substantially outperform the baseline; and in AlloDiverse, multiple structural architectures converge to near-ceiling performance. Full details on the ROC-AUC curves are provided in Supporting Figure S1 and Supporting Table S4.

embedding combination consisting of pocket, sequence, and text representations. The impact of encoder architecture depends strongly on the separability regime (Figure 6). In the PPI-Site data set, all architectures produced ROC curves close to the diagonal and ROC-AUC values near random expectation (AUC ≈ 0.55−0.58) (Figure 6A). Although minor differences between encoder configurations were observed, none consistently outperformed the others. This result confirms that when allosteric and orthosteric sites share structural contexts and are distinguished primarily by conformational differences, architectural modifications alone are insufficient to recover a robust discriminative signal.

More substantial architectural differences emerged in the KinSite and Dual-Site data sets. In KinSite, the Pocket+Text baseline remained among the strongest-performing architectures, suggesting that local pocket geometry combined with functional annotations captures a substantial fraction of the discriminative signal (Figure 6B). The structure-token-only (ST+Pocket+Text) architecture achieved the highest median AUC (≈0.74) and exhibited the most compact performance distribution, indicating that global fold-level organization captured by Foldseek tokens provides complementary information beyond local pocket features. In contrast, the Pocket+Text architecture became the weakest-performing configuration in the Dual-Site data set, where the inclusion of structural encoders consistently improved discrimination (Figure 6C). This reversal reflects the increased structural and functional diversity introduced by combining kinase and nonkinase proteins. Within the relatively homogeneous kinase family, text-derived functional information provides a strong contextual signal. However, in the more heterogeneous Dual-Site regime, explicit structural representations become essential for capturing differences that generalize across distinct protein folds and regulatory mechanisms. Notably, several structurally informed architectures converged to similar ROC-AUC values (0.84−0.87), indicating that the common factor underlying improved performance was the incorporation of structural information itself, rather than the specific choice of structural encoder.

A similar pattern emerged in the AlloDiverse data set (Figure 6D). Architectures incorporating any structural modality consistently achieved near-ceiling ROC-AUC values (AUC ≈0.96−0.98), whereas the Pocket+Text baseline remained substantially lower with considerably greater variability (AUC ≈ 0.93) (Figure 6D). Structure-token, structure-graph, and MD architectures converged to similar

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

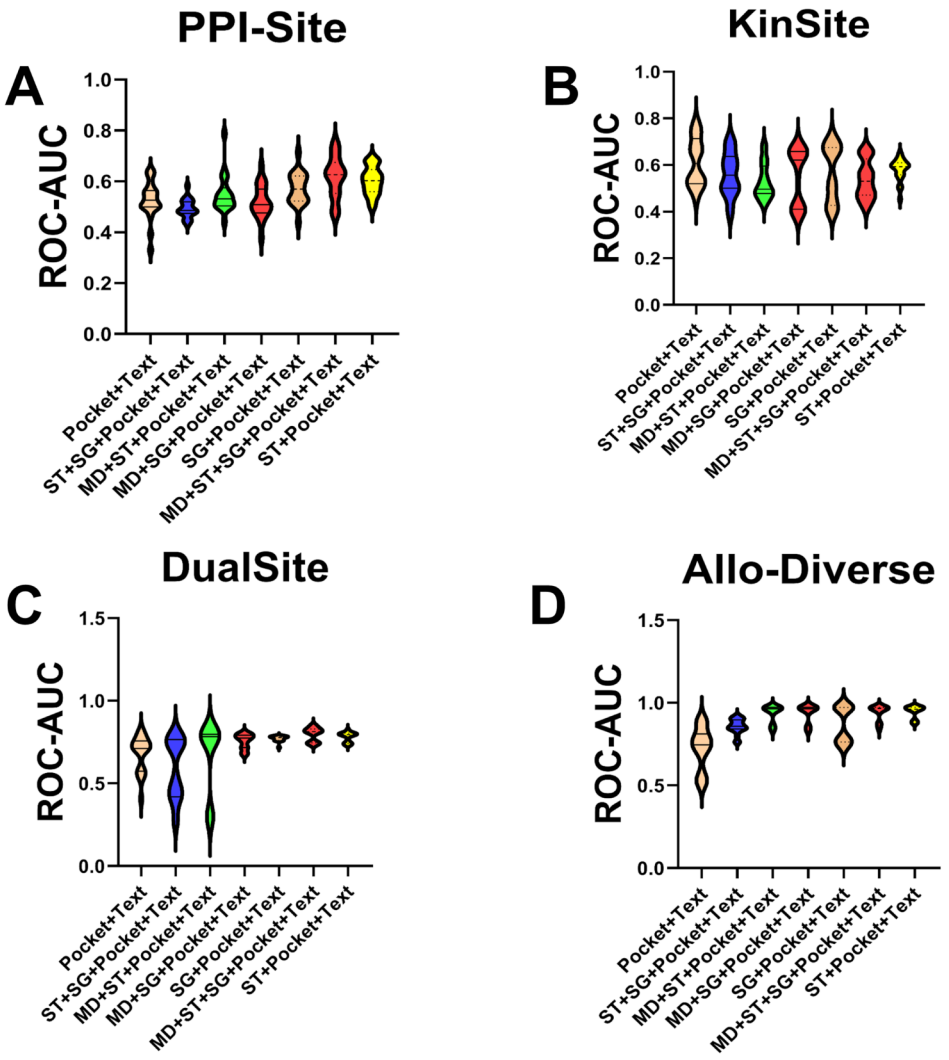


Figure 7. Architecture-dependent ROC-AUC distributions across separability regimes. Violin plots show the distribution of ROC-AUC values across seven OneProt encoder architectures and ten independent random seeds for each of the four data sets. Each panel corresponds to a distinct separability regime: (A) PPI-Site (low-separability), (B) KinSite (intermediate), (C) Dual-Site (intermediate-synergistic), and (D) AlloDiverse (high-separability). The *x*-axis within each panel denotes the seven encoder architectures evaluated: Pocket+Text (baseline, no structural encoders), 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 represent the four downstream embedding combinations used for classification: Pocket only (blue), Pocket+Text (orange), Pocket +Sequence (green), and Pocket+Sequence+Text (red). Each point corresponds to a single independent run, with horizontal bars indicating the median and interquartile range. The dashed horizontal line denotes random performance (ROC-AUC = 0.5).

performance levels despite their distinct representational strategies, reinforcing that the critical factor is the presence of structural information rather than the specific encoder. Notably, the MD encoder—which captures conformational fluctuations and ensemble behavior—did not provide additional gains beyond static structural representations in this regime, suggesting that the discriminative signal in AlloDiverse is sufficiently strong that static features already capture the relevant information. The full set of ROC curves across all seven encoder architectures and embedding combinations is provided in Supporting Figure S1.

The architecture-specific performance distributions revealed several notable trends (Figure 7). The progression from PPI-Site to AlloDiverse reveals a systematic shift in the role of architecture. In PPI-Site, all architectures cluster near random performance (AUC ≈0.55−0.58) regardless of embedding combination (Figure 7A), confirming that when allosteric and orthosteric sites share structural contexts, architectural modifications alone are insufficient to recover a discriminative signal. In PPI-Site, only architectures incorporating structure-token representations (ST+Pocket+Text) or the full multimodal encoder consistently achieved performance above random expectation. This suggests that even in low-separability settings, fold-level structural context provides some weak discriminative signal that local pocket geometry alone cannot capture, although overall performance remains near-random. In KinSite (Figure 7B), the structure-token-only (ST+Pocket +Text) architecture exhibits the most compact distribution and highest median AUC (≈0.74), while architectures incorporating both ST and SG encoders show broader distributions and lower median performance, suggesting redundancy between structural representations. In Dual-Site, architectures incorporating structural information (ST, SG, or MD) consistently outperform the Pocket+Text baseline, with convergence across

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

multiple configurations (AUC ≈0.84−0.87), indicating that the presence of structural information—rather than the specific encoder—drives improved discrimination (Figure 7C). The ST+SG configuration did not exhibit corresponding improvement beyond individual structural encoders, confirming that these representations encode partially overlapping information. In contrast, architectures incorporating MD information formed part of the highest-performing group, indicating that dynamical information provides genuinely complementary signal in data sets where conformational heterogeneity distinguishes regulatory classes. In AlloDiverse, nearly all architectures containing structural information converge to near-ceiling performance (AUC ≈0.96−0.98), with the structure-token architecture remaining among the most stable configurations (Figure 7D).

The narrowing of distributions from PPI-Site to AlloDiverse indicates that as the discriminative signal strengthens, performance becomes less sensitive to both architecture and embedding choices, consistent with the variance decomposition analysis (Figure 4) showing that encoder architecture explains substantially less performance variation than data set composition. The structure-token architecture remained among the most stable configurations, while the full multimodal model achieved a comparable performance without providing a substantial advantage.

Taken together, these results demonstrate that architectural effects are highly context-dependent. In low-separability regimes, performance is constrained by data set characteristics and remains largely insensitive to encoder design. In intermediate regimes, structural and dynamical encoders contribute meaningfully, with different encoders providing complementary benefits depending on the biological characteristics of the data set. In high-separability regimes, multiple architectural solutions achieve near-optimal discrimination and the specific encoder choice becomes less critical. Consistent with the variance decomposition analysis (Figure 4), encoder architecture explains substantially less performance variation than data set composition. Nevertheless, architectural effects become increasingly apparent once the underlying biological signal is sufficiently informative, particularly in intermediate regimes, where the choice of structural representation can meaningfully impact discrimination.

Dual-Site Training Redistributes Allosteric-Site Recovery across Constituent Data Sets

Rather than uniformly increasing recovery across either constituent data set, merging redistributed which individual allosteric sites were preferentially recognized. This effect was particularly evident for PPI-Site. Indeed, although mean allosteric site recovery remained broadly comparable between PPI-Site-only and Dual-Site training (approximately 61% and 57%, respectively), approximately 28% of individual PPI-Site allosteric pockets became more frequently recovered after the addition of KinSite. KinSite exhibited a different response, with the merged classifier becoming substantially more conservative for kinase examples overall: orthosteric recovery increased from approximately 65% to 91%, whereas allosteric recovery decreased from approximately 43% to 9%. Importantly, however, this shift was not equivalent to uniformly assigning KinSite examples to the orthosteric class: the kinase allosteric sites that remained preferentially detectable in Dual-Site differed from those most consistently recovered in KinSite alone, while a distinct subset of PPI-Site allosteric pockets became more frequently recovered. Thus, combining the data sets changes the downstream decision boundary in an asymmetric and example-dependent manner.

Inspection of the individual KinSite allosteric examples further supports this redistribution. Despite the overall reduction in KinSite allosteric site recovery under dual-site training, the sites that remained most consistently detectable were not simply those most readily recognized by the KinSite-only classifier. The three highest-recovery KinSite positives under single-data set training were RIPK1 complexes (PDB 6C4D, 6HHO, and 4ITJ), indicating preferential recognition of a recurrent kinase context (Figure 8). In contrast, the three highest-recovery KinSite positives under dual-site training were kinase complexes ERK5/MAPK7 (PDB 4ZSJ), LIMK2 (PDB 5NXD), and WNK1 (PDB 5WDY). Notably, the latter examples were not among the most consistently recovered positives under KinSite-only training (Figure 8). Thus, the small subset of kinase allosteric sites retained after merged training represents a qualitatively different set of examples rather than simply the most readily classified KinSite positives. This rank redistribution provides further evidence that the addition of PPI-Site changes which regions of the pretrained representation space are preferentially assigned to the allosteric class.

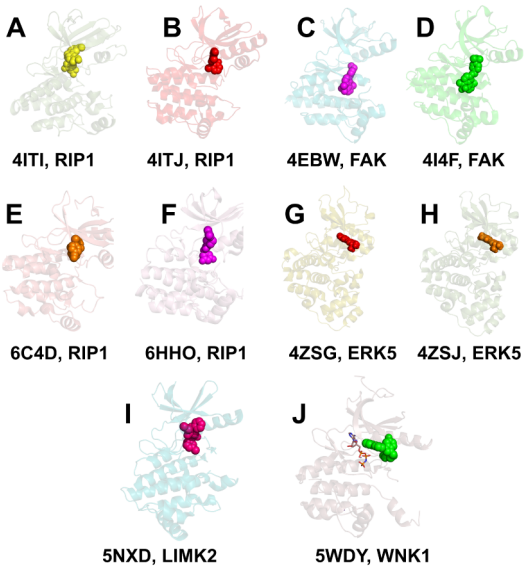


Figure 8. Redistribution of preferentially recovered KinSite allosteric sites following Dual-Site training. Representative structures of the highest-recovery KinSite allosteric pockets under KinSite-only training and Dual-Site training. KinSite-only training preferentially recovered several RIP1 complexes (A, B, E, F) and FAK complexes (C, D) whereas the highest-recovery KinSite positives under Dual-Site training included ERK5/MAPK7 (H), LIMK2 (I), and WNK1 (J). The different sets of preferentially recovered structures illustrate the rank redistribution of KinSite allosteric examples following addition of PPI-Site to the training data set.

The preferential recovery patterns observed for KinSite allosteric sites reveal a fundamental hierarchy of allosteric site detectability. This detectability hierarchy of orthosteric > type III allosteric > type IV allosteric binding reflects the progressive loss of structural definition and evolutionary constraint as sites move from the catalytic core to distal regulatory regions. Our analysis shows that type III allosteric

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

kinase inhibitors, with their well-defined hydrophobic pocket geometry, DFG-out conformational signatures, and moderate sequence conservation, are readily detectable even in the data-limited KinSite setting (Figure 8). This is exemplified by the preferential recovery of RIPK1 complexes under KinSite-only training and the additional type III sites recovered only after PPI-Site regularization, all of which share the defining structural features of type III allosteric inhibition. In contrast, type IV inhibitors, which bind to distal regulatory regions such as myristoyl pockets, PIF pockets, and activation loop docking interfaces, remain beyond the reach of current multimodal representations regardless of data set composition. This persistent failure stems from fundamental biophysical properties such as structural plasticity, evolutionary lability, and lack of recurrent patterns rather than limitations in the model architecture or training data. This finding is significant because it establishes that some allosteric sites are intrinsically more difficult to detect due to their physical and evolutionary properties.

In summary, our systematic profiling of multimodal protein foundation models across four distinct data sets reveals that allosteric-site discrimination is governed primarily by data set composition rather than architectural complexity. Through comprehensive embedding ablations, encoder architecture comparisons, and variance decomposition, we demonstrate that data set identity accounts for the dominant fraction of explainable variance (63.7%), while embedding composition and architectural choice contribute substantially less. The data define three distinct separability regimes: a low-separability PPI-Site regime where allosteric and orthosteric sites share structural and functional contexts, an intermediate KinSite regime characterized by pronounced representation dependence and severe class imbalance, and a high-separability AlloDiverse regime where multiple modalities and architectures converge to near-ceiling performance. Importantly, single-modality and source controls reveal that similar levels of predictive performance can arise from different combinations of signal—including data set provenance, global protein context, coarse compositional features, and local pocket information. The within-protein comparison further demonstrates that pocket-containing representations retain genuine binding-site-specific discriminatory information beyond protein identity. At the encoder level, structural and dynamical representations provide their greatest benefits in intermediate- and high-separability settings, while architectural complexity offers diminishing returns in both low- and high-separability regimes. Finally, the Dual-Site analysis shows that combining data sets restructures the downstream decision boundary in an asymmetric and example-dependent manner, reshaping which individual binding sites are preferentially recognized. Together, these findings establish that the utility of multimodal protein representations is fundamentally context-dependent, and that benchmark performance should not be interpreted simply as a measure of representation quality, but rather as the outcome of interactions between data set composition, pretrained representation content, and model architecture. This framework provides a principled basis for interpreting and improving multimodal approaches to allosteric-site prediction.

DISCUSSION

The systematic characterization of multimodal protein foundation models for allosteric site prediction reveals several principal findings with significant implications for computational biology and drug discovery. First, data set composition is the dominant determinant of predictive performance. Our variance decomposition analysis demonstrates that data set identity accounts for 63.7% of the explainable variance in ROC-AUC, dwarfing the contributions of embedding composition (1.4%) and architectural choice (3.4%). This result challenges the prevailing emphasis on architectural innovation as the primary driver of progress in binding site prediction. The intrinsic characteristics of the training data—structural diversity, conformational heterogeneity, class balance, and the biological relationship between allosteric and orthosteric sites—exert a far greater influence on prediction success than model design choices. This finding does not diminish the importance of model architecture; rather, it reframes the central challenge of allosteric site prediction: the primary obstacle is not architectural sophistication but the availability and composition of training data that adequately capture the biological signals distinguishing regulatory classes.

The identification of three distinct separability regimes provides a framework for understanding how representation utility changes across data sets with different levels of separability. In the low-separability PPI-Site regime, allosteric and orthosteric sites share structural and evolutionary contexts and are distinguished primarily by conformational differences. Here, pocket geometry alone is the most informative modality, and additional modalities introduce noise rather than complementary signal. This counterintuitive finding—that adding more information degrades performance—reflects the biological reality that when regulatory classes share evolutionary and functional contexts, global sequence and text signals provide variance rather than discrimination. In the intermediate KinSite regime, coarse-grained whole-protein sequence properties provide strong discriminative signals, whereas individual learned embedding modalities are less predictive and only selected multimodal configurations approach comparable performance. In the high-separability AlloDiverse regime, allosteric sites span broad structural and functional diversity, enabling multiple modalities to contribute synergistically to near-ceiling performance. This progression from low to high separability reveals that the value of multimodal integration is not absolute but depends critically on the biological distance between the classes being discriminated.

Beyond these empirical observations, our study establishes a general analytical framework for systematically profiling multimodal protein foundation models across distinct biological regimes. Rather than treating multimodal models solely as predictive tools, this approach uses them as scientific instruments to dissect how evolutionary, structural, functional, and dynamical information contributes to discrimination under different biological contexts. By integrating controlled embedding ablations, encoder-level comparisons, and variance decomposition, the framework disentangles the influence of data set characteristics from that of model design, providing a principled basis for interpreting multimodal representations beyond conventional benchmarking. Although demonstrated here using OneProt, the approach is model-agnostic and readily transferable to other multimodal protein foundation models and biological prediction tasks. The framework's modular design—separating encoder-level (what modalities shape the latent space) from embedding-level (what features are used for classification)—enables systematic attribution of performance gains to specific biological signals, a critical

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

capability for building interpretable and trustworthy predictive models.

Another important finding of this study is that context-dependent modality contributions reflect underlying biological organization. Discriminative signals vary systematically with protein family context and binding site architecture. In PPI-Site, allosteric and orthosteric sites share Rossmann-fold architectures and SAM-binding functionality, differing primarily in their spatial relationship to protein−protein interaction interfaces. Pocket geometry captures this conformational signal, while sequence and text modalities provide little additional discrimination. This observation reflects the biological reality that the same protein, the same ligand-binding pocket, and often the same set of residues can be classified as either allosteric or orthosteric depending on subtle conformational arrangements.

In the kinase regime, a different pattern emerges, with discrimination being strongly dependent on the representation used. Unlike PPI-Site, where performance remains weak across representation combinations, KinSite exhibits substantial variation among pretrained representations, consistent with embedding composition accounting for approximately 50% of performance variance in this data set. This pronounced representation dependence indicates that class-associated information is accessible to different extents through different pretrained OneProt representations rather than being uniformly available across the shared representation space. At the same time, the severe class imbalance makes the rare allosteric class particularly challenging, as reflected by the substantially more modest positive-class performance compared to ROC-based discrimination. Together, these results characterize KinSite as an intermediate separability regime in which the choice of representation has a particularly strong influence on downstream discrimination.

The additional baselines further emphasize the data set-specific nature of the KinSite regime. The strong performance of a coarse-grained whole-protein amino acid composition baseline indicates that substantial class-associated signal is accessible from simple sequence composition, despite the considerably weaker performance of the pretrained sequence representation. This contrast highlights that representation complexity does not necessarily translate into improved discrimination when data set-specific compositional differences provide a strong predictive signal.

The model architecture gives an additional, but similarly context-dependent, contribution. Increasing architectural complexity provides diminishing returns once a sufficient biological signal is present. Structural and dynamical encoders contribute meaningfully in intermediate separability regimes, but their impact diminishes in both low separability settings (where the underlying signal is too weak for any architecture to exploit) and high separability settings (where multiple architectures converge to near-ceiling performance). The convergence of different architectural configurations in AlloDiverse indicates that the critical factor is the inclusion of structural information rather than the specific structural encoder used to capture it. This pattern was particularly evident in Dual-site, where the addition of structural encoders consistently improved performance, but no single structural representation strategy dominated. The structure-token encoder, which captures global fold-level organization, provided the most robust performance across regimes, suggesting that broader structural context is more informative than fine-grained local geometry for distinguishing allosteric from orthosteric sites.

Another key finding emerging from this study is that combining data sets with complementary biological characteristics can substantially improve overall discrimination, while also changing which individual examples are preferentially recognized. The Dual-Site data set achieved markedly stronger performance than either PPI-Site or KinSite alone, but the example-level analysis showed that this improvement was not uniform across the constituent data sets. For KinSite, orthosteric-site recovery increased substantially after merging, whereas overall allosteric-site recovery decreased; importantly, the kinase allosteric sites that remained preferentially detectable differed from those most consistently recovered under KinSite-only training. PPI-Site exhibited a different pattern: although mean allosteric-site recovery changed only moderately, a large number of individual allosteric pockets were recovered more frequently after addition of KinSite. Thus, the improved Dual-Site performance reflects an asymmetric and example-dependent redistribution of predictions rather than uniform recovery of previously misclassified sites. Because the OneProt encoders remain frozen during downstream training, this effect is best interpreted as complementary training signals allowing the classifier to establish a more effective decision boundary within the pretrained representation space. The strong architecture × representation interaction observed for Dual-Site further indicates that the ability to exploit these complementary signals depends on the particular OneProt configuration.

These findings have direct implications for the application of multimodal protein representations to allosteric drug discovery. Rather than identifying a single representation or model configuration that is optimal across biological settings, our results show that the information most useful for allosteric-site discrimination depends strongly on the data set and task formulation. PPI-Site illustrates the difficulty of discrimination when information defining the labels—in this case, the spatial relationship to the PPI interface—is not explicitly provided to the classifier. KinSite demonstrates pronounced representation dependence within a shared kinase context, while Dual-Site shows that combining biologically distinct training distributions can enhance the overall discrimination by providing complementary signals to the downstream classifier. In AlloDiverse, strong protein-level and provenance-associated signals coexist with additional pocket-specific discriminatory information, enabling several OneProt configurations to achieve near-ceiling performance.

These results emphasize that progress in allosteric-site discrimination will depend not only on architectural innovation alone but also on systematic data set curation and characterization. The finding that data set identity explains nearly two-thirds of performance variation underscores the importance of training and evaluating models across data sets that capture diverse biological contexts while controlling for unintended source-associated differences. Future work should extend this framework to more diverse and challenging allosteric environments, including cryptic and transient binding sites, and investigate whether conformational ensembles, dynamics-aware representations, and physics-based constraints can provide additional information where static pretrained representations remain insufficient. By establishing how data set characteristics interact with pretrained representation content and model architecture, this study provides a

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

foundation for the development of more robust and generalizable multimodal approaches to allosteric drug discovery.

CONCLUSIONS

Our systematic investigation of multimodal protein foundation models for allosteric-site discrimination establishes three principal conclusions with broad implications for computational drug discovery. First, data set composition is the dominant determinant of predictive performance. This finding fundamentally reframes the central challenge of allosteric-site prediction: the primary obstacle is not architectural sophistication but the availability and composition of training data that adequately capture the biological signals distinguishing regulatory classes. The identification of three distinct separability regimes—low (PPI-Site), intermediate (KinSite, Dual-Site), and high (AlloDiverse)—provides a quantitative framework for understanding how data set composition modulates representation utility and model performance, demonstrating that allostery prediction is not a monolithic problem but comprises qualitatively different learning regimes. Second, the utility of multimodal representations is fundamentally context-dependent. Systematic ablation experiments reveal that the optimal embedding strategy varies systematically across data sets, with biologically interpretable patterns: pocket geometry dominates when regulatory classes are distinguished by conformational differences; text and functional annotations provide essential discriminative power in annotation-rich kinase contexts; sequence-based representations excel when evolutionary signals distinguish regulatory classes; and multiple modalities contribute synergistically when broad structural and functional diversity is present. Performance does not increase monotonically with the number of concatenated embeddings, demonstrating that adding modalities is beneficial only when they provide genuinely complementary information relevant to the classification task. The Data set × Embedding interaction (8.3% of variance) quantifies this context dependence, revealing that modality contributions reflect the underlying biological organization of each data set rather than representing universal properties of multimodal representations.

Third, combining data sets restructures the downstream decision boundary in an asymmetric and example-dependent manner. Prediction-level recovery analysis demonstrates that merging PPI-Site and KinSite does not produce uniform improvement but instead redistributes which individual binding sites are preferentially recognized. For KinSite, orthosteric recovery increases from ∼65% to 91% while allosteric recovery decreases from ∼43% to 9%, with the allosteric sites that remain preferentially detectable representing a qualitatively different set of kinase contexts. For PPI-Site, approximately 28% of individual allosteric pockets become more frequently recovered, with gains distributed across structurally heterogeneous targets. This asymmetric redistribution, occurring despite frozen foundation-model representations, establishes that data set composition fundamentally shapes how the classifier partitions the pretrained representation space, which is a finding with direct implications for training data set design and model interpretation.

Beyond these specific findings, our study establishes a generalizable analytical framework for profiling multimodal protein foundation models across distinct biological regimes. The framework's modular design—separating encoder-level (what modalities shape the latent space) from embedding-level (what features are used for classification)—enables systematic attribution of performance gains to specific biological signals, a critical capability for building interpretable and trustworthy predictive models. Although demonstrated using OneProt, the approach is model-agnostic and readily transferable to other multimodal protein foundation models and biological prediction tasks.

Our study suggests that progress in allosteric-site discrimination will depend not on architectural innovation alone but equally on systematic data set curation and characterization. The finding that data set identity explains nearly two-thirds of performance variation underscores the importance of training and evaluating models across data sets that capture diverse biological contexts while controlling for unintended source-associated differences. The context-dependent utility of modalities highlights the need for adaptive strategies that exploit different representations according to the biological characteristics of the target systems. The demonstration that data set merging restructures the decision boundary in asymmetric, example-dependent ways provides a mechanistic understanding of how training data composition influences model behavior, informing the design of more effective training data sets.

By establishing how data set characteristics interact with pretrained representation content and model architecture across distinct separability regimes, this study provides a foundation for the development of more robust, context-aware, and physically grounded multimodal models for allosteric drug discovery. The quantitative framework presented here moves beyond conventional benchmarking to systematically characterize when and why multimodal integration benefits allostery prediction, providing both conceptual insights and practical guidance for the field. The methodological framework, analytical controls, and data set characterization strategies developed in this study may prove valuable not only for allosteric-site prediction but more broadly for the rigorous evaluation and interpretation of multimodal protein foundation models across diverse biological prediction tasks.

ASSOCIATED CONTENT

Data Availability Statement

All source code, processed data sets, pretrained model checkpoints, and the computational environment required to reproduce the results of this study are publicly available through the OneProt Models GitHub organization (https://github.com/oneprot-models) and the accompanying Zenodo archive (10.5281/zenodo.20997998). The GitHub repository oneprot-embeddings (https://github.com/oneprot-models/oneprot-embeddings) contains the complete software implementation used in this work. The repository provides the OneProt framework for multimodal protein embedding extraction together with utilities for downstream prediction and evaluation. The allosteric-site prediction workflow is organized under src/allostery, where generate_pockets/contains scripts for reconstructing fixed-size binding pocket data sets from the original PPI-Site, KinSite, and ASD resources; extract_embeddings/contains workflows for generating multimodal OneProt embeddings from the reconstructed pockets using the sequence-, structure-, graph-, molecular dynamics-, and multimodal encoder variants evaluated in this study; plotting/contains the scripts used to generate all figures and statistical analyses presented in the manuscript; and utils/

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897149/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

provides shared utilities for data set handling, preprocessing, evaluation, and reproducibility. Environment specifications, including all required software dependencies and package versions, are provided in the directory (https://github.com/oneprot-models/oneprot-embeddings/tree/main/environment), while a fully configured Apptainer/Singularity runtime environment is distributed through the accompanying Zenodo archive as oneprot_container.sif. Complete documentation, configuration files, and example commands required to reproduce the workflow are provided in the repository README. The archive also includes checkpoints.zip, containing the pretrained downstream model checkpoints corresponding to all seven OneProt embedding ablations evaluated in this study. Downstream fine-tuning using the OneProt embeddings is performed via the original unmodified OneProt framework and is executed through the scripts src/saprot_fit_mlp.py and src/saprot_fit_mlp_balanced.py for the case when the data needs to be balanced during fine-tuning. The whole procedure is described in the general OneProt README. The accompanying Zenodo archive contains all processed resources required to reproduce the reported experiments. The HDF5 files (binding_pockets_allosteric.h5, binding_pockets_orthosteric_competitive.h5, allosteric_pockets_kinsite.h5, competitive_pockets_kinsite.h5, and ASD_binding_pockets.h5) contain reconstructed atom-level binding pocket structures used for OneProt pocket embedding extraction for PPI-site allosteric, PPI-site orthosteric, KinSite allosteric, KinSite orthosteric, and ASD-derived allosteric pockets, respectively. Unless stated otherwise, pockets were generated by retaining the 100 residues closest to the annotated ligand or reconstructed pocket center while preserving all atoms belonging to the selected residues. The archive further contains all train, validation, and test partitions used throughout this study. PPI-site_splits.zip and KinSite_splits.zip provide MMseqs2-based data partitions generated using a 30% sequence identity threshold to prevent homologous proteins from appearing across data set splits. AlloDiverse_splits.zip contains the benchmark splits used in this work, where the positive class is derived from the ASD data set and the validation and test sets correspond to a random 50:50 partition of the original ASD test set. ASD_original_split.zip contains the original train/test partition from the published ASD benchmark, while PPI-site_KinSite_original.zip contains the processed PPI-Site and KinSite data sets obtained from the original publications prior to sequence clustering and repartitioning.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.6c03114.

Supporting Texts S1−S4, Figures S1−S5, and Tables S1−S12 provide additional methods, implementation details, data set descriptions, and comprehensive performance analyses supporting the results presented in the main manuscript. Supporting Texts S1−S4: Supporting Text S1 describes the OneProt pretraining architecture, including the structural, pocket, sequence, text, and molecular dynamics encoders, as well as multimodal latent-space alignment. Supporting Text S2 details binding-pocket construction, structural preprocessing, and pocket overlap analyses for the benchmark data sets. Supporting Text S3 describes the downstream allostery classification framework, including hyperparameter optimization, performance metrics, and strategies for handling class imbalance. Supporting Text S4 summarizes implementation details, computational resources, and reproducibility information (PDF)

Figure S1 presents receiver operating characteristic (ROC) curves across OneProt encoder architectures, embedding combinations, and allosteric classification data sets (PDF)

Figure S2 shows precision−recall (PR) curves across encoder architectures under different class-imbalance regimes (PDF)

Figure S3 summarizes mean ROC-AUC values for all combinations of data sets, embedding configurations, and OneProt encoder architectures (PDF)

Figure S4 presents threshold-dependent sensitivity across data sets and embedding combinations (PDF)

Figure S5 presents threshold-dependent specificity across data sets and embedding combinations (PDF)

Table S1 reports the number of binding sites in the train/validation/test split across data sets(PDF)

Table S1 reports the number of binding sites in the train/validation/test split across data sets (PDF)

Table S2 reports the performance results of the LASSO classifier using handcrafted features (PDF)

Table S3 reports the results of within-protein discrimination analysis (PDF)

Table S4 reports the complete mean ROC-AUC values (±standard deviation across independent runs) for all combinations of data sets, embedding configurations, and OneProt encoder architectures (XLSX)

Table S5 reports the corresponding mean PR-AUC values (±standard deviation across independent runs) for all evaluated configurations, providing a quantitative basis for assessing the impact of class imbalance and modality selection on allosteric site discrimination (XLSX)

Table S6 reports the complete Pocket only performance across the same four data sets and seven OneProt encoder architectures (XLSX)

Table S7 reports the corresponding Pocket+Text summary performance across the same four data sets and seven architectures (XLSX)

Table S8 reports detailed Pocket+Sequence+Text performance across data sets and architectures (XLSX)

Table S9 reports detailed Pocket+Sequence performance across the four data sets and seven OneProt encoder architectures (XLSX)

Table S10 reports the complete Sequence-only performance across the same data sets and encoder architectures (XLSX)

Table S11 reports the complete Text-only performance across the same data sets and encoder architectures (XLSX)

Table S12 reports the complete Sequence+Text performance across the same data sets and encoder architectures (XLSX)

Table S13 reports the complete Pocket+Structure Token performance across the same data sets and encoder architectures (XLSX)

Table S14 reports the complete Pocket+Structure Graph performance across the same data sets and encoder architectures (XLSX)

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

https://doi.org/10.1021/acs.jcim.6c03114
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table S15 reports the complete Pocket+Molecular Dynamics performance across the same data sets and encoder architectures (XLSX)

AUTHOR INFORMATION

Corresponding Author

Gennady M. Verkhivker − *Keck Center for Science and Engineering, Schmid College of Science and Technology, Chapman University, Orange, California 92866, United States; Department of Biomedical and Pharmaceutical Sciences, School of Pharmacy, Chapman University, Irvine, California 92618, United States;* orcid.org/0000-0002-4507-4471; Phone: +1-714-516-4586; Email: verkhivk@chapman.edu

Author

Alina Bazarova − *Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich 52428, Germany; Helmholtz AI Munich, Neuherberg 85764, Germany;* orcid.org/0000-0001-8874-1783

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.6c03114

Author Contributions

Conceptualization, G.M.V. and A.B.; methodology, G.M.V. and A.B.; software, A.B.; validation, A.B. and G.M.V.; formal analysis, A.B. and G.M.V.; investigation, A.B. and G.M.V.; resources, A.B.; data curation, A.B. and G.M.V.; writing—original draft preparation, A.B. and G.M.V.; writing—review and editing, A.B. and G.M.V.; visualization, A.B. and G.M.V.; supervision, G.M.V.; project administration, G.M.V.; Funding acquisition, A.B. and G.M.V. All authors have read and agreed to their individual contributions. All authors have approved the submission of this manuscript, and it has been submitted exclusively to the Journal of Chemical Information and Modeling.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.B. is supported by the Helmholtz Association Initiative and Networking Fund in the frame of Helmholtz AI as well as by the Helmholtz Foundation Model Initiative within the project "PROFOUND." A.B. gratefully acknowledges the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC). This research was funded by the National Institutes of Health under Awards 1R01AI181600-01 and 5R01AI181600-02, and Subaward 6069-SC24-11 to G.V. G.V. acknowledges support from Schmid College of Science and Technology at Chapman University for providing computing resources at the Keck Center for Science and Engineering.

REFERENCES

(1) Zhao, J.; Cao, Y.; Zhang, L. Exploring the Computational Methods for Protein-Ligand Binding Site Prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417−426.

(2) Xia, Y.; Pan, X.; Shen, H.-B. A Comprehensive Survey on Protein-Ligand Binding Site Prediction. *Curr. Opin. Struct. Biol.* **2024**, *86*, 102793.

(3) Carpenter, K. A.; Altman, R. B. Databases of Ligand-Binding Pockets and Protein-Ligand Interactions. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 1320−1338.

(4) Utgés, J. S.; Barton, G. J. Comparative Evaluation of Methods for the Prediction of Protein−Ligand Binding Sites. *J. Cheminform.* **2024**, *16*, 126.

(5) Tsai, C. J.; Nussinov, R. A unified view of "how allostery works". *PLoS Comput. Biol.* **2014**, *10*, No. e1003394.

(6) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* **2016**, *12*, No. e1004966.

(7) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, Y.; Dejaegere, A.; Cecchini, M.; Changeux, J. P.; Bolhuis, P. G.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C. J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T. Allostery in Its Many Disguises: From Theory to Applications. *Structure* **2019**, *27*, 566−578.

(8) Nussinov, R.; Tsai, C.-J.; Jang, H. Allostery, and How to Define and Measure Signal Transduction. *Biophys. Chem.* **2022**, *283*, 106766.

(9) Tee, W.-V.; Berezovsky, I. N. Allosteric Solution to the Problems of Undruggable Targets, Drug Toxicity, and Emerging Resistance. *Curr. Opin. Struct. Biol.* **2025**, *95*, 103172.

(10) Zhu, R.; Wu, C.; Zha, J.; Lu, S.; Zhang, J. Decoding Allosteric Landscapes: Computational Methodologies for Enzyme Modulation and Drug Discovery. *RSC Chem. Biol.* **2025**, *6*, 539−554.

(11) Huang, J.; Tang, G.; Liu, N.; Li, X.; Lu, S. Recent Advances in Computational Strategies for Allosteric Site Prediction: Machine Learning, Molecular Dynamics, and Network-Based Approaches. *Drug Discov. Today* **2025**, *30*, 104466.

(12) Zhu, R.; Wu, C.; Li, M.; Liu, X.; Zhang, J. Unlocking Multiscale Allosteric Mechanisms: Advanced Computational Strategies for Drug Discovery. *Med. Res. Rev.* **2026**, *46*, 836−852.

(13) Tran-Nguyen, V.-K.; Camproux, A.-C. Computational Modeling of Protein−Ligand Interactions: From Binding Site Identification to Pose Prediction and Beyond. *Curr. Opin. Struct. Biol.* **2025**, *95*, 103152.

(14) Ekambaram, S.; Arakelov, G.; Dokholyan, N. V. The Evolving Landscape of Protein Allostery: From Computational and Experimental Perspectives. *J. Mol. Biol.* **2025**, *437*, 169060.

(15) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinform.* **2009**, *10*, 168.

(16) Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25*, 3185−3186.

(17) Liu, Y.; Grimm, M.; Dai, W.; Hou, M.; Xiao, Z.-X.; Cao, Y. CB-Dock: A Web Server for Cavity Detection-Guided Protein−Ligand Blind Docking. *Acta Pharmacol. Sin.* **2020**, *41*, 138−144.

(18) Krivák, R.; Hoksza, D. P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. *J. Cheminform.* **2018**, *10*, 39.

(19) Polák, L.; Škoda, P.; Riedlová, K.; Krivák, R.; Novotný, M.; Hoksza, D. PrankWeb 4: A Modular Web Server for Protein−Ligand Binding Site Prediction and Downstream Analysis. *Nucleic Acids Res.* **2025**, *53*, W466−W471.

(20) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. D. Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **2017**, *33*, 3036−3042.

(21) Jiang, M.; Wei, Z.; Zhang, S.; Wang, S.; Wang, X.; Li, Z. FRSite: Protein Drug Binding Site Prediction Based on Faster R−CNN. *J. Mol. Graphics Modell.* **2019**, *93*, 107454.

(22) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Improving Detection of Protein-Ligand Binding Sites with 3D Segmentation. *Sci. Rep.* **2020**, *10*, 5035.

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897149/acs.jcim.6c03114.pdf by guest on 28 October 2026

(23) Kozlovskii, I. P.; Popov, P. Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.* **2020**, *3*, 618.
(24) Pliushcheuskaya, P.; Künze, G. Evaluation of Small-Molecule Binding Site Prediction Methods on Membrane-Embedded Protein Interfaces. *J. Chem. Inf. Model.* **2025**, *65*, 6949−6967.
(25) Škrhák, V.; Novotný, M.; Feidakis, C. P.; Krivák, R.; Hoksza, D. CryptoBench: Cryptic Protein−Ligand Binding Sites Dataset and Benchmark. *Bioinformatics* **2024**, *41*, btae745.
(26) Pryakhin, V.; Smail-Tabbone, M.; Karami, Y. Benchmark Bias and Conformational Dynamics in Allosteric Site Prediction. *bioRxiv* **2026**.
(27) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2016239118.
(28) Carbery, A.; Buttenschoen, M.; Skyner, R.; von Delft, F.; Deane, C. M. Learnt Representations of Proteins Can Be Used for Accurate Prediction of Small Molecule Binding Sites on Experimentally Determined and Predicted Protein Structures. *J. Cheminform.* **2024**, *16*, 32.
(29) Hoksza, D.; Gamouh, H. Exploration of Protein Sequence Embeddings for Protein-Ligand Binding Site Detection. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE: Las Vegas, NV, USA, 2022; pp. 3356−3361. DOI: .
(30) Gamouh, H.; Novotný, M.; Hoksza, D. Hybrid Protein−Ligand Binding Residue Prediction with Protein Language Models: Does the Structure Matter? *Bioinformatics* **2025**, *41*, btaf431.
(31) Riedlová, K.; Škrhák, V.; Gatlin, W. G.; Ludwick, M.; Turano, L.; Novotný, M.; Hoksza, D.; Verkhivker, G. M. Predicting and Decoding Allosteric Binding Sites Using Protein Language Models and Structure-Based Machine Learning: An Energy Landscape-Guided Explainable AI Framework. *J. Chem. Theory Comput.* **2026**, *22*, 5326−5347.
(32) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-Tuning Protein Language Models Boosts Predictions across Diverse Tasks. *Nat. Commun.* **2024**, *15*, 7407.
(33) Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; Yi, J.; Zhao, W.; Wang, X.; Liu, Z.; Zheng, H.-T.; Chen, J.; Liu, Y.; Tang, J.; Li, J.; Sun, M. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nat. Mach. Intell.* **2023**, *5*, 220−235.
(34) Sledzieski, S.; Kshirsagar, M.; Baek, M.; Dodhia, R.; Lavista Ferres, J.; Berger, B. Democratizing Protein Language Models with Parameter-Efficient Fine-Tuning. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121*, No. e2405840121.
(35) Škrhák, V.; Gamouh, H.; Hoksza, D.Hidden in Protein Sequences: Predicting Cryptic Binding Sites*Proceedings Of The 16th ACM International Conference On Bioinformatics, Computational Biology, And Health Informatics*ACM20251−6
(36) Stärk, H.; Ganea, O.-E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. *arXiv* **2022**.
(37) Schneuing, A.; Harris, C.; Du, Y.; Didi, K.; Jamasb, A.; Igashov, I.; Du, W.; Gomes, C.; Blundell, T. L.; Lio, P.; Welling, M.; Bronstein, M.; Correia, B. Structure-Based Drug Design with Equivariant Diffusion Models. *Nat. Comput. Sci.* **2024**, *4*, 899−909.
(38) Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting Locations of Cryptic Pockets from Single Protein Structures Using the PocketMiner Graph Neural Network. *Nat. Commun.* **2023**, *14*, 1177.
(39) Greener, J. G.; Sternberg, M. J. A. Prediction of Allosteric Pockets on Proteins Using Normal Mode Perturbation Analysis. *BMC Bioinform.* **2015**, *16*, 335.
(40) Huang, W.; Lu, S.; Huang, Z.; Liu, X.; Mou, L.; Luo, Y.; Zhao, Y.; Liu, Y.; Chen, Z.; Hou, T.; Zhang, J. Allosite: A Method for Predicting Allosteric Sites. *Bioinformatics* **2013**, *29*, 2357−2359.
(41) Song, K.; Liu, X.; Huang, W.; Lu, S.; Shen, Q.; Zhang, L.; Zhang, J. Improved Method for the Identification and Validation of Allosteric Sites. *J. Chem. Inf. Model.* **2017**, *57*, 2358−2363.
(42) He, J.; Liu, X.; Zhu, C.; Zha, J.; Li, Q.; Zhao, M.; Wei, J.; Li, M.; Wu, C.; Wang, J.; Jiao, Y.; Ning, S.; Zhou, J.; Hong, Y.; Liu, Y.; He, H.; Zhang, M.; Chen, F.; Li, Y.; He, X.; Wu, J.; Lu, S.; Song, K.; Lu, X.; Zhang, J. ASD2023: Towards the Integrating Landscapes of Allosteric Knowledgebase. *Nucleic Acids Res.* **2024**, *52*, D376−D383.
(43) Char, S.; Corley, N.; Alamdari, S.; Yang, K. K.; Amini, A. P. ProtNote: A Multimodal Method for Protein−Function Annotation. *Bioinformatics* **2025**, *41*, btaf170.
(44) Wang, Z.; Ma, Z.; Cao, Z.; Zhou, C.; Zhang, J.; Gao, Y. Q. Prot2Chat: Protein Large Language Model with Early Fusion of Text, Sequence, and Structure. *Bioinformatics* **2025**, *41*, btaf396.
(45) Queen, O.; Huang, Y.; Calef, R.; Giunchiglia, V.; Chen, T.; Dasoulas, G.; Tai, L.; Abbadessa, G.; Howell, O.; Li, M. M.; et al. ProCyon: A Multimodal Foundation Model for Protein Phenotypes. *bioRxiv* **2025**.
(46) Boadu, F.; Wang, Y.; Cheng, J. A Unified Multimodal Model for Generalizable Zero-Shot and Supervised Protein Function Prediction. *Bioinformatics* **2026**, *42*, btag356.
(47) Zhang, L.; Guo, H.; Schaffer, L.; Ko, Y. S.; Singh, D.; Rahmani, H.; Grotjahn, D.; Villa, E.; Gilson, M.; Wang, W.; Ideker, T.; Xing, E.; Xie, P. A Tri-Modal Contrastive Learning Framework for Protein Representation Learning. *Cell Rep. Methods* **2026**, *6*, 101407.
(48) Flöge, K.; Udayakumar, S.; Sommer, J.; Piraud, M.; Kesselheim, S.; Fortuin, V.; Günnemann, S.; van der Weg, K. J.; Gohlke, H.; Merdivan, E.; Bazarova, A. OneProt: Towards Multi-Modal Protein Foundation Models via Latent Space Alignment of Sequence, Structure, Binding Sites and Text Encoders. *PLoS Comput. Biol.* **2025**, *21*, No. e1013679.
(49) van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L. M.; Söding, J.; Steinegger, M. Fast and Accurate Protein Structure Search with Foldseek. *Nat. Biotechnol.* **2024**, *42*, 243−246.
(50) Wang, L.; Liu, H.; Liu, Y.; Kurtin, J.; Ji, S. Learning Hierarchical Protein Representations via Complete 3D Graph Networks. *arXiv* **2022**.
(51) Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **2022**, *3*, 1−23.
(52) Aboelsoud, R.; Kasravi, J.; Kesselheim, S.; Bazarova, A. When Protein Dynamics Matter: Integrating Molecular Dynamics into Protein Foundation Models. In *ICLR 2026 Workshop On Foundation Models For Science*; ICLR, 2026.
(53) Li, Y.; Yi, J.; Li, H.; Li, K.; Kang, F.; Deng, Y.; Wu, C.; Fu, X.; Jiang, D.; Cao, D. Decoding the Limits of Deep Learning in Molecular Docking for Drug Discovery. *Chem. Sci.* **2025**, *16*, 17374−17390.
(54) La Sala, G.; Pfleger, C.; Käck, H.; Wissler, L.; Nevin, P.; Böhm, K.; Janet, J. P.; Schimpl, M.; Stubbs, C. J.; De Vivo, M.; Tyrchan, C.; Hogner, A.; Gohlke, H.; Frolov, A. I. Combining Structural and Coevolution Information to Unveil Allosteric Sites. *Chem. Sci.* **2023**, *14*, 7057−7067.
(55) Ferreiro, D. U.; Hegler, J. A.; Komives, E. A.; Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U. S. A* **2007**, *104*, 19819−19824.
(56) Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M. Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: A Tool to Localize Energetic Frustration in Protein Molecules, Now With Electrostatics. *Nucleic Acids Res.* **2016**, *44* (W1), W356−360.
(57) Freiberger, M. I.; Wolynes, P. G.; Ferreiro, D. U.; Fuxreiter, M. Frustration in Fuzzy Protein Complexes Leads to Interaction Versatility. *J. Phys. Chem. B* **2021**, *125* (10), 2513−2520.
(58) Chen, M.; Chen, X.; Schafer, N. P.; Clementi, C.; Komives, E. A.; Ferreiro, D. U.; Wolynes, P. G. Surveying biomolecular frustration at atomic resolution. *Nat. Commun.* **2020**, *11*, 5944.

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026

(59) Parra, R. G.; Komives, E. A.; Wolynes, P. G.; Ferreiro, D. U. Frustration in Physiology and Molecular Medicine. *Mol. Aspects Med.* **2025**, *103*, 101362.
(60) Dixit, A.; Verkhivker, G. M. The Energy Landscape Analysis of Cancer Mutations in Protein Kinases. *PLoS One* **2011**, *6*, No. e26071.
(61) Raisinghani, N.; Alshahrani, M.; Gupta, G.; Tian, H.; Xiao, S.; Tao, P.; Verkhivker, G. M. Integration of a Randomized Sequence Scanning Approach in AlphaFold2 and Local Frustration Profiling of Conformational States Enable Interpretable Atomistic Characterization of Conformational Ensembles and Detection of Hidden Allosteric States in the ABL1 Protein Kinase. *J. Chem. Theory Comput.* **2024**, *20*, 5317−5336.
(62) Guan, X.; Tang, Q.-Y.; Ren, W.; Chen, M.; Wang, W.; Wolynes, P. G.; Li, W. Predicting Protein Conformational Motions Using Energetic Frustration Analysis and AlphaFold2. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121*, No. e2410662121.
(63) Moine-Franel, A.; Mareuil, F.; Nilges, M.; Ciambur, C. B.; Sperandio, O. A Comprehensive Dataset of Protein-Protein Interactions and Ligand Binding Pockets for Advancing Drug Discovery. *Sci. Data* **2024**, *11*, 402.
(64) Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2017**, *45*, D271−D281.
(65) van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase−Ligand Interaction Space. *J. Med. Chem.* **2014**, *57*, 249−277.
(66) Shen, Z.; Yan, Y.-H.; Yang, S.; Zhu, S.; Yuan, Y.; Qiu, Z.; Jia, H.; Wang, R.; Li, G.-B.; Li, H. ProfKin: A Comprehensive Web Server for Structure-Based Kinase Profiling. *Eur. J. Med. Chem.* **2021**, *225*, 113772.
(67) Modi, V.; Dunbrack, R. L., Jr. Defining a New Nomenclature for the Structures of Active and Inactive Kinases. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 6818−6827.
(68) Modi, V.; Dunbrack, R. L., Jr. Kincore: A Web Resource for Structural Classification of Protein Kinases and Their Inhibitors. *Nucleic Acids Res.* **2022**, *50*, D654−D664.
(69) Shen, Q.; Wang, G.; Li, S.; Liu, X.; Lu, S.; Chen, Z.; Song, K.; Yan, J.; Geng, L.; Huang, Z.; Huang, W.; Chen, G.; Zhang, J. ASD v3.0: Unraveling Allosteric Regulation with Structural Mechanisms and Biological Networks. *Nucleic Acids Res.* **2016**, *44*, D527−D535.
(70) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Garmiri, P.; da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; de Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.
(71) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein−Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.
(72) Roy, A.; Ward, E.; Choi, I.; Cosi, M.; Edgin, T.; Hughes, T. S.; Islam, M. S.; Khan, A. M.; Kolekar, A.; Rayl, M.; Robinson, I.; Sarando, P.; Skidmore, E.; Swetnam, T. L.; Wall, M.; Xu, Z.; Yung, M. L.; Merchant, N.; Wheeler, T. J. MDRepo—an Open Data Warehouse for Community-Contributed Molecular Dynamics Simulations of Proteins. *Nucleic Acids Res* **2025**, *53*, D477−D486.
(73) Rusak, E.; Reizinger, P.; Juhos, A.; Bringmann, O.; Zimmermann, R. S.; Brendel, W. InfoNCE: Identifying the Gap Between Theory and Practice. *arXiv* **2024**.
(74) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR*, 1996; AAAI Press: Menlo Park, CA; pp. 226−231.
(75) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*; SciPy; 2010 pp. 92−96
(76) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng* **2007**, *9*, 90−95.

Downloaded from pubs.acs.org/jcisd8/article-pdf/doi/10.1021/acs.jcim.6c03114/6897649/acs.jcim.6c03114.pdf by guest on 28 October 2026