bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Decoding the Grammar of Protein–Protein Interaction Interfaces with Multimodal Representations

Yuri Gardinazzi[1,2], Edith Natalia Villegas Garcia[1,2], Sergio Senci[1], Davide Di Vora[1,2], Antonio Feltrin[1], and Francesca Cuturello*[1]

[1]Research and Technology Institute, AREA Science Park, Trieste 34149, Italy
[2]University of Trieste, Trieste, Italy

*Corresponding author: francesca.cuturello@areasciencepark.it

Abstract

Protein-protein interactions (PPI) govern essential cellular processes, making the computational identification of interacting sites a central challenge in structural biology, with important implications for protein engineering and the development of targeted therapeutics. Existing prediction algorithms include sequence-based methods, which lack structural information, or structure-based approaches, which often struggle to effectively integrate evolutionary context. Here, we present ESM3-PPISites, a supervised model for residue-level classification of PPI interfaces, leveraging the multimodal representations of the ESM3 Protein Language Model. To ensure a bias-free evaluation, we adopt a stringent redundancy filtering protocol, systematically eliminating latent homology between the training data and a curated benchmark set in both sequence and structural space.

Our findings demonstrate that while ESM3 largest proprietary version yields the highest predictive power, targeted fine-tuning of its small open-weight counterpart significantly narrows the performance gap. Requiring only primary sequence data at inference, ESM3-PPISites achieves unprecedented accuracy, vastly outperforming current approaches. Crucially, we demonstrate the practical impact of these predictions by integrating them as spatial restraints within the HADDOCK3 docking platform. When evaluated on an independent subset of 12 complexes from the Docking Benchmark v5, our prediction-guided pipeline strongly enhances the identification of near-native binding poses over *ab initio* blind docking, while reducing computational runtime by an order of magnitude. This framework establishes a scalable paradigm for high-throughput structural interactomics.

1 Introduction

Protein-protein interactions underlie most cellular processes, mediating from signal transduction and immune responses to large-scale macromolecular assembly. Decades of structural biology studies have established that PPI interfaces are shaped by physical and biochemical constraints that manifest as identifiable structural motifs [1, 2]. Leveraging this inherent predictability, machine learning models can effectively capture the complex, non-linear features that differentiate true interaction sites from the surrounding molecular surface across a wide array of interface configurations. Early systematic analyses of protein complexes revealed the key physical determinants of protein-protein recognition, demonstrating that interfaces are characterized by the burial of accessible surface area, tight atomic packing, and high shape complementarity [3, 4]. Comparative analyses demonstrate that PPI interfaces can be structurally distinguished from non-specific crystallographic contacts by their tighter packing and higher fraction of hydrophobic atoms [5]. Subsequently, the expansion of structural databases enabled a more granular characterization of these interfaces, revealing that they possess specific hydrophobicity profiles, residue propensities, and hydrogen-bonding networks [6]. Crucially, these molecular signatures vary systematically depending on the interaction's biological context [7]. For instance, the transition from large, hydrophobic obligate interfaces to smaller, polar surfaces characteristic of transient complexes reflects a functional 'tuning' of the protein surface [8]. From an energetic standpoint, the binding free energy of a PPI is not evenly distributed across these interfaces. Experimental and computational studies have established the concept of 'hot spots' as small, localized clusters of residues that contribute disproportionately to the binding affinity [9, 10]. Hot spots are typically enriched in bulky or aromatic amino acids, frequently shielded from the solvent by a ring of energetically minor residues. Their evolutionary conservation and specific topological arrangement serve as highly reliable indicators of interaction sites. Because the physicochemical patterns driving interfaces formation are highly complex and context-dependent, deciphering them requires advanced computational frameworks. To meet this challenge, a vast array of machine learning meth-

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ods has been developed, designed to identify PPI sites by capturing their underlying biophysical and evolutionary fingerprints [11, 12, 13]. Based on their input modality, these methods generally fall into two distinct classes: sequence-based and structure-based predictors.

Sequence-based methods attempt to identify binding sites directly from the primary amino acid chain. Historically, these approaches relied on standard classifiers, such as Feed-Forward Neural Networks [14, 15, 16], logistic regression [17, 18], Support Vector Machines [19, 20, 21], probabilistic frameworks that integrate evolutionary information with solvent accessibility [22], and Random Forest [23, 24, 25]. More recently, the field increasingly integrated deep learning architectures, such as Convolutional and Recurrent Neural Networks, facilitating more expressive representation learning [26]. A major paradigm shift occurred with the advent of Protein Language Models (PLMs) [27]: by leveraging high-dimensional embeddings learned from massive sequence databases, these approaches have demonstrated that pre-trained representations can largely bypass the need for explicitly computed evolutionary descriptors [28, 29].

Despite the scalability of sequence-only models, binding interfaces are fundamentally dictated by 3D spatial constraints. Consequently, structure-based methods have evolved from simple classification algorithms using handcrafted descriptors [30, 31, 32] into a diverse methodological spectrum tailored for 3D coordinate data. This evolution spans both the refinement of explicit biophysical potentials [33] and the emergence of geometry-aware architectures, such as Graph Neural Networks, which learn spatial constraints directly from raw coordinates [34, 35]. A major breakthrough in this domain has been the application of Geometric Deep Learning, which enabled to better capture the geometric and chemical features defining interaction sites [36, 37, 38, 39]. The convergence of geometric architectures and biophysical principles has culminated in end-to-end systems like AlphaFold3 [40], predicting structures of macromolecular complexes with unprecedented accuracy.

While sequence-based methods for PPI site prediction often lack explicit structural awareness, structure-based approaches may fail to capture evolutionary information. Protein Language Models provide a powerful framework to bridge these complementary aspects by learning the latent evolutionary patterns underlying molecular recognition through large-scale self-supervised training [41, 42, 43, 44, 45]. Multimodal PLMs, integrating explicit structural and solvent accessibility information, are particularly suited to capture the determinants of binding propensity.

Building on these advances, we leverage the ESM3 model [46] within a supervised residue-level classification framework to identify protein interaction interfaces. By combining ESM3 embeddings with a lightweight dense classifier, the proposed approach significantly outperforms existing methodologies. To prevent data leakage, homology biases were rigorously eliminated across both sequence and structural spaces. Despite substantial differences in training data composition and scale, the model consistently maintained strong predictive performance, highlighting its ability to effectively exploit structural diversity and generalize across unseen proteins. Finally, the predicted interaction sites were integrated into the HADDOCK docking platform [47] as ambiguous interaction restraints, substantially reducing the conformational search space and improving the recovery of near-native protein complexes. Overall, the proposed framework provides an accurate and scalable strategy for interface prediction, supporting large-scale characterization of protein interaction networks.

2 Methods

2.1 Datasets and Curation

We employ two data sources for model training: the first is the **BioLiP** dataset [48], for which we adopt the provided validation set; the second is an in-house dataset derived from **PDBbind** v2020 [49, 50]. For the latter, we select an initial pool of 15,000 protein chains, discarding complexes involving nucleic acids and peptides, and perform MMseqs2 clustering at 25% sequence similarity. We then construct the validation set by sampling 10% of the sequences from the smallest clusters, in order to maximize heterogeneity for grid-search parameter tuning: this ensures that the optimization process is guided by the most diverse and least redundant subset of the sequence space. Residue-level interface labels are obtained from the reference PDB complexes by identifying residue pairs in contact across interacting chains. We define contacts using the standard van der Waals distance–based criterion, as commonly adopted in prior works [11, 48], identifying residue pairs whose atoms fall within a 0.5Å distance threshold.

To evaluate model generalization on strictly unseen test set, we utilize the **ZK448 benchmark** [11]. Originally compiled by Zhang and Kurgan and retrieved via the PIPENN method repository [51], this dataset is widely regarded as a 'gold standard' for protein-protein interaction site prediction. It consists of 336 non-redundant protein chains derived from high-resolution hetero-complexes, specifically curated to ensure a maximum pairwise sequence identity of 25%. We implement a dual-stage redundancy filtering protocol between the training sets and the ZK448 benchmark to ensure that our model learns general biophysical principles, rather than hidden homology:

1. **Sequence-Level Filtering:** MMseqs2 [52] is employed to remove all proteins from the training sets sharing more than 25% sequence identity with any entry in the ZK448 benchmark dataset.

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

2. **Structural-Level Filtering:** Foldseek [53] is employed to remove remaining structural homologs from the training sets, utilizing ESM3-1.4B predicted structures and experimental PDB when availbale. As illustrated in Figure 1, training samples producing structural matches against the benchmark with an E-value below 0.001 and a target/query-normalized TM-score greater than 0.5 are excluded.

The filtering protocol results in final training sets comprising 3,693 proteins for BioLiP and 1,409 for the PDBbind-derived dataset. We also curate an independent dataset derived from the Protein-Protein Docking Benchmark version 5.0 (**DB5**) [54] to evaluate the impact of our predicted interaction sites within HADDOCK-guided protein complex reconstruction. We restrict our selection to DB5 entries with exactly two indicated interacting chains, yielding a starting pool of over 100 dimeric complexes. To guarantee a strictly unbiased evaluation, we apply the dual-stage homology filtering protocol described above against the training sets. Finally, we excluded complexes exhibiting an unbound-to-bound Root Mean Square Deviation (RMSD) greater than 10Å to prevent large structural deviations from confounding our iRMSD-based performance assessment. This procedure results in a robust evaluation set of 12 non-redundant complexes derived from a docking benchmark set. The selected protein complexes are reported in Table 1.

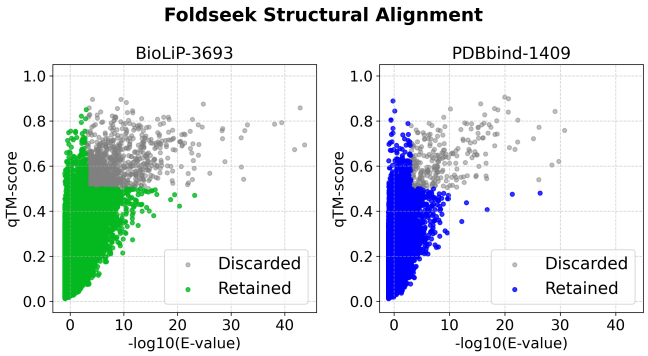


Figure 1: ***Data Leakage Prevention via Structural Homology Filtering.*** *Foldseek alignment to quantify structural similarity between each train-test pair by qTM-score and statistical significance, between the ZK448 test and the two training sets: BioLiP (left) and PDBbind (right). Samples exhibiting significant homology are discarded (gray points). The final training sets names BioLiP-3693 and PDBbind-1409 denote the final count of non-redundant training proteins (colored regions) retained after the filtering process.*

2.2 ESM3 Refinement for Interaction Site Identification

We implement a frozen-weight transfer learning baseline and an end-to-end fine-tuning approach to adapt the ESM3 model [46] for PPI site identification. Relying exclusively on the protein sequence as input, we extract embedding representations from the final hidden layer of the open-weight ESM3 (1.4B), and the large-scale, proprietary version (98B). We classify interface residues by passing representations through a lightweight downstream head, consisting of a single fully connected layer. The network is optimized using a binary cross-entropy loss applied to the resulting classifier logits. To account for the severe class imbalance, the loss incorporates a dynamic positive class weighting scheme. Training was performed using the AdamW optimizer with early stopping based on the validation loss, using a tolerance of 5 epochs without improvement. In the baseline setting, the models were kept frozen and L2 normalization was applied to each residue embedding prior to classification. In this setting, we adopt a learning rate of $5 \cdot 10^{-4}$ and a batch size of 96. When finetuning the smaller model, gradients are backpropagated through the transformer blocks of the ESM3 models, allowing the representations to adapt to the downstream task. To stabilize training, Layer Normalization was applied across the hidden dimension. For both the PDBbind and BioLip datasets, we apply a learning rate of $2 \cdot 10^{-5}$ and a weight decay of 0.5, utilizing a batch size of 1 with gradient accumulation over 4 steps. The larger ESM3 variant was used in baseline mode only, since this model is not publicly available and does not permit end-to-end finetuning. All computations were conducted using NVIDIA DGX A100 and DGX H100 systems.

2.3 Translation of Interface Predictions into Docking Restraints

We evaluate the possible application of our model in a protein-protein interaction prediction context by performing HADDOCK3 experiments [55] on a set of 12 complexes with experimentally determined structures. The performance of our prediction-guided docking is compared against both an ab initio blind docking control and a positive control using native-contact restraints through the HADDOCK CAPRIeval protocol, calculating the interface iRMSD relative to the native complex. For the unrestrained **ab initio** runs, we employ a rigid-body sampling of 10,000 models followed by flexible refinement of the best 200 models. While extended sampling ranges were explored, increasing to 15,000 trials provides no performance improvement, and halving to 5,000 trials proved insufficient to sample high quality configurations. This confirms that 10,000 trials represent an optimal balance of computational efficiency and accuracy.

The **oracle** upper bound for docking performance is established using a ground-truth reference utilizing native interfacial contacts extracted from the experimental complexes. These interactions are defined as $C\alpha$ pairs on opposing chains within an 8Å distance threshold, and modeled as unambiguous distance restraints with a tar-

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Complex	Protein 1	Protein 2	Protein names
1US7_A:B	2FXS_A	2W0G_A	Heat shock protein 82 N-ter domain / HSP 90 co-chaperone CDC37 C-ter domain
1Z5Y_D:E	1L6P	2B1K_A	N-term of DsbD E.coli / CCMG protein
2YVJ_A:B	2YVF_A	2E4P_A	Ferredoxin reductase BPHA4 / Biphenyl dioxygenase ferredoxin subunit
3PC8_A:C	3PC6_A	3PC7_A	DNA repair protein XRCC1 / DNA ligase III-alpha BRCT domain
4H03_A:B	1GIQ_A	1IJJ_A	Iota toxin component IA / Alpha actin
1KXP_A:D	1IJJ_B	1KW2_B	Actin / Vitamin D binding protein
1ZHH_A:B	1JX6_A	2HJE_A	Autoinducer 2-binding periplasmic protein LuxP / Autoinducer 2 sensor kinase / phosphatase LuxQ
1ZHI_A:B	1M4Z_A	1Z1A_A	BAH domain of Orc1 / Sir Orc-interaction domain
2A5T_A:B	1Y20_A	2A5S_A	NMDA receptor R1-4A subunit ligand-binding core / NMDA receptor R2A subunit ligand-binding core
2AJF_A:E	1R42_A	2GHV_E	ACE2 / SARS spike protein receptor binding domain
1JK9_B:A	1QUP_A	2JCW_A	CCS metallochaperone / SOD1 superoxide dismutase
2C0L_A:B	1FCH_A	1C44_A	PTS1 and TRP region of PEX5 / SCP2

Table 1: ***Protein complexes for leakage-free evaluation obtained from the DB5 database.***

get of 10.0Å (6.0Å lower and 4.0Å upper bounds). For this reference run, we generate 100 rigid-body models and refine the top 10 candidates via energy minimization.

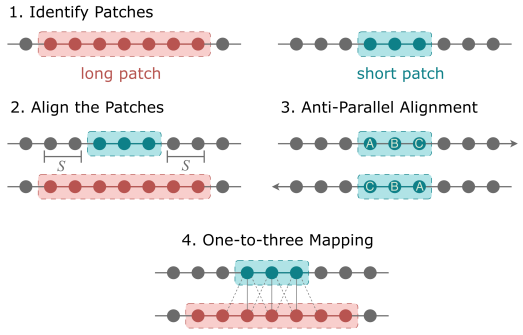


Figure 2: ***Mapping of predictions onto docking restraints.*** *Patches are identified as contiguous predictions (1) and aligned by centering the shorter patch (2). Both parallel and anti-parallel orientations are evaluated (3), and a one-to-three residue mapping to the longer patch defines ambiguous interaction restraints.*

To drive the docking process, we implement a **prediction-guided** HADDOCK protocol that maps predicted residue-level interface of single proteins onto spatial distance restraints between the interacting partners. First, we cluster predicted residues into 'patches', defined as the subsets separated by a maximum of three positions in the sequence. We select the two patches per chain with the highest mean predicted probability and systematically evaluate all patch-wise combinations between the two chains. For each pair, the shorter patch (L_{short}) is centered relative to the longer one (L_{long}) using an offset $s = (L_{\text{long}} - L_{\text{short}})/2$. This alignment provides the anchor for our restraints; specifically, to account for registration uncertainties, we map each residue $r_{\text{short},i}$ to a three-residue window on the longer patch, spanning the aligned position $r_{\text{long},s+i}$ and its flanking neighbors $r_{\text{long},s+i\pm1}$ (Figure 2). To accommodate both parallel and anti-parallel binding modes, the shorter patch is evaluated in both forward and reverse orientations. Contacts are then enforced via Ambiguous Interaction Restraints, which use the same distance parameters as the oracle reference. We process each set of constraints through an independent docking run consisting of 100 rigid-body models, followed by the selection and energy minimization of the best-performing model. The resulting configurations are then ranked by HADDOCK score for final evaluation.

3 Results

3.1 Performance Evaluation

In the context of strongly unbalanced classifications, where positive interface residues are vastly outnumbered by non-interacting residues, we employ a comprehensive suite of metrics. The F1-score isolates success on the critical minority class by disregarding the massive true-negative rate. Complementing this, the MCC provides a symmetric, balanced summary of the entire confusion matrix, ensuring a high score is achieved only if the model performs well on both the minority and majority classes. As illustrated in Figure 3, the baseline 1.4B parameter model (ESM3-1.4B) establishes a solid foundation, while scaling to the 98B parameter model (ESM3-98B) consistently delivers superior performance. However, supervised fine-tuning (FT-ESM3-1.4B) significantly narrows the gap with the larger 98B variant. This finding is particularly relevant for the community, as it suggests that targeted optimization on domain-specific data can robustly compensate for the lower native capacity of smaller models, providing an efficient alternative to massive, proprietary architectures. Notably, the models achieve strong predictive performance across both training sets, despite a significant size disparity (3,693 versus 1,409 proteins). While the larger dataset yields higher

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

absolute scores, the consistent performance on an independent smaller set suggests that the model is effectively capturing recurrent and relevant interface patterns.

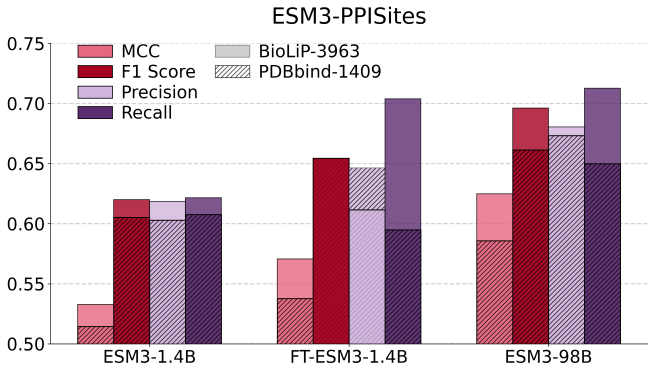


Figure 3: ***Performance evaluation of ESM3-based architectures.*** *Baseline (ESM3-1.4B), scaled (ESM3-98B) and fine-tuned (FT-ESM3-1.4B) models evaluation on the ZK448 test set (through MCC, F1 Score, Precision, and Recall metrics). Results are stratified by training source, comparing BioLiP-3963 (solid) aginst PDBbind-1409 (hatched) datasets.*

3.2 Comparison with Existing Methods

We designate the best-performing model as **ESM3-PPISites** and compare its performance against a diverse array of established sequence- and structure-based methods. Metrics for sequence-based methods are obtained from [28], while PeSTo is evaluated using the official implementation released by the authors. As summarized in Table 2, ESM3-PPISites achieves state-of-the-art results across the considered metrics, significantly outperforming all existing approaches. Specifically, ESM3-PPISites attains an F1-score of 0.696 and an MCC of 0.625, representing a substantial margin over PeSTo [38], a high-performance geometric deep learning architecture. This gap suggests that the multimodality of ESM3, which integrates sequence and structure during generative pre-training, captures more nuanced biophysical features than models relying exclusively on the protein geometry. The performance advantage is even more pronounced when compared to purely sequence-based models, where ESM3-PPISites nearly doubles the F1-score and MCC of the top-performing baseline. The lower performance of sequence-based models underscores the inherent difficulty of interface prediction when spatial context is absent. Overall, ESM3-PPISites demonstrates superior discriminative power, maintaining high specificity while identifying interaction sites with high confidence.

3.3 Analysis of Predictions

A detailed analysis of the predictive behavior of ESM3-PPISites on the independent ZK448 benchmark, both at the residue and protein levels, focuses on the physicochemical properties of the identified residues and the consistency of the model's performance across individual protein targets. As illustrated by the log-odds ratio of predicted interactions grouped by amino acid class (Figure 4A), the model successfully recovers the established signatures of protein-protein interfaces. The predictions demonstrate a marked enrichment for hydrophobic and aromatic residues, which are well known to dominate binding interfaces and serve as energetic hot spots. Conversely, polar and charged residues display lower predicted interaction frequencies, consistent with the biophysical constraints governing protein-protein molecular recognition. Furthermore, the distribution of Relative Solvent Accessibility for the predicted interface residues (Figure 4B) shows that the model accurately targets surface-exposed regions. The density sharply increases and peaks above the standard 0.2 consensus cutoff for solvent accessibility[56], indicating that the model hardly predicts buried core residues as interaction sites. Remarkably, when quantifying this surface-exposed population, the ratio of predicted interacting residues to the total number of exposed amino acids perfectly matches the ground-truth proportion of truly interacting residues within the exposed surface (26%). This alignment reveals two critical insights: first, it confirms that the model is not simply a surface-exposure detector that over-predicts accessible sites, but instead remains highly selective; second, it demonstrates that the model preserves the native interface density observed in experimentally resolved complexes, accurately reproducing the true spatial footprint of interaction sites across the protein surface.

At the protein level, the model exhibits highly robust generalization across the diverse benchmark. The scatter plot comparing the F1-score and MCC shows that the majority of the proteins achieve scores well above 0.5 in both metrics, underscoring stable performance across different structural topologies without severe pathological failures (Figure 4C). Finally, when comparing the true against the predicted fraction of interacting residues per protein (Figure 4D), we observe a strong correlation

Table 2: ***Performance on the ZK448 test set across models for interface residue classification.***

Model	Ref.	F1-score	MCC	AUC
Sequence-based:				
ESM3-PPISites		**0.696**	**0.625**	**0.913**
PIPENN-EMB	[28]	0.385	0.254	0.729
SCRIBER	[18]	0.333	0.230	0.715
SSWRF	[21]	0.287	0.178	0.687
CRFPPI	[23]	0.266	0.154	0.681
LORIS	[17]	0.263	0.151	0.656
SPRINGS	[16]	0.229	0.111	0.625
PSIVER	[22]	0.191	0.066	0.581
SPRINT	[20]	0.183	0.057	0.570
SPPIDER	[19]	0.198	0.071	0.517
Structure-based:				
Pesto	[38]	0.317	0.144	0.67

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

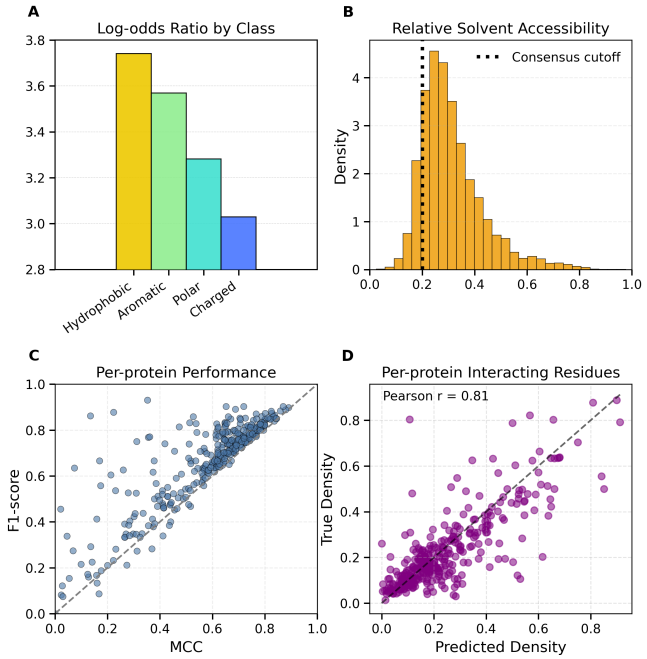


Figure 4: ***Analysis of predictions.(A)*** *Log-odds ratio of predicted interacting residues categorized by biochemical class.* ***(B)*** *Density distribution of Relative Solvent Accessibility for predicted interface residues.* ***(C)*** *Per-protein performance evaluation.* ***(D)*** *Correlation between the true and predicted fraction of interacting residues per protein.*

(Pearson r = 0.81), indicating that the model accurately scales its predictions to the expected size of the binding patch, without a systematic bias toward over- or under-predicting the interface area.

3.4 Guiding PPI Docking with ESM3-PPISites Restraints

When integrating the ESM3-PPISites output as spatial restraints in HADDOCK3, we evaluate whether it can successfully guide complex assembly and reduce the conformational search space. We measure the interface Root Mean Square Deviation (iRMSD) relative to the experimental structures of the complexes and compare the results of the prediction-guided protocol against a computationally intensive *ab initio* blind docking baseline and an *oracle* reference utilizing true native contacts. The *oracle* set-up serves as a positive control, and the procedure consistently yields configurations with an iRMSD below 5Å for all evaluated complexes.

As illustrated in Figure 5, we perform a side-by-side comparison for each complex, evaluating the lowest iRMSD configurations generated by both the ESM3-PPISites-guided protocol and the *ab initio* baseline, reporting their HADDOCK score-based rank on the y-axis. We define successful docking runs as those sampling at least one configuration with an iRMSD below a strict 10Å cutoff. Under these criteria, the ESM3-PPISites protocol achieves a high success rate, yielding low iRMSD models for 9 out of the 12 evaluated complexes. While two of these nine represent borderline cases (around 10Å iRMSD), the overall performance far exceeds the computationally exhaustive *ab initio* approach, which manages to sample acceptable configurations for only 4 complexes, half of which are themselves borderline successes. Remarkably, our prediction-guided approach achieves a high-resolution iRMSD lower than 5Å in 4 successful runs (1Z5Y, 1ZHH, 2YVJ, 3PC8), and these highly accurate configurations place within the top three models ranked by HADDOCK score. Taken together, these results demonstrate that leveraging predicted interface patches as Ambiguous Interaction Restraints facilitates successful docking.

Beyond improving the quality of the resulting structural ensembles, the integration of ESM3-PPISites predictions significantly accelerates the docking process. Figure 6 details the computational runtime across the three benchmarking strategies carried out on a 64-cores AMD EPYC 9374F node. The unguided ab initio protocol exhibits massive computational overhead, requiring between 10 and 40 minutes per complex due to the broad rigid-body sampling required to blindly explore the interaction landscape. In contrast, restricting the search space using our predicted patches drastically truncates the required runtime to under 10 minutes for all evaluated complexes, mirroring the optimal computational efficiency of the oracle reference. Taken together, these findings highlight that translating our ESM3-PPISites predictions into spatial restraints for Haddock3 simultaneously enhances the resolution of the predicted macromolecular complexes and eliminates the severe computational bottlenecks associated with traditional blind docking.

4 Discussion

The results presented in this work support a central claim: PPI interfaces possess a learnable 'grammar' effectively captured through multimodal representation learning. By leveraging ESM3 embeddings within a supervised framework, we establish a new state-of-the-art for residue-level interface prediction, consistently outperforming established benchmarks while relying solely on primary sequence at inference time. Our results indicate that the model's multimodal generative pretraining fosters the emergence of an expressive, high-dimensional feature space that captures not only local geometry but also evolutionary and long-range dependencies that govern protein interactions. The superior predictive accuracy of ESM3-PPISites serves as direct evidence that these embeddings have successfully internalized the biophysical rules of interface formation and functional binding potential.

A notable finding is the efficiency-performance trade-off highlighted by the comparison between model scales and training regimes. While the large ESM3-98B model achieves the highest absolute performance, fine-tuning the smaller 1.4B variant substantially narrows this gap.

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

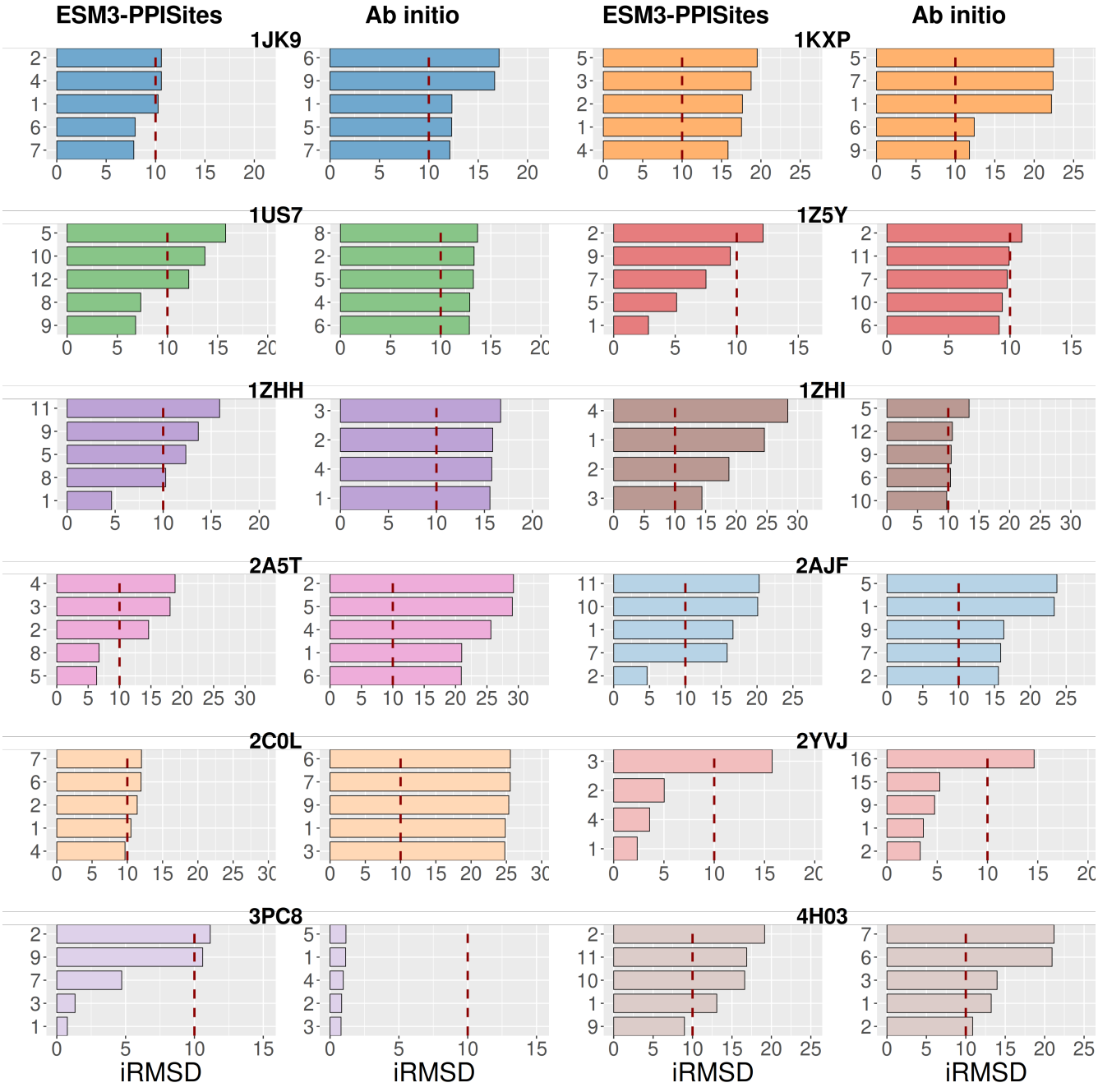


Figure 5: ***Comparison of docking performance between ESM3-PPISites-guided and ab initio protocols.*** *The x-axis represents the iRMSD relative to the experimental structures. Models are ordered on the y-axis by their iRMSD values, with their corresponding HADDOCK score-based ranks displayed as labels. The dashed line indicates the strict 10Å cutoff used to define a successful docking run.*

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

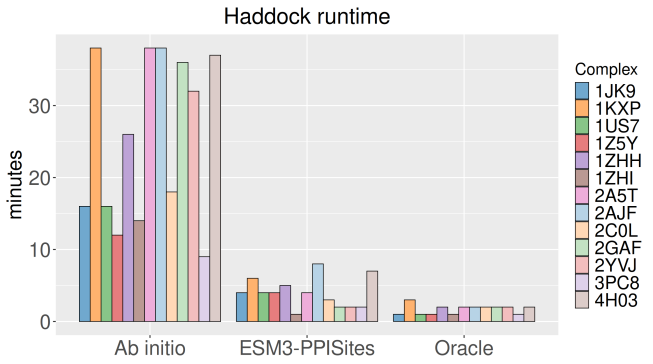


Figure 6: ***HADDOCK3 runtimes.*** *ab initio, ESM3-PPISites-guided, and oracle protocols for each complex.*

This shows that task-specific adaptation can compensate for reduced model capacity, making high-quality interface prediction more accessible without reliance on proprietary large-scale models.

Finally, the integration of predicted interfaces into docking workflows provides strong evidence of their practical utility. In this context, the prediction-guided protocol offers a substantial improvement over the sampling without guidance: by incorporating residue-level interface predictions as ambiguous interaction restraints, we achieve near-native docking solutions in 9 out of 12 complexes, representing a threefold increase in success rate over the unrestrained baseline. Importantly, this gain is accompanied by a significant reduction in computational cost: the guided protocol effectively reduces the conformational search space, enabling an order-of-magnitude speed-up compared to exhaustive ab initio sampling.

In conclusion, this work demonstrates that multimodal protein language models provide a powerful framework for decoding the rules of protein-protein recognition, combining strong predictive performance and tangible improvements in docking applications. Building upon these findings, future work will exploit the high-confidence interface predictions generated by ESM3-PPISites to design models capable of directly inferring protein-protein interactions. Ultimately, these framework holds the potential to decode interaction specificity and map partner compatibility at the proteome scale.

5 Availability of Data and Software

All code used in this work is publicly available in a GitHub repository, which also includes a user-friendly implementation for blind prediction: `https://github.com/RitAreaSciencePark/ESM3-PPISites`. ESM3 representations and PDB files are store in Zenodo at `https://doi.org/10.5281/zenodo.20366947`. The trained model are accessible via Hugging Face at `https://huggingface.co/area-science-park/ESM3-PPISites`. To facilitate ease of use, a web-based inference dashboard is also provided, enabling real-time predictions without local environment configuration: `https://huggingface.co/spaces/area-science-park/esm-ppisites`.

6 Authors' Contribution

All authors contributed equally to this work. Y.G. developed and optimized the fine-tuning pipeline for the ESM3-1.4B model, implemented the Foldseek-based data curation strategy, and managed the GitHub repository. E.N.G.V. focused on the finetuning of the ESM3-1.4B and ESM2 models, the extraction of representations from the ESM3-98B model, and the Hugging Face deployment. S.S. carried out predictions' analyses on the ZK448 test set, performed the PeSTo model evaluation and supported data curation. D.D. curated and optimized the code for baseline ESM3, contributed to predictions' analysis, and managed the Zenodo repository. A.F. assembled the PDBbind-derived dataset, performed data curation, and contributed to prediction analysis on the ZK448 test set. F.C. supervised and coordinated the study and led the writing of the manuscript. All authors contributed to the manuscript writing and revision.

7 Acknowledgments & Fundings

The authors acknowledge the Area Science Park supercomputing platform ORFEO made available for conducting the research reported in this article and the technical support from the team of the Laboratory of Data Engineering. Y.G. is supported by the Programma Nazionale della Ricerca (PNR) grant J95F21002830001 with the title 'FAIR-by-design'. E.N.V.G. was supported by the project PON 'BIO Open Lab (BOL) - Rafforzamento del capitale umano' (CUP J72F20000940007). S.S. is supported by I-HUB FVG - INNOVATION HUB FVG (CUP J23B24000030002). D.D. is supported by Progetto 'QuB - Quantum Behavior in Biological Functions' (CUP J95F21002820001) financed by 'Fondo ordinario per gli enti e le istituzioni di ricerca' (FOE) 2021.

8 Competing Interests

The authors declare no competing interests.

9 Use of Large Language Models

Large language models were used in an open-access setting to provide limited support for language editing. Their use was restricted to auxiliary tasks such as improving text clarity, without involvement in the scientific design of this work.

References

[1] Cyrus Chothia and Joël Janin. Principles of protein–protein recognition. *Nature*, 256(5520):705–708, 1975.

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[2] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[3] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198, 1999.

[4] Teresa A Larsen, Arthur J Olson, and David S Goodsell. Morphology of protein–protein interfaces. *Structure*, 6(4):421–427, 1998.

[5] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. A dissection of specific and non-specific protein–protein interfaces. *Journal of molecular biology*, 336(4):943–955, 2004.

[6] Changhui Yan, Feihong Wu, Robert L Jernigan, Drena Dobbs, and Vasant Honavar. Characterization of protein–protein interfaces. *The protein journal*, 27(1):59–70, 2008.

[7] Mu Gao and Jeffrey Skolnick. Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107(52):22517–22522, 2010.

[8] Irene MA Nooren and Janet M Thornton. Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492, 2003.

[9] Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

[10] Irina S Moreira, Pedro A Fernandes, and Maria J Ramos. Hot spots—a review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4):803–812, 2007.

[11] Jian Zhang and Lukasz Kurgan. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Briefings in bioinformatics*, 19(5):821–837, 2018.

[12] Farzan Soleymani, Eric Paquet, Herna Viktor, Wojtek Michalowski, and Davide Spinello. Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, 20:5316–5341, 2022.

[13] Minhyeok Lee. Recent advances in deep learning for protein-protein interaction analysis: A comprehensive review. *Molecules*, 28(13):5169, 2023.

[14] Yanay Ofran and Burkhard Rost. Predicted protein–protein interaction sites from local sequence information. *FEBS letters*, 544(1-3):236–239, 2003.

[15] Yanay Ofran and Burkhard Rost. Protein–protein interaction hotspots carved into sequences. *PLoS computational biology*, 3(7):e119, 2007.

[16] Gurdeep Singh, Kaustubh Dhole, Priyadarshini P Pai, and Sukanta Mondal. Springs: prediction of protein-protein interaction sites using artificial neural networks. Technical report, PeerJ PrePrints, 2014.

[17] Kaustubh Dhole, Gurdeep Singh, Priyadarshini P Pai, and Sukanta Mondal. Sequence-based prediction of protein–protein interaction sites with l1-logreg classifier. *Journal of theoretical biology*, 348:47–54, 2014.

[18] Jian Zhang and Lukasz Kurgan. Scriber: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*, 35(14):i343–i353, 2019.

[19] Aleksey Porollo and Jarosław Meller. Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 66(3):630–645, 2007.

[20] Ghazaleh Taherzadeh, Yuedong Yang, Tuo Zhang, Alan Wee-Chung Liew, and Yaoqi Zhou. Sequence-based prediction of protein–peptide binding sites using support vector machine. *Journal of computational chemistry*, 37(13):1223–1229, 2016.

[21] Zhi-Sen Wei, Ke Han, Jing-Yu Yang, Hong-Bin Shen, and Dong-Jun Yu. Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests. *Neurocomputing*, 193:201–212, 2016.

[22] Yoichi Murakami and Kenji Mizuguchi. Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15):1841–1848, 2010.

[23] Zhi-Sen Wei, Jing-Yu Yang, Hong-Bin Shen, and Dong-Jun Yu. A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE transactions on nanobioscience*, 14(7):746–760, 2015.

[24] Qingzhen Hou, Paul FG De Geest, Christian J Griffioen, Sanne Abeln, Jaap Heringa, and K Anton Feenstra. Serendip: Sequential remastering to derive profiles for fast and accurate predictions of ppi interface positions. *Bioinformatics*, 35(22):4794–4796, 2019.

[25] Qingzhen Hou, Bas Stringer, Katharina Waury, Henriette Capel, Reza Haydarlou, Fuzhong Xue, Sanne Abeln, Jaap Heringa, and K Anton Feenstra. Serendip-ce: sequence-based interface predic-

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

tion for conformational epitopes. *Bioinformatics*, 37(20):3421–3427, 2021.

[26] Yiwei Li, G Brian Golding, and Lucian Ilie. Delphi: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics*, 37(7):896–904, 2021.

[27] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

[28] David PG Thomas, Carlos M Garcia Fernandez, Reza Haydarlou, and K Anton Feenstra. Pipenn-emb ensemble net and protein embeddings generalise protein interface prediction beyond homology. *Scientific Reports*, 15(1):4391, 2025.

[29] SeyedMohsen Hosseini, G Brian Golding, and Lucian Ilie. Seq-insite: sequence supersedes structure for protein interaction site prediction. *Bioinformatics*, 40(1):btad738, 2024.

[30] Piero Fariselli, Florencio Pazos, Alfonso Valencia, and Rita Casadio. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, 269(5):1356–1361, 2002.

[31] Thomas C Northey, Anja Barešić, and Andrew CR Martin. Intpred: a structure-based predictor of protein–protein interaction sites. *Bioinformatics*, 34(2):223–229, 2018.

[32] Subhrangshu Das and Saikat Chakrabarti. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific Reports*, 11(1), January 2021.

[33] Bowen Dai and Chris Bailey-Kellogg. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17):2580–2588, 2021.

[34] Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Manon Réau, Nicolas Renaud, Li C Xue, and Alexandre MJJ Bonvin. Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759, 2023.

[36] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature methods*, 17(2):184–192, 2020.

[37] Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, 2023.

[38] Lucien F Krapp, Luciano A Abriata, Fabio Cortés Rodriguez, and Matteo Dal Peraro. Pesto: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature communications*, 14(1):2175, 2023.

[39] Carlos HM Rodrigues and David B Ascher. Csm-potential2: A comprehensive deep learning platform for the analysis of protein interacting interfaces. *Proteins: Structure, Function, and Bioinformatics*, 93(1):209–216, 2025.

[40] Josh Abramson, Jonah Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024.

[41] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

[42] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

[43] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):723, 2019.

[44] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pages 2020–12, 2020.

[45] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

[46] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[47] Cyril Dominguez, Rolf Boelens, and Alexandre MJJ Bonvin. Haddock: A protein–protein docking ap-

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.29.728739; this version posted June 2, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

proach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

[48] Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

[49] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004.

[50] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015. Updated version 2020 accessed from http://www.pdbbind-cn.org/.

[51] Bas Stringer, Hans de Ferrante, Sanne Abeln, Jaap Heringa, K Anton Feenstra, and Reza Haydarlou. Pipenn: protein interface prediction from sequence with an ensemble of neural nets. *Bioinformatics*, 38(8):2111–2118, 2022.

[52] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[53] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

[54] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, Alexandre MJJ Bonvin, and Zhiping Weng. Updates to the integrated protein–protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, 427(19):3031–3041, 2015.

[55] Marco Giulini, Victor Reys, João M. C. Teixeira, Brian Jiménez-García, Rodrigo V. Honorato, Anna Kravchenko, Xiaotong Xu, Raphaëlle Versini, Anna Engel, Stefan Verhoeven, and Alexandre M. J. J. Bonvin. Haddock3: A modular and versatile platform for integrative modeling of biomolecular complexes. *Journal of Chemical Information and Modeling*, 65(13):7315–7324, 2025.

[56] Castrense Savojardo, Matteo Manfredi, Pier Luigi Martelli, and Rita Casadio. Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Frontiers in molecular biosciences*, 7:626363, 2021.