CSC2457 3D & Geometric Deep Learning

Fast end-to-end learning on protein surfaces

Freyr Sverrisson, Jean Feydy, Bruno E. Correia, Michael M. Bronstein

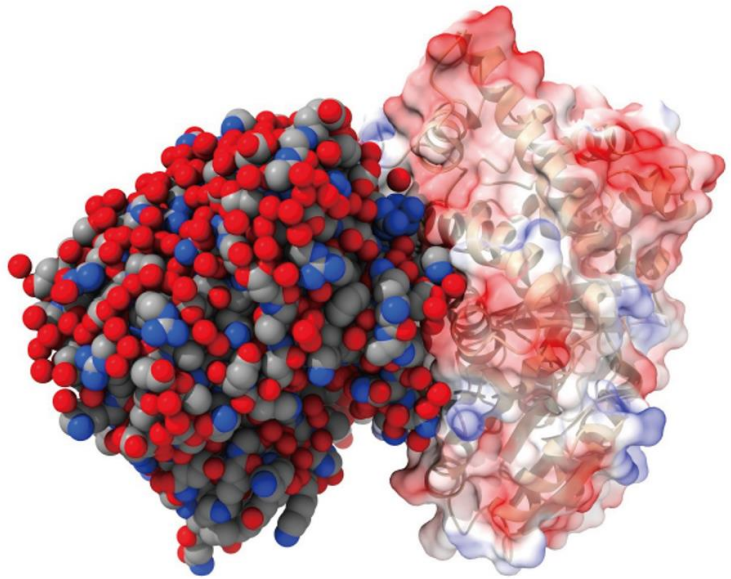
Date: March 30, 2021

Presenter: Youheng Ge

Instructor: Animesh Garg

UNIVERSITY OF TORONTO

Motivation and Main Problem



Structural Biology

- Analysis interacting surfaces
- Identify binding site
- Help to develop new drug therapies

Contributions

Prior work: Molecular Surface Interaction Fingerprinting (MaSIF)

- Reliance on precomputed meshes and handcrafted features
- Significant computational time and memory requirements

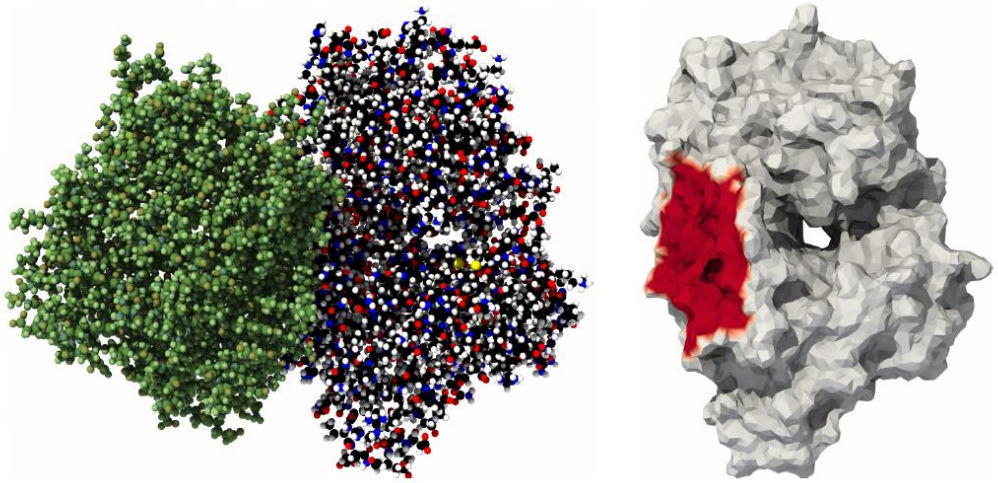
This work: Differentiable Molecular Surface Interaction Fingerprinting (dMaSIF)

- Free of any precomputed features
- Computations are performed on the fly, with a small memory footprint

Problem Setting

Interaction prediction: Take as inputs two surface patches and predict if these locations are likely to come into close contact in the protein complex.

Binding site identification: Classify the surface of a given protein into interaction sites and non-interaction sites.



Input Data Notation

- Cloud of atoms $\{\mathbf{a}_1, \ldots, \mathbf{a}_A\} \subset \mathbb{R}^3$
- Chemical types $\{\mathbf{t}_1, \ldots, \mathbf{t}_A\} \subset \mathbb{R}^6$

Sample Surface Points and Normals

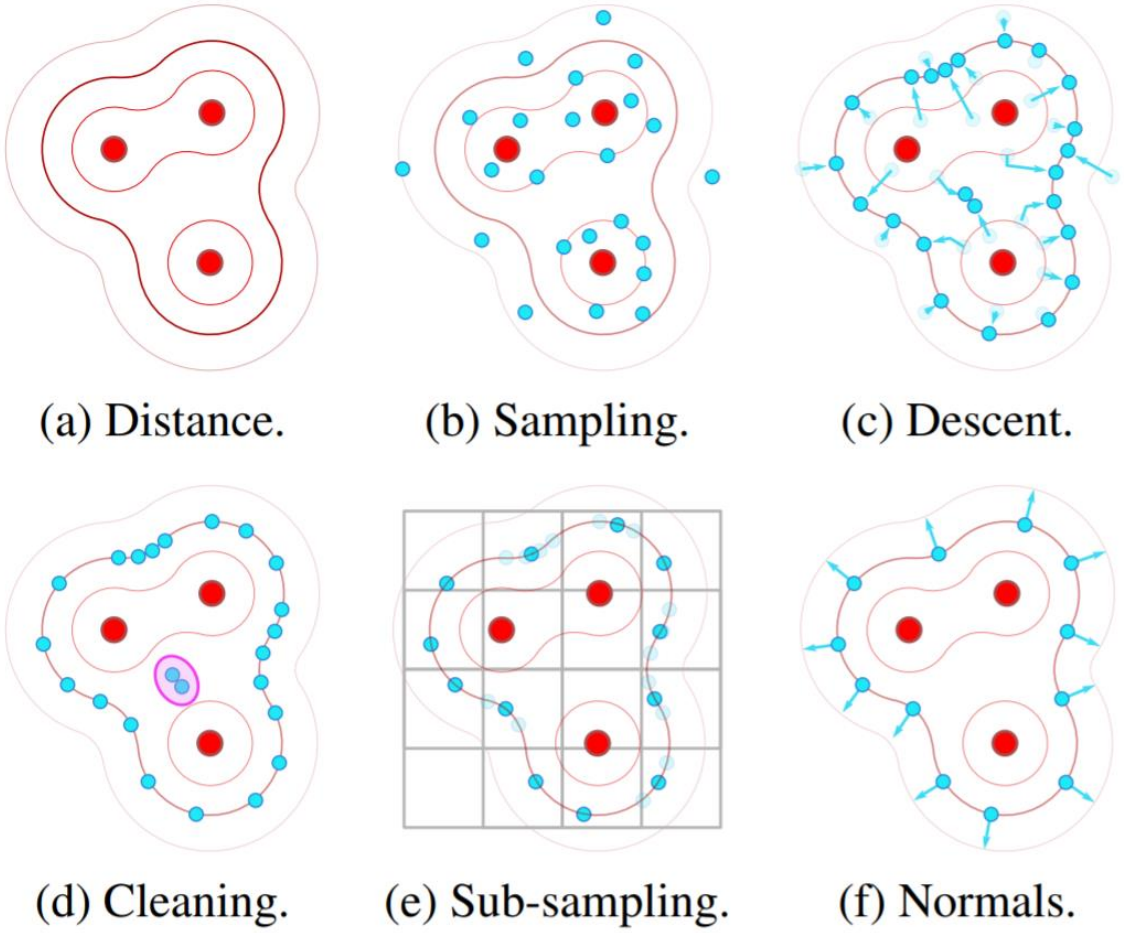
Smooth Distance Function

$$\mathrm{SDF}(\mathbf{x}) = -\sigma(\mathbf{x}) \cdot \log \sum_{k=1}^{\mathrm{A}} \exp(-\|\mathbf{x} - \mathbf{a}_k\| \,/\, \sigma_k)$$

$$\sigma(\mathbf{x}) \quad = \quad \sum_{k=1}^{\mathrm{A}} \exp(-\|\mathbf{x} \; - \; \mathbf{a}_k\|)\, \sigma_k \,/\, \sum_{k=1}^{\mathrm{A}} \exp(-\|\mathbf{x} - \mathbf{a}_k\|)$$

Associate an atomic radius σ_k to each atom

Sample Surface Points and Normals



(a) Distance. (b) Sampling. (c) Descent.

(d) Cleaning. (e) Sub-sampling. (f) Normals.

Compute Chemical Features

- For each point x_i, find 16 nearest atom centers.
- Apply a MLP to the 16 vectors:

$$[\mathbf{t}_k^i, 1/\|\mathbf{x}_i - \mathbf{a}_k^i\|]$$

- Sum over the output vectors and apply a second MLP to the result.
- Concatenate these 6D chemical features to the 5+5 mean and Gaussian curvatures to create a full feature vector of size 16.

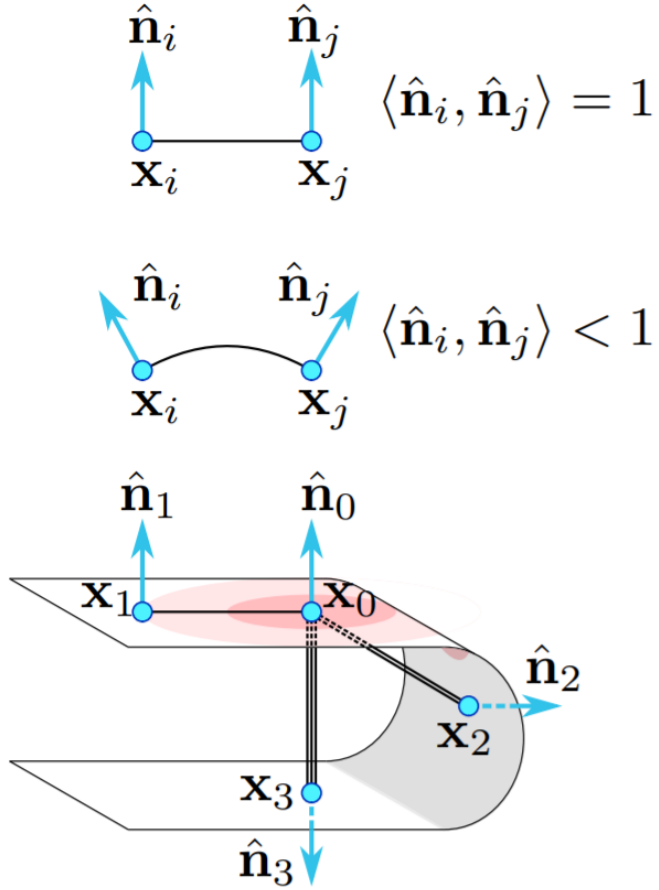
Apply Quasi-Geodesic Convolution

Geodesic distance

$$\mathrm{d}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| \cdot (2 - \langle \hat{\mathbf{n}}_i, \hat{\mathbf{n}}_j \rangle)$$

Smooth Gaussian window

$$w(\mathrm{d}_{ij}) = \exp(-\mathrm{d}_{ij}^2 \,/\, 2\sigma^2)$$



Apply Quasi-Geodesic Convolution

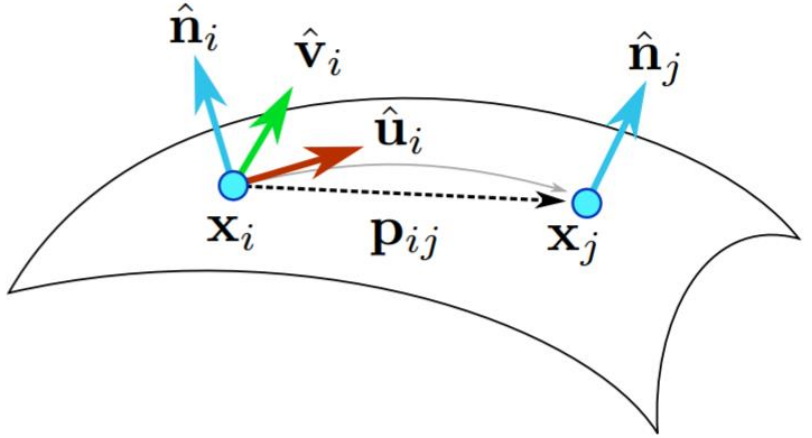
Geodesic convolution

$$\mathbf{f}_i' \leftarrow \sum_{j=1}^{\mathrm{N}} w(\mathrm{d}_{ij})\,\mathrm{MLP}(\mathbf{p}_{ij})\,\mathbf{f}_j$$

$$\mathbf{p}_{ij} = \left[\ (\mathbf{x}_j - \mathbf{x}_i)^\top\ \right] \cdot \left[\ \hat{\mathbf{n}}_i \ \middle|\ \hat{\mathbf{u}}_i \ \middle|\ \hat{\mathbf{v}}_i\ \right]$$

$$\mathbf{q}_{ij} = \left[\ (\hat{\mathbf{n}}_j - \hat{\mathbf{n}}_i)^\top\ \right] \cdot \left[\ \hat{\mathbf{n}}_i \ \middle|\ \hat{\mathbf{u}}_i \ \middle|\ \hat{\mathbf{v}}_i\ \right]$$

$$\mathrm{Conv}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{f}_j) = \mathrm{Window}(\mathrm{d}_{ij}) \cdot \mathrm{Filter}(\mathbf{p}_{ij}\ ,\ \mathbf{q}_{ij}) \cdot \mathbf{f}_j$$



Final Step

Binding Site Identification:

- Apply an MLP to the output of the convolutions to produce the final site/non-site binary output.

Interaction Prediction:

- Compute dot products between the feature vectors of both proteins.
- Use the products as interaction scores between pairs of points.

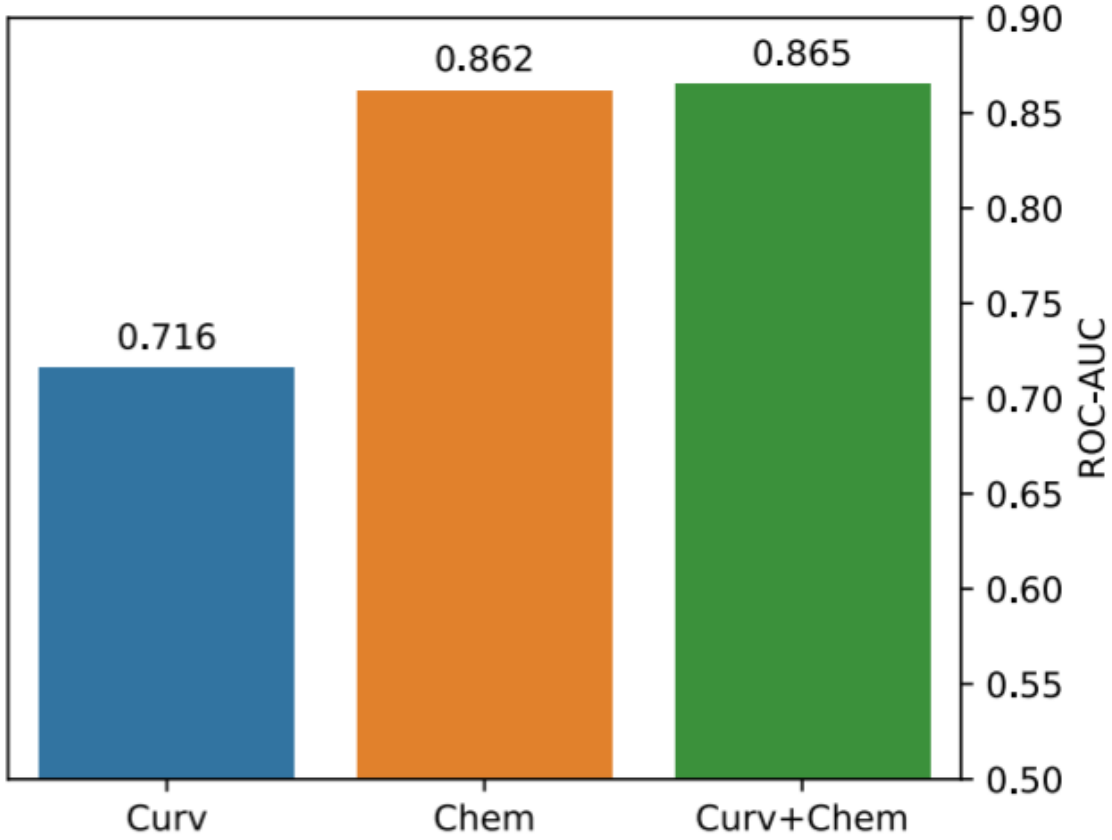
Results and Discussion

Precomputation time

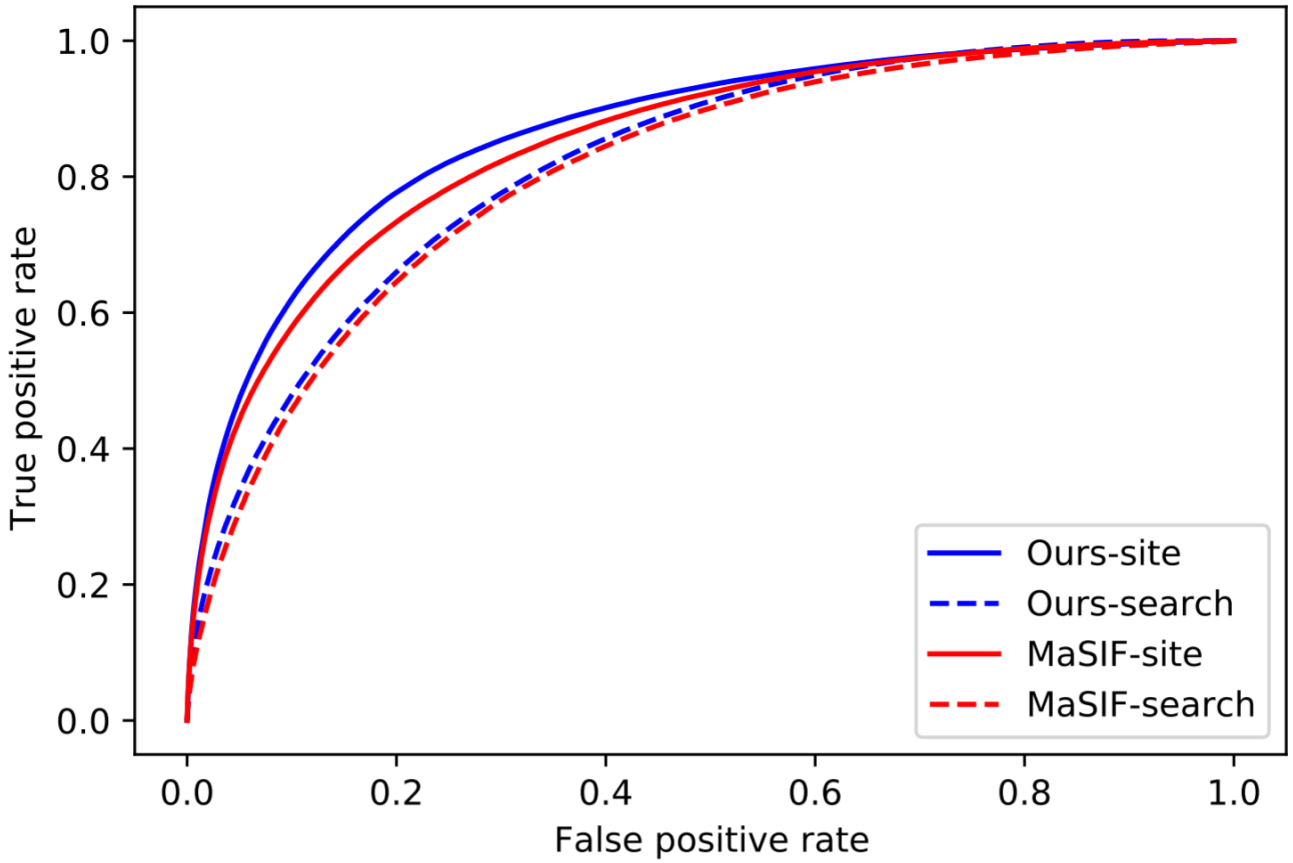
Computation	MaSIF	Ours
Surface generation	6.11±6.18 s	59.0±15.2 ms*
Input features	19.69±16.08 s	6.59±1.22 ms*
Local coordinates	50.65±45.15 s	0.46±0.09 ms*

Results and Discussion

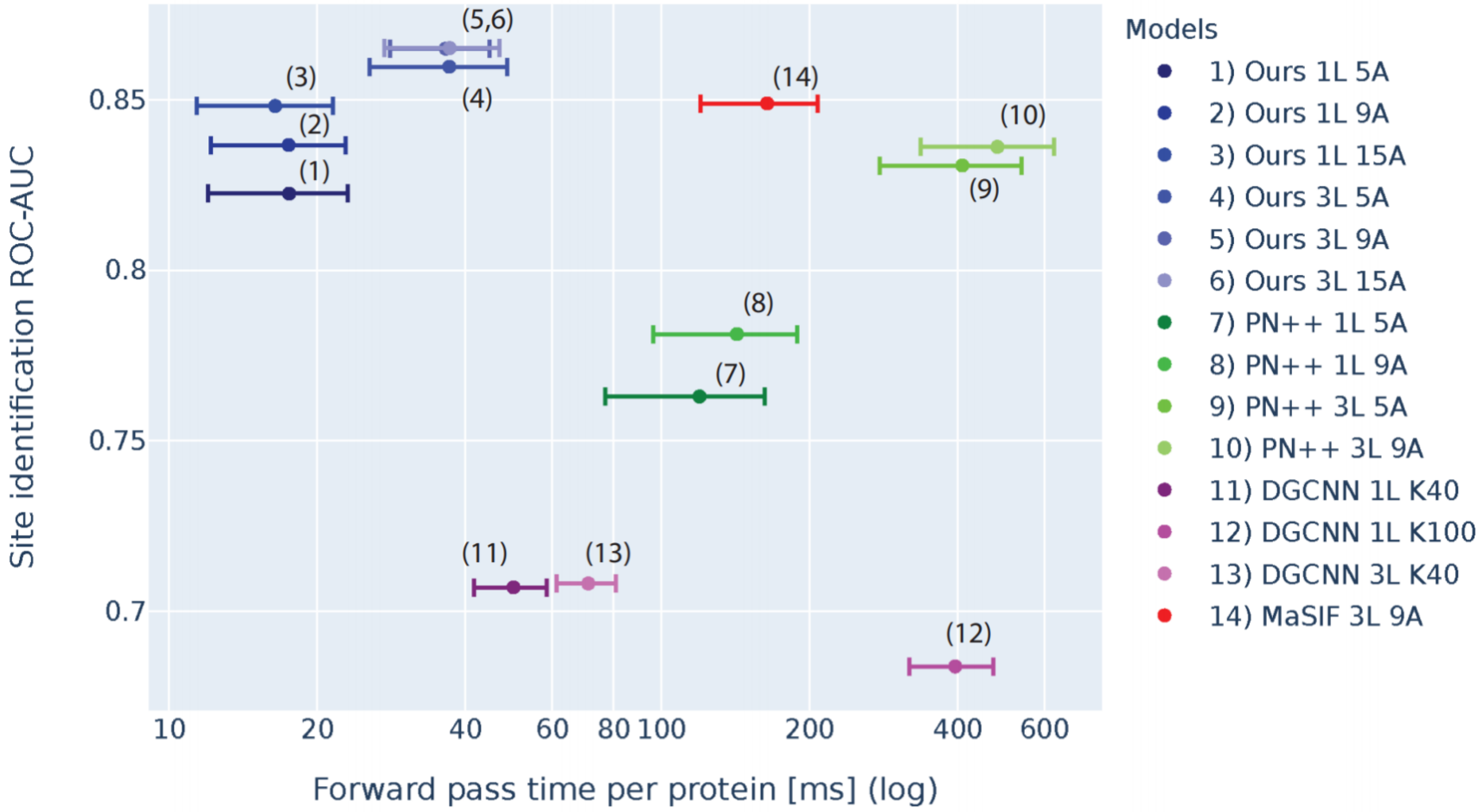
Ablation study



Results and Discussion



Results and Discussion



Results and Discussion

Limitations

- The concatenation of geometric curvatures to the vector of learned chemical features does not significantly improve the performance.
- Ignores the atoms inside the proteins, which may cause the loss of binding information.

Thanks for listening!