Learning interaction patterns from surface representations of protein structure.

Pablo Gainza* Institute of Bioengineering EPFL, Lausanne, Switzerland pablo.gainza@epfl.ch

Emanuele Rodolà Dept. of Computer Science Sapienza U, Rome e.rodola@gmail.com **Freyr Sverrisson*** Institute of Bioengineering EPFL, Lausanne, Switzerland freyr.sverrisson@epfl.ch

Davide Boscaini Fondazione Bruno Kessler Trento, Italy dboscaini@fbk.eu

Bruno E. Correia* Institute of Bioengineering EPFL, Lausanne, Switzerland Federico Monti[†] Faculty of Informatics, USI Lugano, Switzerland

fmonti210gmail.com

Michael Bronstein[†] Department of Computing Imperial College London, UK michael.bronstein@gmail.com

Abstract

bruno.correia@epfl.ch

Predicting interactions between proteins and other biomolecules purely based on structure is an unsolved problem in biology. The protein molecular surface, a high-level representation of protein structure, displays the chemical and geometric features that can enter in contact with other biomolecules. This representation abstracts underlying details such as the precise arrangement of atoms and the amino acid sequence. Here we hypothesize that features in the molecular surface representation arrange in patterns, and proteins that perform similar interactions may display similar patterns. We model the discretized molecular surface as a graph and use data-driven geometric-deep learning tools to learn these patterns, and exploit them through three prediction challenges: (a) pocket similarity comparison, (b) protein-protein interaction site prediction and (c) prediction of interaction patterns based on surface patterns.

1 Introduction

Predicting interactions interactions between proteins and other biomolecules purely from protein structures remains one of the most important challenges in computational structural biology, with broad relevance for biomedical research [1, 2]. A high-level representation of protein structure, the *molecular surface* [3], describes proteins as continuous shapes. These shapes solely display the geometric and chemical features that can enter in contact with other biomolecules (Fig. 1a). The molecular surface's usefulness for many tasks involving protein interactions has long been known, and has been the preferred structural description to study the electrostatic interactions between proteins and water [5]. Some more recent efforts have attempted to embed molecular surface features using handcrafted techniques such as 3D Zernike descriptors [6, 7, 8, 9] and geometric invariant fingerprint

^{*}and Swiss Institute of Bioinformatics, Switzerland.

[†]and Twitter, UK



Figure 1: **Overview of the MaSIF conceptual framework.** a. Left, conceptual representation of a protein surface engraved with an interaction fingerprint, surface features that may reveal their potential biomolecular interactions. Right, surface segmentation into overlapping radial patches of a fixed geodesic radius used in MaSIF. b. Patches comprise geometric and chemical features mapped on the protein surface. c. Polar geodesic coordinates used to map the position of the features within the patch. d. MaSIF uses geometric deep learning tools to apply convolutional neural networks to the data. Fingerprint descriptors are computed for each patch using application-specific neural network architectures, which contain reusable building blocks (geodesic convolutional layers).

descriptors (GIF) [12] for fast searches and functional classification. The scope of these handcrafted approaches is limited as it is hard, if not impossible, to determine a priori the right set of features for a given prediction task.

We hypothesize that molecular surfaces are fingerprinted with patterns of features (termed *interaction fingerprints*), and that proteins that perform similar interactions may display similar fingerprints. We present MaSIF (Molecular Surface Interaction Fingerprinting), a geometric deep learning [4] method that exploits the graph structure of the discretized molecular surface to learn embeddings of the interaction fingerprints. MaSIF is showcased with three proof-of-concept applications (Fig. 1e): a) protein binding pocket similarity comparison (MaSIF-ligand); b) protein-protein interaction (PPI) site prediction in protein surfaces, to predict which surface patches on a protein are more likely to interact with other proteins (MaSIF-site); c) predicting the interaction likelihood of two surface patches for fast searching of potential PPI binding partners (MaSIF-search).

2 Methods

MaSIF decomposes a protein surface into overlapping radial patches with a fixed geodesic radius (Fig. 1a-b). All discretized protein surfaces are computed from protein structures using the MSMS program [10] and regularized using PyMesh [11]. Essential to our approach is to describe the molecular surface data in geodesic space. Geodesic distances are approximated using the Dijkstra algorithm on the edge weighted graph (graph geodesics) between the points along the surface. Around each vertex of the mesh, we extract a patch with geodesic radius of r=9 Å (MaSIF-site) or r=12 Å (MaSIF-site and MaSIF-ligand) (Fig. 1b). For each vertex within the patch, we compute two geometric features (shape index [15] and distance-dependent curvature [12]) and three chemical features (hydropathy

index [16], continuum electrostatics [17], and the location of free electrons and proton donors [18]) (Fig. 1b). The vertices within a patch are assigned geodesic polar coordinates (Fig. 1c): the radial coordinate, representing the geodesic distance to the center of the patch; and the angular coordinate, computed with respect to a random direction from the center of the patch, as the patch lacks a canonical orientation.

The geodesic polar coordinates allow MaSIF to spatially localize features within each patch. In these coordinates, we then construct a system of local Gaussian kernels for which the parameters are learnable, based on the MoNET architecture [13]. Specifically, our local polar system contains θ angular bins, and ρ polar bins, for a total of $J = \rho \cdot \theta$ bins. For each vertex in the discretized molecular surface x, with neighbors N(x), and each vertex $y \in N(x)$, we define the coordinates u(x, y), the radial and angular coordinates of y with respect to x. The mapping of each grid cell j for feature vector f and the patch centered at x, $D_j(x)$ f is defined as:

$$D_j(x)f = \sum_{y \in N(x)} w_j(u(x,y))f(y), \quad j = 1, ..., J_y$$

where w_i is a weight function, and f(y) are the features at vertex y.

Since the radial coordinates are computed with respect to a random direction, it becomes essential to compute information that is invariant to different directions (rotation invariance). To this end, we perform K rotations on the patch and compute the maximum over all rotations [14], producing a vectorized output for the patch location. The procedure is repeated for different patch locations similar to a sliding window operation on images, producing the surface fingerprint at each point, in the form of a vector that embeds information about the surface patterns of the center point and its neighborhood. The learning procedure consists of minimizing the parameter set of the local kernels and filter weights with respect to the application-specific training data and cost function. The embedded data can then be further processed in neural network architectures.

2.1 Dataset generation and splits.

For all three applications we assembled databases of proteins from co-crystal structures (holo structures). For MaSIF-ligand, a database of 11685 structures binding the small molecules ADP, COA, FAD, NADP, and SAM was built. Structures were clustered based on sequence identity (30% sequence threshold), reducing the number of structures to 1459. Structures were then randomly split into training (72%), validation (8%) and testing (20%). We also selected a test set with an additional structural split, where pockets in the test set were removed if they aligned to at least one training set pocket with a TM-align score [26] of more than 0.5. For MaSIF-site a database of 6001 PPI pairs was used as the training set, with the ground truth labeled based on the surface areas that become inaccessible in co-crystal structure. The dataset was split into the interacting subunits, resulting in 12002 proteins. Proteins were then clustered by sequence identity (30% threshold) and one representative was selected from each cluster. The training and test set were then further split using scikit-learn's hierarchical clustering, based on a structural alignment using TM-align scores [26], resulting in 3004 training and 358 test set proteins. 53 proteins from the test set were further selected for comparison with other baseline methods. For MaSIF-search only a structural split was performed. The interfaces of the 6001 PPIs used in MaSIF-site were aligned all-against-all and and split into training and testing based on scikit-learn's hierarchical clustering. After removing cases that failed with TM-align, this resulted in 4944 training and 957 testing PPIs. Details on the architectures and training are shown in Fig. S1-S3.

3 Results

In our first application, MaSIF-ligand, we explore whether the interaction fingerprints in protein surfaces hold sufficient information to decipher the small-molecule-binding preference of protein pockets. For this proof-of-concept we selected pockets that recognize six different small molecules (Fig 2a), all of which have abundant structural data available and bind to proteins with little sequence identity. We trained The classification task consisted of classifying a pocket into six different classes. In our two datasets, MaSIF-ligand's outperformed two other state-of-the-art (benchmarked in [21]) methods: ProBIS [19] and KRIPO [20] (Fig. 2b-c).



Figure 2: MaSIF proof-of-concept applications. (a) MaSIF-ligand. MaSIF-ligand receives as input a protein pocket and predicts a score for 6 ligands. (b-c) Performance of MaSIF-ligand and two baseline methods on a test set with amino acid (b) sequence-redundancy removed, and (c) sequence+structure redundancy removed. (d) MaSIF-site. MaSIF-site receives receives as input a single protein surface and computes a score on each vertex for its propensity to form PPIs. (e) Example of the effect of multiple layers of convolution on MaSIF-site's predictions. (f) Performance of MaSIF-site vs. two baseline methods on a test set of 53 proteins involved in transient PPIs. (g) MaSIF-search. MaSIF-search is trained to produce descriptors that are similar (close in Euclidean space) for interacting patches and dissimilar (far in Euclidean space) for non-interacting patches. (h) Performance of MaSIF-search and the baseline method GIF [12].

Our second application, MaSIF-site, receives a protein as input and outputs a predicted score for each vertex of the surface mesh on the likelihood of this vertex being part of the interface with another protein. MaSIF site does not use explicit knowledge of the interacting partner (Fig. 2d). We find that in this task using multiple layers of convolution results in a much better performance, illustrated in Figure 2e with an example comparing the prediction of a network with one layer of convolution with a network with 3 layers of convolution. We compared the performance of MaSIF-site with that of the established, state-of-the-art [24] predictors: SPPIDER [22] and PSIVER[23]. MaSIF-site significantly outperforms these established tools.

As a last example of MaSIF, we exploit MaSIF's embedding as vectors to predict specific interactions between proteins. Vector embedding, inspired by earlier work on GIF descriptors, is attractive because, once the vectors are precomputed, nearest-neighbor techniques can scan billions of vectors in milliseconds. Thus, we introduce MaSIF-search a new paradigm for the fast search of protein binding partners based on surface fingerprints. MaSIF-search produces similar descriptors for pairs of interacting patches, and dissimilar descriptors for non-interacting patches (Fig. 2g). To test this, we assembled a database with >100K pairs of interacting protein surface patches with high shape complementarity, as well as a set of randomly chosen surface patches, to be used as non-interacting patches. A trio of protein surface patches termed *binder*, *target*, and *random* patches were fed into the network of MaSIF-search (Fig. 2g). The neural network is simultaneously trained to minimize the Euclidean distance between the fingerprint descriptors of binders and targets, while maximizing the dissimilarity between targets and random. In order to help the network learn faster and to improve performance in general, we *invert* the features of one of the patches (the binder patch), by multiplying all features by -1 (with the exception of hydropathy features). For this task we compare our results to GIF descriptors, which were originally proposed for fast searching of protein surfaces (Fig. 2h).

MaSIF-search's descriptors, which vastly outperform GIF descriptors, could be coupled with an alignment tool, such as RANSAC, for further refinement of results for fast PPI prediction.

4 Conclusions

In summary, we present a conceptual framework to decipher interaction fingerprints, leveraging the representation of protein structures as molecular surfaces, together with powerful geometric deep learning techniques. Thanks to our data-driven approach, high-level features are learned for each task, which may be impossible with handcrafted approaches. In all three applications presented here, MaSIF outperforms baseline methods, consistent with the emergent trends in the field of computer vision where learned features outperform handcrafted features. Overall, MaSIF search provides a new tool for the study of protein interactions, with broad applications in biomedical research.

References

- [1] Donald, B. R. Algorithms in Structural Molecular Biology. The MIT Press (2011).
- [2] Zhang, Q. C. et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490, 556–560 (2012)
- [3] Richards, F. M. Areas, Volumes, Packing, and Protein Structure. Annual Review of Biophysics and Bioengineering 6, 151–176 (1974).
- [4] Bronstein, M. M., Bruna, J., Lecun, Y., Szlam, A. & Vandergheynst, P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4), 18-42. (2017).
- [5] Sharp, K. Electrostatic Interactions In Macromolecules: Theory And Applications. *Annual Review of Biophysics and Biomolecular Structure* 19, 301–332 (1990).
- [6] Daberdaku, S. & Ferrari, C. Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics* (Oxford, England) 35, 1870–1876 (2019).
- [7] Kihara, D., Sael, L., Chikhi, R. & Esquivel-Rodriguez, J. Molecular Surface Representation Using 3D Zernike Descriptors for Protein Shape Comparison and Docking. *Current Protein & Peptide Science* 12, 520–530 (2011).
- [8] Zhu, X., Xiong, Y. & Kihara, D. Large-scale binding ligand prediction by improved patch-based method Patch-Surfer 2.0. *Bioinformatics* 31, 707–713 (2015).
- [9] Venkatraman, V., Yang, Y. D., Sael, L. & Kihara, D. Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics* 10, (2009).
- [10] Sanner, M. F., Olson, A. J. & Spehner, J. Reduced surface: An efficient way to compute molecular surfaces. Biopolymers 38, 305–320 (1996).
- [11] Zhou, Q. PyMesh Geometry Processing Library for Python. Available at: https://github.com/PyMesh/PyMesh (2019).
- [12] Yin, S., Proctor, E. A., Lugovskoy, A. A. & Dokholyan, N. V. Fast screening of protein surfaces using geometric invariant fingerprints. *Proceedings of the National Academy of Sciences* 106, 16622–16626 (2009).
- [13] Monti, F. et al. Geometric deep learning on graphs and manifolds using mixture model CNNs. in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 5425–5434 (2017).
- [14] Masci, J., Boscaini, D., Bronstein, M. M. & Vandergheynst, P. Geodesic Convolutional Neural Networks on Riemannian Manifolds. *Proceedings of the IEEE International Conference on Computer Vision* 2015-February, 832–840 (2015).
- [15] Koenderink, J. J. & van Doorn, A. J. Surface shape and curvature scales. *Image and Vision Computing* 10, 557–564 (1992).
- [16] Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157, 105–132 (1982).
- [17] Jurrus, E. et al. Improvements to the APBS biomolecular solvation software suite. *Protein Science* 27, 112–128 (2018).

- [18] Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology* 326, 1239–1259 (2003).
- [19] Konc, J. et al. ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in Protein Binding Sites. *Journal of Chemical Information and Modeling* 55, 2308–2314 (2015).
- [20] Ritschel, T., Schirris, T. J. & Russel, F. G. KRIPO a structure-based pharmacophores approach explains polypharmacological effects. *Journal of Cheminformatics* 6, (2014).
- [21] Ehrt, C., Brinkjost, T. & Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Computational Biology* 14, (2018).
- [22] Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins: Structure, Function and Genetics* 66, 630–645 (2007).
- [23] Murakami, Y. & Mizuguchi, K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* 26, 1841–1848 (2010).
- [24] Xue, L. C., Dobbs, D., Bonvin, A. M. J. J. & Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters* 589, 3516–3526 (2015).
- [25] Perez, L. & Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv preprint arXiv:1712.04621 (2017).
- [26] Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33, 2302–2309 (2005).

5 Supplementary Figures



Figure S1: **Network architecture for MaSIF-ligand.** 32 randomly sampled pocket patches are fed through convolutional layers followed by a fully connected layer (FC80). Descriptors are combined in a 80x80 covariance matrix followed by two fully connected layers (FC64 and FC7) and then softmax cross-entropy loss. The network was trained for 40 'wall-clock' hours, and killed after 40 hours, which allowed for 335000 iterations.



Figure S2: Network architecture for MaSIF-site. Patches are fed through convolutional layers followed by a series of fully connected layers (FC5, FC4, FC2), and finally a sigmoid cross-entropy loss. MaSIF-site was trained for 40 'wall clock' hours (43 epochs).



Figure S3: **Network architecture for MaSIF-search.** Patches from the target and the corresponding binder or a random patch are fed through convolutional layers, followed by a fully connected layer (FC80). The L2-distance between the resulting descriptors is computed and the neural network is optimized to minimize this distance with respect to binder and maximize it with respect to the random patch. MaSIF-search was trained for 40 'wall-clock' hours, which allowed for 335000 iterations.