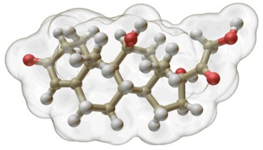
Full-Atom Peptide Design with Geometric Latent Diffusion

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, Yang Liu

Contents

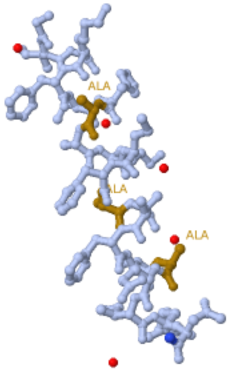
1. Introdcution
2. Task definition
3. Method
4. Experiments
5. Conclusion

Why Peptide?



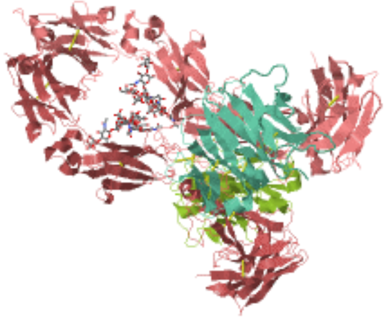
Small Molecule

V.S.



Peptide

V.S.



Antibody

Small Molecule – Drawbacks

- Low specificity
- Toxicity
- Low synthesizability

Peptide – Advantages

- High specificity
- Good safety
- High synthesizability

Peptide – Advantages

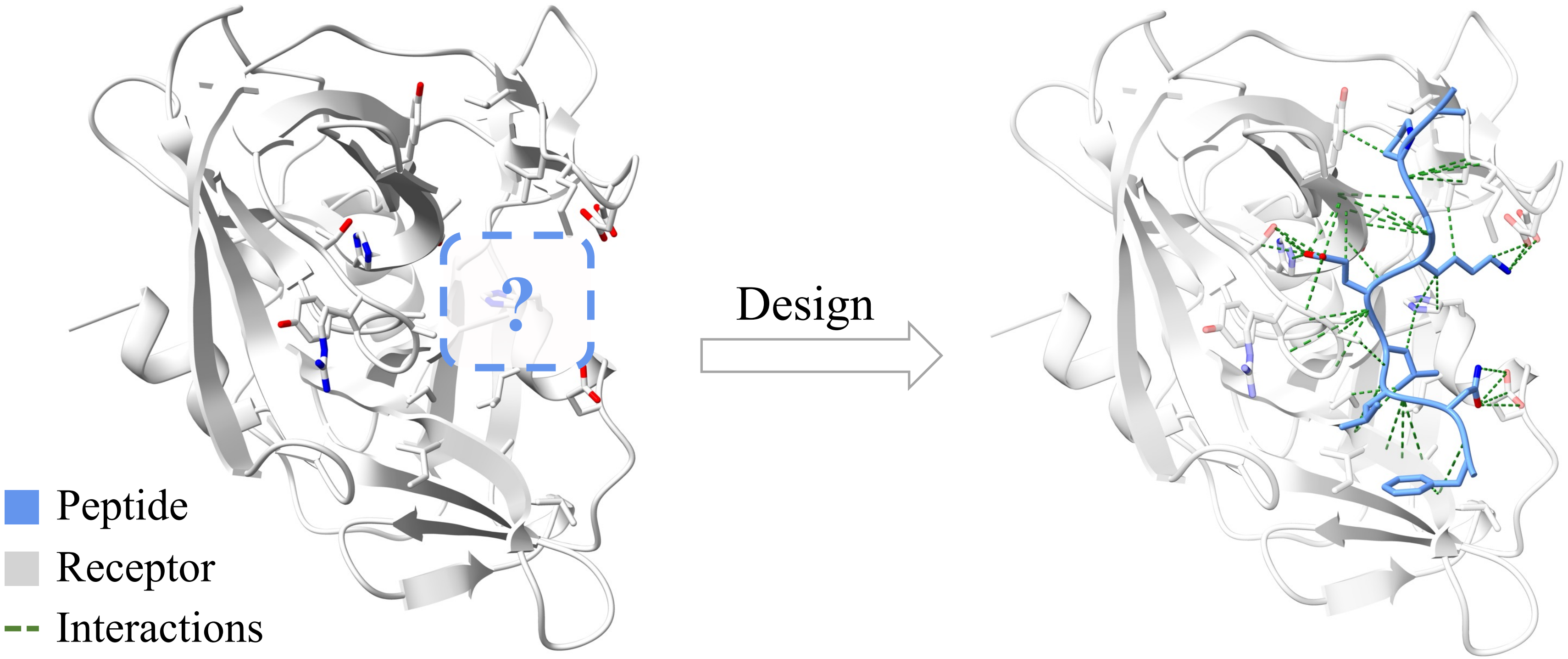
- High cell permeability
- Oral availability
- Low cost

Antibody – Drawbacks

- Low cell permeability
- Injection
- High cost

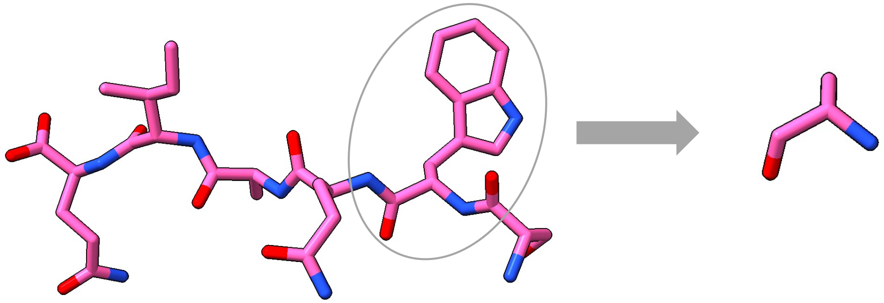
Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy* 7.1 (2022): 48.

Peptide Design: Given the binding site $\mathcal{G}_b = \{(x_i, \overrightarrow{X_i})\}$, the model is required to generate the full-atom structure of a peptide binder $\mathcal{G}_p = \left\{\left(x_j, \overrightarrow{X_j}\right)\right\}$, where x and $\vec{X}$ denote the amino acid type and the coordinates of all atoms in the amino acid.



Challenges

Variable Data Length

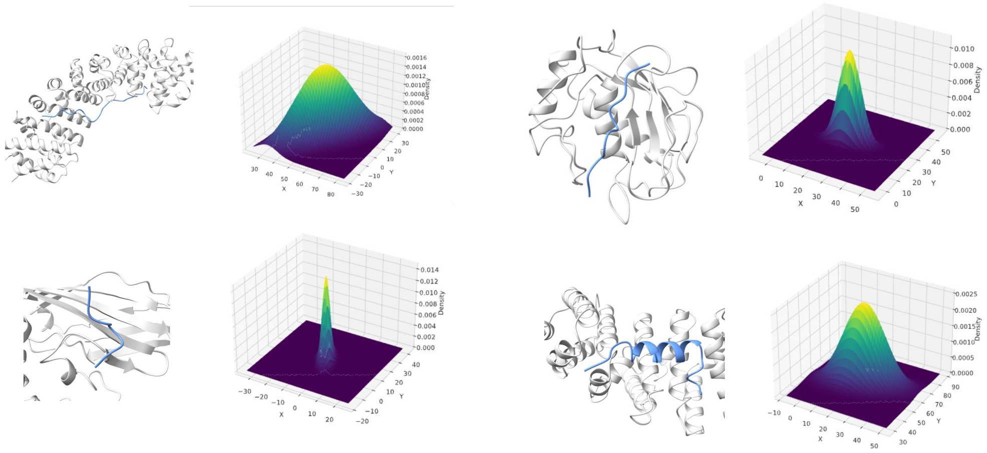


$$x_j^t = W \quad \longrightarrow \quad x_j^{t-1} = A$$

$$\vec{X}_j^t \in \mathbb{R}^{14\times3} \quad \longrightarrow \quad \vec{X}_j^{t-1} \in \mathbb{R}^{5\times3}$$

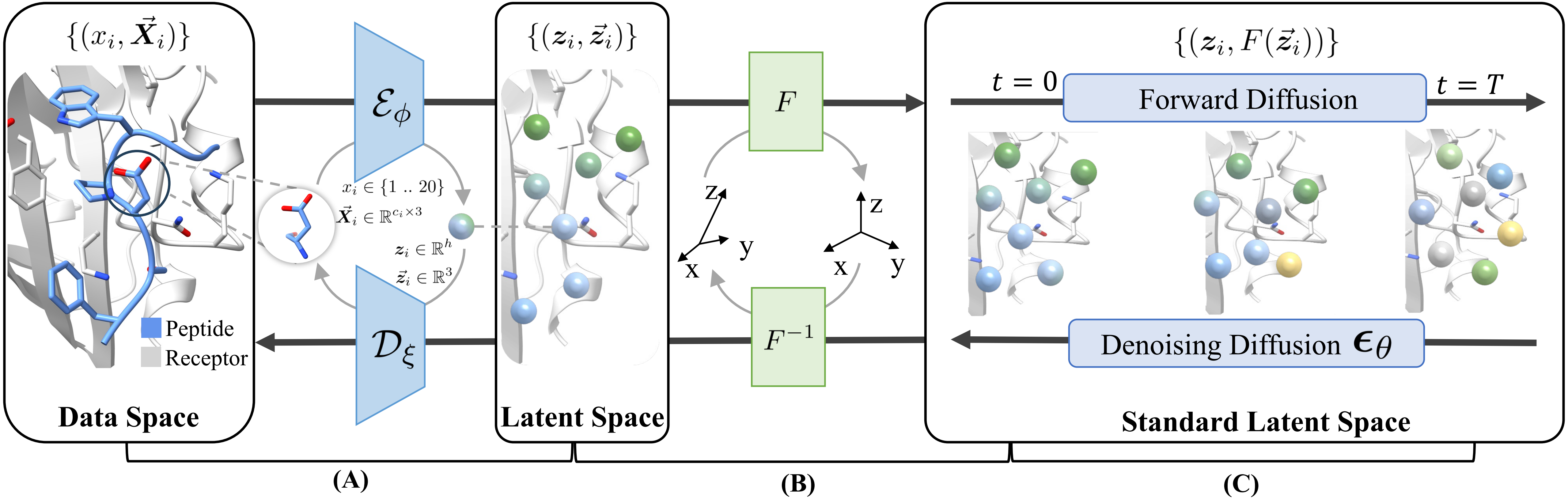
- Different amino acids have different number of atoms
- Denosing amino acid types result in abrupt changes in the number of atoms (i.e. data length), which is not compatible with current diffusion framework.

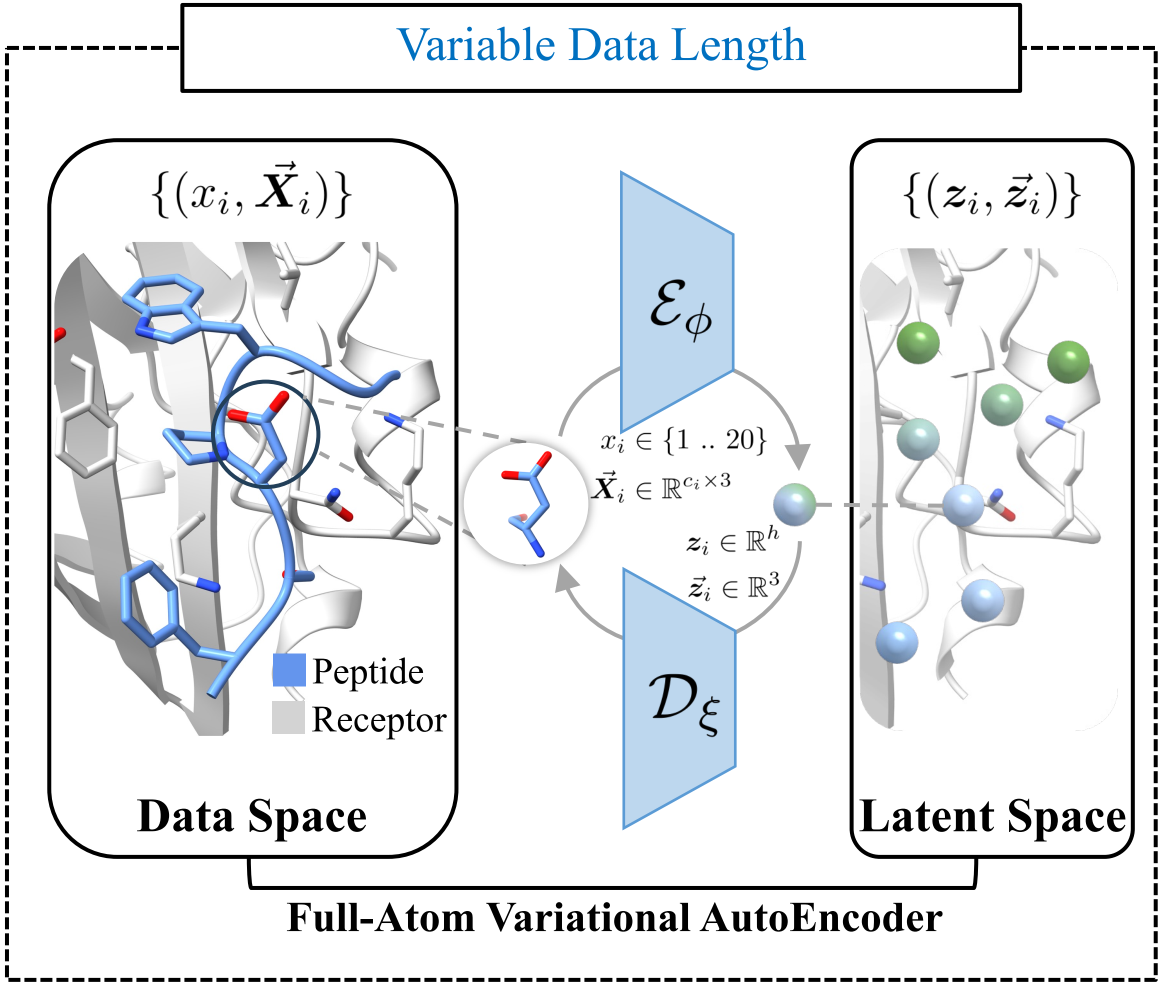
Diverse Pocket Geometry



- $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ with diverse expectation and covariance, leading to poor generalizability
- Current diffusion models tend to produce exploding coordinates for some binding sites in the test set.

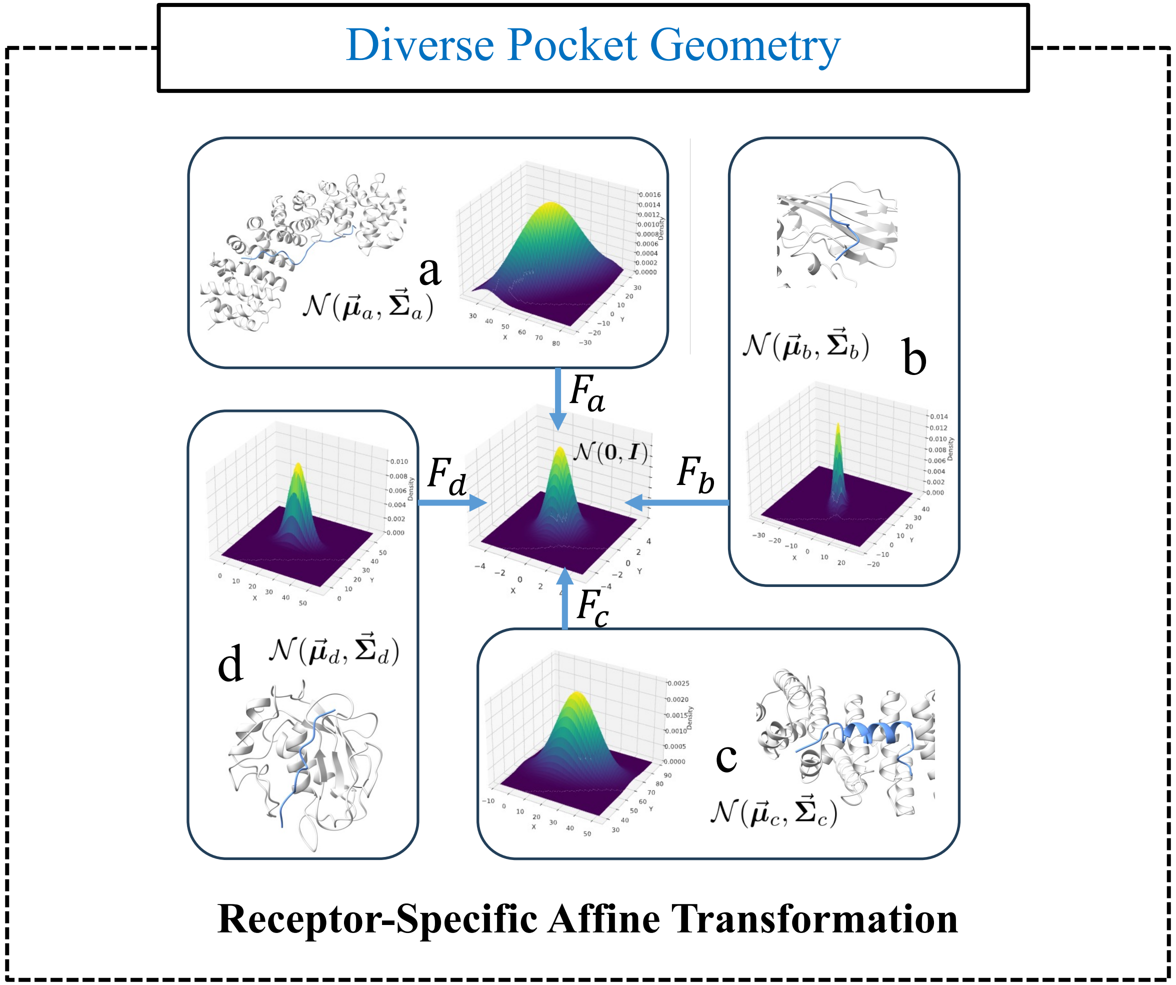
Peptide Design with Geometric LAtent Diffusion (PepGLAD)


$\{(x_i, \vec{\boldsymbol{X}}_i)\}$
Peptide
Receptor
Data Space
$\mathcal{E}_\phi$
$\mathcal{D}_\xi$
$x_i \in \{1 \,..\, 20\}$
$\vec{\boldsymbol{X}}_i \in \mathbb{R}^{c_i \times 3}$
$\boldsymbol{z}_i \in \mathbb{R}^h$
$\vec{\boldsymbol{z}}_i \in \mathbb{R}^3$
$\{(\boldsymbol{z}_i, \vec{\boldsymbol{z}}_i)\}$
Latent Space
F
F^{-1}
z
y
x
z
x
y
$\{(\boldsymbol{z}_i, F(\vec{\boldsymbol{z}}_i))\}$
$t = 0$
Forward Diffusion
$t = T$
Denoising Diffusion $\boldsymbol{\epsilon}_\theta$
Standard Latent Space
(A)
(B)
(C)


Variable Data Length

$\{(x_i, \vec{\boldsymbol{X}}_i)\}$
$\mathcal{E}_\phi$
$\{(\boldsymbol{z}_i, \vec{\boldsymbol{z}}_i)\}$
$x_i \in \{1 \; .. \; 20\}$
$\vec{\boldsymbol{X}}_i \in \mathbb{R}^{c_i \times 3}$
$\boldsymbol{z}_i \in \mathbb{R}^h$
$\vec{\boldsymbol{z}}_i \in \mathbb{R}^3$
$\mathcal{D}_\xi$
Peptide
Receptor
Data Space
Latent Space

Full-Atom Variational AutoEncoder


Diverse Pocket Geometry

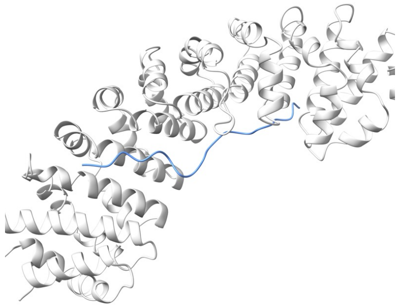
a
$\mathcal{N}(\vec{\boldsymbol{\mu}}_a, \vec{\boldsymbol{\Sigma}}_a)$
F_a
$\mathcal{N}(\vec{\boldsymbol{\mu}}_b, \vec{\boldsymbol{\Sigma}}_b)$
b
F_d
$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
F_b
F_c
d
$\mathcal{N}(\vec{\boldsymbol{\mu}}_d, \vec{\boldsymbol{\Sigma}}_d)$
c
$\mathcal{N}(\vec{\boldsymbol{\mu}}_c, \vec{\boldsymbol{\Sigma}}_c)$

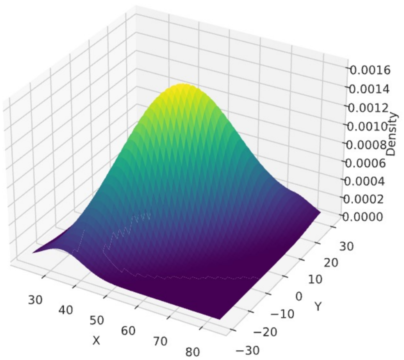
Receptor-Specific Affine Transformation

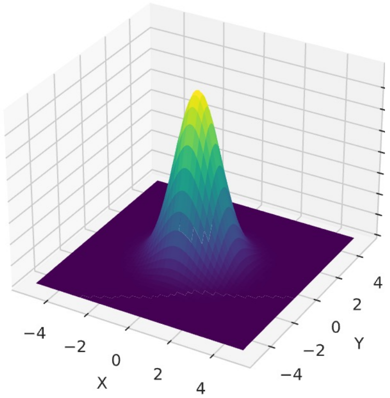
Receptor-Specific Affine Transformation

easily invertible normalization trick

$$\begin{tikzcd} \mathcal{N}(\vec{\mu}, \vec{\Sigma}) \arrow[r, bend left, "F(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{L}}^{-1}(\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}})"] & \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \arrow[l, bend left, "F^{-1}(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{L}}\vec{\boldsymbol{x}} + \vec{\boldsymbol{\mu}}"] \end{tikzcd}$$







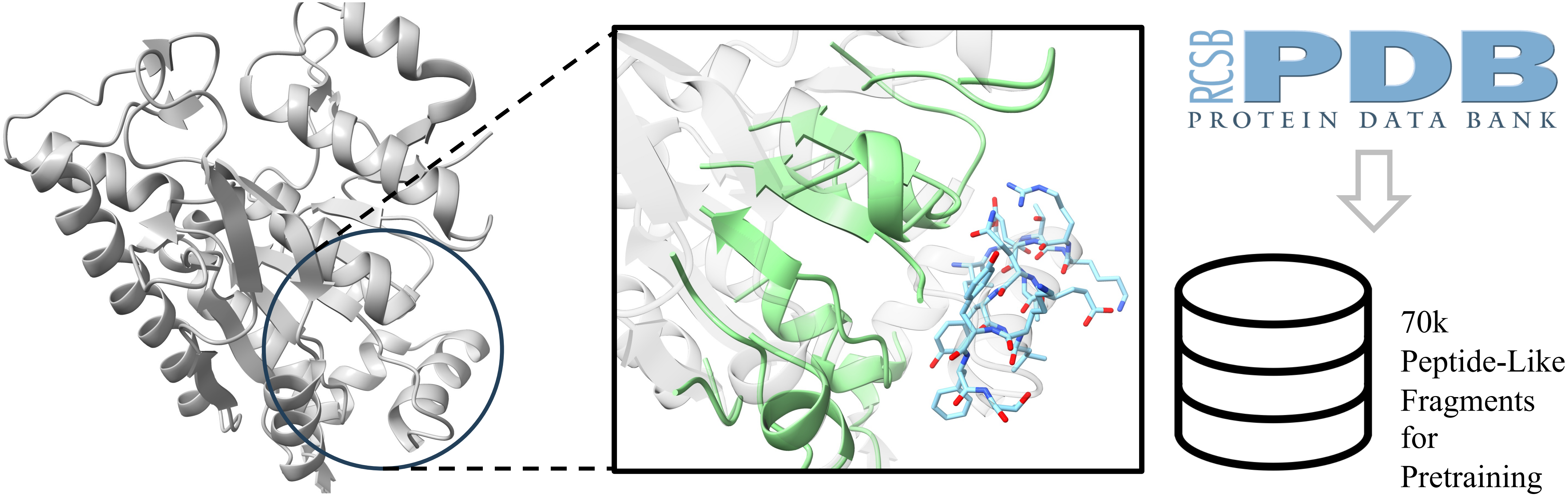
Cholesky Decomposition

$$\vec{\boldsymbol{\Sigma}} = \vec{\boldsymbol{L}}\vec{\boldsymbol{L}}^{\top}, \vec{\boldsymbol{L}} \in \mathbb{R}^{3\times 3}$$

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}\begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

Dataset: ProtFrag

70K peptide-like fragments within monomers for training the full-atom variational autoencoder



70k Peptide-Like Fragments for Pretraining

Dataset: PepBench

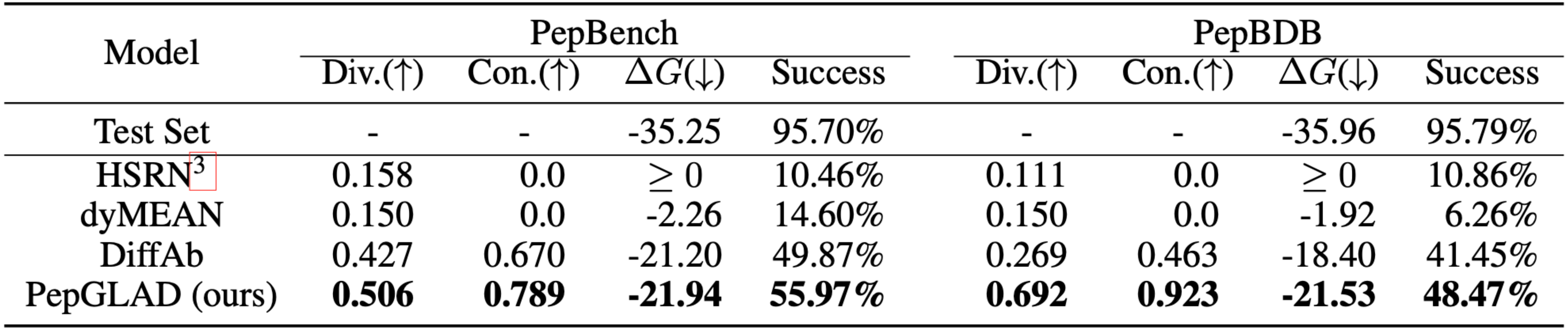
- **Training/Validation:** 6K cleaned non-redundant peptides (4-25 residues) from PDB
- **Test:** 93 complexes curated by experts from existing literature[1]
- **Split:** Cluster all complexes with target proteins sequence identity above 40%, and remove the complexes sharing the same clusters with those from the test set. Such split test the generalization ability of the generative models with respect to different target proteins.
- Url: https://doi.org/10.5281/zenodo.13358010

[1] Tsaban, Tomer, et al. "Harnessing protein folding neural networks for peptide–protein docking." *Nature communications* 13.1 (2022): 176.

Exp1: Sequence-Structure Co-Design

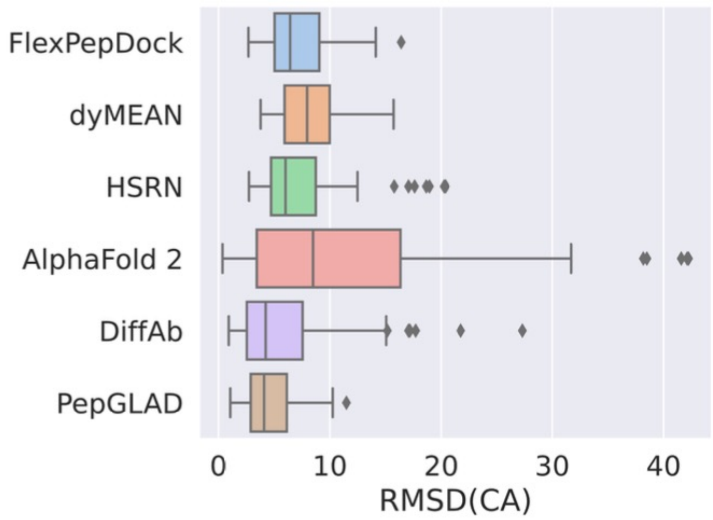
Metrics:

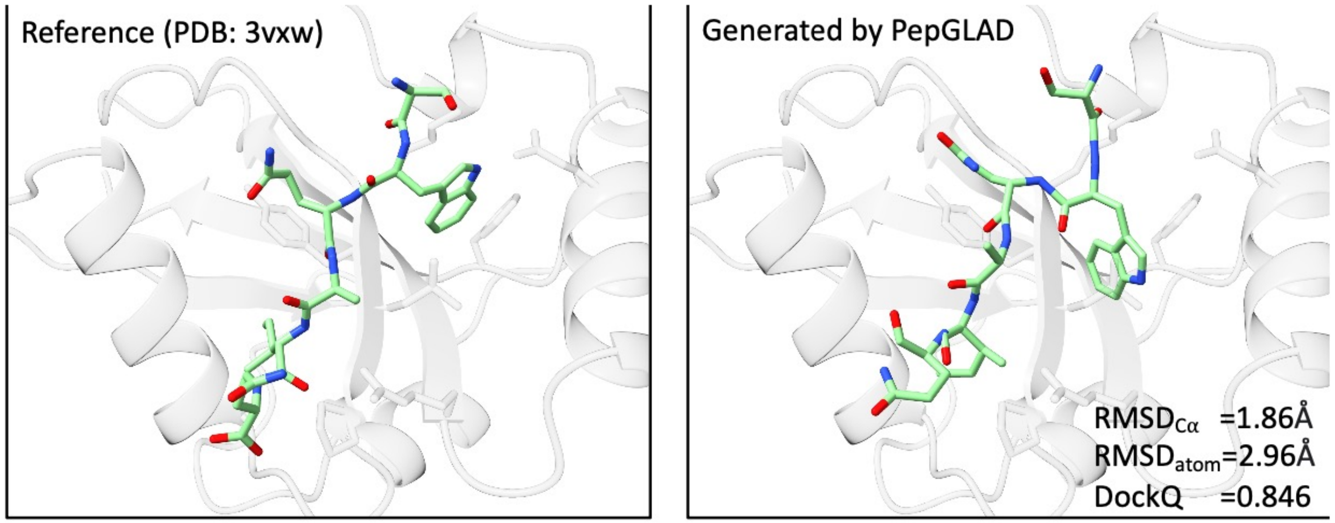
- **Diversity:** Ratio of unique clusters of sequence-structure clustering
- **Consistency:** Association between sequence clusters and structure clusters (similar sequences should lead to similar structures)
- **ΔG**: Binding energy measured by Rosetta
- **Success:** Ratio of ΔG < 0



Model	PepBench				PepBDB			
	Div.(↑)	Con.(↑)	ΔG(↓)	Success	Div.(↑)	Con.(↑)	ΔG(↓)	Success
Test Set	-	-	-35.25	95.70%	-	-	-35.96	95.79%
HSRN[3]	0.158	0.0	≥ 0	10.46%	0.111	0.0	≥ 0	10.86%
dyMEAN	0.150	0.0	-2.26	14.60%	0.150	0.0	-1.92	6.26%
DiffAb	0.427	0.670	-21.20	49.87%	0.269	0.463	-18.40	41.45%
PepGLAD (ours)	**0.506**	**0.789**	**-21.94**	**55.97%**	**0.692**	**0.923**	**-21.53**	**48.47%**

Exp2: Binding Conformation Generation




Reference (PDB: 3vxw)

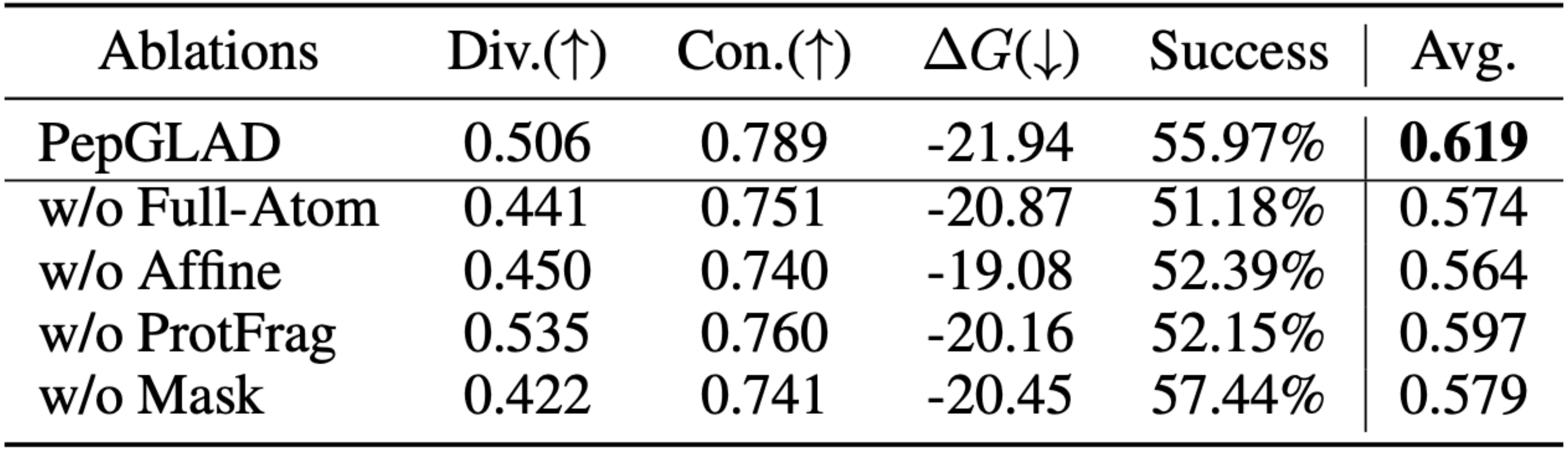
Generated by PepGLAD

Model	PepBench			PepBDB		
	RMSD$_{C_\alpha}$(↓)	RMSD$_{atom}$(↓)	DockQ(↑)	RMSD$_{C_\alpha}$(↓)	RMSD$_{atom}$(↓)	DockQ(↑)
FlexPepDock	6.43	7.52	0.393	-	-	-
AlphaFold 2	8.49	9.20	0.355	-	-	-
dyMEAN	7.96	8.35	0.374	17.64	17.56	0.142
HSRN	6.02	7.59	0.508	9.28	9.72	0.394
DiffAb	4.23	7.60	0.586	13.96	13.12	0.236
PepGLAD (ours)	**4.09**	**5.30**	**0.592**	**8.87**	**8.62**	**0.403**

Ablation Study

Significance:

Affine Transformation > Full-Atom Modeling > Masked Autoencoder > Protein Fragments Training



Ablations	Div.(↑)	Con.(↑)	$\Delta G(\downarrow)$	Success	Avg.
PepGLAD	0.506	0.789	-21.94	55.97%	**0.619**
w/o Full-Atom	0.441	0.751	-20.87	51.18%	0.574
w/o Affine	0.450	0.740	-19.08	52.39%	0.564
w/o ProtFrag	0.535	0.760	-20.16	52.15%	0.597
w/o Mask	0.422	0.741	-20.45	57.44%	0.579

Conclusion

- PepGLAD: full-atom model for peptide design given the binding site on the target protein
- We curate PepBench with carefully selected test complexes and split criterion to test the generalization ability across different target proteins
- We curate ProtFrag of 70K peptide-like fragments for data augmentation, which may facilitate future research on peptide design
- PepGLAD surpasses state-of-the-art models in terms of sequence-structure co-design and binding conformation generation

Thank you for your attention!

Paper Link

Code Link