HelixDiff: Hotspot-Specific Full-atom Design of Peptides Using Diffusion Models

Xuezhi Xie[12] **Pedro A Valiente**[1] **Jisun Kim**[1] **Philip M Kim**[123*]
[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
[2]Department of Computer Science, University of Toronto.
[3]Department of Molecular Genetics, University of Toronto.
* Corresponding email: pm.kim@mail.utoronto.ca.

Abstract

Peptide engineering has emerged as a critical discipline within biomedicine, finding applications in therapeutics, diagnostics, and synthetic biology. Despite their prevalence in biological processes, pursuing de novo therapeutic peptide design remains a formidable challenge. We here focus on generating helical peptides and present HelixDiff, a score-based diffusion model to learn and generate all-atom helical structures. We incorporate a hotspot-specific inpainting mechanism for the conditional design of α -helix structures that align with critical residues at protein-peptide interfaces. The results of our model showcase the production of helix structures with near-native geometries for a substantial portion of the test scenarios, showing root mean square deviations (RMSDs) less than 1Å. HelixDiff has shown better sequence recovery and Rosetta scores for unconditional and conditional generations than HelixGAN, our previous gan-based model. The case study involving glucagon-like peptide-1 (GLP-1) underscored HelixDiff's exceptional capacity to generate therapeutic D-peptides. The HelixDiff D-GLP-1 design is more stable than our earlier HelixGAN design when both D-peptides are bound to the GLP-1 receptor according to molecular dynamics simulations.

1 Introduction

Computational peptide design holds tremendous potential within biomedicine, spanning diagnostics, biosensors, therapeutics, and synthetic biology. At the core of its significance are protein-peptide interactions, pervasive within molecular pathways, deeply influencing cellular functions by orchestrating up to 40% of protein-protein interactions Petsalaki and Russell [2008]. Peptides, marked by attributes conducive to therapeutic progress, boast heightened specificity and reduced toxicity relative to small molecules. Notably, top-selling drugs, like glucagon-like peptide-1 (GLP-1) analogs, underscore the pharmaceutical value of peptides Wang et al. [2022]. The impetus for computational tools to streamline peptide modeling and engineering arises from their pervasive roles in biology and their potential for therapeutic advancement.

α-helix peptides, constituting the most prevalent secondary structure in proteins, play a pivotal role in conferring stability, comprising about 30% of the average globular protein's architecture Lau and Dunn [2018]. α-helical peptides are involved in nearly 40% of homodimeric and 26% of heterodimeric protein-protein interfaces Guharoy and Chakrabarti [2007]. Despite their prominence, native α-helical peptides pose challenges as drug candidates due to their diminished conformational stability in the absence of a protein scaffold and their susceptibility to proteolysis. In contrast, peptides composed of D-amino acids exhibit distinctive advantages, including low immunogenicity, cost-effective manufacturing, and robust proteolytic stability Kreil [1997], Rabideau and Pentelute [2015], Uppalapati et al. [2016].

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

We previously developed an in-house methodology that converts (L)-peptides into highly stable D-analogs through a mirror-image search in the protein data bank (D-PDB) Garton et al. [2018] and a deep learning approach called HelixGAN Xie et al. [2023] for de novo design. Using both methods, we have designed D-peptide analogs capable of activating the GLP-1, PTH, and GLP-2 receptors Garton et al. [2018], Xie et al. [2023], Valiente et al. [2023] while also inhibiting SARS-CoV-2 infections in vitro Valiente et al. [2021, 2022]. However, the D-PDB database encompasses a relatively substantial collection of native helical structures, representing a fraction of the possible stable helices, while the sample generation and indirect search for hotspot design limit the HelixGAN model.

Recently, diffusion models have surged to prominence in the realm of biology Watson et al. [2022], Lee et al. [2022], Luo et al. [2022] . Diffusion models excel at detecting intricate patterns within expansive datasets and crafting synthetic data infused with desired attributes, surpassing the performance of GAN - the leading generative model of the past decade Dhariwal and Nichol [2021]. In this study, we developed HelixDiff, a method that utilizes diffusion to estimate the distribution of helical structures, ensuring invariance to translation and rotation. HelixDiff can directly sample all-atom helical conformations while adhering to peptide design constraints. We integrated a hotspot-specific inpainting mechanism into HelixDiff, allowing for the direct generation of target helices with desired hotspot configurations.

2 Methods

The PDB, housing over 100,000 protein structures, is vital for structural biologists. Our approach extracts all PDB helices, resulting in our helical database. This database was split into training and test sets. We meticulously filtered the test set by sequence similarities cut-off to ensure no more than 42.8% sequence identity with the training data. To evaluate our model, we randomly selected 3 or 4 hotspot residues from the test set, yielding 270 test samples with L-type hotspots. Among these samples, one-third (90 test samples) were randomly chosen to assess the model's performance in generating 3-hotspot L-helical, 3-hotspot D-helical, and 4-hotspot D-helical structures.

We developed a score-based diffusion model named helixdiff to generate full-atom helices (Fig 1). Helices are represented using image-like representations, and it was encoded as a concatenated vector of one-hot encoded sequences and structure information using angles. As shown in Fig S1, those structural features include torsional angles, bond angles and planar angles, as well as one-hot encoded sequences, fully defining the helices with all side chains. We illustrate our framework that learns to generate full-atom peptides in Fig 1 A. We use the U-Net architecture Lin et al. [2017] for the score network.

We leverage the score-based generative modeling framework of Song et al. [2020], modeling the perturbation process with the following SDE:

$$\mathrm{d}x_t = \mu(x_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t, \tag{1}$$

where x_t is the perturbed sample at time t, $\mu(\cdot, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and W_t is the standard Wiener process.

The reverse-time diffusion process can be derived as follows [Anderson, 1982, Song et al., 2020]:

$$\mathrm{d}x_t = \left[\mu(x_t, t) - g(t)^2 \nabla_{x_t} \log p_t(x_t)\right] \mathrm{d}\bar{t} + g(t)\mathrm{d}\bar{W}_t, \tag{2}$$

where p_t is the marginal density of the perturbation process, $\bar{t}$ is an infinitesimal negative time step. In order to use this process as a generative model, we train a score network to approximate the score function with the score matching objective [Hyvärinen and Dayan, 2005, Song et al., 2020]. The details regarding our diffusion model is decribed in Appendix 1.1

To be useful, we developed a conditional hotpots-specific inpainting module that is specifically tailored to the receptor of interest in Fig 1 B,C. Here, we constrained the structural generation of the novel peptides to a set of identified hotspots to produce functional designs. Like in image inpainting, the hotspot residue information is given as surrounding context to the module to fulfill the rest of the data. Our model could generate different plausible estimates given specific hotspot residues.

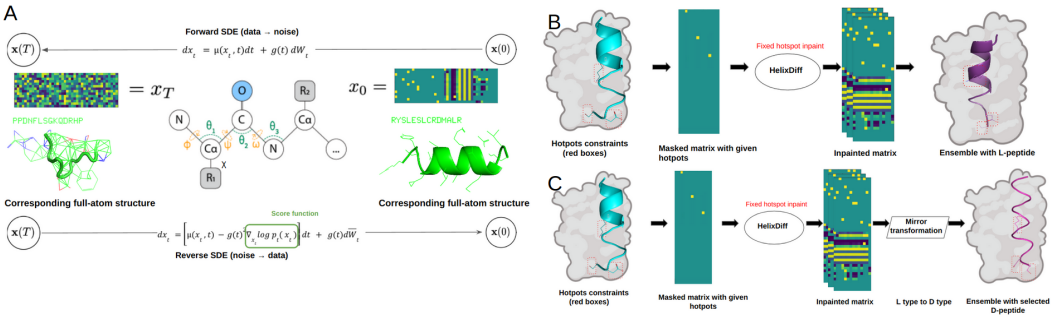

Figure 1: **Flowchart of the de novo methodology for generating helical structures coined as HelixDiff. A).**Flowchart for the score-based diffusion model. The model is trained to generate realistic helical structures from noise by learning a reverse "denoising" process given the forward diffusion process that maps data to Gaussian noise. The feature map in the middle shows the encoded information to generate the full-atom structures, including the sequence and the structural information (angle) vectors, respectively. **B).** The hotspot-specific inpainting module flowchart to generate L-type peptides. The module is given only the sequence information for the target hotspots as shown. Then the hotspot-specific inpainting module generated all the rest regions as shown in the inpainted matrix. Please note, all generated regions are different. This model directly generates L helices since the training data were L-peptides. **C).** Hotspot-specific inpainting module flowchart to generate D-type peptides. The flowchart is similar to B) but with an extra step called the mirror transformation, which could transform the generated L-helices into D-peptides.

3 Results and discussion

Unconditional sampling analysis. We unconditionally generated synthetic helical structures using HelixDiff and compared our results with HelixGAN. We tested the structure's quality by randomly unconditionally sampling 2,000 helices from HelixDiff and HelixGAN. The generated data contain a similar range of physical structural features as the training data (Fig 2 A-D). Notably, the majority of the generated data (77.7%) exhibited a sequence identity with the training data exceeding 50%, outperforming HelixGAN (Fig 2E, Fig S2). We also compared the Rosetta scores distribution obtained with HelixDiff and HelixGAN, to the training data. HelixDiff showed a more similar distribution to training data than HelixGAN, indicating a better performance for α-helix structure generation.

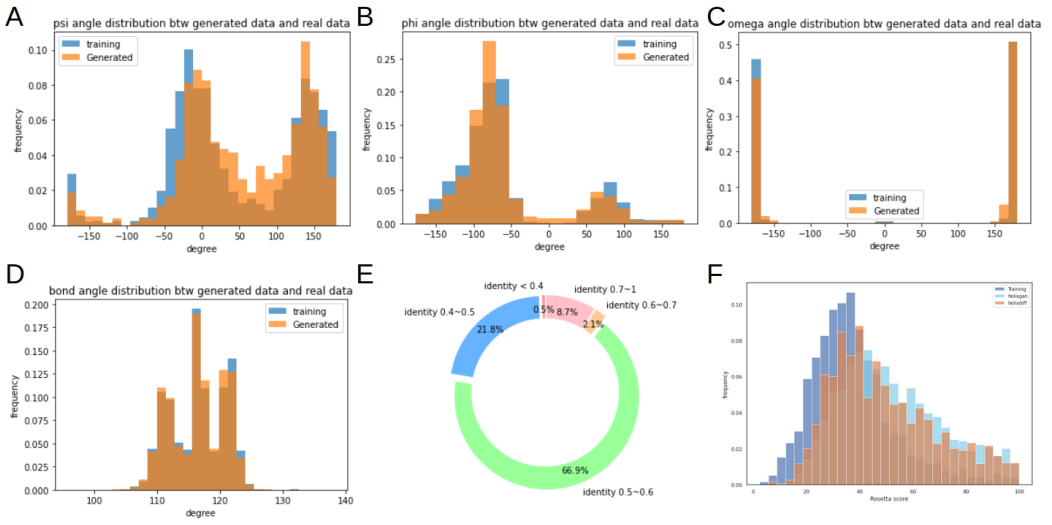

Figure 2: **Unconditional sampling analysis**. **A-D)** Structural features in terms of psi,phi,omega and bond angles between 2k generated samples and real data. **E)** Sequence identities compared with training data. **F)** Rosetta score distribution regarding training data, HelixGAN and HelixDiff.

Hotspots-specific peptide generation analysis We assessed the performance of our hotspot-specific inpainting mechanism on the test set, generating L-type helices that matched the desired hotspot residues (Table 1). The evaluation involved calculating the RMSD between the target and generated helix structures via partial alignment of the hotspots and matched residue atoms. Impressively, in 54.5% of the 3-hotspot test cases, the RMSDs of matched residues in the newly generated helices were less than 1 Å for the target hotspots. At the same time, 31.1% fell between 1 and 1.5 Å (Table 1). The Rosetta scoring function classified most of the generated helix structures as reasonable (Fig S3), which is better than HelixGAN.

One of our primary objectives is to use our model for designing D-peptides, mimicking bioactive L-peptides. We applied a mirror conversion step to transform the helices generated with HelixDiff into D-peptides. Given a set of hotspots in a known L-peptide, our model conditionally generates novel D-helix peptide structures using the inpainting mechanism and mirror conversion. In 58.9% of the 3-hotspot test cases, the RMSDs of matching residues were below 1.5 Å, while 10% fell below 1.0 Å (Table 1). Most novel D-helix structures received favorable classifications based on their Rosetta scores (Fig S4). We also tested our model to generate D-peptides using 4-hotspot test sets. HelixDiff could generate D-peptides with matching residues below 1.5 Å in 36.1% of the 4-hotspot test cases. This new model significantly outperformed HelixGAN, which could not support the 4-hotspot D-peptide design due to the inefficiencies of the indirect search.

Testcases	Hotspots	Method	rmsd (< 1Å)	rmsd (1 $\sim$ 1.5Å)	rmsd (1.5 $\sim$ 2Å)	rmsd (2$\sim$ 5Å)	NA
L type	3	HelixGAN	7.8%	23.3%	34.4%	22.2%	12.2%
L type	3	HelixDiff	54.5%	31.1%	11.1%	3.3%	0
D type	3	HelixGAN	0	32.6%	40.4%	3.4%	23.6%
D type	3	HelixDiff	10%	48.9%	37.8%	3.3%	0
D type	4	HelixDiff	1.4%	34.7%	55.6%	8.3%	0

Table 1: **Evaluation of the hotspot-specific conditional generation by helixGAN and our helixDiff.** For every two rows, the same testcases were performed between HelixGAN and HelixDiff. The lowest RMSD for each testcase is calcualted and summerized as the table. The hotspots columns indicate the random-selected hotpot residues in the testcases. Noted, HelixGan could not perform 4 hotpsot conditional generations.

De novo design of a D-peptide analog of GLP-1. Here, we targeted a set of well-known hotspot residues of GLP-1 to design a novel GLP-1 D-peptide analog using HelixDiff. To generate a D-GLP1 analog, we used the ligand structure in complex with the full-length GLP-1 receptor as a starting point (PDB code: 5vai). Before designing the D-peptide analog, we divided the GLP-1 structure into three overlapping fragments named helix1, helix2, and helix3. Helix1 extends from H7 to L20, helix2 runs from F12 to A25, while helix3 runs from Q23 to R36. Positions H7, E9, and F12 were designated as hotspots in helix1. For helix2, T13, D15, and Y19 were selected as hotspots, while F28, I29, and L32 were chosen in helix3 (Fig 3A). Previous computational and experimental studies supported our hotspot residue selection Garton et al. [2018], Manandhar and Ahn [2015]. Several D-helix structures were generated after running HelixDiff independently for each helix. The root mean square deviation (RMSD) between the specific atoms in each starting helix structure and the generated D-helix structures was used to evaluate the match quality. We next joined the best-matched peptides generated to design D-GLP1-diff (Fig 3 A).

We then superimposed the D-GLP-1-diff structure onto the Cryo-EM structure of GLP1R bound to GLP1 to build the 3D structure of the GLP1 receptor (GLP1R) in complex with the novel D-peptide analog (Fig 3 B). The GLP1R+D-peptide complex was then embedded in a POPC: PSM (1:1) bilayer before evaluating its binding mode stability using 200 ns MD simulations (Fig 3 B). We also simulated the GLPR1 bound to D-GLP-1-diff acetylated in the N-terminal, the wild-type L-GLP-1, and the D-GLP-1 peptide designed earlier using HelixGAN as controls. The peptides designed in the current work (D-GLP-1-diff and D-GLP-1-diff-Acetylated) share the same orientation with the L-peptide, similar to D-GLP-1-HelixGAN. Like the L-GLP1, all D-peptides quickly stabilized in a new equilibrium position close to the initial structure, according to the RMSD profiles calculated for the peptide's heavy atoms (Fig 3 C). The calculated RMSF profiles revealed that both D-GLP-1-diff analogs were more stable than the one created using HelixGAN (Fig 3 D).

Comparing our GAN model, HelixGAN, with the diffusion model, HelixDiff, reveals distinct strengths and performance differences. While both models are geared towards peptide generation, HelixDiff, as a score-based generative diffusion model, notably outshines HelixGAN in several aspects. HelixDiff

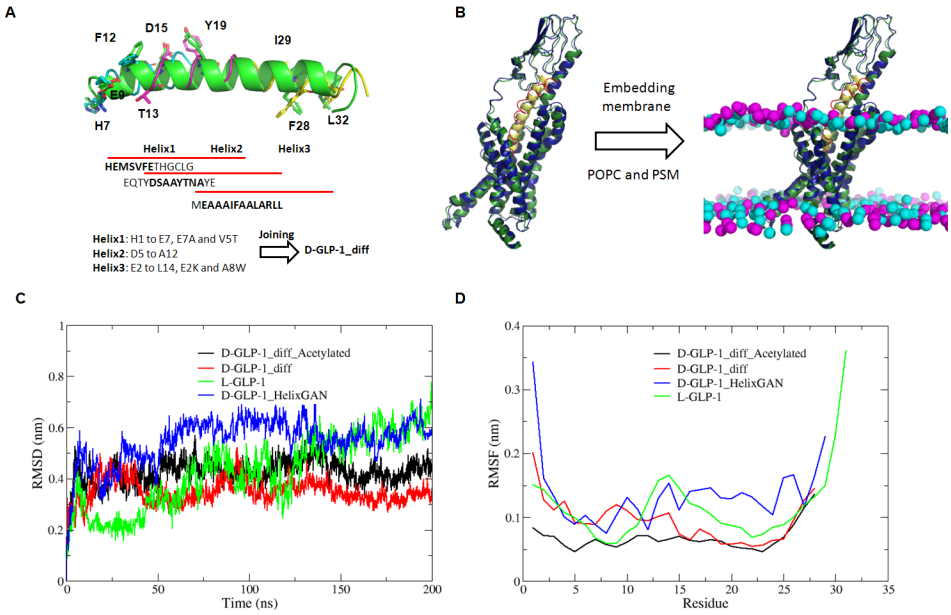

Figure 3: **Design and modeling the 3D structure of D-GLP-1-diff analogs bound to the GLP-1 receptor.**. A) Strategy for design two novel D-GLP-1 analogs using HelixDiff. The GLP1 structure was divided into three overlapping peptides named helix1, helix2, and helix3. Helix1 extends from H7 to L20, helix2 runs from F12 to A25 while helix3 extends from Q23 to R36. Critical hotspots for the GLP1 function chosen for the design were colored as licorice. B) Structural superposition of D-GLP-1-diff analog (red) over the L-GLP1 (yellow) structure bound to the GLP1 receptor (pdb code: 5vai). In dark green is shown the GLP1R coupled to GLP1, while in dark blue is represented the receptor bound to the D-peptide analog. The magenta and cyan spheres represent the phosphate groups of the lipids polar heads. C) Root mean square deviation (RMSD) of the heavy atoms of L-GLP-1 and the D-GLP-1 analogs bound to the GLP1R. D) Root mean square fluctuation (RMSF) per residue of the heavy atoms of L-GLP1 and the D-GLP-1 analogs bound to the GLP1R.

consistently exhibits superior performance in unconditional generation tasks, generating peptide structures with higher fidelity and improved structural quality. This advantage is further accentuated in conditional generation scenarios, where HelixDiff excels in producing peptides tailored to specific requirements, such as matching critical hotspots in known structures. The key to HelixDiff's success lies in its ability to harness the power of stochastic differential equations, enabling smoother data-to-noise transformations and more accurate synthesis through reverse diffusion. In contrast, HelixGAN, while effective, may struggle with maintaining the same level of precision and structural fidelity. This comparison underscores the potential of score-based generative diffusion models like HelixDiff as formidable tools for peptide design. It highlights their capacity to outperform traditional GAN-based approaches in this domain.

D-peptide analogs of bioactive peptides offer compelling therapeutic advantages, such as reduced immunogenicity, cost-effectiveness, and enhanced protease stability, as established in previous studies Kreil [1997], Rabideau and Pentelute [2015], Uppalapati et al. [2016]. HelixDiff has demonstrated superior performance than our previously developed methods in generating D-peptides that match specific hotspots within known L-peptides. This discovery is relevant for expanding the structural diversity available for designing D-peptides from scratch. Our selection of GLP1 as a case study is particularly noteworthy, as its potential in treating diabetes mellitus and obesity Prasad-Reddy and Isaacs [2015], Sonne et al. [2014]. GLP-1 is a helical GPCR agonist with distinct hotspot residues characterized by unique chemical and physical properties Donnelly [2012]. Molecular dynamics (MD) simulations have unveiled a stable binding mode for the D-GLP1 analog with GLP1R. This novel D-peptide analog exhibits greater stability than our previous D-GLP1 design throughout MD simulations. This heightened stability holds critical implications for enhancing the peptide's capacity to activate GLP1R.

HelixDiff consistently performs better in unconditional generation tasks than HelixGAN, generating higher fidelity and improved quality structures. This benefit is emphasized in conditional generation scenarios, such as matching crucial hotspots in bioactive peptides. Our findings illustrate the benefits of using a score-based generative diffusion model, including improved sample generation quality, a valuable inpainting pipeline, and improved stability in constructing D-GLP-1 analogs. We will undertake experimental assays in the future to evaluate the biological function of the designed D-GLP-1 peptide.

References

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

D. Donnelly. The structure and function of the glucagon-like peptide-1 receptor and its ligands. *British journal of pharmacology*, 166(1):27–41, 2012.

M. Garton, S. Nim, T. A. Stone, K. E. Wang, C. M. Deber, and P. M. Kim. Method to generate highly stable d-amino acid analogs of bioactive helical peptides using a mirror image of the entire pdb. *Proceedings of the National Academy of Sciences*, 115(7):1505–1510, 2018.

M. Guharoy and P. Chakrabarti. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics*, 23(15): 1909–1918, 2007.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

G. Kreil. D-amino acids in animal peptides. *Annual review of biochemistry*, 66(1):337–345, 1997.

J. L. Lau and M. K. Dunn. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry*, 26(10):2700–2707, 2018.

J. S. Lee, J. Kim, and P. M. Kim. Proteinsgm: Score-based generative modeling for de novo protein design. *bioRxiv*, pages 2022–07, 2022.

G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems*, 2022.

B. Manandhar and J.-M. Ahn. Glucagon-like peptide-1 (glp-1) analogs: recent advances, new possibilities, and therapeutic implications. *Journal of medicinal chemistry*, 58(3):1020–1037, 2015.

E. Petsalaki and R. B. Russell. Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology*, 19(4):344–350, 2008.

L. Prasad-Reddy and D. Isaacs. A clinical review of glp-1 receptor agonists: efficacy and safety in diabetes and beyond. *Drugs in context*, 4, 2015.

A. E. Rabideau and B. L. Pentelute. A d-amino acid at the n-terminus of a protein abrogates its degradation by the n-end rule pathway. *ACS central science*, 1(8):423–430, 2015.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

D. P. Sonne, M. Hansen, and F. K. Knop. Mechanisms in endocrinology: bile acid sequestrants in type 2 diabetes: potential effects on glp1 secretion. *European journal of endocrinology*, 171(2): R47–R65, 2014.

M. Uppalapati, D. J. Lee, K. Mandal, H. Li, L. P. Miranda, J. Lowitz, J. Kenney, J. J. Adams, D. Ault-Riché, S. B. Kent, et al. A potent d-protein antagonist of vegf-a is nonimmunogenic, metabolically stable, and longer-circulating in vivo. *ACS chemical biology*, 11(4):1058–1065, 2016.

P. A. Valiente, H. Wen, S. Nim, J. Lee, H. J. Kim, J. Kim, A. Perez-Riba, Y. P. Paudel, I. Hwang, K.-D. Kim, et al. Computational design of potent d-peptide inhibitors of sars-cov-2. *Journal of Medicinal Chemistry*, 64(20):14955–14967, 2021.

P. A. Valiente, S. Nim, J. Lee, S. Kim, and P. M. Kim. Targeting the receptor-binding motif of sars-cov-2 with d-peptides mimicking the ace2 binding helix: lessons for inhibiting omicron and future variants of concern. *Journal of Chemical Information and Modeling*, 62(15):3618–3626, 2022.

P. A. Valiente, S. Nim, J. Kim, and P. M. Kim. Computational design of potent and selective d-peptide agonists of the glucagon-like peptide-2 receptor. *Journal of Medicinal Chemistry*, 66(15): 10342–10353, 2023.

L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, and C. Fu. Therapeutic peptides: Current applications and future directions. *Signal Transduction and Targeted Therapy*, 7 (1):48, 2022.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. Dec. 2022.

X. Xie, P. A. Valiente, and P. M. Kim. Helixgan a deep-learning methodology for conditional de novo design of α-helix structures. *Bioinformatics*, 39(1):btad036, 2023.

Appendix - HelixDiff

1.1 the Score-based generative modeling

Score-based generative modeling (SBGM) was initially introduced by Song et al. We also utilize it for the HelixDiff. In the forward or sampling process, a random variable x_t is generated by simulating the Stochastic Differential Equation (SDE) over time, commencing from an initial value x0. This initial value, x0, undergoes perturbation through the addition of Gaussian noise during the forward process, yielding perturbed samples x_t. The forward process within the context of SBGM can be succinctly characterized by the following SDE:

$$dx_t = \mu(x_t, t)dt + g(t)\, dW_t$$

where x_t is the perturbed sample at time t, $\mu(x_t, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and W_t is a standard Wiener process that represents the random fluctuations in the system.

The backward process pertains to the generation of a sequence of purified samples in a reverse sequence, conditioned on a target observation sequence. This reverse process is derived from the forward process through the utilization of gradients of the log-likelihood concerning the model parameters, often referred to as the score function. When provided with a forward Stochastic Differential Equation (SDE), an analogous SDE in reverse time can be formulated as follows:

$$dx_t = \left[\mu(x_t, t) - g(t)^2 \nabla_{x_t} log\, p_t(x_t)\right] dt + g(t) d\overline{W}_t$$

Where $\nabla_{x_t} log\, p_t(x_t)$ is the score function, σ(t) is the diffusion coefficient, and $\overline{W}_t$ is the standard Wiener process.

The score function is a measure of how the log-likelihood of the perturbed sample, denoted as x_t, evolves throughout the forward diffusion process. By calculating this score function, we can ascertain both the direction and magnitude of the gradient with respect to x_t, which subsequently influences the drift term in the reverse-time Stochastic Differential Equation (SDE). Typically, this score function is estimated using a neural network. In our study, we employed a UNet-based architecture featuring an attention module to estimate this score.

The Variance Exploding Stochastic Differential Equation (VESDE) represents a novel type of SDE introduced as a more efficient alternative to conventional SDEs within score-based generative models. The primary objective of this diffusion process is to augment the noise variance within the SDE to prevent sample collapse into a lower-dimensional subspace. Notably, VESDE differs from the original SDE by having a tractable reverse process in which the drift coefficient of the forward process does not exert an influence. The forward process of VESDE is defined as follows:

$$dx_t = \sqrt{\frac{d[\sigma^2(t)]}{dt}}\, dW_t$$

To generate samples from the reverse-time process, we employ a numerical solution approach to solve the reverse-time Stochastic Differential Equation (SDE). Specifically, we utilize a predictor-corrector solver method. This method involves a sequential process where a predictor, such as the Euler-Maruyama method, is initially applied to estimate the state for the next time step. Subsequently, a corrector method, like Langevin Markov Chain Monte Carlo (MCMC), is employed to adjust the predicted state, ensuring that it adheres to the desired marginal distribution:

Predictor: $x_{t-\mathrm{d}t} = x_t + (\sigma_t^2 - \sigma_{t-\mathrm{d}t}^2) s_{\theta^*}(x_t, t) + \sqrt{\sigma_t^2 - \sigma_{t-\mathrm{d}t}^2}\, z_p$

Corrector: $x_{t-\mathrm{d}t} = x_{t-\mathrm{d}t} + \epsilon_t s_{\theta^*}(x_{t-\mathrm{d}t}, t) + \sqrt{2\epsilon_t}\, z_c$

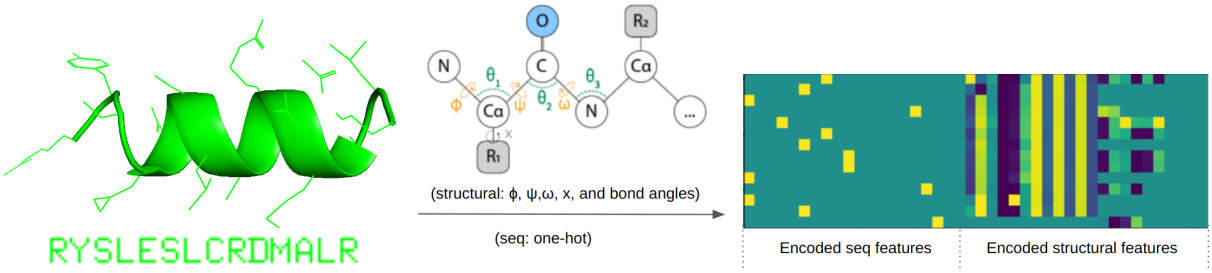

Figure S1: Encoding illustration: the peptide is encoded as the combined vectors of the primary sequence information and structural information. The structural information is further divided into sidechain and mainchain. The mainchain is represented as phi (ϕ), psi (ψ), omega (ω) and bond angles while the sidechain is represented using 5 chi (χ) angles.

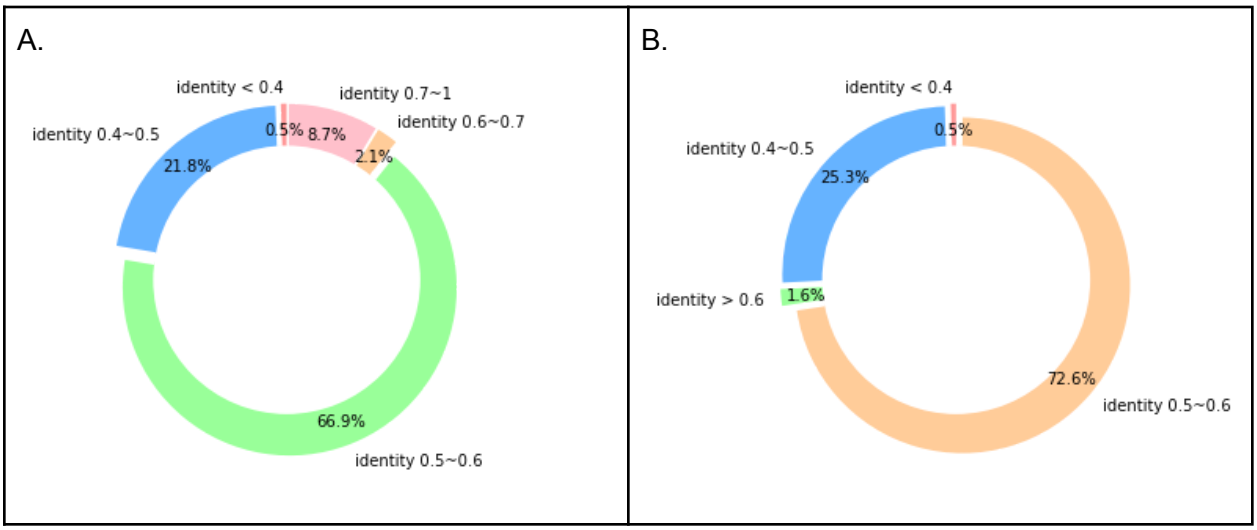

Figure S2: Seq identities for unconditional generations regarding helixDiff (A) and HelixGan (B).

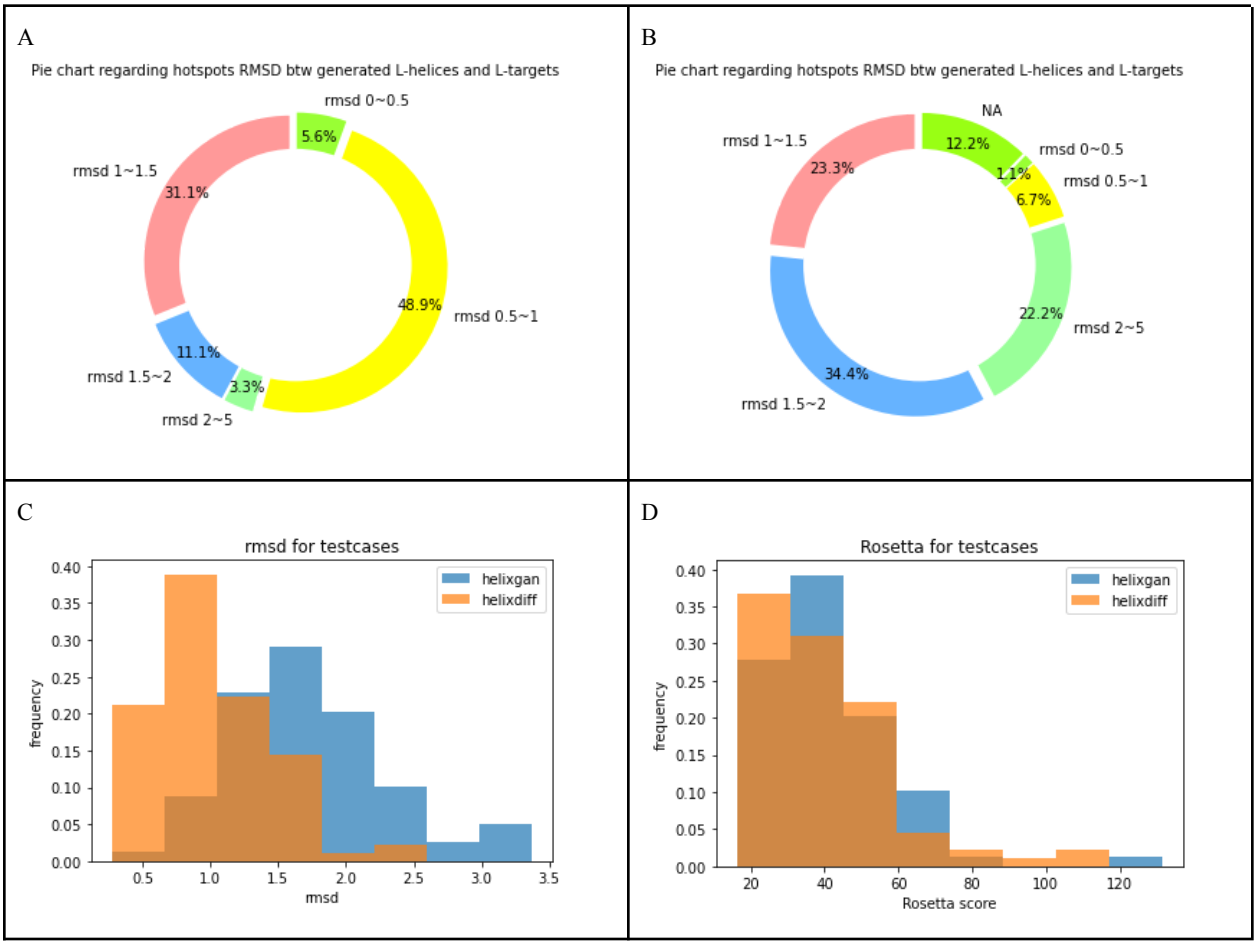

FIgure S3; De novo design of L-helical peptides with constrained hotspots using the hotspot-specific inpainting method implemented in HelixDiff. (A) Assessment of the performance of HelixDiff to generate L-helical peptides with constrained hotspots in a test set of 90 samples. The RMSD was calculated between the matched and hotspots atoms in each generated and target helix. B) RMSD regarding the same assessment of the performance of HelixGan (C) Histogram regarding RMSD between Helixgan and HelixDiff (D.) Histogram regarding Rosetta score distribution across the novel generated D-helices between Helixgan and HelixDiff.

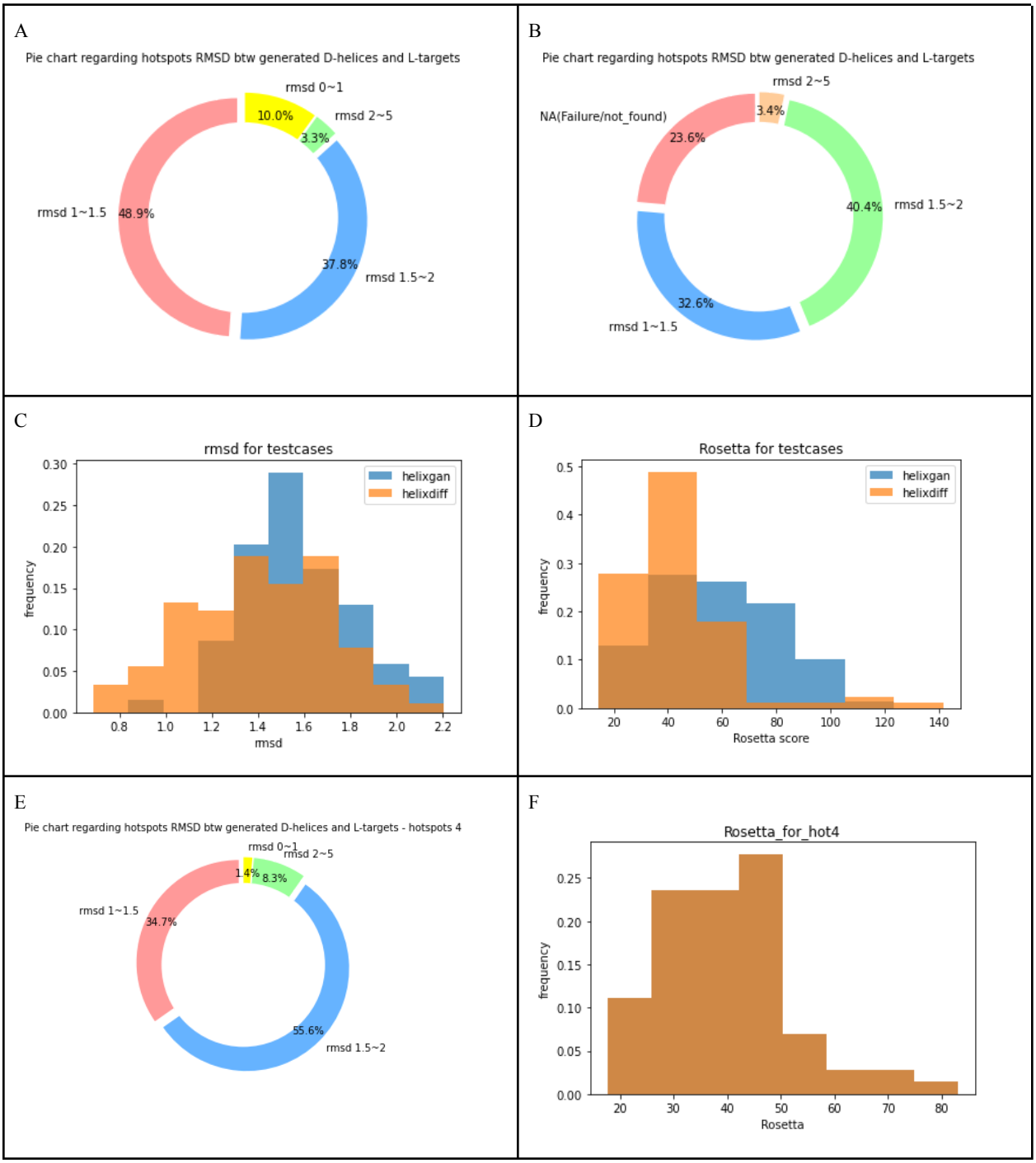

Figure S4: De novo design of D-helical peptides with constrained hotspots using the hotspot-specific inpainting method implemented in HelixDiff. (A) Assessment of the performance of HelixDiff to generate D-helical peptides with constrained hotspots in a test set of 90 samples where 3 random hotspots are selected. The RMSD was calculated between the matched and hotspot atoms in each generated and target helix. (B) RMSD regarding the same assessment of the performance of HelixGan (C) Histogram regarding RMSD between HelixGAN and HelixDiff (D.) Histogram regarding Rosetta score distribution across the novel generated D-helices. between HelixGAN and HelixDiff. (E) Assessment of the performance of HelixDiff to generate D-helical peptides with constrained hotspots in a test set of 90 samples where 4 random hotspots are selected. (F) Rosetta score distribution across the novel generated D-helices.