Unified Structure-Aware Sequence Framework for Protein-Peptide Interaction

Aanisha S[1], Arjun N[2], Priyadharshini U[3], Mohamed Jamal Basha K[4], Dr. P. Kola Sujatha[5]
[1,2,3,4] *Department of Information Technology Madras Institute of Technology Campus, Anna University Chennai, India*
[5] *Associate Professor & Supervisor Department of Information Technology Madras Institute of Technology Campus, Anna University Chennai, India*

***Abstract*—Protein–peptide interactions play a fundamental role in numerous biological processes including cell signaling, immune response, molecular recognition, and drug targeting. Accurate identification of peptide-binding regions in proteins is essential for understanding biological mechanisms and designing therapeutic molecules. Existing computational approaches primarily focus on interaction prediction using sequence-based learning techniques, while providing limited structural interpretation and lacking mutation-aware analysis. This paper proposes a unified structure-aware sequence framework for protein–peptide interaction analysis using geometric deep learning and structural topology modeling. The framework integrates Evolutionary Scale Modeling (ESM-2) embeddings with Graph Neural Networks (GNNs) to capture both evolutionary and spatial characteristics of protein structures. A structure-aware groove localization module is introduced to identify potential peptide-binding grooves using residue-level geometric representations. To improve structural understanding, a Concavity-Gated Equivariant Graph Neural Network (CG-EGNN) is proposed, incorporating surface concavity and orientation information for geometry-aware message passing.**

In addition, a mutation sensitivity analysis framework is developed to evaluate the effect of amino acid substitutions on groove probability, concavity, and residue orientation. A pocket classification module using DBSCAN clustering and an improved 1D CNN is employed to identify exact binding pocket regions from protein surface structures. Finally, an LLM-based explanation module using Mistral-7B-Instruct generates human-readable interpretations for mutation-induced structural changes.

Experimental evaluation on PepBDB and SKEMPI datasets demonstrates that the proposed framework effectively localizes peptide-binding grooves, predicts pocket regions with high accuracy, and provides meaningful structural interpretation of mutation effects.

I. INTRODUCTION

A. Protein–Peptide Interactions and Their Biological Importance

Protein–peptide interactions are central to numerous biological and cellular processes, including intracellular signaling, immune regulation, enzyme inhibition, and molecular recognition. Peptides bind to specific regions on protein surfaces, enabling communication and regulatory mechanisms essential to cellular function. Accurate identification of the location, strength, and stability of such interactions is therefore critical for understanding disease mechanisms and for guiding the rational design of peptide-based therapeutics. Recent advances in deep learning have substantially improved computational modeling of protein-peptide interactions by enabling models to learn discriminative structural and sequence-derived features directly from protein data, in place of traditional handcrafted, energy-based, or docking-driven approaches.

B. Limitations of End-to-End Interaction Prediction

Despite these advances, most existing deep learning models for protein–peptide interaction prediction adopt a single-stage, end-to-end formulation that maps entire protein structures directly to interaction outcomes. This formulation places an unrealistic burden on the model, which must simultaneously localize relevant surface regions and learn fine-grained interaction patterns without any intermediate structural guidance. Because binding is an inherently

localized phenomenon, operating over the full protein surface results in diluted learning, unstable predictions, and reduced sensitivity to biologically relevant regions. Furthermore, end-to-end models typically lack intermediate supervision, making early representation errors difficult to isolate or correct, and are commonly restricted to a single predictive objective, limiting their extensibility to related tasks such as mutation impact analysis.

C. Need for Structural Localization Prior to Interaction Learning

Protein–peptide interactions are governed jointly by spatial feasibility and biochemical compatibility. Structural representations define the geometric arrangement of residues and determine physical accessibility, but geometry alone cannot explain interaction specificity, since spatially proximal residues may fail to interact due to unfavorable chemical properties. Conversely, sequence-derived or physicochemical features capture interaction tendencies but lack awareness of spatial constraints, which can produce predictions that are chemically plausible yet structurally infeasible. This motivates a modeling strategy in which structural features determine *where* interaction should be evaluated, while functional features determine *whether* that interaction is favorable requiring their explicit alignment rather than independent or loosely coupled treatment.

D. Motivation for a Structured, Mutation-Aware Framework

Most contemporary models used for protein–peptide interaction prediction operate as black boxes, providing outputs without exposing the structural or biochemical reasoning behind them. This limits their usefulness in scientific and therapeutic design contexts, where interpretability is essential for experimental validation. In addition, protein interactions are highly sensitive to residue-level perturbations: small amino acid substitutions can significantly alter binding affinity, destabilize existing interactions, or create new ones. Existing computational frameworks rarely incorporate such mutation-level effects directly into their prediction pipeline. This motivates the development of a structured, multi-stage, and mutation-aware framework that progressively refines interaction understanding from coarse structural localization to fine-grained, interpretable, and quantitative interaction analysis.

E. Contributions of This Paper

The main contributions of this work are summarized as follows:

- A unified, multi-stage framework for protein–peptide interaction analysis that decomposes the problem into groove localization, pocket classification, binding affinity prediction, and mutation sensitivity analysis, in place of conventional single-stage, end-to-end prediction.
- A Concavity-Gated Equivariant Graph Neural Network (CG-EGNN) that incorporates surface concavity and residue orientation into geometry-aware message passing, improving groove localization.
- A structure-aware integration strategy that aligns ESM-2 sequence embeddings with geometric graph features.
- A mutation sensitivity and recommendation module that quantifies the impact of residue-level substitutions on groove probability, concavity, and orientation, and ranks candidate mutations.
- A comprehensive experimental evaluation on the PepBDB and SKEMPI datasets, demonstrating effective groove localization, high-accuracy pocket classification, and reliable binding affinity prediction.

II. RELATED WORK

A. Sequence-Based Deep Learning for Interaction Prediction

CAMP employs convolutional neural networks over amino acid composition, physicochemical properties, and evolutionary information to identify interacting pairs, while TPepPro adopts a transformer-based architecture to capture longer-range sequence dependencies. These approaches are computationally efficient and scalable; however, they lack explicit three-dimensional structural awareness, which limits their ability to localize precise binding regions and restricts generalization to novel protein–peptide pairs with limited evolutionary information.

B. Structure-Based and Geometric Deep Learning Approaches

The availability of large-scale structural data has enabled models that incorporate spatial and geometric information directly. MaSIF applies geometric deep learning to protein surfaces to identify potential interaction fingerprints, effectively capturing surface-level and spatial features. Graph Neural Networks (GNNs) and Equivariant Graph Neural Networks (EGNNs) have further advanced structural biology applications by modeling residues as graph nodes and learning spatial relationships through message passing, while preserving equivariance to rotation and translation. Despite these advances, most such models process the entire protein surface without an explicit intermediate localization stage, which introduces noise and reduces both efficiency and interpretability.

C. Protein Language Models and Sequence Embeddings

Protein language models such as ESM (Evolutionary Scale Modeling) have enabled the use of sequence embeddings that encode rich evolutionary and biochemical context learned from large-scale unsupervised pretraining. These embeddings improve generalization and capture functional patterns beyond what geometric features alone can represent. However, when used independently, such embeddings lack explicit spatial grounding, limiting their effectiveness for precise interaction localization.

D. Mutation-Aware and Affinity Prediction Models

Mutation-aware models have been developed to study how amino acid substitutions influence binding affinity, typically using datasets such as SKEMPI and regression-based techniques to estimate changes in interaction strength. While these approaches provide valuable insight into interaction stability, they are generally employed in isolation from structural localization pipelines, meaning mutation effects are evaluated without direct reference to groove- or pocket-level geometric context. Transformer-based sequence models, similarly, lack the spatial grounding required to connect mutation effects to specific structural regions.

E. Research Gap

From the above analysis, several key limitations are identified in existing approaches:

- Lack of structured, multi-stage modeling: Most methods treat interaction prediction as a single-step task, without separating coarse region localization from fine-grained residue-level analysis.
- Fragmented integration of structural and functional features: Structural and sequence-derived representations are typically used independently or combined only superficially, without spatial alignment.
- Limited mutation-aware integration: Existing frameworks do not incorporate mutation effects directly within the interaction localization and prediction pipeline.
- Absence of design-oriented capability: Current methods are predominantly predictive rather than prescriptive, offering no mechanism for suggesting beneficial mutations.

These gaps motivate the proposed unified, structure-aware, mutation-sensitive framework described in this work.

III. PROBLEM FORMULATION

Accurate characterization of protein–peptide interactions requires learning a mapping from a receptor's three-dimensional structure to residue-level binding annotations, discrete pocket locations, and mutation-sensitivity scores, without requiring the peptide structure at inference time.

A. Structural Representation

A protein receptor is represented as a set of residues

$$P = \{ r_1, r_2, \ldots, r_n \}, x_i \in \mathbb{R}^3$$

where x_i denotes the Cα coordinate of residue r_i. Each residue is additionally associated with a sequence-derived embedding $f_i \in \mathbb{R}^d$ obtained from a pretrained protein language model (ESM-2). During training, the bound peptide is represented as a set of atoms $Q = \{ a_1, a_2, \ldots, a_m \}$. The protein is further encoded as a residue graph $G = (V, E)$, where $V = \{ r_i \}_{i=1}^{n}$ and each node carries the pair (x_i, f_i) as its coordinate and feature attributes, respectively. An edge $(i, j) \in$ E exists if atoms $\|x_i - x_j\| \leq 8\dot{A}$, capturing the local spatial neighborhood of each residue.

B. Groove Localization

The first objective is to learn a function $F_\theta(\cdot)$, parameterized by θ, that maps the residue graph to a per-residue groove probability:

$$\hat{p} = F_\theta(G), \qquad \hat{p} \in [0,1]^n$$

where $\hat{p}_i$ represents the predicted probability that residue i lies within the protein–peptide binding groove. The corresponding ground-truth label is derived from the protein–peptide distance as

$$y_i = \begin{cases} 1, & min_{a_k \in Q} ||x_i - a_k|| \leq 5 \text{ Å} \\ 0, & otherwise \end{cases}$$

and θ is optimized by minimizing the binary cross-entropy loss between $\hat{p}_i$ and y_i across all residues. While groove localization identifies residues likely to participate in binding, it does not by itself resolve where on the protein surface a druggable pocket is centered, nor how sensitive that surface is to point mutations. These requirements motivate the two sub-problems that follow.

C. Pocket Identification

The second objective is to learn a function $H_\emptyset(\cdot)$ that classifies residues within the predicted groove, $S = \{ i \in V : \hat{p} \geq \tau \}$, as belonging to a pocket or not, and to localize the resulting pocket regions $C = \{c_1, \ldots, c_k\}$ via their geometric centroids:

$$center(c_j) = \frac{1}{c_j} \sum_{i \in c_j} x_i$$

D. Mutation sensitivity and Affinity prediction

The third objective is to quantify the effect of residue-level mutations on both the predicted binding surface and the experimentally measured binding affinity. For a residue *i* mutated from its wild-type amino acid to a substituted type a' , $Gi \rightarrow a'$ denote the resulting mutated graph, obtained by updating the physicochemical features (charge, hydrophobicity, side-chain orientation) of residue *i*. The mutation's impact on the predicted groove is defined as

$$\Delta i(a') = F_\theta(Gi \rightarrow a') - F_\theta(G)$$

and, independently, a regression function $R_\psi(\cdot)$ is learned to estimate the corresponding change in binding affinity directly from mutation and interface features, such that

$$R_\psi(P, i, a') \approx \Delta log\ K_d$$

where $\Delta log\ K_d$ denotes the experimentally observed shift in dissociation constant reported in the SKEMPI database.

E. Unified Objective

The proposed framework must jointly satisfy three requirements: accurate identification of groove residues from structure alone, reliable localization of discrete binding pockets within the identified groove, and faithful estimation of how individual mutations perturb both the predicted binding surface and the underlying binding affinity.

IV. DATASET DESCRIPTION AND PREPROCESSING

Three publicly available structural and mutational databases were used to construct the training data for the groove localization, pocket identification, and mutation sensitivity modules, respectively. Each dataset required a distinct preprocessing pipeline to convert raw structural or tabular records into the feature representations defined in Section III.

A. Protein–Peptide Structural Dataset (PepBDB)

The groove localization module was trained on PepBDB (v20200318), containing ~13,000 protein–peptide complexes in PDB format. For each complex, atomic records were

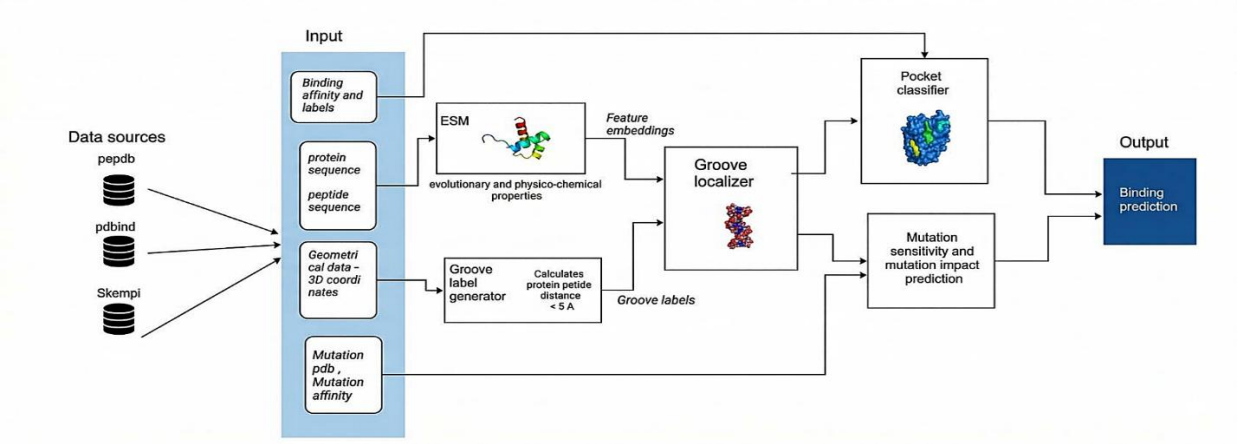

Figure 1 - System Architecture

aggregated to the residue level, retaining the Cα coordinate x_i each residue as defined in Eq. (1).

B. Feature Engineering

Since raw PDB records carry no physicochemical information, five descriptors were derived per residue and combined into the node feature vector f_i (Eq. 2). The residue-level charge takes values of −1, 0, +0.5, or +1, assigned as −1 for ASP/GLU, +1 for LYS/ARG, +0.5 for HIS, and 0 for all other residues. A binary hydrophobicity indicator is set to 1 for the hydrophobic residues and 0 otherwise. The side-chain orientation vector is computed as the difference between the side-chain centroid and the Cα position, and is expressed as a floating-point value. The minimum peptide distance is the shortest Euclidean distance, in Ångströms, between a residue and any peptide atom. Finally, the binary contact-residue indicator is set to 1 whenever the minimum peptide distance is at most 5 Å, and this indicator is used directly as the groove label y_i in Eq. (4).

For the concavity-gated model (CG-EGNN), a sixth feature — surface concavity was appended to f_i.

C. Graph Construction and Splits

A residue graph $G = (V, E)$ was built per complex, connecting residues within 8 Å (Section III-A), split 80:20 at the graph level.

TABLE I . GROOVE DATASET SPLIT

Model	Graphs	Train	Test
GCN-GNN	500	400	100
GVP-GNN	19,944	15,955	3,989
EGNN	19,944	15,955	3,989
CG-EGNN	19,699	15,759	3,940

D. Pocket Classification Dataset

Labels for the pocket classifier H_ϕ were derived, not manually annotated, by applying DBSCAN to predicted groove residues dense clusters were labeled pocket, sparse/noise points non-pocket yielding ~146K non-pocket and ~96K pocket points, with a held-out test split of 48,485 points.

E. Mutation - Affinity Dataset (SKEMPI)

The affinity regressors and mutation-sensitivity module were trained on SKEMPI, which provides wild-type and mutant KdK_d Kd values, mutation identity, interface location and experimental temperature per record. The regression target $\Delta log\ K_d$ (Eq. 12) was computed from the wild-type and mutant affinity columns; mutation records were encoded using substitution identityinterface location.

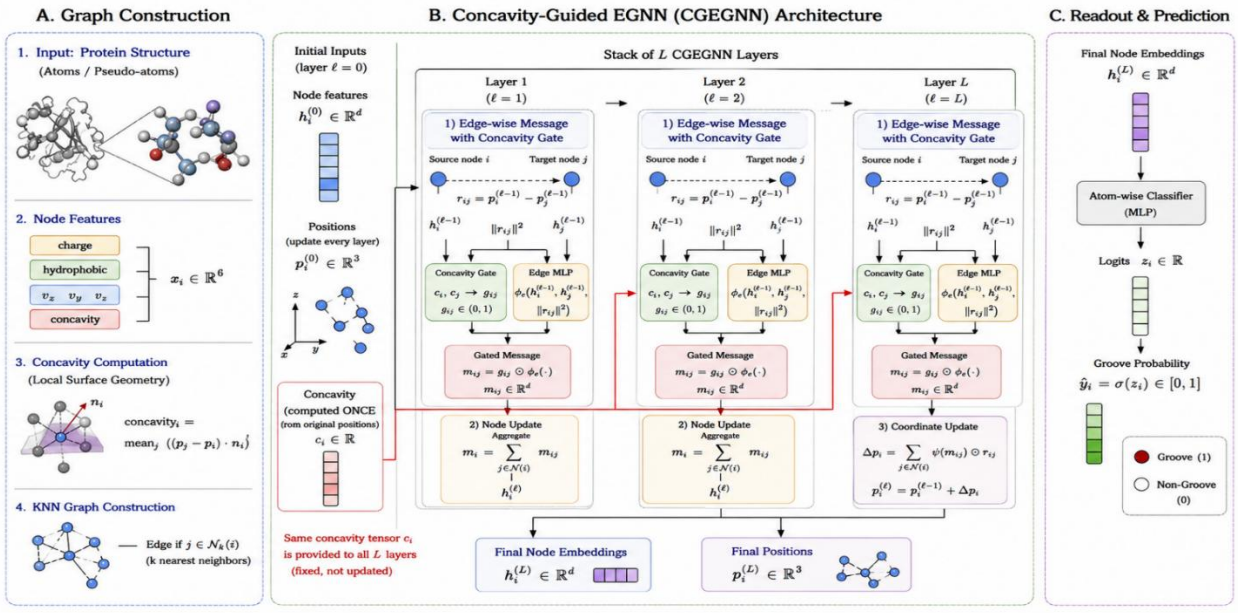

Figure 2 - Overview of the CG-EGNN architecture

V. PROPOSED METHODOLOGY

The proposed framework consists of three sequential modules operating on a single receptor structure: a structure-aware groove localizer, a pocket identification stage, and a mutation sensitivity and affinity-prediction stage. The output of each stage feeds into the next, progressively refining groove residues into druggable pockets and mutation-sensitive hotspots.

A. Structure-Aware Groove Localization via CG-EGNN

Residues are encoded with a six-dimensional physicochemical feature vector that explicitly preserves static local surface concavity to prevent representational drift during training. A nearest-neighbor graph connects these spatial coordinates. We introduce a Concavity-Gated Equivariant Graph Neural Network (CG-EGNN) where edge messages are dynamically modulated by a learned concavity gate. This gating mechanism selectively amplifies information flow within concave, groove-like topological regions while suppressing signals from convex surfaces.

Aggregated spatial messages update node embeddings equivariantly across network layers, strictly preserving rotational and translational symmetries. A terminal multi-layer perceptron classifies the probability of each residue belonging to a groove. Trained via binary cross-entropy, the CG-EGNN significantly outperforms baseline graph convolutional networks, geometric vector perceptrons, and concavity-agnostic equivariant models.

B. Pocket Identification

Predicted groove residues are grouped using density-based spatial clustering to algorithmically generate training labels, eliminating the need for manual boundary annotation. A one-dimensional convolutional neural network subsequently processes these localized spatial descriptors to yield pocket probabilities. During inference, high-confidence candidate points are re-clustered, and their geometric centroids are extracted to designate precise predicted binding site coordinates.

C. Mutation Sensitivity Analysis

To evaluate potential binding modulators, isolated groove residues undergo exhaustive *in silico* substitution across all standard alternative amino acids. The resulting mutated spatial graphs are evaluated by the pre-trained CG-EGNN to quantify shifts in groove probability, local surface concavity, and side-chain angular deviation. Substitutions are ranked by their aggregate structural impact to identify the highest-scoring mutation candidates.

D. Binding Affinity Regression

Operating independently of the structural sensitivity module, a regression framework trained on the SKEMPI database predicts experimental binding affinity variations directly from mutation identities and baseline physicochemical features. Among evaluated tree-based and support vector models, a hyperparameter-tuned XGBoost architecture achieved the highest predictive accuracy across standard error and correlation metrics. To provide post hoc interpretability, a lightweight large language model generates concise rationales for top-ranked mutations. Based strictly on local topography classifications and calculated sensitivity scores, the model outputs human-readable explanations of likely structural consequences, serving exclusively as a supplementary analytical aid for researchers without altering the underlying predictive pipeline.

VI. RESULTS AND CONCLUSION

A. Groove Prediction Performance

An evaluation of the discriminative capabilities across the four graph-based architectures reveals comparable baseline classification accuracies of approximately 0.94 (Table I). However, significant variance is observed in the ROC-AUC metric, which more robustly captures the models' ability to distinguish between groove and non-groove regions amidst class imbalance. The proposed CG-EGNN achieves the highest ROC-AUC of 0.7759, significantly outperforming the baseline GCN-GNN (0.5912), GVP-GNN (0.4906), and the standard EGNN (0.6712). The suboptimal performance of GVP-GNN suggests that explicitly representing vector features is insufficient without an accompanying geometric prior. While the standard EGNN improves discrimination

via equivariant message passing, the superior performance of the CG-EGNN confirms that explicitly gating information flow based on local surface concavity provides a highly effective inductive bias for structural groove localization.

TABLE II. GROOVE PREDICTION RESULTS

Model	Test Accuracy	ROC-AUC
GCN-GNN	0.9421	0.5912
GVP-GNN	0.9401	0.4906
EGNN	0.9410	0.6712
CG-EGNN	0.9421	0.7759

B. Binding Affinity Regression Performance

Predictive performance metrics for the evaluated regression models, following hyperparameter optimization via randomized grid search cross-validation, establish XGBoost as the superior architecture (Table II). Prior to tuning, XGBoost provided the strongest baseline performance and maintained this superiority post-tuning, achieving an R^2 of 0.851 alongside the lowest error rates (MAE: 0.585, RMSE: 0.876). Random Forest demonstrated competitive but slightly inferior metrics (R^2: 0.828), whereas SVR failed to capture the underlying variance effectively (R^2: 0.472). These empirical results confirm that hyperparameter-tuned XGBoost is the optimal regression engine for estimating binding affinity fluctuations within this pipeline.

TABLE III. REGRESSION PERFORMANCE (AFTER TUNING)

Model	MAE	RMSE	R^2
Random Forest	0.630	0.939	0.828
XGBoost	0.585	0.876	0.851
SVR	1.131	1.648	0.472

C. Pocket Classification Performance

The pocket classification module demonstrates a robust overall accuracy of 0.92 across 48,485 evaluated surface points, as detailed in Table III. Despite the inherent spatial class imbalance between pocket and non-pocket regions, the 1D-CNN maintains a strong macro-average F1-score of 0.91. Specifically, it achieves a precision of 0.92 for both classes and an acceptable recall of 0.87 for the minority pocket class. These metrics indicate that the classification module, when paired with DBSCAN-based regional clustering, reliably discriminates functional binding pockets from extraneous surface topographies without defaulting to majority-class prediction.

TABLE IV. POCKET CLASSIFICATION

Class	Precision	Recall	F1-Score	Support
Non-Pocket (0)	0.92	0.95	0.93	29,353
Pocket (1)	0.92	0.87	0.89	19,132
Accuracy	-	-	0.92	48,485
Macro Average	0.92	0.91	0.91	48,485
Weighted Average	0.92	0.92	0.92	48,485

D. Mutation Sensitivity Analysis

Following exhaustive *in silico* residue substitution, the aggregated shifts across key geometric and probabilistic metrics reveal distinct structural sensitivities (Table IV). While the majority of the interface remains structurally stable under perturbation (reflected by near-zero mean shifts), a specific subset of residues exhibits high variance. Mutations yielding positive topological shifts indicate enhanced structural suitability for binding, whereas negative shifts flag potentially destabilizing substitutions. By grounding these probabilistic shifts in explicit geometric parameters (ΔConcavity, ΔOrientation), the module successfully isolates mutation-critical hotspots that remain undetectable by sequence-only predictive models.

TABLE V. RESULTS OF MUTATION SENSITIVITY PREDICTION

Metric	Mean	Min	Max
ΔGroove	−0.0254	−0.0670	0.0029
ΔConcavity	−0.0158	−0.0404	0.0088
ΔOrientation	0.0551	0.0227	0.1120

VII. CONCLUSION AND FUTURE WORKS

Our multi-stage framework for protein-peptide interaction analysis successfully bridges the gap between sequence-centric and structure-centric models. By pairing sequence embeddings with a Concavity-Gated Equivariant Graph Neural Network (CG-EGNN), we demonstrated that explicitly gating for surface geometry fundamentally improves groove

localization. Downstream from this core architecture, the pipeline reliably maps pocket boundaries, estimates binding affinity shifts, and extracts clear structural rationales for specific mutation sensitivities. To transition these computational predictions into applied drug discovery, several methodological gaps must be addressed in future iterations to ensure scientific integrity and predictive robustness. Specifically, the framework currently requires *in vitro* laboratory assays, such as surface plasmon resonance, to empirically verify the algorithmically defined mutation hotspots and their true binding kinetics. Additionally, the model lacks molecular dynamics simulations to capture the solvent effects and transient conformational flexibility such as induced-fit side-chain adjustments that rigid-body structural assumptions inherently miss. The predicted binding interfaces must also be confirmed through fully flexible docking protocols rather than static representations. Finally, while currently optimized for peptides, adapting the core pipeline to evaluate broader protein-small-molecule interactions against larger benchmark datasets will be essential for real-world therapeutic design.

Despite these current limitations, the proposed architecture establishes a highly scalable foundation for high-throughput screening. By localizing critical binding geometries and evaluating mutation sensitivities early in the computational pipeline, the framework drastically reduces the processing overhead typically associated with exhaustive docking protocols. As this methodology matures, deploying such stage-wise, interpretable models ensures that the early discovery phases remain grounded in verifiable structural data, ultimately accelerating therapeutic design by allowing researchers to prioritize only the most viable candidates for costly laboratory testing.

REFERENCES

[1] R. Wang, X. Yang, C. Pang, and L. Wei, "A Multi-Objective Comprehensive Framework for Predicting Protein-Peptide Interactions and Binding Residues," IEEE J. Biomed. Health Inform., vol. 29, no. 11, pp. 7884-7893, 2025, doi: 10.1109/JBHI.2025.3539313.

[2] A. Abid et al., "PepNN: Deep Attention Model for Peptide Binding Site Identification," Nat. Commun., vol. 13, p. 504, 2022.

[3] R. Evans et al., "Protein complex prediction with AlphaFold-Multimer," bioRxiv, 2021, doi: 10.1101/2021.10.04.463034.

[4] B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. O. Dror, "Learning from Protein Structure with Geometric Vector Perceptrons," in Proc. 9th Int. Conf. Learn. Represent. (ICLR), 2021.

[5] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) Equivariant Graph Neural Networks," in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021, pp. 9323–9332.

[6] Z. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science, vol. 379, no. 6637, pp. 1123–1130, 2023.

[7] J. Moal and I. H. Moal, "SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation," Bioinformatics, vol. 35, no. 2, pp. 318–320, 2019.

[8] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures," J. Med. Chem., vol. 47, no. 12, pp. 2977–2980, 2004.

[9] L. Lei et al., "PepBDB: A comprehensive structural database of biological peptide-protein interactions," Bioinformatics, vol. 37, no. 22, pp. 4272–4274, 2021.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD), 1996, pp. 226–231.

[12] H. M. Berman et al., "The Protein Data Bank," Nucleic Acids Res., vol. 28, no. 1, pp. 235–242, 2000.

[13] A. Q. Jiang et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

[14] B. Huang, M. Yang, and S. Shen, "DeepPeptide: A deep learning approach to predicting peptide-protein binding interfaces," Brief. Bioinform., vol. 23, no. 1, 2022.

[15] J. Meier et al., "Language models enable zero-shot prediction of the effects of mutations on protein function," Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, pp. 29287–29303, 2021.
[16] X. Ren, L. Yan, K. Liu, and Y. Zhao, "Graph neural networks in structure-based drug design," Brief. Bioinform., vol. 23, no. 4, 2022.
[17] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, no. 7873, pp. 583–589, 2021.
[18] Y. Wang et al., "Graph attention networks for protein-protein interface prediction," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 19, no. 3, pp. 1432–1442, 2021.
[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
[20] J. Schymkowitz et al., "The FoldX web server: An online force field," Nucleic Acids Res., vol. 33, no. Web Server issue, pp. W382–W388, 2005.