Surface-based Molecular Design with Multi-modal Flow Matching

Authors: Fang Wu, Zhengyuan Zhou, Shuting Jin, Xiangxiang Zeng,  Jure Leskovec, Jinbo Xu | Authors Info & Claims

KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2
Pages 3192 - 3203 • https://doi.org/10.1145/3711896.3737139

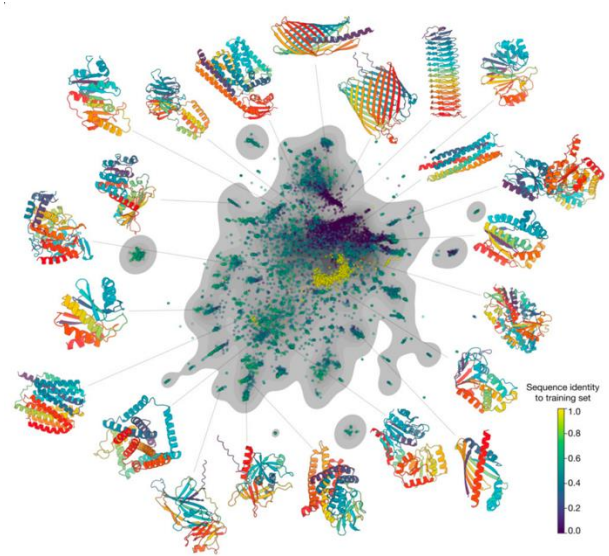
Speakers: Tian Zhu & Adrian Dobbelstein

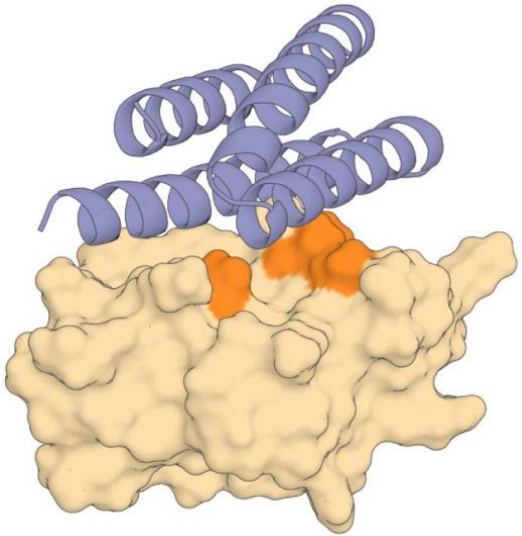
EE-262 Graph representations for biology and medicine

Supervisors: Bruno Correia & Michael Bronstein

Background: Protein

- As the central executioners of genetic information, **proteins govern complex cellular functions**, including molecular recognition, signal transduction, and immune defense.

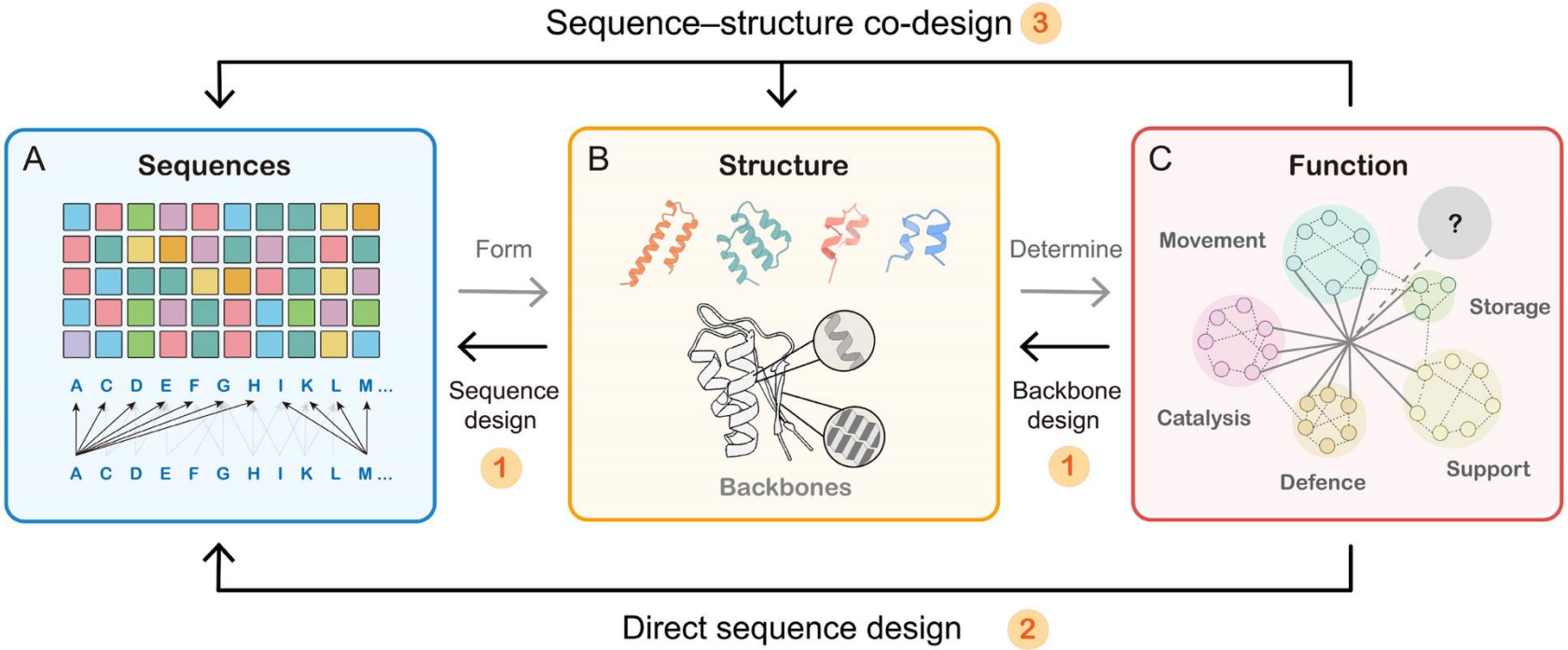




Zhang, G., Liu, C., Lu, J., Zhang, S. and Zhu, L., 2025.. Biology, 14(9), p.1268.

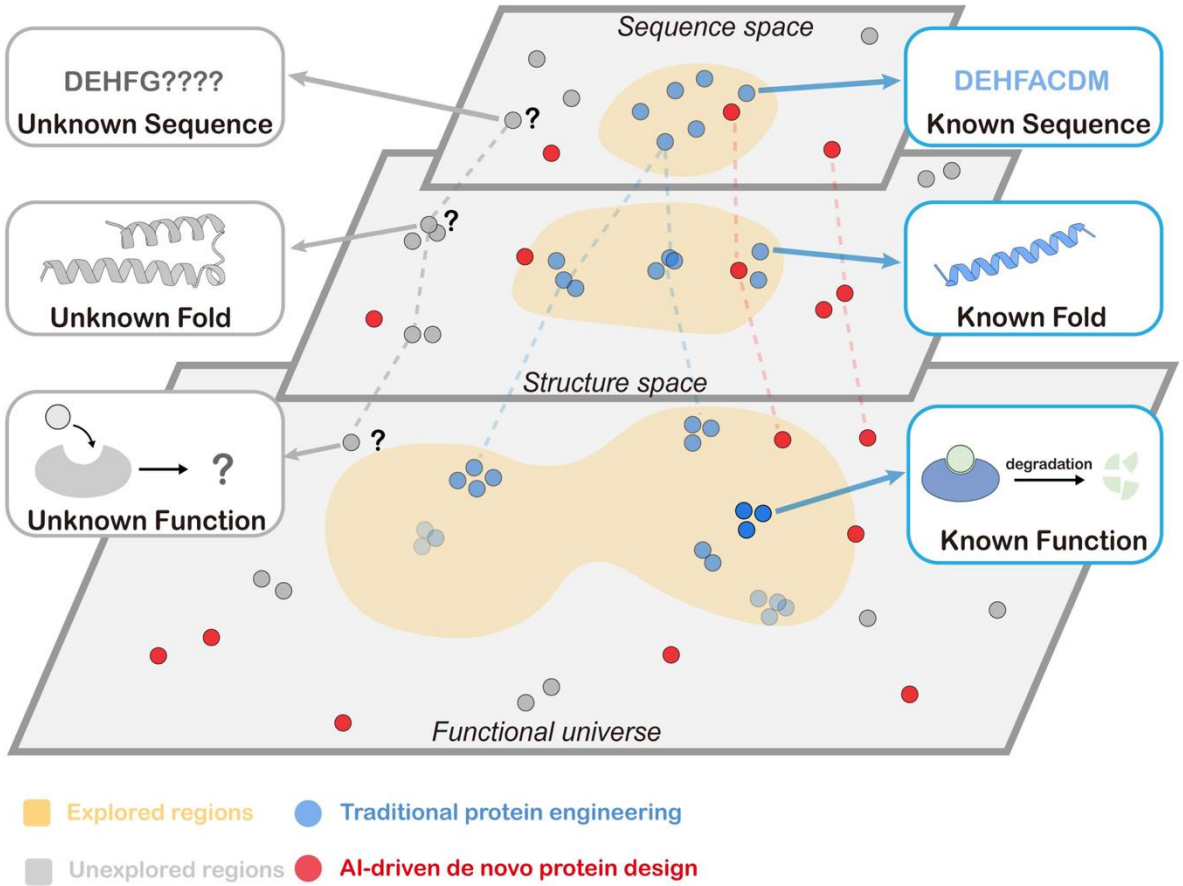
Background

Protein Anfinsen's dogma



Zhang, G., Liu, C., Lu, J., Zhang, S. and Zhu, L., 2025.. Biology, 14(9), p.1268.

Background: *de novo* Design

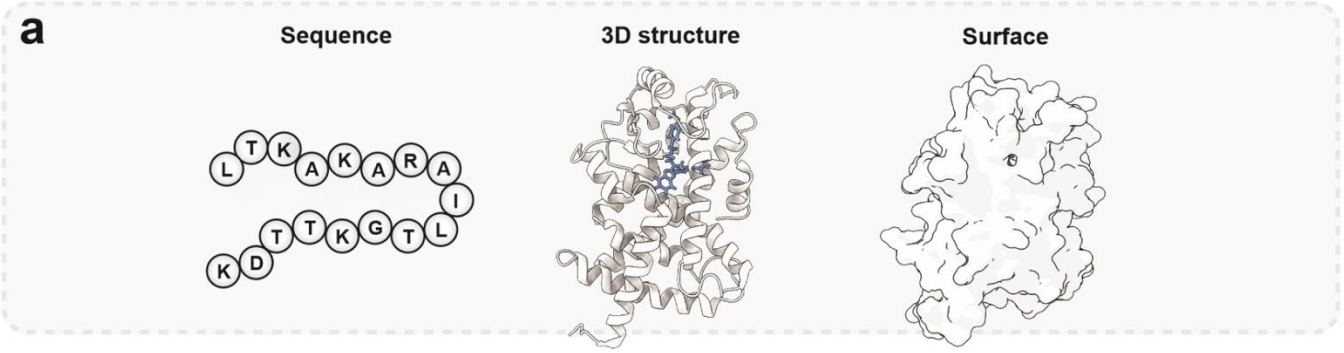


Zhang, G., Liu, C., Lu, J., Zhang, S. and Zhu, L., 2025.. Biology, 14(9), p.1268.

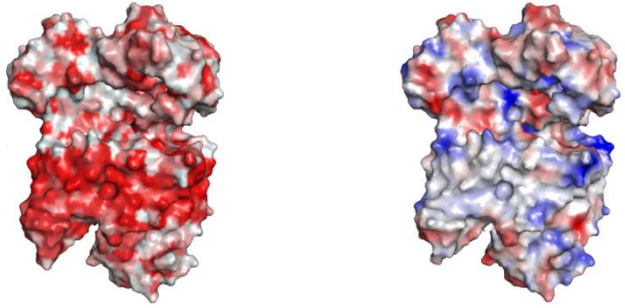
Adrian Dobbelstein & Tian Zhu

Background: Protein Surface

- Beyond primary sequence and 3D topology, the **molecular surface** encodes critical physicochemical properties that dictate binding behaviour.



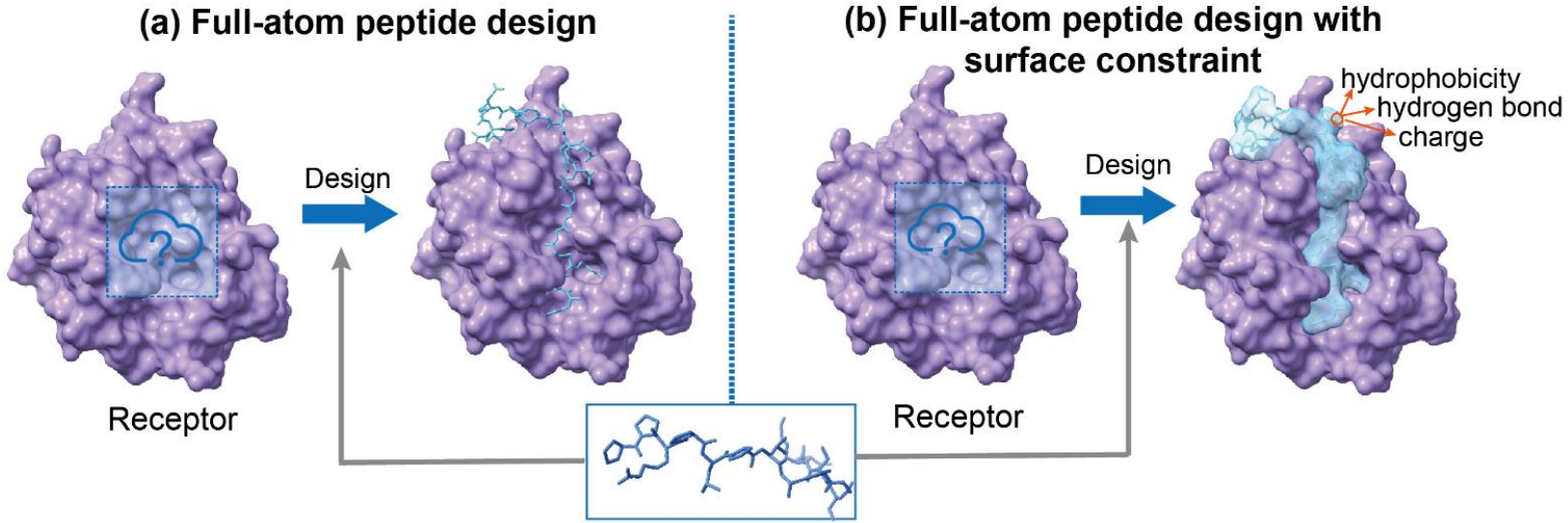
hydrophobicity **Electrostatic potential**



Özçelik, R., van Tilborg, D., et al., 2023. ChemBioChem, 24(13), p.e202200776.

Task: Peptide Design

- **Goal:** Generate a full-atom peptide binder that complements the surface properties of a given target protein receptor.
- **Key Innovation:** Unlike standard design (a), the approach **(b)** incorporates **surface representations**—hydrophobicity, hydrogen bonding, and charge—to ensure optimal binding.
- **Definition:** Peptide refers to a short oligomer of fewer than 20 amino acids.



Wu, F., Zhou, Z., et al., 2025, August. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining

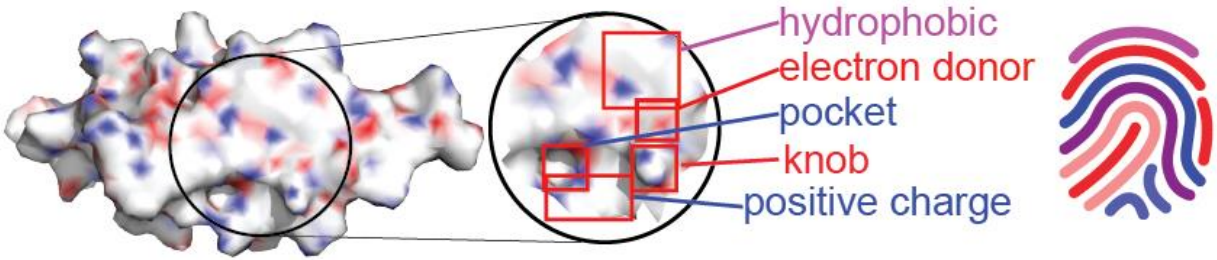
Task: Motivation

- Protein-protein Interactions (PPIs) are largely dictated by **how complementarily the surfaces of interacting proteins fit together**.
- The surface's electrostatic potential and hydrophobicity are key determinants of PPIs' strength and specificity.
- Surfaces geometries, such as protrusions, grooves, and clefts enable lock-and-key or induced-fit mechanisms essential for specific binding.

MaSIF

Protein molecular surface

Interaction fingerprint



Gainza, P., Sverrisson, F., et al. Nat Methods 17, 184–192 (2020)

Adrian Dobbelstein & Tian Zhu

Multimodal Space for Protein

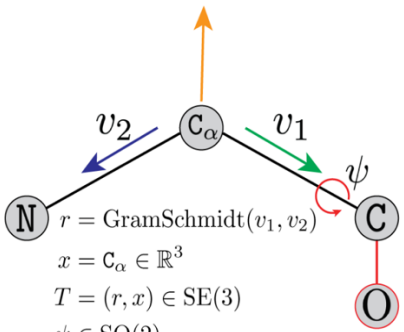
Sequence

A G V A S K S

a_i

Discrete words
(Discrete/Simplex Space)

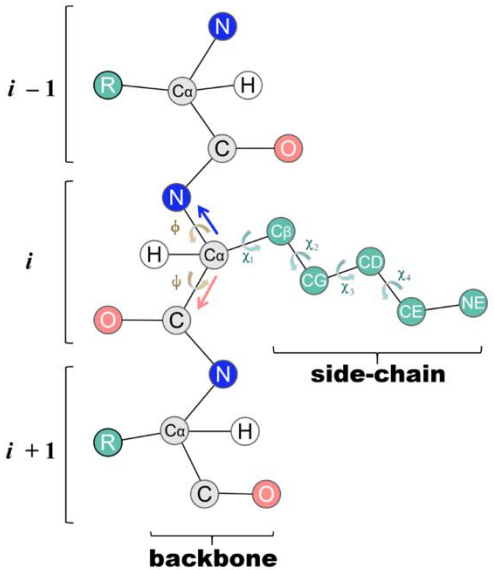
Backbone



$x_i + O_i$

Translation + Rotation
(SE(3) Space)

Side-chain



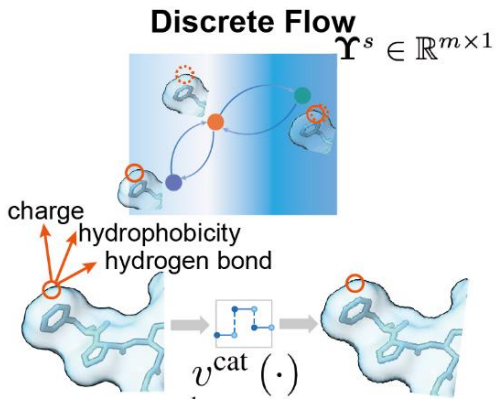
χ_i

Torsion angles
(T^4 Space)

Yim, J., Trippe, B.L., et al., 2023, July. In Proceedings of the 40th International Conference on Machine Learning (pp. 40001-40039).
Li, J., Cheng, C., et al., 2024, July. In Proceedings of the 41st International Conference on Machine Learning (pp. 27615-27640).

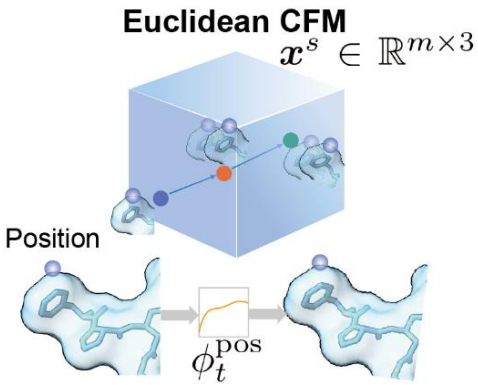
Multimodal Space for Surface

Physicochemical Properties



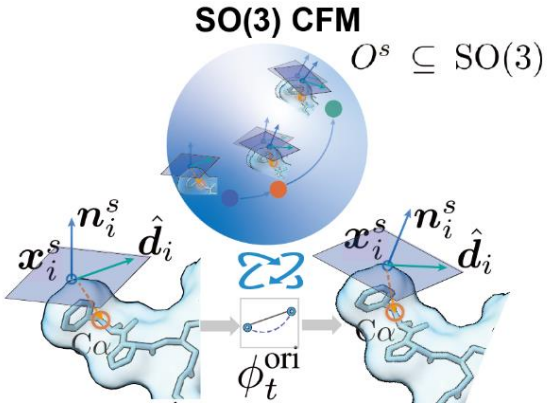
γ_i^s, τ_i^s

Points



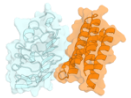
x_i^s

Normals



n_i^s

Problem Formulation



Goal: This work designs peptides based on target proteins, so we aim to model the **conditional joint distribution**

$$p\left(C^{\text{pep}} \mid C^{\text{rec}}\right)$$

where

$$C^{\text{rec}} = \{(a_i, O_i, \boldsymbol{x}_i, \boldsymbol{\chi}_i)\}_{i=1}^{n_{\text{rec}}}$$

$$C^{\text{pep}} = \left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}} \cup \left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \boldsymbol{\Upsilon}_i^s\right)\right\}_{i=1}^{m} \quad m \gg n_{\text{pep}}$$

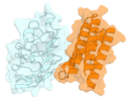
Structure representation

$a_i \in \{1, \ldots, 20\}$ denotes the residue type
$x_i \in \mathbb{R}^3$ and $O_i \in \mathrm{SO}(3)$ represents the backbone frame
$\chi_i \in [0, 2\pi)^4$ denotes the sidechain torsion angles

Surface representation

x_i^s is the coordinate of the surface point cloud
n_i^s is the unit normal vector of the surface point cloud
τ_i^s and γ_i^s capture its continuous and categorical physicochemical properties

Problem Formulation



Goal: This work designs peptides based on target proteins, so we aim to model the **conditional joint distribution**

$$p\left(C^{\text{pep}} \mid C^{\text{rec}}\right)$$

where

$$C^{\text{rec}} = \{(a_i, O_i, \boldsymbol{x}_i, \boldsymbol{\chi}_i)\}_{i=1}^{n_{\text{rec}}}$$

$$C^{\text{pep}} = \left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}} \cup \left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^{m} \quad m \gg n_{\text{pep}}$$

Then we can get

$$p\left(C^{\text{pep}} \mid C^{\text{rec}}\right) \propto p\left(\left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}} \middle| C^{\text{rec}}\right) p\left(\left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^{m} \middle| C^{\text{rec}}\right),$$

Introduction: Flow Matching

Goal: Find a velocity field $u_\theta(x_t, t)$ that, when followed/integrated, transforms p_0 into p_{data}.
The *particularity of CFM* is how the velocity field is learned, as we will detail below.

For unknown target data distribution p_{data}, it is hard to choose a priori a probability path or velocity field. CFM core idea is to choose a conditioning variable z and a conditional probability path $p(x|t,z)$ such that

- the induced global probability path $p(x|t)$ transforms p_0 into p_{data}
- the associated velocity field $u_\theta(x_t, t|z)$ has an analytic form.

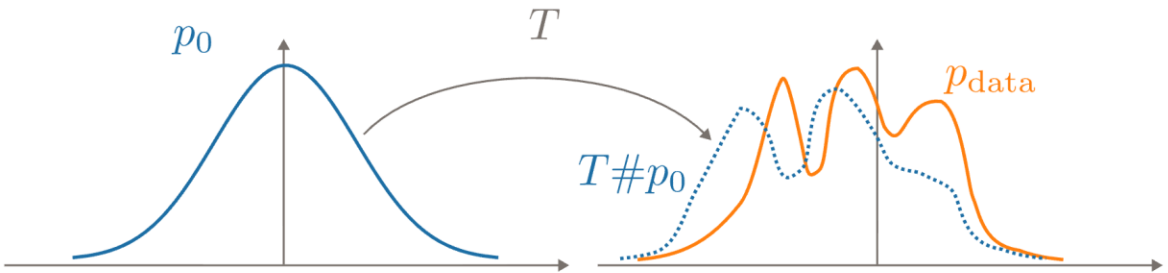


Figure 1. Modern generative modelling principle: trying to find a map $\boldsymbol{T}$ that sends the base distribution $\boldsymbol{p_0}$ as close as possible to the data distribution $\boldsymbol{p_{\text{data}}}$.

Campbell, A., Yim, J., et al. In International Conference on Machine Learning (pp. 5453-5512). PMLR.

Introduction: Flow Matching

- A first choice is to condition on the base points and the target points, i.e., z is a random variable defined as:

$$z \overset{\text{choice}}{=} (x_0, x_1) \sim p_0 \times p_{\text{data}} \,.$$

- Among all the possible probability paths, one can choose to use very concentrated Gaussian distributions and simply interpolate between x_0 and x_1 in straight line: for some fixed standard deviation σ, it writes as

$$p\big(x|t, z = (x_0, x_1)\big) \overset{\text{choice}}{=} \mathcal{N}\big((1-t)\cdot x_0 + t \cdot x_1, \sigma^2 \mathrm{Id}\big) \,.$$

- To recover the correct distributions p_0 at t=0 (resp. p_{data} at t=1), one must enforce σ=0, finally leading to, where δ denotes the Dirac delta distribution.

$$p\big(x|t, z = (x_0, x_1)\big) \overset{\text{choice}}{=} \delta_{(1-t)\cdot x_0 + t \cdot x_1}(x) \,,$$

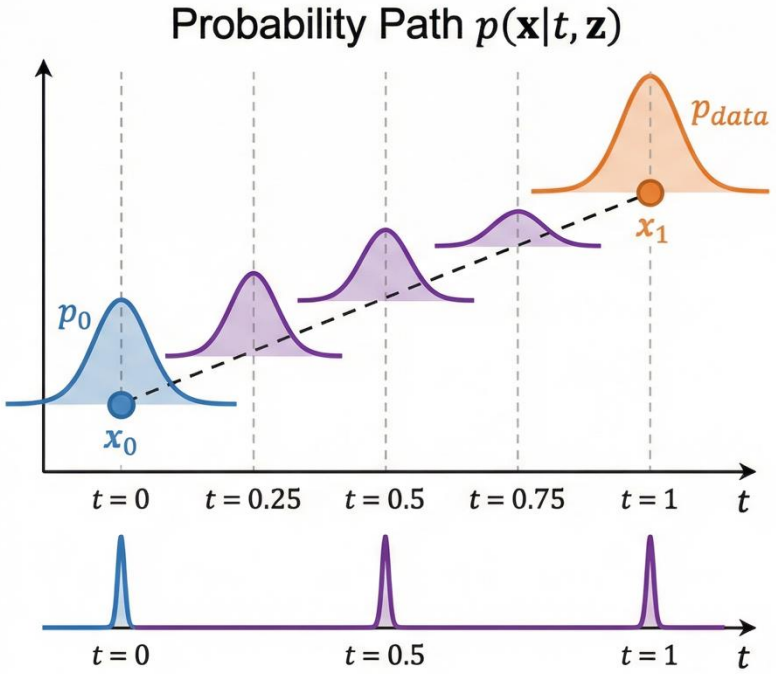
- Conditional vector field:

$$u^{\text{cond}}(x, t, z = (x_0, x_1)) = x_1 - x_0$$

Campbell, A., Yim, J., et al. In International Conference on Machine Learning (pp. 5453-5512). PMLR.

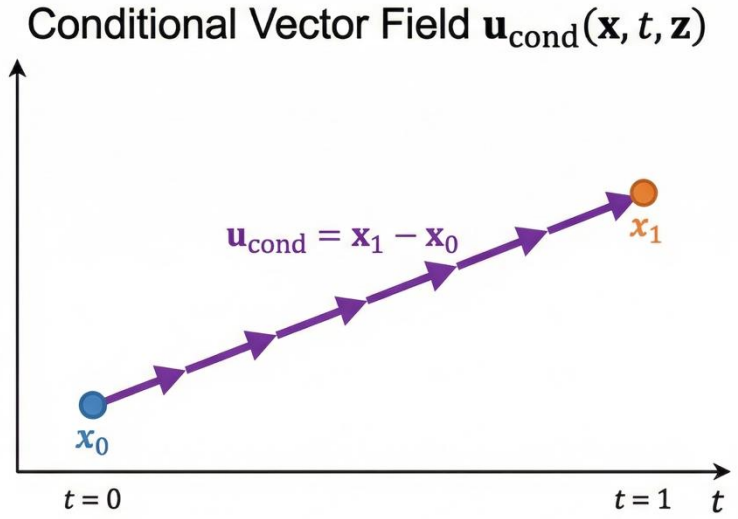
Adrian Dobbelstein & Tian Zhu

Introduction: Flow Matching



Path: $\mathbf{p}(\mathbf{x}|t,z) = \mathcal{N}((1-t)\mathbf{x}_0 + tx_1, \sigma^2\mathbf{Id})$

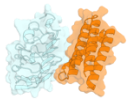
$\rightarrow \delta((1-t)\mathbf{x}_0 + tx_1)(\mathbf{x})$ as $\sigma \rightarrow 0$



Constant Velocity:

$\mathbf{u}_{\text{cond}}(\mathbf{x},t,z) = \mathbf{x}_1 - \mathbf{x}_0$

Campbell, A., Yim, J., et al. In International Conference on Machine Learning (pp. 5453-5512). PMLR.



Flow Matching in SurfFlow

Position

Interpolation and vector fields:

$$\phi_t^{\text{pos}}\left(x_0^s, x_1^s\right) = t x_1^s + (1-t) x_0^s,$$

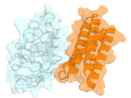
$$u_t^{\text{pos}}\left(x_t^s \mid x_1^s, x_0^s\right) = x_1^s - x_0^s = \frac{x_1^s - x_t^s}{1-t}.$$

Training objectives:

$$\mathcal{L}_{\text{pos}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1), p(x_1^s), p(x_0^s), p(x_t^s \mid x_0^s, x_1^s)} \left\| v^{\text{pos}}\left(x_t^s, t, C^{\text{rec}}\right) - \left(x_1^s - x_0^s\right) \right\|_2^2,$$

Sampling:

$$x_{t+\frac{1}{N}}^s = x_t^s + \frac{1}{N} v^{\text{pos}}\left(x_t^s, t, C^{\text{rec}}\right).$$



Flow Matching in SurfFlow

Normal Vector Orientations

Interpolation and vector fields:

$$\phi_t^{\text{ori}}\left(O_0^s, O_1^s\right) = \exp_{O_0^s}\left(t \log_{O_0^s}\left(O_1^s\right)\right),$$

$$u_t^{\text{ori}}\left(O_t^s \mid O_0^s, O_1^s\right) = \frac{\log_{O_t^s}\left(O_1^s\right)}{1-t},$$

Training objectives:

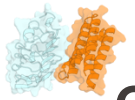
$$\mathcal{L}_{\text{ori}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1), p(O_1^s), p(O_0^s), p(O_t^s \mid O_0^s, O_1^s)}$$

$$\left\| v^{\text{ori}}\left(O_t^s, t, C^{\text{rec}}\right) - \frac{\log_{O_t^s}\left(O_1^s\right)}{1-t} \right\|_{\text{SO}(3)}^2,$$

Sampling:

$$O_{t+\frac{1}{N}} = \exp_{O_t^s}\left(\frac{1}{N} v^{\text{ori}}\left(O_t^s, t, C^{\text{rec}}\right)\right).$$

Introduction: Discrete Flow Matching (CTMC)



Countinuous Time Markov Chain (CTMC)

- A sequence trajectory x_t over time t ∈ [0, 1] that follows a CTMC alternates between **resting in its current state and periodically jumping to another randomly chosen state**.
- The frequency and destination of the jumps are **determined by the rate matrix** $\mathrm{R_t}$ with the constraint its off-diagonal elements are non-negative.
- We can write the transition probability as

$$p_{t+\mathrm{d}t|t}(j|x_t) = \begin{cases} R_t(x_t, j)\mathrm{d}t & \text{for } j \neq x_t \\ 1 + R_t(x_t, x_t)\mathrm{d}t & \text{for } j = x_t \end{cases} \quad (1)$$

$$= \delta\{x_t, j\} + R_t(x_t, j)\mathrm{d}t \quad (2)$$

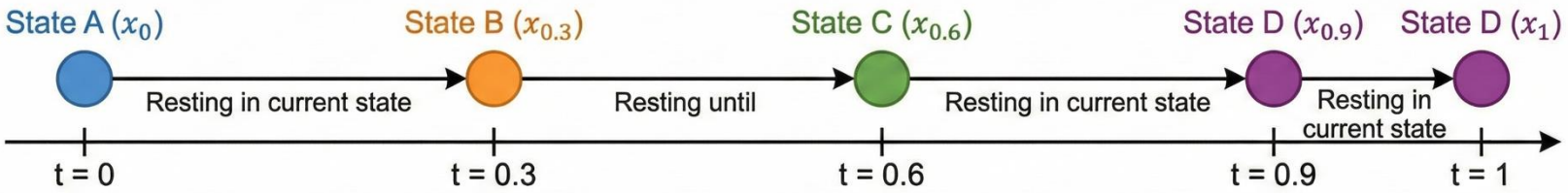
$$x_{t+\Delta t} \sim \mathrm{Cat}(\delta\{x_t, x_{t+\Delta t}\} + R_t(x_t, x_{t+\Delta t})\Delta t), \quad (3)$$

Campbell, A., Yim, J., et al. In International Conference on Machine Learning (pp. 5453-5512). PMLR.

Adrian Dobbelstein & Tian Zhu

Introduction: Discrete Flow Matching (CTMC)

CTMC Trajectory over Time $t \in [0, 1]$



The sequence trajectory x_t alternates between resting and periodically jumping to another randomly chosen state.

Determining Jumps with Rate Matrix R_t

Rate Matrix R_t (at time t)

	State A	State B	State C	State D
State A	$-r_{AA}$	$r_{AB} > 0$	0	0
State B	0	$-r_{BB}$	$r_{BC} > 0$	0
State C	0	0	$-r_{CC}$	0
State D	0	0	0	$-r_{DD}$

Determines Frequency and Destination

Jumps are determined by the rate matrix R_t. The transition probability over a small time dt is $p_{t+dt|t}(j|x_t) = \delta\{x_t, j\} + R_t(x_t, j)dt$ (for $j \neq x_t$, prob is $R_t(x_t, j)dt$).

Campbell, A., Yim, J., et al. In International Conference on Machine Learning (pp. 5453-5512). PMLR.

Discrete Flow Matching in SurfFlow

Interpolation Stage:

$$p_{t|1}^{\text{mask}}\left(\Upsilon_t^s \mid \Upsilon_1^s\right) = \text{Cat}(t\delta\{\Upsilon_1^s, \Upsilon_t^s\} + (1-t)\delta\{\text{M}, \Upsilon_t^s\}).$$

Rate Matrix:

$$R_t^*\left(\Upsilon_t^s, j \mid \Upsilon_1^s\right) := \frac{\text{ReLU}\left(\partial_t p_{t|1}\left(j \mid \Upsilon_1^s\right) - \partial_t p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)\right)}{S \cdot p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)},$$

Training Objective:

$$\mathcal{L}_{\text{cat}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1), p(\Upsilon_1^s), p_{t|1}(\Upsilon_t^s|\Upsilon_1^s)}\left[\log v^{\text{cat}}\left(\Upsilon_t^s, t, C^{\text{rec}}\right)\right].$$

Wu, F., Zhou, Z., et al., 2025, August. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Overall Training Objective

$$\mathcal{L}_{\text{CFM}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{cat}}\mathcal{L}_{\text{cat}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}}$$

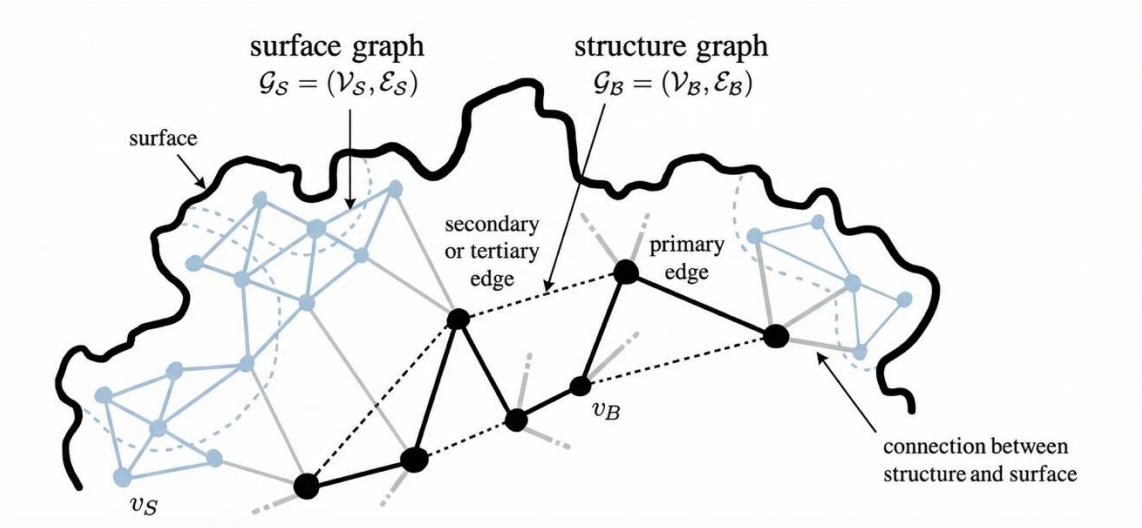
λ_*: Hyperparameters that balance the contribution of different objective terms.
$\mathcal{L}_{\text{pos}}$: Supervises the spatial coordinates of the generated surface point cloud.
$\mathcal{L}_{\text{ori}}$: Aligns the generated surface normals with the target orientation.
$\mathcal{L}_{\text{cat}}$: Enforces consistency in discrete surface properties.
$\mathcal{L}_{\text{con}}$: Enforces consistency in continuous surface properties.
$\mathcal{L}_{\text{str}}$: Flow Matching objective modeling the joint distribution of residue positions (a), orientations (O), amino acid types (x), and torsion angles (χ).

Equivariant Surface Geometric Network

Graph Construction: we can build a dynamic heterogeneous surface graph

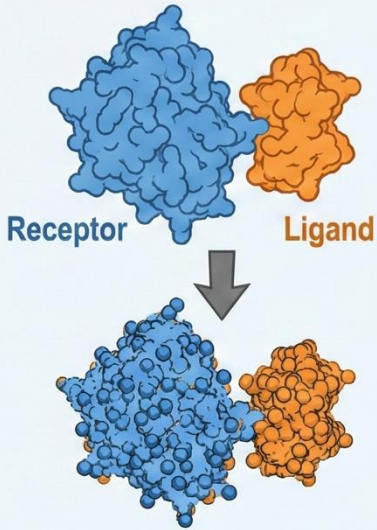
$$\mathcal{G} = (\mathcal{G}_{\text{rec}}, \mathcal{G}_{\text{pep}})$$

at different timestep t, where the intra and inter-point connectivities $(\mathcal{E}_{\text{rec}}, \mathcal{E}_{\text{pep}}, \mathcal{E}_{\text{inter}})$ are determined based on a spatial distance threshold cutoff r .



Graph Construction

1. Input & Discretization



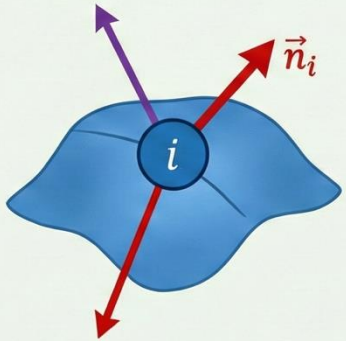
Input: 3D Protein Complex Structure

Process: Generate discrete surface nodes ($\mathcal{V}$) based on residue positions or surface sampling (e.g., Connolly surface).

2. Node Feature Initialization

Scalar Features ($h_i^{(0)}$)

Invariant properties (e.g., Amino acid type, Charge, Hydrophobicity). Cannot rotate.



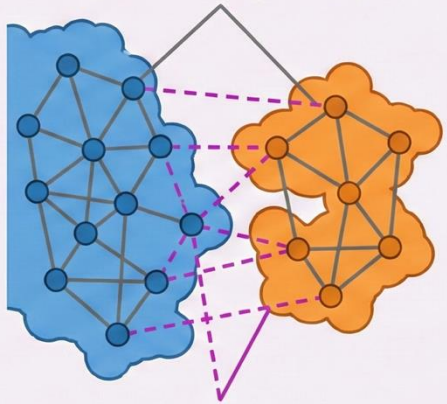
Vector Features ($x_i^{(0)}, \vec{n}_i$)

Geometric properties: 3D Coordinate (x_i) & Surface Normal vector ($\vec{n}_i$). Rotates with structure.

3. Dual-Graph Connectivity

Intra-graph Edges ($\mathcal{E}_{intra}$)

Connect nodes within SAME protein based on local distance ($d < R_{intra}$). Captures local geometry.

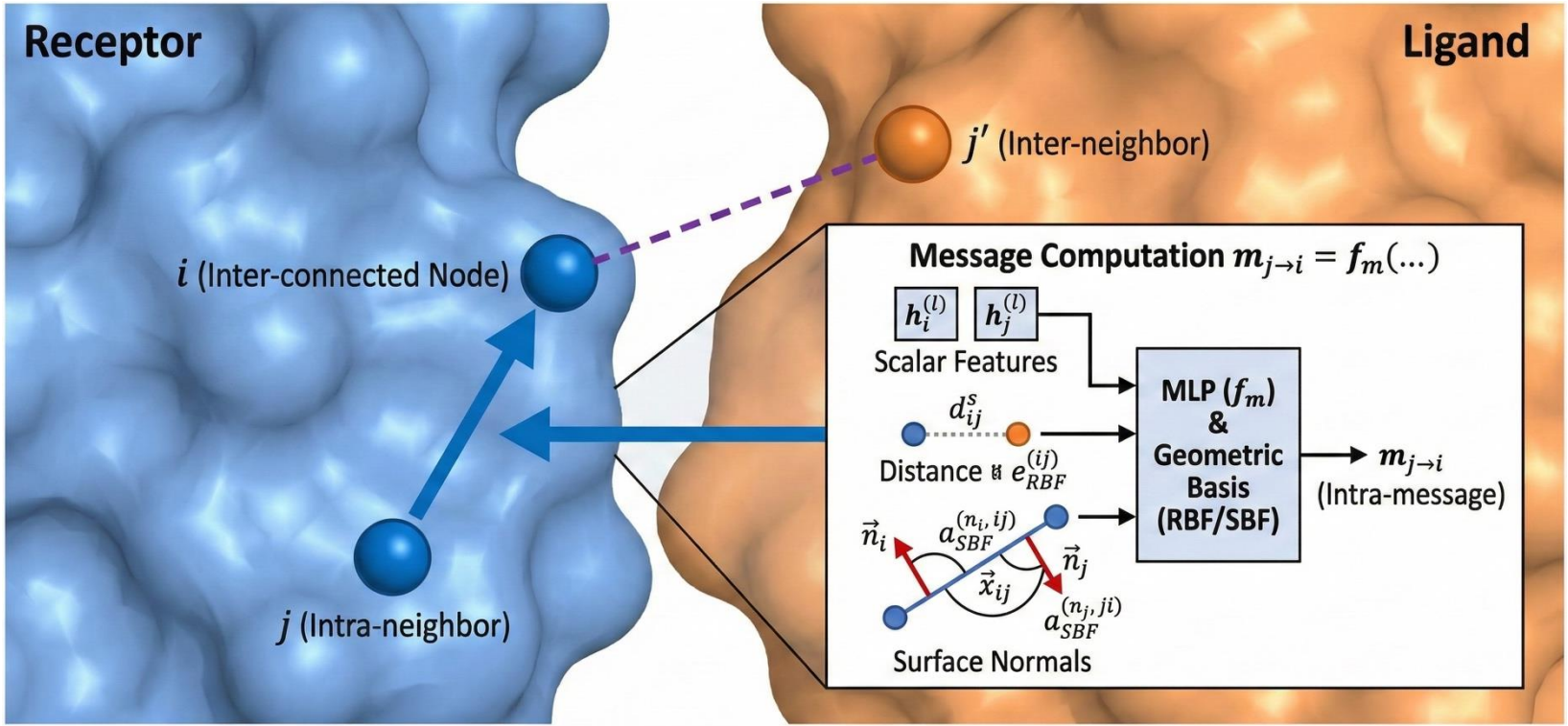


Inter-graph Edges ($\mathcal{E}_{inter}$)

Connect nodes across INTERFACE based on proximity ($d < R_{inter}$). Captures interaction.

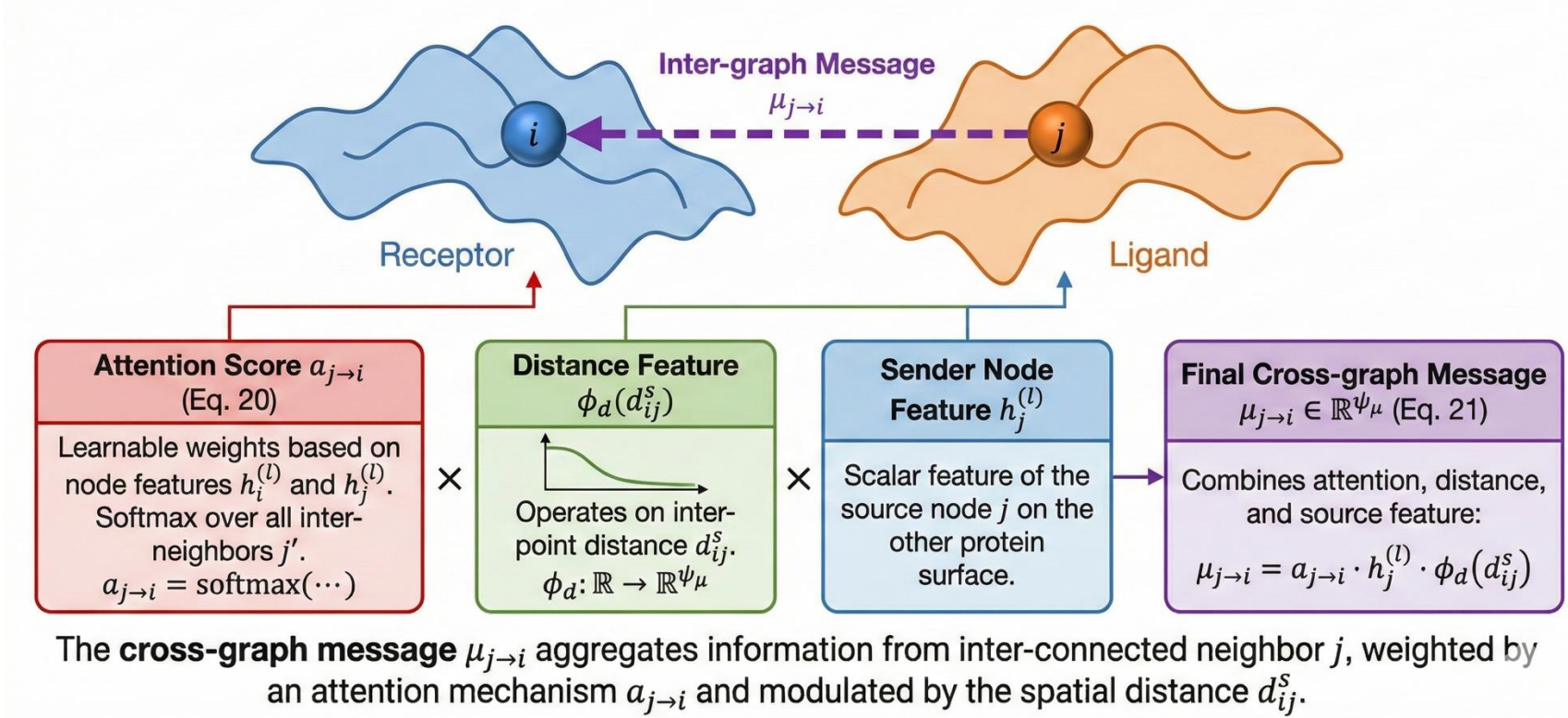
Wu, F., Zhou, Z., et al., 2025, August. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Intra-graph Message Passing



Node i, located at the interface, receives a message from its internal neighbor j. This message $m_{j\to i}$ encodes local geometry (distance d_{ij}, angles with normals $\vec{n}_i, \vec{n}_j$) processed by an MLP, as defined in Eq. (18).

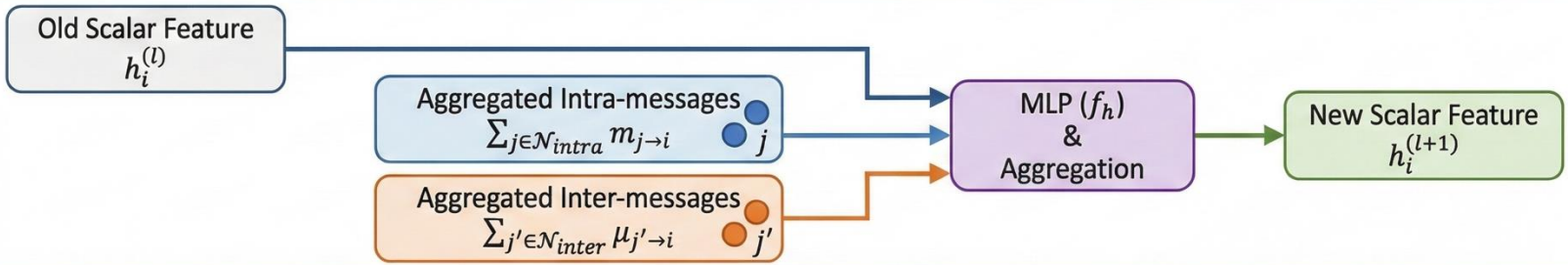
Cross-graph Message Passing



Aggregation for intra- and inter-messages

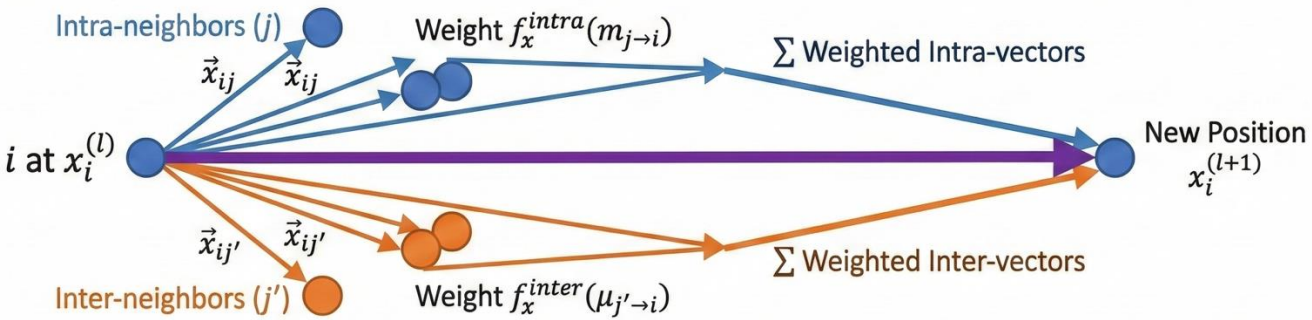
Scalar Feature Update ($h_i^{(l+1)}$) - Eq. (22)

$$h_i^{(l+1)}(h_{h\to i}) + ai(\mathcal{N}_{intra}(\mu_{i'j'\to i})m_{j'\to i}) \quad (22)$$

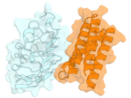


Coordinate Update ($x_i^{(l+1)}$) - Eqs. (23, 24)

$$i - \vec{x}_i(x_i^{(l+1)}) \quad (23) \qquad \vec{x}_i - f_{intl}(x_i^{(l+1)}) \quad (24)$$



Coordinates are updated by adding weighted relative vectors from both intra- and inter-neighbors, ensuring E(3)-equivariance.



Experiments

Training set: 8,365 non-redundant protein-peptide complexes across 292 clusters.
Test set: 10 clusters and 158 complexes.

Evaluation Metrics

Geometry:

- *Amino Acid Recovery rate (AAR):* Quantifies sequence identity between the generated peptides and the native reference.
- *Root-mean-square deviation (RMSD) of Cα atoms:* Assesses structural deviation of the peptide backbone after alignment with the native complex.
- *Secondary-structure similarity ratio (SSR):* Measures the consistency of secondary structure elements between generated and native conformations.
- *Binding Site Ratio (BSR):* Evaluates the spatial overlap of interfacial residues between the generated and native ligands.

Energy (Using Rosetta):

- *Affinity:* The percentage of generated peptides exhibiting lower binding energy than the native ligand.
- *Stability:* he proportion of complexes achieving a lower total energy score compared to the native state.

Design:

- *Designability:* The fraction of sequences that fold into their designed structures Cα RMSD < 2 Å , validated via ESMFold.
- *Diversity:* The average pairwise structural dissimilarity (calculated as TM-score) within the generated samples.

Experiments

Table 1: Evaluation of different methods in the sequence-structure co-design task and ablation studies on key components of SurfFlow. The best and suboptimal results are labeled boldly and underlined.

	Geometry				Energy		Design	
	AAR % ↑	RMSD Å ↓	SSR % ↑	BSR % ↑	Stb. % ↑	Aff. % ↑	Des. % ↑	Div. ↑
RFdiffusion [61]	40.14	4.17	63.86	26.71	**26.82**	16.53	**78.52**	0.38
ProteinGen [37]	45.82	4.35	29.15	24.62	23.48	13.47	71.82	0.54
Diffusion [39]	47.04	3.28	74.89	49.83	15.34	17.13	48.54	0.57
PepGLAD [26]	50.43	3.83	80.24	19.34	20.39	10.47	75.07	0.32
PPIFlow [34]	48.35	3.59	68.13	25.94	15.77	12.08	46.53	0.51
PepFlow [33]	51.25	2.07	83.46	86.89	18.15	21.37	65.22	0.42
SurfFlow (w/o ESGN)	52.59	2.05	83.77	86.91	19.42	19.82	68.41	0.60
SurfFlow (w/o Position)	53.26	1.99	84.79	87.15	21.30	22.38	72.09	0.60
SurfFlow (w/o Orientation)	53.04	2.00	84.60	87.04	20.79	22.46	72.36	0.60
SurfFlow (w/o Biophysical Prop.)	52.31	2.03	83.96	86.98	19.55	20.25	70.83	0.58
SurfFlow	**54.07**	**1.96**	**85.11**	**87.38**	22.46	**22.51**	73.60	**0.61**

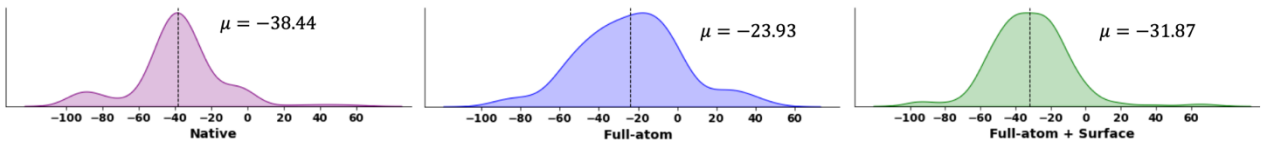
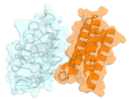


Figure 3: Binding energy distributions of designed and native peptides, where the lower is better.

Conclusion: SurfFlow can achieve better performance than other peptide design models



Takeaway

Pros.:

- This work presents SurfFlow, a novel model that produces all protein modalities – sequence, structure, and surface – concurrently.
- Empirical results prove the reasonability and promise of considering molecular surfaces for protein discovery

Cons.:

- No experiments to ablate all peptide surface component in its entirety and fair comparision to other seq-structure codesign methods.
- The surface of a peptide is determined entirely by its amino acid sequence. Therefore, it is no sense to jointly generating both the amino acid sequence and surface shape

Thank you!

Any questions?