Peptune

De Novo Design of Therapeutic Peptides

MULTI- OBJECTIVE DISCRETE DIFFUSION

Sophia Tang* • Yinuo Zhang* • Pranam Chatterjee

Penn | DukeNUS Medical School

Can we generate *valid* therapeutic peptides simultaneously optimized across *multiple* properties?

Lack of Multi-Objective Guidance Strategies for Discrete Generation

Previous discrete guidance methods rely on projecting to and from the continuous latent space or gradient estimation. Multi-objective guidance strictly in the discrete state space remains underexplored.

Lack of generative models non-natural and cyclic peptides

Existing sequence-based models represent peptides as sequences of the 20 natural amino acids, but fail to represent the diverse non-natural amino acids and cyclic peptides that can improve therapeutic properties.

BOND-DEPENDENT MASKED DISCRETE DIFFUSION

Bond-Dependent Masking Schedule

$$q(\mathbf{z}_t|\mathbf{x}_0) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t(\mathbf{x}_0)\mathbf{x}_0 + (1-\alpha_t(\mathbf{x}_0))\mathbf{m})$$

$$\alpha_t(\mathbf{x}_0) = \begin{cases} 1-t^w & \mathbf{x}_0 = \mathbf{b} \\ 1-t & \mathbf{x}_0 \neq \mathbf{b} \end{cases}$$

Late masking of peptide bond tokens

Enforces early unmasking of peptide bond tokens

Bond-Dependent NELBO Loss

The loss for peptide bond tokens is weighted heavier by the exponent

$$\mathcal{L}_{\text{NELBO}}^{\infty} = \mathbb{E}_{t,q(\mathbf{z}_t|\mathbf{x}_0)}\left[-\sum_{\ell:\mathbf{x}_0^{(\ell)}=\mathbf{b}} \frac{w}{t}\log\langle \mathbf{x}_0^{(\ell)}, \mathbf{x}_\theta^{(\ell)}(\mathbf{z}_t,t)\rangle - \sum_{\ell:\mathbf{x}_0^{(\ell)}\neq\mathbf{b}} \frac{1}{t}\log\langle \mathbf{x}_0^{(\ell)}, \mathbf{x}_\theta^{(\ell)}(\mathbf{z}_t,t)\rangle \right]$$

Peptide Bond Tokens — Side-Chain Tokens

Bond-Dependent Reverse Posterior

$$q(\mathbf{z}_s|\mathbf{z}_t,\mathbf{x}_0) = \begin{cases} \left\langle \left(\frac{s}{t} - \frac{s^w}{t^w}\right)\mathbf{b} + \frac{t-s}{t}\mathbf{1}, \mathbf{x}_0 \right\rangle \mathbf{x}_0 + \left\langle \left(\frac{s^w}{t^w} - \frac{s}{t}\right)\mathbf{b} + \frac{s}{t}\mathbf{1}, \mathbf{x}_0 \right\rangle \mathbf{m} & \mathbf{z}_t = \mathbf{m} \\ \mathbf{z}_t & \mathbf{z}_t \neq \mathbf{m} \end{cases}$$

Parameterize — Parameterize

$$p_\theta(\mathbf{z}_s|\mathbf{z}_t) = \begin{cases} \left\langle \left(\frac{s}{t} - \frac{s^w}{t^w}\right)\mathbf{b} + \frac{t-s}{t}\mathbf{1}, \mathbf{x}_\theta(\mathbf{z}_t,t) \right\rangle \mathbf{z}_s + \left\langle \left(\frac{s^w}{t^w} - \frac{s}{t}\right)\mathbf{b} + \frac{s}{t}\mathbf{1}, \mathbf{x}_\theta(\mathbf{z}_t,t) \right\rangle \mathbf{m} & \mathbf{z}_t = \mathbf{m} \\ \mathbf{z}_t & \mathbf{z}_t \neq \mathbf{m} \end{cases}$$

Unconditional Peptide SMILES Generator

Selection

Start from a fully masked sequence (root node) and follow a sequence of *optimal* unmasking steps to a leaf node (unexpanded partially masked sequence)

Expansion

From the probability distribution generated from the trained diffusion model, apply Gumbel noise and sample *M* distinct partially unmasked sequences.

$$\log \tilde{p}_{\theta,i}(\mathbf{z}_{s,i}|\mathbf{z}_t) = \log p_\theta(\mathbf{z}_{s,i}|\mathbf{z}_t) + \mathbf{G}_i$$

MONTE-CARLO TREE GUIDANCE (MCTG)

Membrane Permeability
Validation Spearman: **0.943**

Binding Affinity
Validation Spearman: **0.630**

We trained XGBoost classifiers for key therapeutic properties, including binding affinity, membrane permeability, solubility, non-hemolysis, and non-fouling for multi-objective guidance

	Solubility		Hemolysis		Non-fouling	
Metric	Ours	PeptideBERT	Ours	PeptideBERT	Ours	PeptideBERT
F1	**0.660**	0.597	**0.846**	0.483	**0.768**	0.699
Accuracy	**0.661**	0.651	**0.846**	0.823	0.766	**0.873**

Iteration

EXPERIMENTS

Multi-Objective Peptide SMILES Generation

	Permeability Data	Binding Data	PepMDLM
Mean nAAs Per Peptide	2.215	2.150	2.940
Cyclic Peptides (%)	0.467	0.027	0.100

Model	Validity (↑)	Uniqueness (↑)	Diversity (↑)	SNN (↓)	Randomness (↑)	KL-Divergence (↑)
Data	1.000	1.000	0.885	1.000	4.55	0 (Reference)
PepMDLM	0.450	1.000	0.705	0.513	4.11	0.174
PepTune	1.000	1.000	0.677	0.486	4.12	0.173

Membrane Permeability Density Plot

GFAP Binding Affinity Density Plot

Dual-Targeting Peptides to TfR and GLAST Protein

Average Binding Affinity to GLAST Over Iterations

Average Binding Affinity to TfR Over Iterations

Average Solubility Over Iterations

Average Non-Hemolysis Over Iterations

Average Non-Fouling Over Iterations

Rollout

From each partially unmasked child sequence, use greedy unmasking to fully unmask the sequence for scoring. Feed the unmasked sequences into a set of *K* classifiers to determine Pareto optimality

Backpropogation

Calculate a reward vector of the fraction of the Pareto-optimal set that a sequence has a greater or equal score. Add the rewards across child nodes and add to the rewards of all predecessor nodes

$$\mathbf{r}_k(\mathbf{x}_{s,i}) = \frac{1}{|\mathcal{P}^*|}\sum_{n=1}^{|\mathcal{P}^*|}\mathbb{I}\left[\mathbf{s}_k(\mathbf{x}_{s,i}) \succeq \mathbf{s}_k(\bar{\mathbf{x}}_n)\right]$$

All property scores are optimized simultaneously over iterations of MCTG

* average scores are calculated from rolled out child sequences

Paper

Our Lab