Fast End-to-End Learning on Protein Surfaces

Freyr Sverrisson[1]*, Jean Feydy[2]*, Bruno E. Correia[1], Michael M. Bronstein[2,3]
[1]École polytechnique fédérale de Lausanne, [2]Imperial College London, [3]Twitter, *Equal contribution

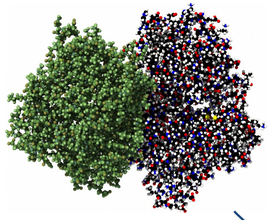
CVPR VIRTUAL JUNE 19-25

Highlights

- We expand upon our previous work on characterizing **protein fingerprints**: functional interaction patterns on protein surfaces.
- Our previous method MaSIF relied on **mesh** convolutions. It was limited by high computational requirements and costly precomputations.
- We present **dMaSIF**: a **quasi-geodesic point neural network** which is precomputation-free and orders of magnitude faster than MaSIF.
- dMaSIF is **fully differentiable** back to the atomic structure. This opens the door to end-to-end approaches in protein modeling and design.

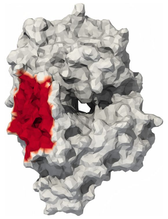
Introduction

1. A protein's amino acid sequence determines its three-dimensional structure (**black**). This three-dimensional structure determines which other molecules (**green**) the protein can bind to.



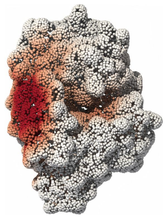
2. A continuous **protein surface** can be defined as those parts of the protein that are accessible by other molecules.

   We use known interactions to **label** the surface points (**red**).

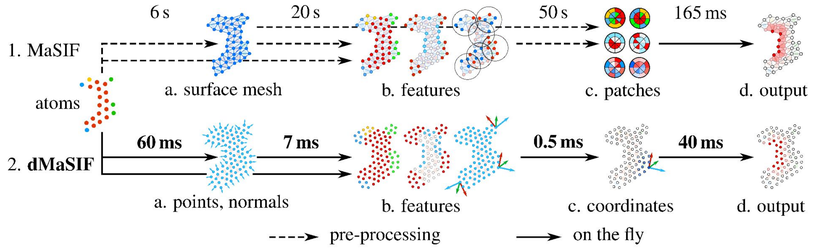


3. We use protein atoms to generate a **point cloud** representation of the protein surface.

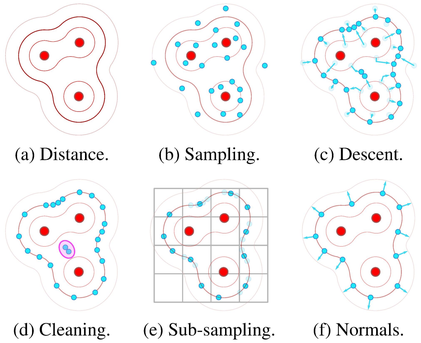
   We **predict** interaction patterns (**red**) based on the surface features.



Method overview



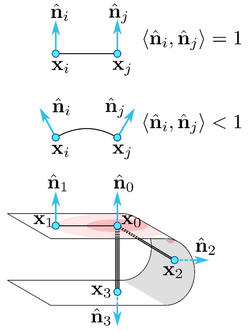
1. We **generate** the protein surface on the fly by:
   a) Defining a smooth **distance function** based on the atom coordinates;
   b) Randomly sampling points around the atoms;
   c) Using the gradient of the distance function to walk the sampled points towards a **level set**.



2. We approximate **geodesic distances** on the fly using the point normals:

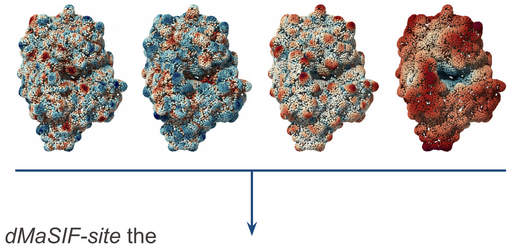
$$d_{ij} = \|x_i - x_j\|_2 \cdot (2 - \langle n_i, n_j \rangle)$$

   We leverage the KeOps library to implement **quasi-geodesic convolution** layers. They scale up to high-resolution point clouds and are easy to interface with PyTorch_Geometric.

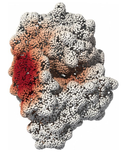


Results

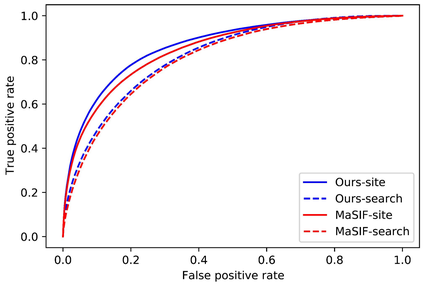
1. Our **input features**: data-driven **chemical** features + Gaussian, mean **curvatures** at different scales. These are computed directly from the atom types and positions.



2. In *dMaSIF-site* the task is to **segment** the surface into interacting / non-interacting sites. *dMaSIF-search* is trained to find point **correspondences** between two interacting proteins.



3. dMaSIF runs orders of magnitude **faster** than MaSIF and has much **lower memory** requirements. Moreover, dMaSIF slightly outperforms MaSIF on the two tasks we tested on.



Acknowledgements

SDSC
NCCR Molecular Systems Engineering
amazon research awards
erc European Research Council