STRUCTURE-AWARE LANGUAGE MODELING ENHANCES PEPTIDE CONDITIONAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have demonstrated that protein language models can effectively drive receptor-conditioned peptide generation. However, existing language-model-based approaches rely almost exclusively on sequence context to learn compatibility between peptide binders and their target proteins. It remains unclear whether incorporating target protein structural information can provide additional benefits, particularly regarding the rationality of spatial conformations. Furthermore, most language-model-based methods do not explicitly account for the receptor site where peptides are expected to bind, yet such binding sites are critical for functional peptides. To bridge this gap, we investigate how structure-aware protein language models contribute to receptor-conditioned peptide design and introduce binding-site constraints within the language-model paradigm to enable site-specific generation. Experimental results show that combining receptor sequences with structural conditions effectively improves the spatial conformational rationality of peptide–protein complexes and the complementarity of binding interfaces. At the local scale, introducing binding-site constraints shifts generated candidate peptides toward the target region, indicating that binding-site prompts exert a soft directional bias. Code and data will be released at https://github.com/anonymous/VersaPep.

1 INTRODUCTION

Peptides can engage protein surfaces that are difficult to address with small molecules, making them attractive candidates for modulating protein interactions and developing therapeutics (Muttenthaler et al., 2021; Nevola & Giralt, 2015; Fosgerau & Hoffmann, 2015; Craik et al., 2013; Wang et al., 2022). Designing a peptide for a particular receptor, however, requires more than producing a plausible amino-acid sequence. The peptide must adopt a compatible conformation, form an appropriate interface with the receptor, and, when a functional region is known, bind near that region (Abdin et al., 2022; Petsalaki et al., 2009; Trabuco et al., 2012). These requirements raise two connected questions for receptor-conditioned generation: how to use structural information to improve predicted complex quality, and how to guide the location of the predicted interaction.

Existing methods approach these questions from different directions. PepMLM (Chen et al., 2026) demonstrates that a protein language model (PLM) can generate peptide binders conditioned on a target sequence without requiring its structure. Geometry-based methods such as DiffPepBuilder (Wang et al., 2024) instead use three-dimensional information to design peptide sequences and structures together. The success of sequence-conditioned generation makes it valuable to ask whether explicit receptor structure can provide additional information within a PLM. It also leaves open a more local question: whether specifying a receptor binding site can shift where generated peptides are predicted to bind.

Structure-aware PLMs offer a way to investigate both questions. SaProt (Su et al., 2024) represents amino-acid identity alongside discrete 3Di tokens derived from protein structure (Van Kempen et al., 2024). Yet adding structural tokens does not, by itself, establish which input is responsible for a change in generation quality. Receptor sequence, receptor structure, and peptide-side structure may contribute differently; peptide-side tokens may also carry information about the intended peptide pose. Likewise, an improvement in overall complex confidence would not show whether a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

binding-site hint has actually changed peptide placement. Resolving these effects requires controlled comparisons of the available structural and site information.

In this study, we introduce a **Ver**satile **s**tructure-**a**ware receptor-conditioned **Pep**tide generation framework (**VersaPep**). It uses gated cross-chain attention (CCA) to connect receptor and peptide representations, and binding-site feature-wise linear modulation (BS-FiLM) to incorporate optional site hints. Its conditioning settings allow us to mask the receptor sequence or 3Di tokens, supply or withhold peptide 3Di tokens, and specify true or predicted receptor binding sites. We evaluated the generated peptides on an independent test set of 300 targets using predicted complex confidence and spatial localization. Computational evaluations shows that incorporation of receptor sequence information and structural conditional constraints can optimize the spatial conformation of peptide-protein complexes and enhance the complementarity of their binding interfaces. From a local structural perspective, the introduction of binding site constraints can guide generated candidate peptides to converge more accurately on the target binding region, showing that binding site guidance can provide a subtle yet effective directional bias for peptide-generation optimization.

In summary, the contributions of this study are threefold:

- Introduces a receptor-conditioned peptide generation framework that accepts the global receptor context and optional local binding-site hints based on structure-aware protein language modeling.
- Disentangles the contributions of the receptor sequence, receptor 3Di, and peptide-side 3Di, showing that the receptor sequence and structure provide complementary information for the predicted complex quality.
- Shows through true-site, predicted-site, and incorrect-site guidance that local hints can alter the spatial preference of generated peptides while broadly preserving predicted complex confidence.

We believe the findings presented in this work can inspire the rational design of functional peptides and may shed light on developing optimization approaches for other protein-targeting small binders.

2 Related Work

2.1 Geometry-based peptide generation

Geometry-based peptide generation methods model receptor–peptide interactions directly in three-dimensional space and condition generation on receptor structures, binding regions, or interface geometry. Related binding-site predictors identify candidate receptor regions (Saladin et al., 2014; Wardah et al., 2020; Chandra et al., 2023). Geometric design methods can explicitly constrain peptide conformation and placement relative to the receptor (Watson et al., 2023; Cao et al., 2022); structure-conditioned sequence design can then optimize candidate sequences (Dauparas et al., 2022). A representative approach, **DiffPepBuilder** (Wang et al., 2024), employs an SE(3)-equivariant diffusion model to jointly generate peptide sequences and structures conditioned on receptor geometry and binding regions. Other geometric approaches include torsional flow matching and full-atom latent diffusion (Lin et al., 2024; Kong et al., 2024); peptide docking methods address peptide placement or pose refinement (Raveh et al., 2011; Alam et al., 2017; Zhou et al., 2018; Lee et al., 2015; Kurcinski et al., 2015; Wang et al., 2025). These methods operate primarily on geometric representations and address questions distinct from how structural inputs interact with sequence priors learned by pretrained PLMs.

2.2 Language-modeling-based peptide generation

Protein language models (PLMs) provide a natural framework for conditional peptide sequence generation by leveraging sequence priors learned from large-scale protein pretraining (Rives et al., 2021; Rao et al., 2020; Ahmed et al., 2020; Lin et al., 2023; Bepler & Berger, 2021). **PepMLM** (Chen et al., 2026), for example, demonstrates that receptor sequence context can guide target-conditioned peptide generation. Protein language models have also supported function prediction and generation in other protein-design settings (Meier et al., 2021; Madani et al., 2023). Existing PLM-based approaches, however, rely mainly on sequence conditioning, with receptor structure

rarely represented as an explicit and controllable input. This limits direct analysis of how structural signals affect generation quality and spatial behavior.

2.3 Structure-aware protein language models

Structure-aware PLMs further incorporate discrete structural information into protein representations. **SaProt** (Su et al., 2024), for instance, jointly encodes amino-acid identities and 3Di structure tokens, allowing explicit structural information to participate in protein representation learning (Van Kempen et al., 2024). Other structure-aware language models incorporate structural tokens or structure-informed objectives (Li et al., 2024; Heinzinger et al., 2024; Zhang et al., 2024; Hsieh et al., 2025; Zheng et al., 2023). Related inverse-folding and structure-conditioned sequence-design methods use structural context in different ways (Hsu et al., 2022; Gao et al., 2022; Anand et al., 2022). However, in receptor-conditioned peptide generation, it remains unclear which structural signals are effective, whether their contributions depend on the available input context, and whether local structural hints can alter the spatial preference of generated candidates.

3 Methods

3.1 Problem Formulation and Multimodal Encoding

Given a receptor chain $R = \{r_i\}_{i=1}^{L_R}$ and a specified peptide length L_P, VersaPep generates a peptide $P = \{p_j\}_{j=1}^{L_P}$ conditioned on the available receptor sequence, structural representation, and optional binding-site information. The receptor and peptide chains are concatenated into a single sequence with special boundary tokens.

As illustrated in Fig. 1a, each residue is represented by amino-acid identity, 3Di structural token, binding-site (BS) label, chain identity, and positional information. The initial representation of residue i is

$$H_i^{(0)} = E_{\mathrm{AA}}(a_i) + E_{\mathrm{3Di}}(s_i) + E_{\mathrm{BS}}(b_i) + E_{\mathrm{chain}}(c_i) + E_{\mathrm{pos}}(i),$$

where a_i and s_i denote the amino-acid and 3Di tokens, b_i denotes the optional binding-site label, and $c_i \in \{\mathrm{R}, \mathrm{P}\}$ identifies the receptor or peptide chain. Missing modalities are represented using modality-specific mask tokens. We zero-initialize the added BS and chain embeddings to preserve the pretrained SaProt representation at the start of training.

Peptide amino-acid identities are masked for sequence generation, while peptide 3Di tokens are retained or masked according to the generation setting (Devlin et al., 2019). Receptor sequence and receptor 3Di tokens provide global structural context, whereas the BS mask specifies a local receptor region without providing peptide coordinates or a docking pose.

3.2 Multi-Task VersaPep Architecture

VersaPep uses a pretrained SaProt encoder as its backbone and introduces two modules at different stages of the network (Fig. 1c). Gated cross-chain attention (CCA) is inserted at an intermediate encoder depth to transfer receptor-wide information to peptide representations through an input-dependent gate (Vaswani et al., 2017). After the final encoder layer, BS-FiLM injects local binding-site information into the peptide representations before prediction (Perez et al., 2018).

The final SaProt representations are layer-normalized and separated into receptor and peptide components. When binding-site information is available, BS-FiLM modulates the peptide representations while leaving the receptor representations unchanged.

VersaPep employs three task-specific prediction heads. The sequence head predicts peptide amino-acid identities, the 3Di head predicts peptide structural tokens, and the binding-site head predicts binary receptor-site labels. The model is trained with the multi-task objective

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seq}} + \lambda_{\mathrm{struct}}\mathcal{L}_{\mathrm{struct}} + \lambda_{\mathrm{bind}}\mathcal{L}_{\mathrm{bind}},$$

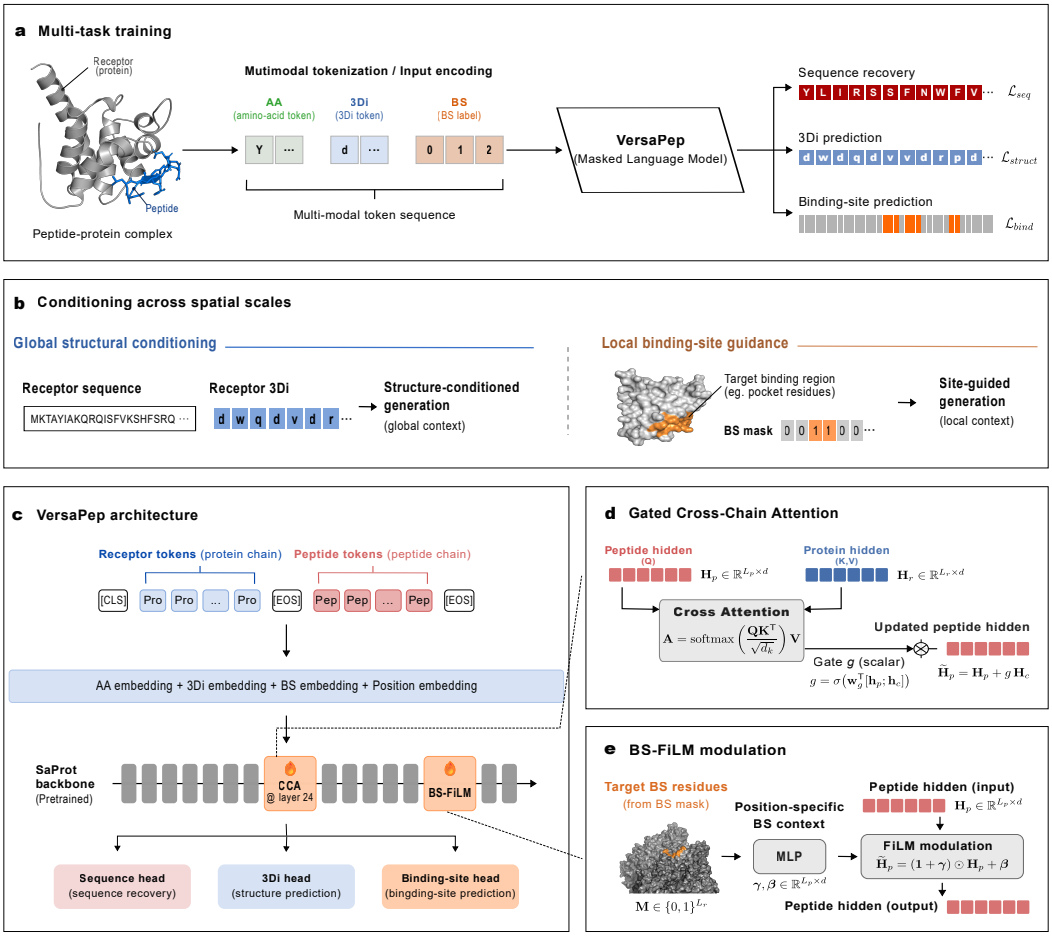

Figure 1: Overview of VersaPep for receptor-conditioned peptide generation. **(a)** Multi-task training with amino-acid, 3Di, binding-site (BS), and chain-identity inputs, supervised by peptide sequence recovery, peptide 3Di prediction, and receptor binding-site prediction. **(b)** Two generation settings across spatial scales: global structural conditioning using receptor sequence and 3Di tokens, and local binding-site guidance using a residue-level BS mask. **(c)** VersaPep encodes receptor and peptide tokens with a pretrained SaProt backbone augmented by gated cross-chain attention (CCA) and BS-FiLM, followed by three task-specific prediction heads. **(d)** CCA captures receptor–peptide interactions and uses a learnable gate to update peptide representations. **(e)** BS-FiLM converts the specified receptor-site context into FiLM parameters γ and β for position-specific modulation of peptide representations.

where $\mathcal{L}_{\text{seq}}$ and $\mathcal{L}_{\text{struct}}$ supervise amino-acid and 3Di prediction over the corresponding masked peptide positions, respectively, and $\mathcal{L}_{\text{bind}}$ supervises receptor binding-site prediction. The coefficients λ_{struct} and λ_{bind} control the contributions of the auxiliary objectives.

Training proceeds in stages. The pretrained backbone is initially frozen while the newly introduced modules and prediction heads are optimized. The complete model is subsequently jointly fine-tuned to adapt the pretrained representation to receptor-conditioned peptide generation.

3.3 Gated Cross-Chain Attention

CCA explicitly models receptor–peptide interactions at a selected intermediate encoder depth (Fig. 1d). Let $H_R \in R^{L_R \times d}$ and $H_P \in R^{L_P \times d}$ denote the receptor and peptide representations

at this depth. Peptide representations serve as queries, while receptor representations provide keys and values:

$$A = \mathrm{MHA}\left(\mathrm{LN}_P(H_P), \mathrm{LN}_R(H_R), \mathrm{LN}_R(H_R)\right),$$

where A contains receptor-derived context for each peptide position.

To adaptively control the contribution of this context, pooled peptide and cross-chain representations are used to predict an input-dependent scalar gate g. The peptide states are then updated through

$$g = \sigma(W_g[\mathrm{Pool}(H_P); \mathrm{Pool}(A)] + b_g), \qquad \widetilde{H}_P = H_P + g \cdot \mathrm{Dropout}(A).$$

Because g depends on the current peptide representation and receptor-derived context, the strength of cross-chain information transfer varies across input complexes. Only peptide representations are updated by CCA, while receptor representations remain unchanged before the remaining SaProt layers.

3.4 Binding-Site FiLM Modulation

BS-FiLM operates after the final encoder normalization and before the sequence and 3Di prediction heads (Fig. 1e). Given a binary receptor binding-site mask, each peptide position attends only to receptor residues marked as belonging to the selected site:

$$C^{\mathrm{BS}} = \mathrm{MHA}_{\mathrm{BS}}\left(\mathrm{LN}(U_P), \mathrm{LN}(U_{R,\mathrm{BS}}), \mathrm{LN}(U_{R,\mathrm{BS}})\right),$$

where U_P denotes the final peptide representation and $U_{R,\mathrm{BS}}$ contains the selected receptor-site representations.

A two-layer MLP combines the peptide representation with the corresponding binding-site context and predicts position-specific scale and shift parameters. The resulting modulation is

$$(\gamma_i, \beta_i) = f_\theta\big([U_{P,i}; C_i^{\mathrm{BS}}]\big), \qquad \widehat{H}_{P,i} = (1 + \gamma_i) \odot U_{P,i} + \beta_i.$$

The output layer of f_θ is zero-initialized, such that BS-FiLM initially reduces to an identity transformation and does not perturb the pretrained representation. When no valid binding-site residue is provided, BS-FiLM is bypassed and the original peptide representation is used directly. The modulated peptide representations are passed to the sequence and 3Di prediction heads, whereas the receptor binding-site head operates directly on the receptor representations.

4 Results

4.1 Overall performance benchmarking

On an independent test set of 300 targets (see the Appendix for dataset details), we benchmarked VersaPep against two baselines, PepMLM (Chen et al., 2026) and DiffPepBuilder (Wang et al., 2024), under a unified evaluation protocol. Peptide–receptor complexes were predicted with AlphaFold-Multimer (Jumper et al., 2021; Evans et al., 2021; Mirdita et al., 2022), with DockQ (Basu & Wallner, 2016) and Rosetta (Alford et al., 2017; Chaudhury et al., 2010) serving as supplementary metrics (results in the Appendix).

As reported in Table 1, and further corroborated by the success rates stratified across ipTM thresholds in Fig. 2, VersaPep outperformed both baselines across these metrics under the unified evaluation protocol; the consistent trend indicates that the gains stem from both a more accurate predicted peptide conformation and a more reliable placement of the peptide at the receptor interface.

To probe how much of this performance relies on externally supplied peptide structural information, we evaluated an ablation variant of VersaPep without 3Di tokens. In this setting, the receptor sequence and structure are provided, but no peptide 3Di tokens are given: the model predicts the

Table 1: AlphaFold-Multimer confidence metrics for generated peptide–receptor complexes.

Method	ipTM (max) ↑	ipTM (mean) ↑	Interface PAE (mean) ↓	Interface pLDDT (max) ↑	Peptide pLDDT (max) ↑
DiffPepBuilder	0.525 ± 0.196	0.408 ± 0.151	11.19 ± 3.66	0.745 ± 0.114	0.505 ± 0.169
PepMLM	0.605 ± 0.220	0.446 ± 0.181	10.40 ± 4.12	0.785 ± 0.128	0.589 ± 0.209
VersaPep	$\mathbf{0.623 \pm 0.212}$	$\mathbf{0.469 \pm 0.190}$	$\mathbf{9.88 \pm 4.33}$	$\mathbf{0.801 \pm 0.119}$	$\mathbf{0.615 \pm 0.211}$
VersaPep w/o 3Di	$\underline{0.615 \pm 0.217}$	$\underline{0.449 \pm 0.185}$	$\underline{10.21 \pm 4.23}$	$\underline{0.789 \pm 0.121}$	$\underline{0.599 \pm 0.208}$

Note: Values are mean ± standard deviation. Bold and underlined values indicate the best and second-best mean values, respectively.

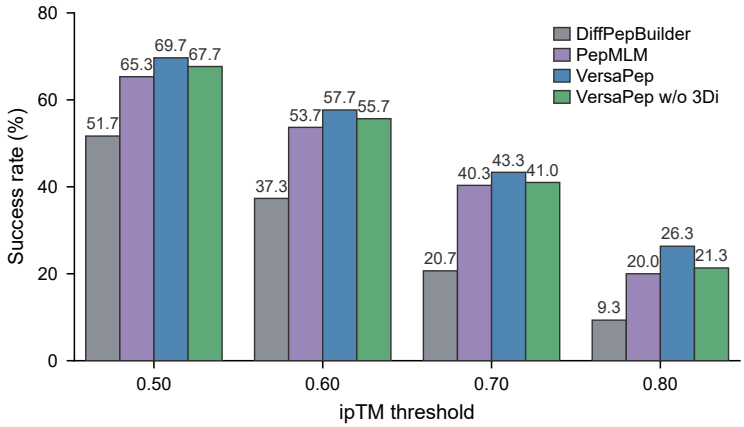

Figure 2: Overall structural performance of generated peptide–receptor complexes evaluated by AlphaFold-Multimer. For each method, bars show the percentage of targets whose highest ipTM score across generated candidates meets or exceeds the indicated threshold. Values above the bars denote the corresponding success rates.

peptide 3Di tokens, generates a peptide sequence from them, and repeats this self-consistency loop for five rounds. Even without pre-supplied peptide structural context, this variant still outperformed both baselines on every metric. Moreover, when candidate peptides were stratified by ipTM, its top-ranked peptides achieved a higher success rate.

Taken together, these results suggest that receptor sequence and structural context alone are sufficient to drive the generation of peptides whose bound conformations are predicted with higher confidence, and that explicit structural guidance for the peptide further improves the quality of the predicted binding interface.

4.2 Effects of global structural conditioning

We next examined how receptor sequence and structural information contribute to the quality of VersaPep's predicted complexes. Starting from the full model, we ablated either the receptor 3Di tokens (Seq only) or the receptor sequence tokens (3Di only). Both restricted settings yielded lower best ipTM and higher interface PAE than full receptor conditioning (Fig. 3a,b), so neither source alone matched the performance obtained when both were available. The 3Di-only setting performed slightly better than sequence-only on both metrics, suggesting that receptor structure makes the larger individual contribution, with receptor sequence providing a complementary benefit.

We suspected, however, that supplied peptide 3Di tokens might partly mask this receptor contribution by compensating for missing receptor information. To isolate it, we repeated the receptor ablations without peptide 3Di tokens, generating the peptide sequence directly rather than through iterative prediction or refinement (Fig. 3d,e). Under these stricter conditions, full receptor conditioning again gave the best predicted complex quality, but masking either receptor input caused a substantially larger performance loss than when peptide 3Di was available (Fig. 3c). This pattern

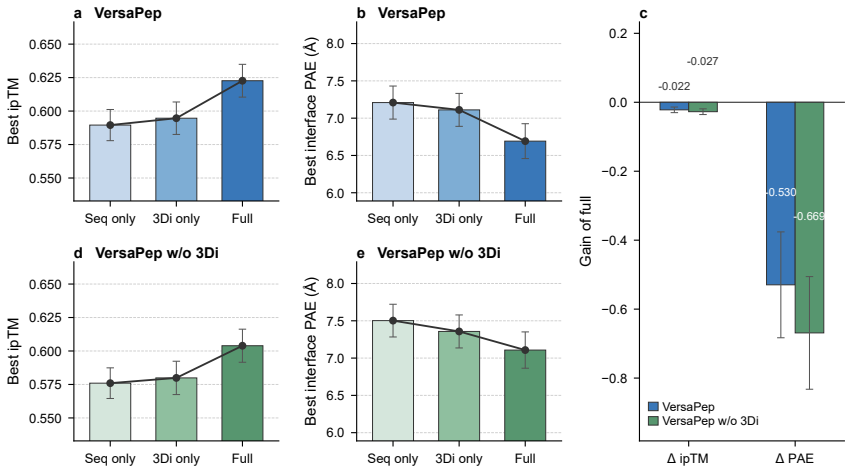

Figure 3: Effects of global structural conditioning on predicted complex quality. For VersaPep (**a**,**b**) and VersaPep w/o 3Di (**d**,**e**), receptor sequence-only, receptor 3Di-only, and combined (Full) conditioning are compared in terms of ipTM and best interface PAE. Panel (**c**) summarizes the changes in ipTM and interface PAE under Full conditioning relative to sequence-only conditioning for both models. Bars show the mean values; points and error bars indicate the means and target-level bootstrap 95% confidence intervals, respectively.

indicates a greater reliance on combined receptor sequence and structure when no peptide structural context is supplied.

Together, these comparisons suggest that the two receptor inputs carry complementary information for peptide generation. Receptor 3Di tokens contribute more strongly on their own, but their benefit is greatest when combined with receptor sequence, particularly when the model must generate without a supplied peptide structural context.

4.3 Local binding-site guidance

The preceding analyses showed that global structural conditioning improves predicted complex quality, particularly when receptor sequence and structure are used together. We then examined whether a more localized input could guide peptide placement toward a specified region of the receptor. To isolate the effect of the binding-site hint, we did not supply peptide 3Di tokens, which can encode information about peptide pose, and generated sequences directly without iterative refinement. The influence of peptide 3Di tokens on site guidance is examined in the Appendix.

We compared generation without a site hint (No BS) against guidance from the binding site in the native complex (True BS), a site predicted by VersaPep (Pred BS), or an incorrect site sampled from non-interface receptor residues (Wrong BS). Best ipTM and interface PAE were broadly similar across these settings (Fig. 4a), suggesting that site guidance did not substantially alter overall predicted complex confidence. Its effect was instead apparent in peptide localization. Distances to the true pocket followed the ordering True BS $<$ Pred BS $<$ No BS $<$ Wrong BS, whereas distances to the wrong pocket showed the opposite trend (Fig. 4b). Pred BS fell between True BS and No BS, indicating that the model's own site predictions can also bias generation toward the intended region. This localization pattern became more apparent among candidates passing higher ipTM thresholds, with True BS yielding the largest fraction of peptides closer to the true pocket (Fig. 4c).

As shown in Fig. 5, the representative complexes for 1A5H and 1EKB illustrate the localization effect of binding-site conditioning. Conditioning on the true binding pocket places the generated peptide close to the target site, reducing the minimum peptide-to-pocket distance relative to unconstrained generation. These examples demonstrate that VersaPep's global structural context supports complex formation, whereas explicit local binding-site information controls where the generated peptide is positioned on the receptor surface.

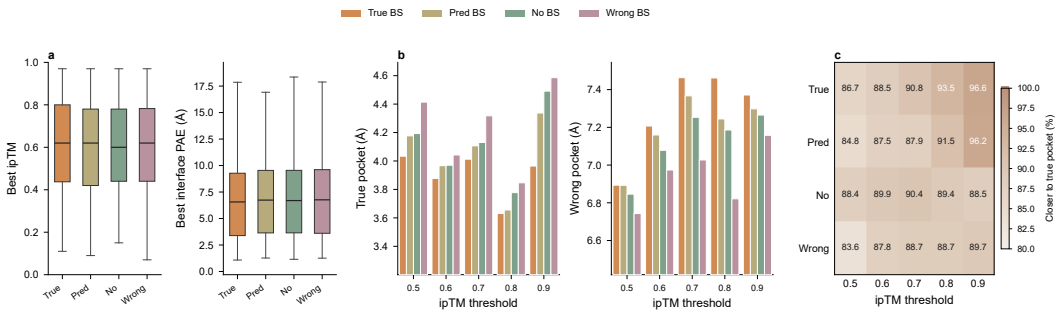

Figure 4: Effect of local binding-site guidance on generated peptide–receptor complexes. **(a)** Best ipTM (higher is better) and best interface PAE (lower is better) across the four settings. **(b)** Distances from the predicted peptide to the true and wrong pockets, in ångströms, among candidates meeting each ipTM threshold. **(c)** Percentage of cases in which the predicted peptide is closer to the true pocket than to the wrong pocket at each ipTM threshold.

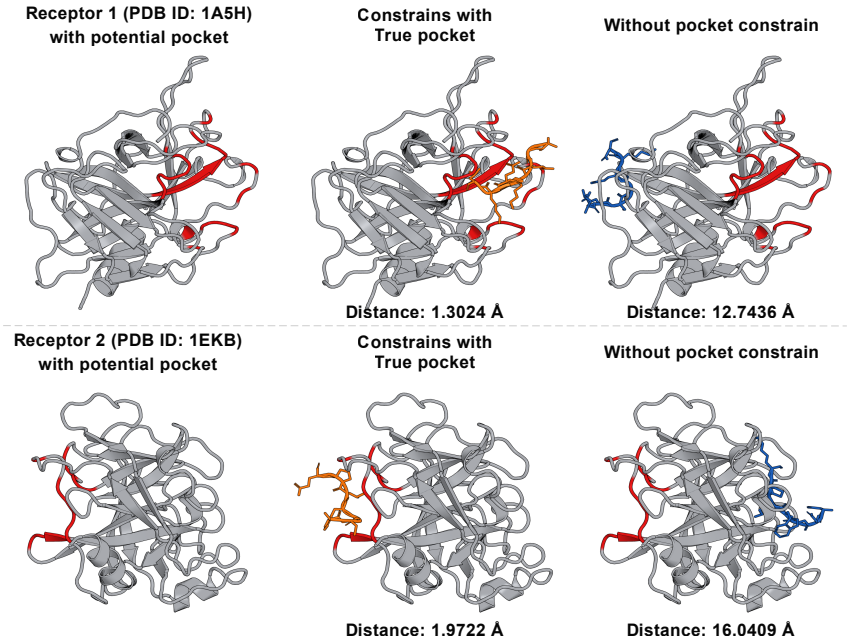

Figure 5: AlphaFold-Multimer-predicted complexes for targets 1A5H and 1EKB under local binding-site guidance. Receptors are shown as gray cartoons, with the designated binding-pocket residues highlighted in red. The middle panels show complexes generated with a binding-pocket constraint, with generated peptides colored orange, whereas the right panels show peptides generated without a pocket constraint, colored blue. Distances represent the minimum heavy-atom distance between the generated peptide and the designated binding-pocket residues.

4.4 Architectural Ablations

The ablations examined whether VersaPep's architectural components and binding-site injection contribute to predicted complex quality (Table 2). In the module ablation, combining CCA with BS-FiLM achieved the highest mean ipTM and lowest interface PAE, although neither component improved every metric on its own. In the binding-site injection ablation, incorporating the site signal through an embedding, with or without the prediction head, generally improved ipTM, interface PAE, and peptide pLDDT relative to the base configuration. These trends support the use of the

combined architecture and explicit binding-site injection, while showing that their contributions vary across metrics.

Because AlphaFold-Multimer evaluation of all ablation variants was computationally costly, these experiments used a subset of the independent test set. Their absolute values may therefore differ from those in the full-set comparison above; the relevant comparisons are between variants within each ablation panel.

Table 2: AlphaFold-Multimer confidence metrics for module and binding-site injection ablations.

Variant	BS input	Max ipTM ↑	Mean ipTM ↑	Min interface PAE ↓	Max peptide pLDDT ↑
A. Module ablation					
SaProt	No BS	0.563 ± 0.206	0.445 ± 0.178	8.05 ± 4.29	0.528 ± 0.211
+ CCA	No BS	0.560 ± 0.199	$\underline{0.448 \pm 0.174}$	7.95 ± 4.30	0.517 ± 0.202
+ BS-FiLM	No BS	$\underline{0.567 \pm 0.201}$	0.440 ± 0.171	$\underline{7.88 \pm 4.26}$	$\underline{0.537 \pm 0.200}$
+ CCA + BS-FiLM	No BS	$\mathbf{0.570 \pm 0.205}$	$\mathbf{0.454 \pm 0.179}$	$\mathbf{7.79 \pm 4.03}$	$\mathbf{0.536 \pm 0.199}$
B. Binding-site injection ablation					
Base	True BS	0.557 ± 0.209	0.444 ± 0.179	8.11 ± 4.49	0.528 ± 0.203
Embed only	True BS	$\underline{0.569 \pm 0.211}$	$\mathbf{0.455 \pm 0.178}$	7.86 ± 4.43	0.537 ± 0.214
Embed + head	True BS	$\mathbf{0.573 \pm 0.213}$	$\underline{0.452 \pm 0.181}$	$\mathbf{7.80 \pm 4.50}$	$\mathbf{0.539 \pm 0.218}$

Note. Values are reported as mean ± standard deviation. Bold and underlined values indicate the best and second-best mean values, respectively, within each ablation panel.

5 Discussion and Conclusion

The results highlight two roles for structural information in receptor-conditioned peptide generation. On the independent test set, VersaPep produced complexes with stronger AlphaFold-Multimer confidence than PepMLM and DiffPepBuilder across the principal interface and peptide metrics. Its performance without supplied peptide 3Di tokens shows that receptor context alone can support generation of plausible predicted complexes; providing peptide 3Di tokens improved the results further. This distinction matters because peptide 3Di tokens can carry information about the desired pose. The two settings therefore represent different levels of structural guidance, rather than interchangeable measures of de novo generation.

The receptor-input ablations help explain the contribution of global context. Sequence-only and 3Di-only conditioning each fell short of their combination, with 3Di alone providing a slightly stronger signal. Removing peptide-side structural guidance made the loss from masking receptor information more apparent. Together, these observations suggest that receptor sequence and structure provide complementary constraints, while peptide-side structural information can partly offset missing receptor context. The architectural ablations support this interpretation: combining CCA and BS-FiLM gave the most balanced module-level result, although individual components did not improve every metric uniformly.

Local binding-site guidance addressed a different question: whether generation could be directed toward a specified receptor region. Without supplied peptide 3Di tokens, true-site and predicted-site hints shifted predicted peptide placement toward the true pocket while leaving overall complex confidence broadly similar. The effect was visible in both pocket-distance analyses and the 1A5H and 1EKB examples. Predicted-site guidance was weaker than guidance from the native site, and the observed shift should be understood as a preference in the generated distribution rather than precise control of every candidate.

These conclusions remain based primarily on predicted structures and confidence measures, with complementary DockQ and Rosetta analyses reported in the Supplementary Materials. They do not establish experimental binding or affinity (Bret et al., 2024; Terwilliger et al., 2024). Moreover, the peptide-3Di-guided and unguided settings differ in the structural information available at generation time. Within these limits, VersaPep shows that global structural context can improve predicted peptide–receptor complex quality, while local site information can influence where generated peptides are placed. Their distinct effects offer a practical basis for choosing the form of guidance according to the information available and the design objective.

AI USE STATEMENT

AI Use Statement. We used ChatGPT to provide feedback on experimental and evaluation designs, assist in interpreting results, support code development and debugging, identify and format relevant references, and improve the clarity and readability of the manuscript. All AI-assisted outputs, including text, code, references, and interpretations, were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The Appendix provides details of the independent test set, experimental settings, implementation, metric definitions, additional evaluations, and ablation analyses to support reproduction of the reported results. The data and source code supporting this study will be released upon publication: https://github.com/anonymous/VersaPep.

REFERENCES

Osama Abdin, Satra Nim, Han Wen, and Philip M Kim. Pepnn: a deep attention model for the identification of peptide binding sites. *Communications biology*, 5(1):503, 2022.

Elnaggar Ahmed, M Heinzinger, C Dallago, G Rihawi, Yu Wang, L Jones, T Gibbs, T Feher, C Angerer, S Martin, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *BioRxiv*, 2020.

Nawsad Alam, Oriel Goldstein, Bing Xia, Kathryn A Porter, Dima Kozakov, and Ora Schueler-Furman. High-resolution global peptide-protein docking using fragments-based piperflexpepdock. *PLoS computational biology*, 13(12):e1005905, 2017.

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031, 2017.

Namrata Anand, Raphael Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Po-Ssu Huang. Protein sequence design with a learned potential. *Nature communications*, 13(1):746, 2022.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

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

Hélène Bret, Jinmei Gao, Diego Javier Zea, Jessica Andreani, and Raphaël Guerois. From interaction networks to interfaces, scanning intrinsically disordered regions using alphafold2. *Nature communications*, 15(1):597, 2024.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

Abel Chandra, Alok Sharma, Iman Dehzangi, Tatsuhiko Tsunoda, and Abdul Sattar. Pepcnn deep learning tool for predicting peptide binding residues in proteins using sequence, structural, and language model features. *Scientific reports*, 13(1):20882, 2023.

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, 2010.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, 44(6):1002–1010, 2026.

David J Craik, David P Fairlie, Spiros Liras, and David Price. The future of peptide-based drugs. *Chemical biology & drug design*, 81(1):136–147, 2013.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Richard Evans, Michael O'neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Keld Fosgerau and Torsten Hoffmann. Peptide therapeutics: current status and future directions. *Drug discovery today*, 20(1):122–128, 2015.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

Cheng-Yen Hsieh, Xinyou Wang, Daiheng Zhang, Dongyu Xue, Fei Ye, Shujian Huang, Zaixiang Zheng, and Quanquan Gu. Elucidating the design space of multimodal protein language models. *arXiv preprint arXiv:2504.11454*, 2025.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. *Advances in Neural Information Processing Systems*, 37:74808–74839, 2024.

Mateusz Kurcinski, Michal Jamroz, Maciej Blaszczyk, Andrzej Kolinski, and Sebastian Kmiecik. Cabs-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic acids research*, 43(W1):W419–W424, 2015.

Hasup Lee, Lim Heo, Myeong Sup Lee, and Chaok Seok. Galaxypepdock: a protein–peptide docking tool based on interaction similarity and energy optimization. *Nucleic acids research*, 43(W1): W431–W435, 2015.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726, 2024.

Haitao Lin, Odin Zhang, Huifeng Zhao, Dejun Jiang, Lirong Wu, Zicheng Liu, Yufei Huang, and Stan Z Li. Ppflow: Target-aware peptide design with torsional flow matching. *bioRxiv*, pp. 2024–03, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41 (8):1099–1106, 2023.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. *Nature reviews Drug discovery*, 20(4):309–325, 2021.

Laura Nevola and Ernest Giralt. Modulating protein–protein interactions: the potential of peptides. *Chemical Communications*, 51(16):3302–3315, 2015.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Evangelia Petsalaki, Alexander Stark, Eduardo García-Urdiales, and Robert B Russell. Accurate prediction of peptide binding sites on protein surfaces. *PLoS computational biology*, 5(3): e1000335, 2009.

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

Barak Raveh, Nir London, Lior Zimmerman, and Ora Schueler-Furman. Rosetta flexpepdock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PloS one*, 6 (4):e18934, 2011.

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.

Adrien Saladin, Julien Rey, Pierre Thévenet, Martin Zacharias, Gautier Moroy, and Pierre Tufféry. Pep-sitefinder: a tool for the blind identification of peptide binding sites on protein surfaces. *Nucleic acids research*, 42(W1):W221–W226, 2014.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pp. 6987–7009, 2024.

Thomas C Terwilliger, Dorothee Liebschner, Tristan I Croll, Christopher J Williams, Airlie J McCoy, Billy K Poon, Pavel V Afonine, Robert D Oeffner, Jane S Richardson, Randy J Read, et al. Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature methods*, 21(1):110–116, 2024.

Leonardo G Trabuco, Stefano Lise, Evangelia Petsalaki, and Robert B Russell. Pepsite: prediction of peptide-binding sites from protein surfaces. *Nucleic acids research*, 40(W1):W423–W427, 2012.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Fanhao Wang, Yuzhe Wang, Laiyi Feng, Changsheng Zhang, and Luhua Lai. Target-specific de novo peptide binder design with diffpepbuilder. *Journal of Chemical Information and Modeling*, 64(24):9135–9149, 2024.

Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy*, 7(1):48, 2022.

Yuzhe Wang, Fanhao Wang, Laiyi Feng, Changsheng Zhang, and Luhua Lai. Diffpepdock: Efficient protein–peptide docking and binder screening via se (3)-equivariant diffusion. *Protein Science*, 34(11):e70338, 2025.

Wafaa Wardah, Abdollah Dehzangi, Ghazaleh Taherzadeh, Mahmood A Rashid, Mohammad GM Khan, Tatsuhiko Tsunoda, and Alok Sharma. Predicting protein-peptide binding sites with a deep convolutional neural network. *Journal of Theoretical Biology*, 496:110278, 2020.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zuobai Zhang, Jiarui Lu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Structure-informed protein language model. *arXiv preprint arXiv:2402.05856*, 2024.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In *International conference on machine learning*, pp. 42317–42338. PMLR, 2023.

Pei Zhou, Bowen Jin, Hao Li, and Sheng-You Huang. Hpepdock: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic acids research*, 46(W1):W443–W450, 2018.

A Additional Experimental Details

A.1 Dataset Construction

We constructed the peptide-receptor complex dataset from experimentally resolved structures in the RCSB PDB. Candidate complexes were first selected using PDB structure metadata and chain-level sequence annotations. We retained structures solved by X-ray crystallography or electron microscopy with a resolution of 4.0 A or better. Within each PDB entry, protein chains of 4-50 amino acids were treated as candidate peptide chains, while protein chains of at least 50 amino acids were treated as candidate receptor chains. This initial filtering yielded 18,007 candidate PDB entries, including 14,912 X-ray structures and 3,095 electron microscopy structures.

We then downloaded and parsed the corresponding mmCIF files, removing water molecules, ligands, and non-protein residues, and retaining only amino-acid residues with valid Cα atoms. For each candidate peptide chain, the receptor chain with the largest number of contacts in the same structure was selected as its paired receptor. Residue-level contacts were defined by the minimum heavy-atom distance between two residues, with a cutoff of 5.0 A. Peptide-receptor pairs with fewer than five contacts were discarded. Based on this contact definition, binary residue-level interface labels were assigned to both peptide and receptor chains.

To ensure structural quality, complexes with a high fraction of missing backbone atoms were removed: samples were discarded if more than 20% of peptide residues or more than 30% of receptor residues had incomplete backbone atoms. Each final sample contains the PDB ID, experimental method, resolution, peptide and receptor chain IDs, amino-acid sequences, residue-level interface labels, and contact statistics. In addition, Foldseek was used to generate 3Di structural tokens for each chain, which were aligned with the corresponding amino-acid sequences and used as auxiliary structural input features.

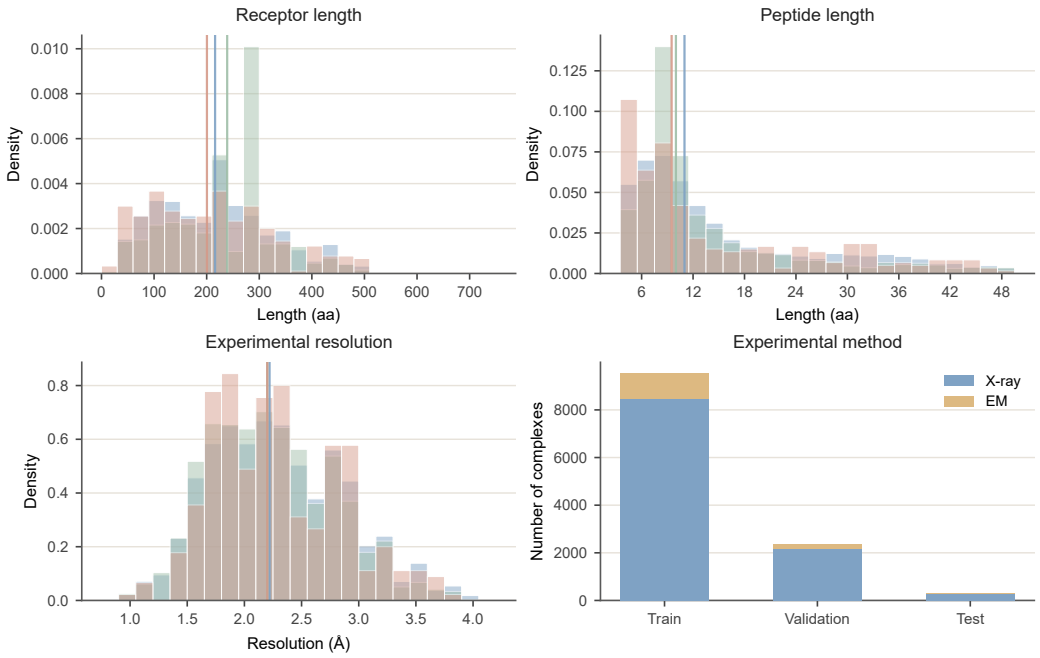

Figure 6: Core composition of the cluster-aware data split. Distributions are shown for receptor length, peptide length, experimental resolution, and experimental method. Vertical lines in the continuous panels mark split-specific medians.

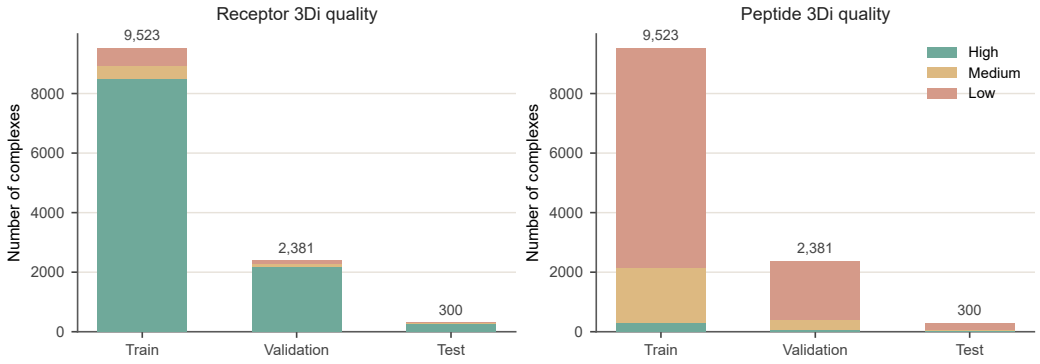

Figure 7: 3Di quality distributions for receptor and peptide chains. Bars show counts in each partition; labels above the bars show the total number of examples.

A.2 Dataset composition and interface statistics

We summarize the composition of the cluster-aware data split used throughout the experiments (Supplementary Fig. 6–8). The split contains 12,204 non-overlapping complexes, including 9,523 training, 2,381 validation, and 300 test examples. The split was generated with a fixed random seed (42) after sequence-cluster grouping at 80% identity and 80% coverage; consequently, no complex path is shared between any two partitions.

The core distributions are broadly comparable across partitions. Receptor lengths have medians of 216, 239, and 201 residues in the training, validation, and test sets, respectively, while peptide lengths have medians of 11, 10, and 10 residues. The median experimental resolutions are 2.22, 2.20, and 2.20 Å. X-ray diffraction is the dominant experimental method in all partitions. The test

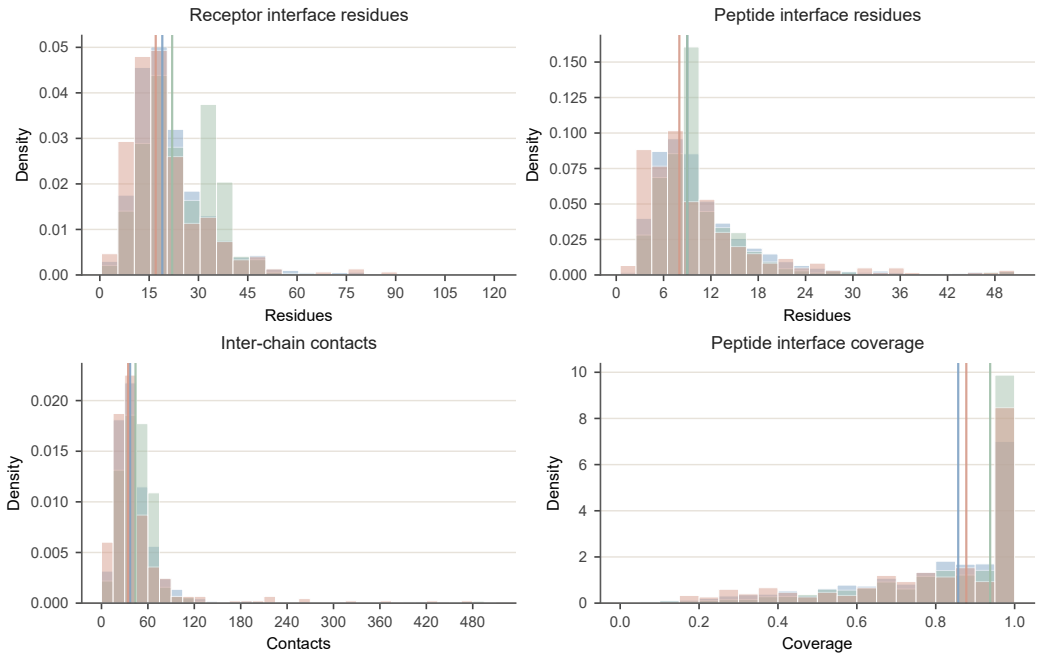

Figure 8: Interface-related distributions across the three partitions, including receptor-side interface residues, peptide-side interface residues, inter-chain contacts, and peptide interface coverage.

set contains only one electron-microscopy example, so method-specific comparisons involving the test set should be interpreted cautiously.

The 3Di quality profiles are also similar between partitions, with high-quality receptor 3Di tokens predominating and peptide 3Di tokens more frequently labeled medium or low quality. This asymmetry reflects the relative difficulty of assigning reliable structural tokens to the shorter peptide chains and is preserved by the split rather than introduced by a post-hoc filtering step.

Interface statistics indicate that the examples represent compact receptor–peptide interfaces. Across the three partitions, the median numbers of receptor-side interface residues are 19, 22, and 17, the median numbers of peptide-side interface residues are 9, 9, and 8, and the median inter-chain contact counts are 37, 44, and 35. The median peptide interface coverage is 0.857, 0.938, and 0.877, respectively. Thus, most peptide residues participate in the annotated interface, while the overall interface-size distributions remain close across partitions. These diagnostics support the use of the split for evaluating generalization under cluster-level separation, while highlighting the limited diversity of experimental methods in the held-out test set.

B Additional Evaluation Metrics

B.1 Sequence and Generation Quality Metrics

To complement the confidence-related metrics reported in the main text, we summarized the remaining evaluation metrics for the five primary systems in Table 3. These metrics include pTM, amino-acid entropy, hydrophobicity, net charge, diversity, and novelty. VersaPep w/o 3Di(Iter) achieved the highest mean diversity (0.840) and mean novelty (0.903), whereas VersaPep obtained the highest pTM maximum (0.821) and mean pTM (0.799). Across the four systems, the mean amino-acid entropy ranged from 2.507 to 2.613, the mean hydrophobicity ranged from -0.419 to -0.288, and the mean net charge ranged from -0.344 to 0.450.

Table 3: Additional evaluation metrics for the four primary systems. All values are averaged over the evaluated samples and rounded to three decimal places.

Model	pTM		Entropy	Hydrophobicity	Net charge	Diversity			Novelty		
	Max.	Mean				Max.	Mean	Min.	Max.	Mean	Min.
pepMLM	0.812	0.791	2.507	−0.288	−0.344	0.944	0.738	0.454	0.901	0.753	0.587
VersaPep	0.821	0.799	2.613	−0.415	0.228	0.860	0.674	0.405	0.894	0.799	0.660
VersaPep w/o 3Di	0.814	0.791	2.508	−0.418	0.450	0.942	0.739	0.440	0.962	0.870	0.737
VersaPep w/o 3Di(Iter)	0.816	0.793	2.601	−0.419	0.194	0.956	0.790	0.504	0.961	0.871	0.731

B.2 Additional Binding-Site Localization Analysis

For the binding-site analysis, the peptide-to-pocket distances are summarized in Table 4. Across the four binding-site conditions, the minimum distance to the true pocket ranged from 3.920 to 4.132 Å, whereas the minimum distance to the wrong pocket ranged from 5.795 to 5.951 Å. This separation was also maintained across the confidence-stratified subsets shown in Table 5.

Table 4: Peptide-to-pocket distances under different binding-site conditions. Values are reported in Å and averaged over the evaluated samples.

Binding-site input	Predicted pocket		True pocket		Wrong pocket	
	Mean	Min	Mean	Min	Mean	Min
True-bs	5.192	3.096	6.462	3.920	8.197	5.951
Pred-bs	5.167	3.130	6.510	4.037	8.142	5.916
None-bs	5.233	3.161	6.561	4.120	8.166	5.846
Wrong-bs	5.279	3.186	6.574	4.132	8.068	5.795

Table 5: Minimum peptide-to-pocket distances in confidence-stratified subsets.

Threshold	Binding-site input	True pocket min.	Wrong pocket min.
> 0.5	True-bs	4.034	6.894
	Pred-bs	4.177	6.893
	None-bs	4.193	6.846
	Wrong-bs	4.414	6.743
> 0.6	True-bs	3.877	7.207
	Pred-bs	3.967	7.160
	None-bs	3.970	7.078
	Wrong-bs	4.042	6.974
> 0.7	True-bs	4.013	7.463
	Pred-bs	4.106	7.367
	None-bs	4.150	7.253
	Wrong-bs	4.313	7.027
> 0.8	True-bs	3.631	7.462
	Pred-bs	3.686	7.245
	None-bs	3.777	7.186
	Wrong-bs	3.846	6.831
> 0.9	True-bs	3.965	7.371
	Pred-bs	4.382	7.298
	None-bs	4.481	7.266
	Wrong-bs	4.564	7.158

C Complementary Evaluation Using Independent Structure-Based Metrics

To further evaluate the generated peptide–receptor complexes independently of the AF-based confidence analysis, we performed complementary assessments using DiffPepDock, DockQ, and PyRosetta. DiffPepDock evaluation included 273 targets that passed receptor-alignment quality control, whereas DockQ and PyRosetta were evaluated on a strict common set of 280 targets, with five candidate structures retained for each model.

Table 6: DiffPepDock evaluation results. Pocket distance is reported in Å. Lower values indicate better performance for pocket distance and clash fraction, whereas higher values indicate better performance for the remaining metrics.

Method	Pocket F1 Mean@5	Pocket F1 Best@5	Distance Mean@5	Distance Best@5	Contact Fraction Mean@5	Contact Fraction Best@5	Residue Contacts Mean@5	Residue Contacts Best@5	Clash Fraction Mean@5
PepMLM	0.278	0.330	10.405	9.857	0.534	0.591	18.719	22.414	0.109
DiffPepBuilder	0.281	0.329	10.395	9.848	0.535	0.594	19.014	22.700	0.107
VersaPep	0.279	0.331	10.379	9.803	0.532	0.596	18.778	22.729	0.109
VersaPep w/o 3Di	0.279	0.329	10.379	9.796	0.536	0.596	18.989	23.275	0.108
VersaPep w/o 3Di(Iter)	0.281	0.332	10.384	9.799	0.534	0.598	19.192	23.645	0.109

As shown in Table 6, VersaPep w/o 3Di(Iter) achieved the highest mean and best Pocket F1 scores, while VersaPep w/o 3Di obtained the lowest best-pocket distance and the highest mean contact fraction. VersaPep w/o 3Di(Iter) also produced the largest numbers of residue-level contacts and the highest best contact fraction. The lowest mean clash fraction was observed for DiffPepBuilder. However, none of the pairwise differences in the DiffPepDock evaluation remained significant after BH-FDR correction.

Table 7: DockQ evaluation results on the strict common target set. Selected metrics were calculated using the candidate-selection protocol described in the evaluation procedure.

Method	Mean DockQ	Best DockQ	Selected DockQ	Selected F1	Selected F_{nat}	Selected iRMSD	Selected LRMSD	Best Acceptable Rate	Best Medium Rate	Best High Rate
PepMLM	0.246	0.360	0.287	0.221	0.225	6.426	15.362	0.611	0.300	0.093
DiffPepBuilder	0.205	0.295	0.218	0.146	0.147	6.978	16.333	0.561	0.196	0.018
VersaPep	0.254	0.365	0.297	0.235	0.243	6.116	14.747	0.593	0.300	0.093
VersaPep w/o 3Di	0.243	0.353	0.269	0.200	0.204	6.514	15.491	0.579	0.311	0.082
VersaPep w/o 3Di(Iter)	0.239	0.342	0.265	0.195	0.201	6.698	15.704	0.579	0.275	0.057

For DockQ evaluation, S1 achieved the highest mean DockQ, best DockQ, selected DockQ, selected F1, and selected F_{nat} values, together with the lowest selected iRMSD and LRMSD among the five systems. In total, 25 of 63 main-benchmark pairwise tests remained significant after multiple-testing correction, with the significant differences mainly involving comparisons against DiffPepBuilder.

Table 8: PyRosetta interface analysis results. Interface energies are reported in Rosetta energy units (REU), and interface dSASA is reported in $\mathrm{Å}^2$. Lower values are preferred for interface ΔG, ΔG/dSASA, and unsatisfied hydrogen bonds.

Method	Mean Interface ΔG	Best Interface ΔG	Selected Interface ΔG	Selected ΔG/dSASA	Shape Complementarity	Packstat	Interface H-bonds	Unsatisfied H-bonds	Interface dSASA
PepMLM	-73807.3	-241866.9	-206315.3	-189.762	0.561	0.628	2.911	11.529	1249.589
DiffPepBuilder	-101337.5	-345553.1	-275449.7	-299.799	0.523	0.622	2.371	11.889	1230.821
VersaPep	-150144.4	-380676.7	-276124.2	-290.492	0.552	0.623	3.200	11.818	1276.921
VersaPep w/o 3Di	-239779.9	-519489.2	-275935.1	-231.446	0.551	0.630	3.268	11.732	1265.105
VersaPep w/o 3Di(Iter)	-135959.6	-380402.6	-241275.8	-236.788	0.548	0.626	3.089	11.971	1254.400

The PyRosetta results showed complementary trends. VersaPep w/o 3Di obtained the most favorable mean and best interface ΔG, whereas VersaPep achieved the most favorable selected interface ΔG and the largest selected interface dSASA. VersaPep w/o 3Di produced the highest interface packing score and the largest number of interface hydrogen bonds. Six of the 81 corrected pairwise tests were significant, mainly in comparisons of VersaPep or VersaPep w/o 3Di with DiffPepBuilder. These results provide additional structure-based evidence while preserving the distinction between confidence-based evaluation and independent structural assessment.

D TRAINING SCHEDULE AND OPTIMIZATION DYNAMICS

The model was trained using four sequential optimization phases, as summarized in Table 9. Phase 0 optimized the initial training objective while monitoring hash recovery, validation loss, and validation perplexity. Phase 1 introduced the G1, G3, and binding-site objectives. Phase 2 further incorporated the S and R objectives. Phase 3 introduced the G2 objective, with separate monitoring of amino-acid and 3Di recovery.

Table 9: Training schedule reconstructed from the recorded history files.

Phase	Epochs	Logged objectives	Samples per epoch	Main monitored metrics
Phase 0	17	Initial objective	2372	Hash recovery, validation loss, validation PPL
Phase 1	25	G1, G3, B	2372	Hash recovery, binding-site AUROC, binding-site AP
Phase 2	33	G1, G3, S, B, R	2372	Recovery metrics, binding-site AUROC/AP, validation loss/PPL
Phase 3	15	G2, G1, G3, S	Not recorded	G2 amino-acid/3Di recovery, recovery metrics, binding-site AUROC/AP