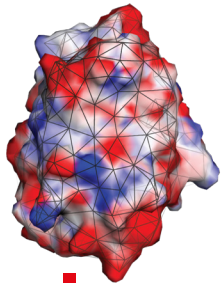
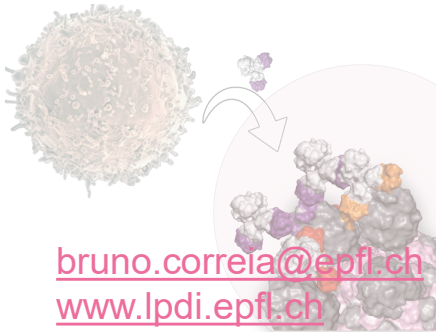
Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning



Bruno Correia
Laboratory of Protein Design and Immunoengineering
Institute of Bioengineering - STI

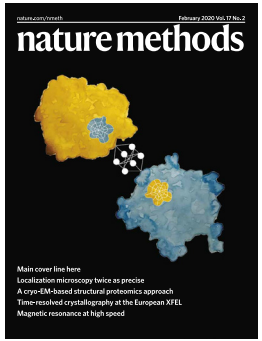


bruno.correia@epfl.ch
www.lpdi.epfl.ch

A "Split Personality" Protein Engineering Lab

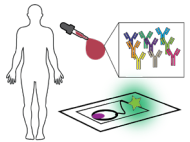
Protein Modeling
-Methods developments
-Protein function prediction
-Protein design

Experimental Characterization
-Biochemistry and biophysics
-High-throughput screening
-*In vivo* testing



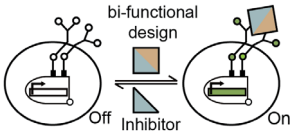
by Laura Persat

Gainza,…, Correia, *Nature Methods 2020*
Bonet,…, Correia, *Plos Comp Bio 2018*
Sesterhenn ,…, Correia, *Science 2020*
Yang ,…, Correia, *Nat Chem Bio 2021*

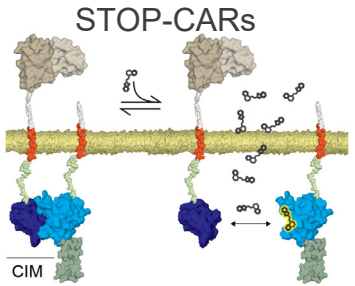


Biosensors

Vaccines



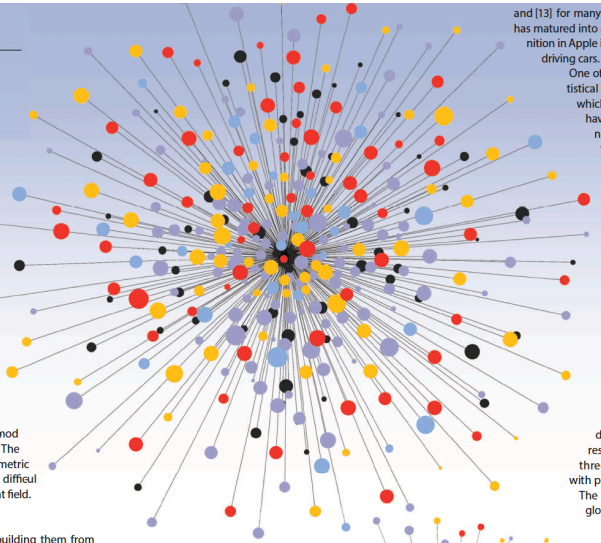
Synthetic Cells



Giordano, …, Correia, *Nature Biotech 2020*
Sesterhenn ,…, Correia, *Plos Bio 2018*
Mathony,…, Correia*, Niopek*, *Nat Chem Bio 2020*

On the importance of getting to know your neighbors

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst



Geometric Deep Learning

Going beyond Euclidean data

Digital Object Identifier 10.1109/MSP.2017.2693418
Date of publication: 11 July 2017

Many scientific fields study data with an underlying structure that is non-Euclidean. Some examples include social networks in computational social sciences, sensor networks in communications, functional networks in brain imaging, regulatory networks in genetics, and meshed surfaces in computer graphics. In many applications, such geometric data are large and complex (in the case of social networks, on the scale of billions) and are natural targets for machine-learning techniques. In particular, we would like to use deep neural networks, which have recently proven to be powerful tools for a broad range of problems from computer vision, natural-language processing, and audio analysis. However, these tools have been most successful on data with an underlying Euclidean or grid-like structure and in cases where the invariances of these structures are built into networks used to model them.

Geometric deep learning is an umbrella term for emerging techniques attempting to generalize (structured) deep neural models to non-Euclidean domains, such as graphs and manifolds. The purpose of this article is to overview different examples of geometric deep-learning problems and present available solutions, key difficulties, applications, and future research directions in this nascent field.

Overview of deep learning

Deep learning refers to learning complicated concepts by building them from simpler ones in a hierarchical or multilayer manner. Artificial neural networks are popular realizations of such deep multilayer hierarchies. In the past few years, the growing computational power of modern graphics processing unit (GPU)-based computers and the availability of large training data sets have allowed successfully training neural networks with many layers and degrees of freedom (DoF)[1]. This has led to qualitative breakthroughs on a wide variety of tasks, from speech recognition[2], [3] and machine translation[4] to image analysis and computer vision[5]–[11] (see [12] and [13] for many additional examples of successful applications of deep learning). Today, deep learning has matured into a technology that is widely used in commercial applications, including Siri speech recognition in Apple iPhone, Google text translation, and Mobileye vision-based technology for autonomously driving cars.

One of the key reasons for the success of deep neural networks is their ability to leverage statistical properties of the data, such as stationarity and compositionality through local statistics, which are present in natural images, video, and speech[14], [15]. These statistical properties have been related to physics[16] and formalized in specific classes of convolutional neural networks (CNNs) [17]–[19]. In image analysis applications, one can consider images as functions on the Euclidean space (plane), sampled on a grid. In this setting, stationarity is owed to shift invariance, locality is due to the local connectivity, and compositionality stems from the multiresolution structure of the grid. These properties are exploited by convolutional architectures[20], which are built of alternating convolutional and downsampling (pooling) layers. The use of convolutions has a twofold effect. First, it allows extracting local features that are shared across the image domain and greatly reduces the number of parameters in the network with respect to generic deep architectures (and thus also the risk of overfitting), without sacrificing the expressive capacity of the network. Second, the convolutional architecture itself imposes some priors about the data, which appear very suitable especially for natural images [17]–[19], [21].

While deep-learning models have been particularly successful when dealing with speech, image, and video signals, in which there are an underlying Euclidean structure, recently there has been a growing interest in trying to apply learning on non-Euclidean geometric data. Such kinds of data arise in numerous applications. For instance, in social networks, the characteristics of users can be modeled as signals on the vertices of the social graph[22]. Sensor networks are graph models of distributed interconnected sensors, whose readings are modeled as time-dependent signals on the vertices. In genetics, gene expression data are modeled as signals defined on the regulatory network[23]. In neuroscience, graph models are used to represent anatomical and functional structures of the brain. In computer graphics and vision, three-dimensional (3-D) objects are modeled as Riemannian manifolds (surfaces) endowed with properties such as color texture.

The non-Euclidean nature of such data implies that there are no such familiar properties as global parameterization, common system of coordinates, vector space structure, or shift invariance. Consequently, basic operations like convolution that are taken for granted in the Euclidean case are even not well defined on non-Euclidean domains. The purpose of this article is to show different methods of translating the key ingredients of successful deep-learning methods, such as CNNs, to non-Euclidean data.

Geometric learning problems

Broadly speaking, we can distinguish between two classes of geometric learning problems. In the first class of problems, the goal is to characterize the structure of the data. The second class of problems deals with analyzing functions defined on a given non-Euclidean domain. These two classes are related, because understanding the properties of functions defined on a domain conveys certain information about the domain, and vice versa, the structure of the domain imposes certain properties on the functions on it.

Structure of the domain

As an example of the first class of problems, assume to be given a set of data points with some underlying low-dimensional structure embedded into a high-dimensional Euclidean space. Recovering that low-dimensional structure is often referred to as *manifold learning* or nonlinear dimensionality reduction and is an instance of unsupervised learning (note that the notion of manifold in this setting can be considerably more general than a classical smooth manifold; see, e.g.,

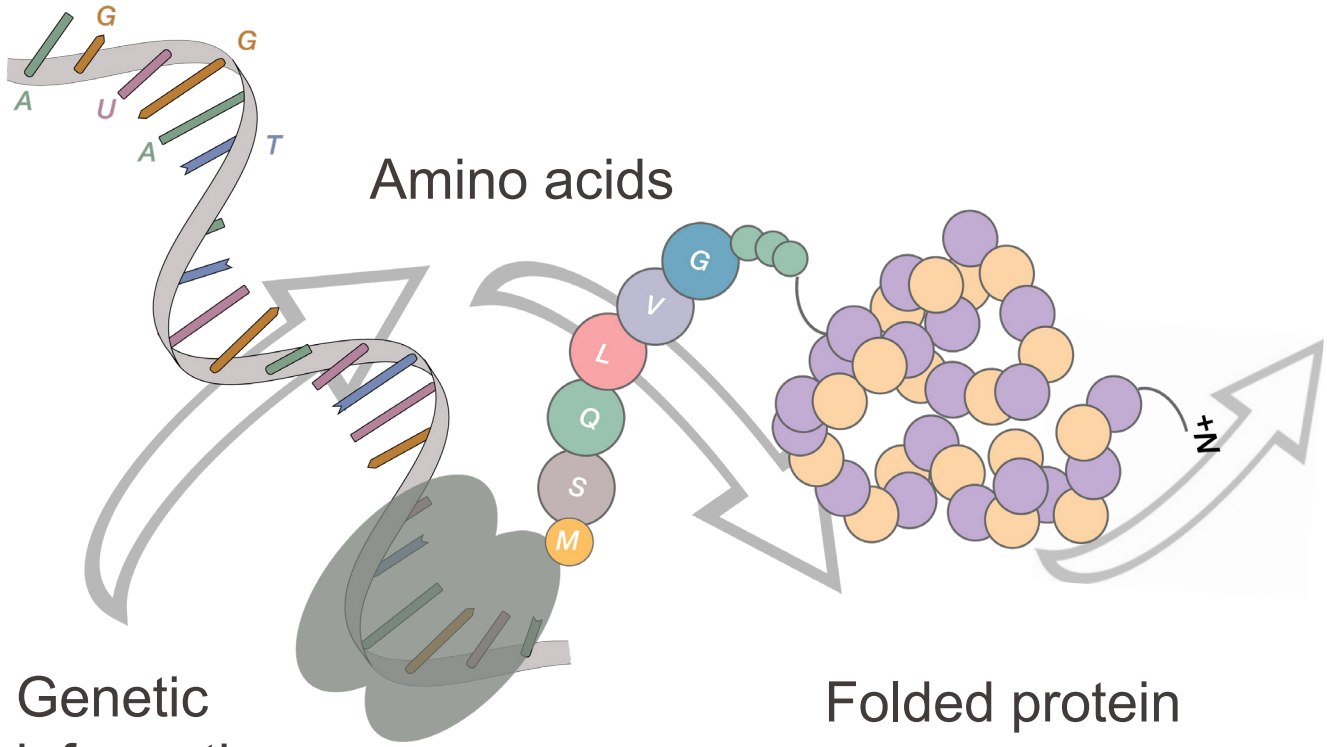
Outline

I) Brief intro to protein structure and function

II) Deciphering surface fingerprints for protein functional assignment

III) Fingerprint-driven design of de novo protein-protein interactions

Proteins are a fundamental molecular unit of life



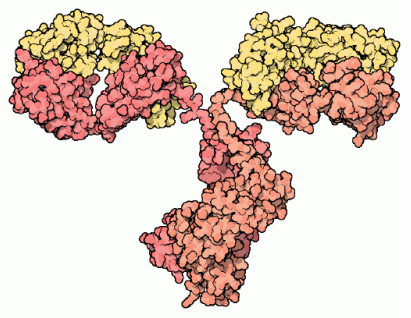
Molecular function(s)

- Binding/recognition
- Catalysis
- Mechanical functions
- ….

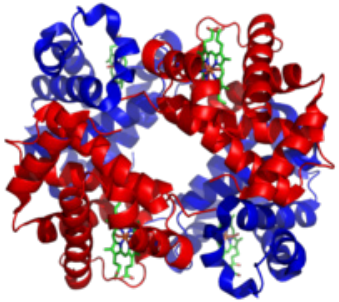
Biological function(s)

- Proliferation
- Metabolic processes
- Host defense
- ….

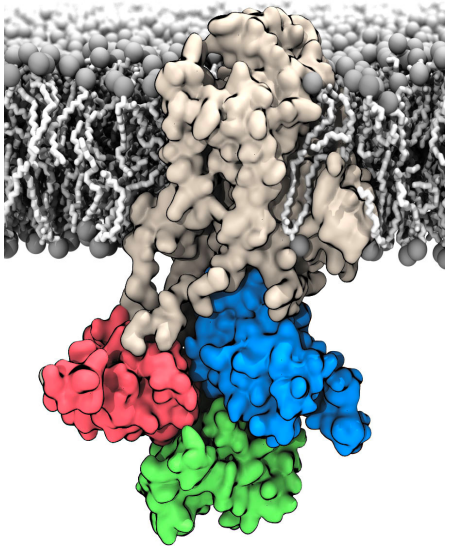
Protein function



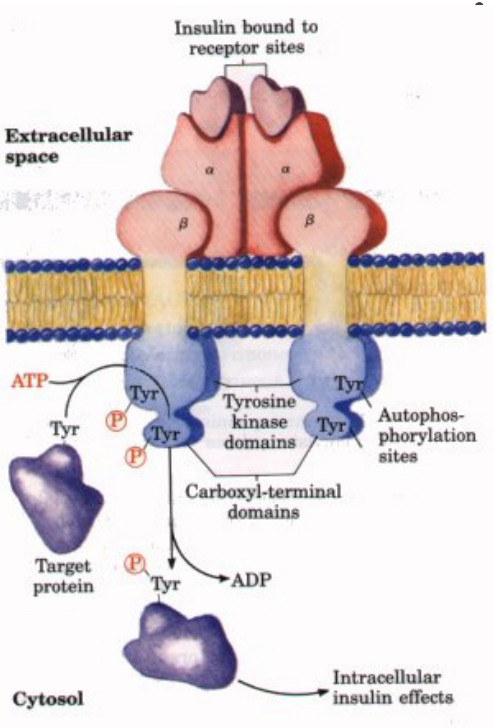
Defense (antibody)



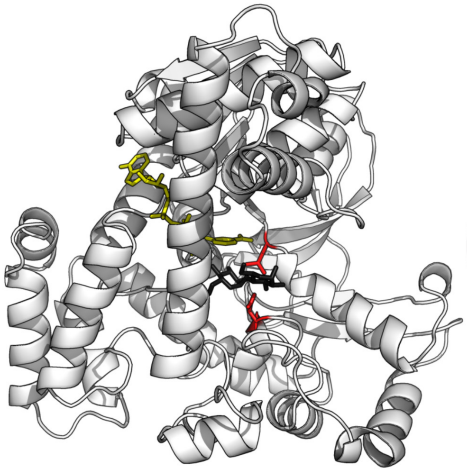
Storage (haemoglobin)

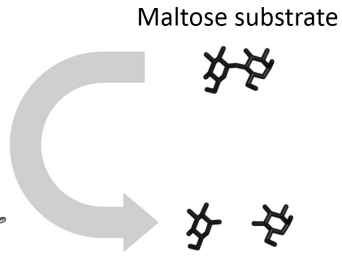


Transport (calcium pump)

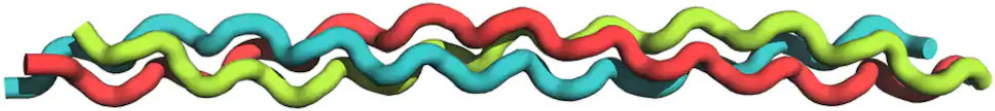


Communication (insulin)





Catalysis (enzyme)



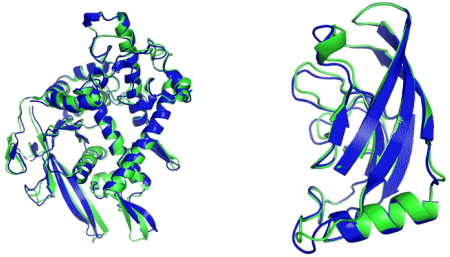
Structure (collagen)

Pre-emptively addressing a common point !!!

AlphaFold2 from Google DeepMind did not solve all the scientific questions in protein science.

Structure prediction problem ✓

Sequence ⟶ Structure

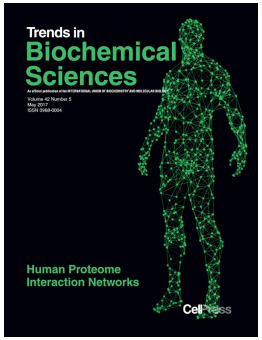


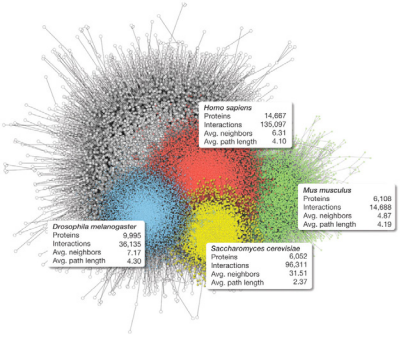
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

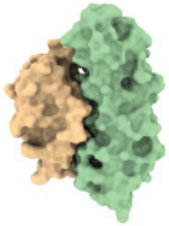
T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

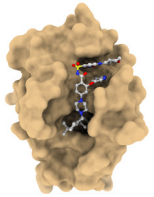
Function prediction problem ✗







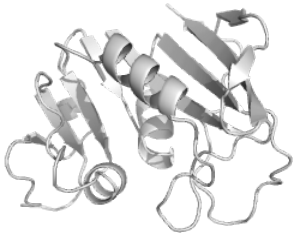
Protein-protein interactions



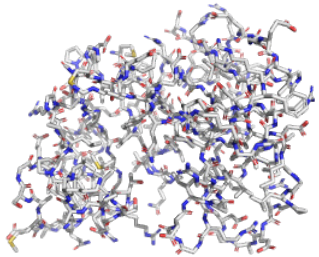
Protein-metabolite interactions

“If you want to understand function – study structure”(Crick)

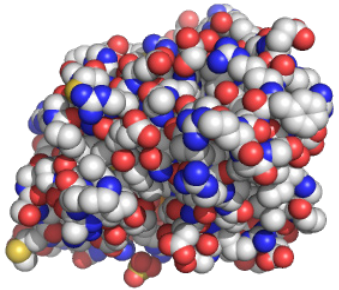
Protein structures are studied at different levels



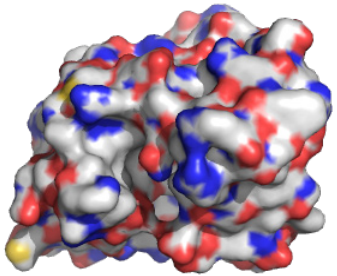
Secondary structures (ribbon diagram)



Graph (stick diagram)

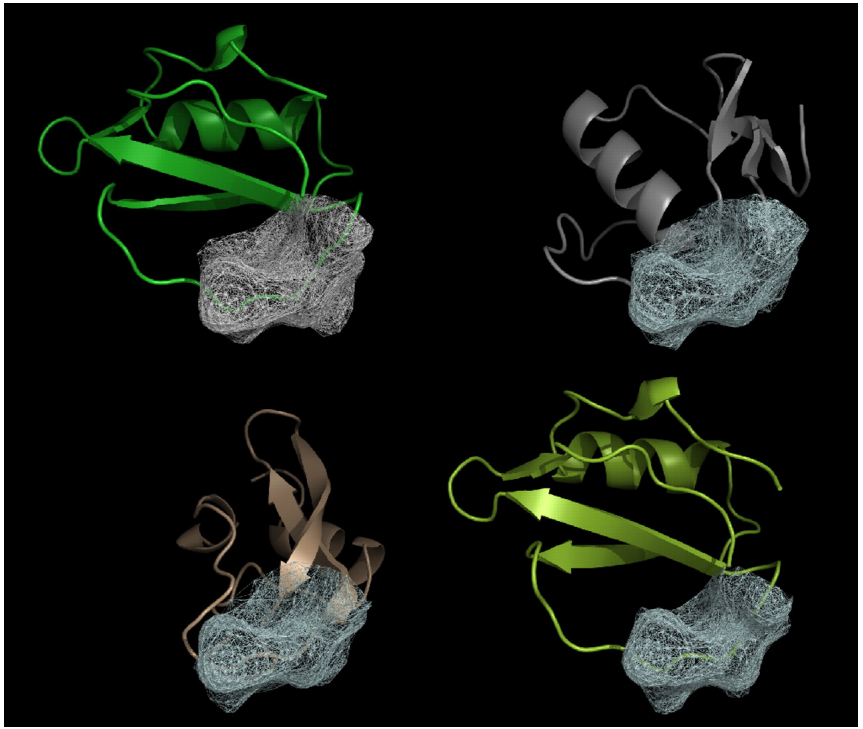


Point cloud (atomic diagram)



Molecular surface

Dissimilar sequence, dissimilar structural architecture, but similar function

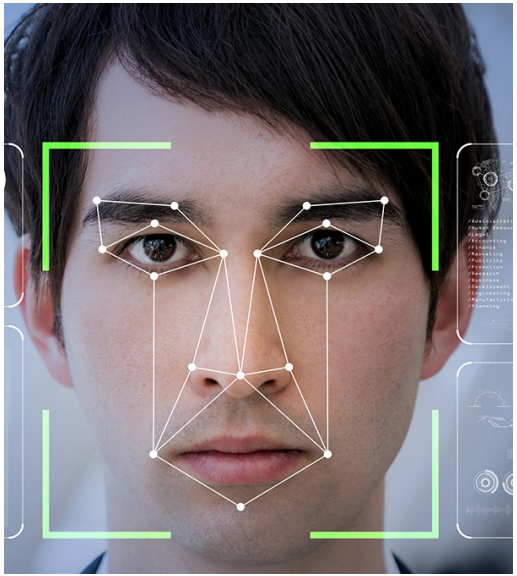


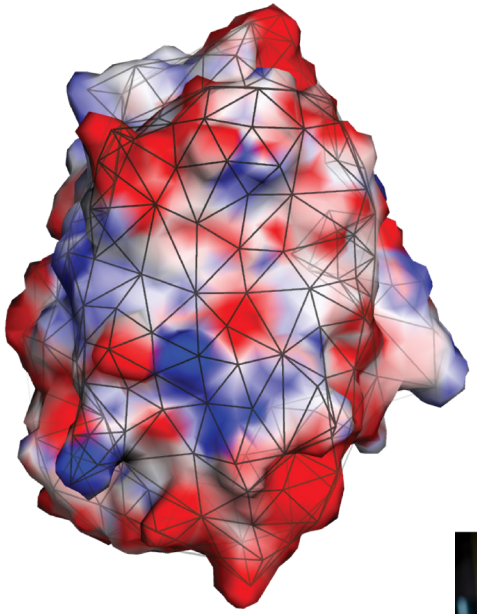
Yin et al. 2009

-Some similarities can be observed at the surface level.

The many (sur)faces of protein structures

Can we identify surface patterns that reveal functional features of proteins ?

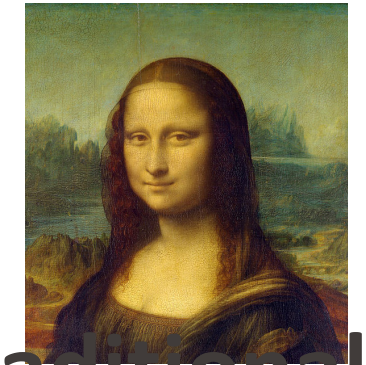




Gainza,…, **Correia**
Nature Methods, 2020

Pablo Gainza & Freyr Sverrisson

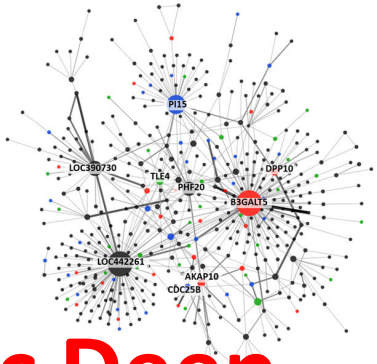
Which data science framework to use ?



Traditional Deep Learning

Images

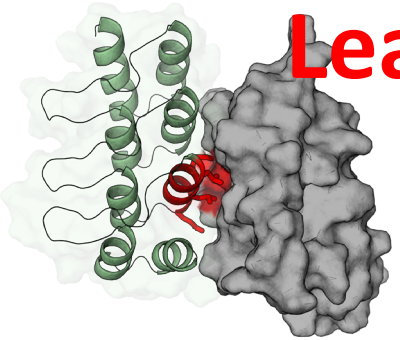
Acoustic signals



Social networks

Interaction networks

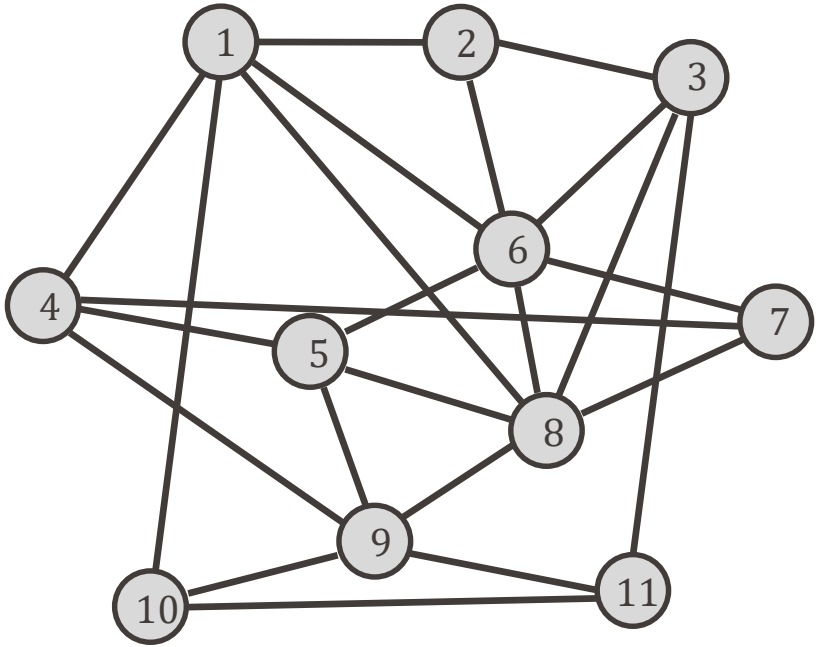
Geometric Deep Learning



Molecules

3D geometric data

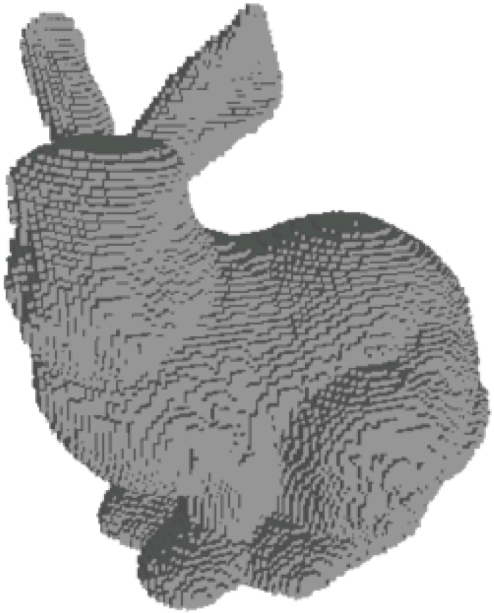
Prototypical objects



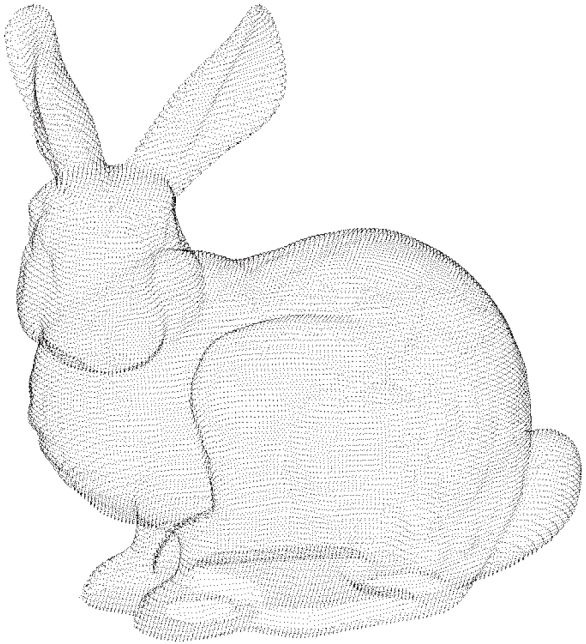
Graphs

Surfaces

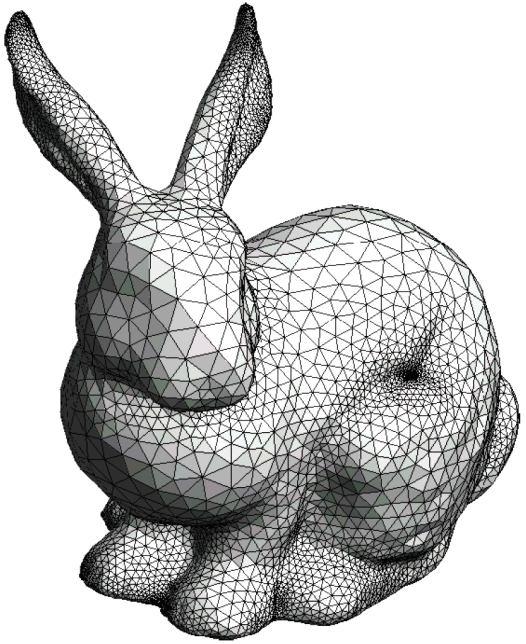
Representation



Volumetric



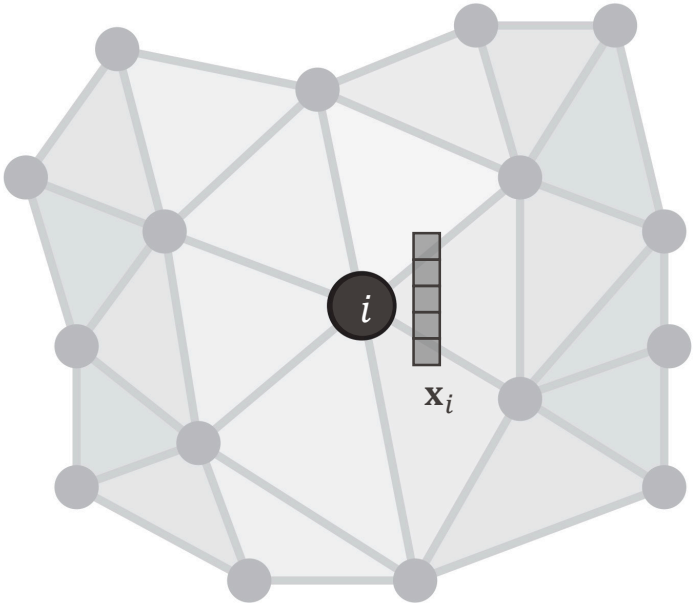
Point cloud



Surface / mesh

MoNet architecture

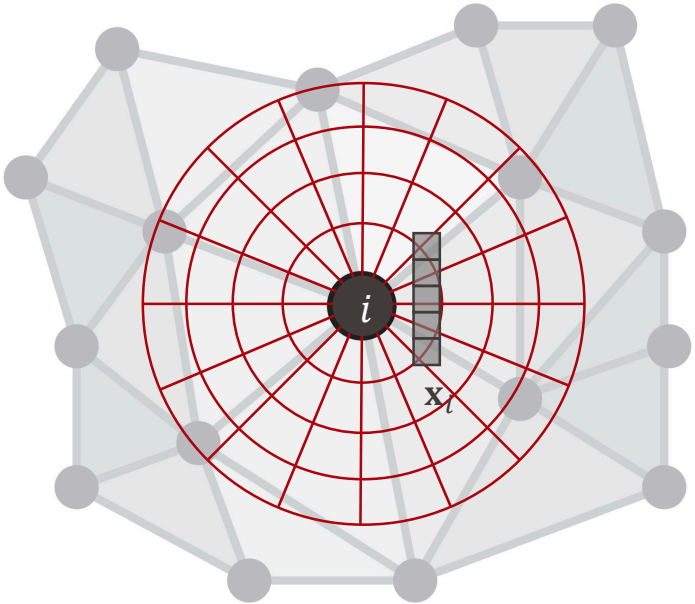
- **Vertex-wise d-dimensional features:** $n \times d$ matrix $\mathbf{X}$



Monti et Bronstein 2017

MoNet architecture

- **Vertex-wise d-dimensional features:** $n \times d$ matrix $\mathbf{X}$
- **Local geodesic polar coordinates** $\mathbf{u}_{ij}$ around i



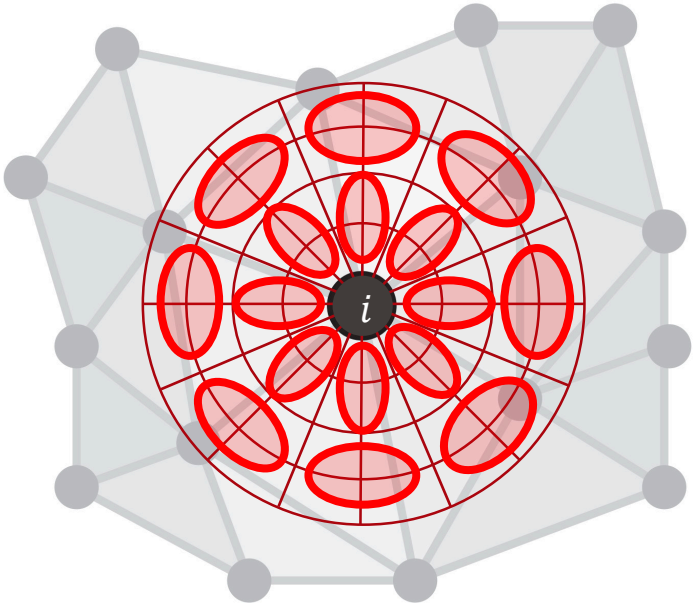
Monti et Bronstein 2017

MoNet architecture

- **Vertex-wise d-dimensional features:** $n \times d$ matrix $\mathbf{X}$
- **Local geodesic polar coordinates** $\mathbf{u}_{ij}$ around i
- **Local weights** $w_1(\mathbf{u}), \dots, w_L(\mathbf{u})$ w.r.t.$\mathbf{u}$, e.g. Gaussians:

$$w_\ell(\mathbf{u}) = \exp\left(-(\mathbf{u} - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}_\ell^{-1} (\mathbf{u} - \boldsymbol{\mu}_\ell)\right)$$

'soft pixels'



Monti et Bronstein 2017

MoNet architecture

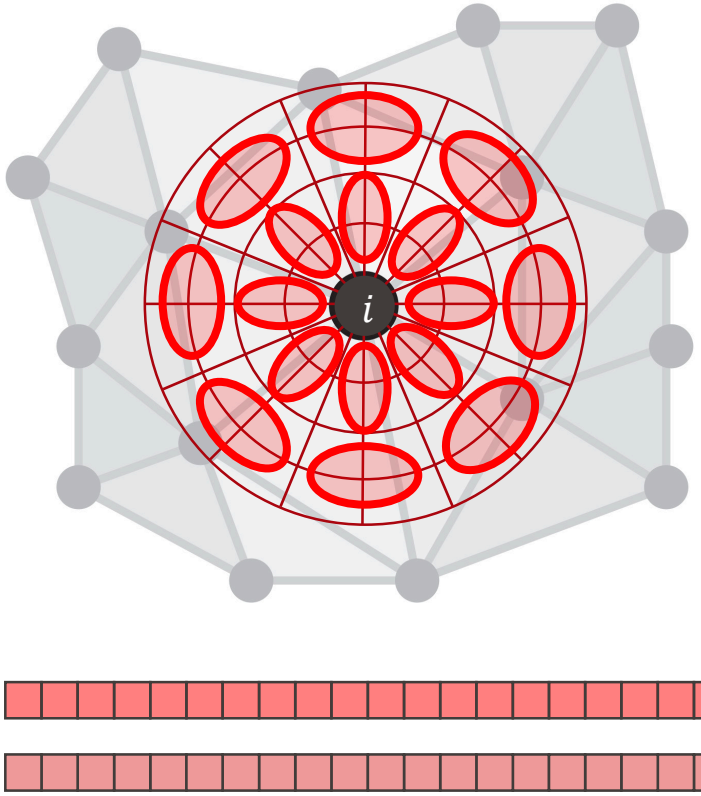
- **Vertex-wise d-dimensional features:** $n \times d$ matrix $\mathbf{X}$
- **Local geodesic polar coordinates** $\mathbf{u}_{ij}$ around i
- **Local weights** $w_1(\mathbf{u}), \dots, w_L(\mathbf{u})$ w.r.t.$\mathbf{u}$, e.g. Gaussians:

$$w_\ell(\mathbf{u}) = \exp\left(-(\mathbf{u} - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}_\ell^{-1} (\mathbf{u} - \boldsymbol{\mu}_\ell)\right)$$

  'soft pixels'

- **Spatial convolution** with filter g:

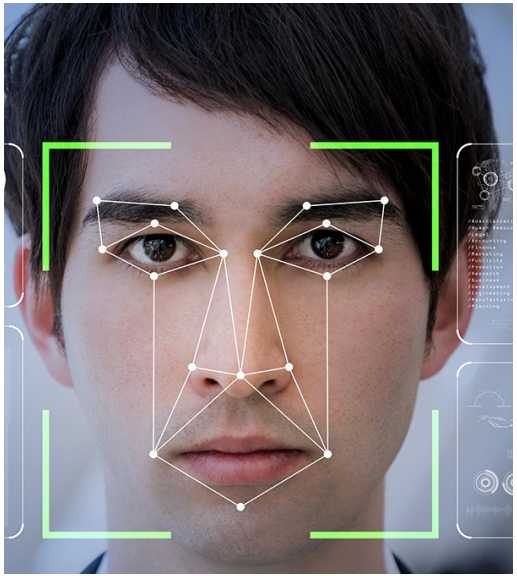
$$\mathbf{x}'_i = \frac{\sum_{\ell=1}^{L} g_\ell \sum_{j=1}^{n} w_\ell(\mathbf{u}_{ij}) \mathbf{x}_j}{\sum_{\ell=1}^{L} g_\ell \sum_{j=1}^{n} w_\ell(\mathbf{u}_{ij})}$$

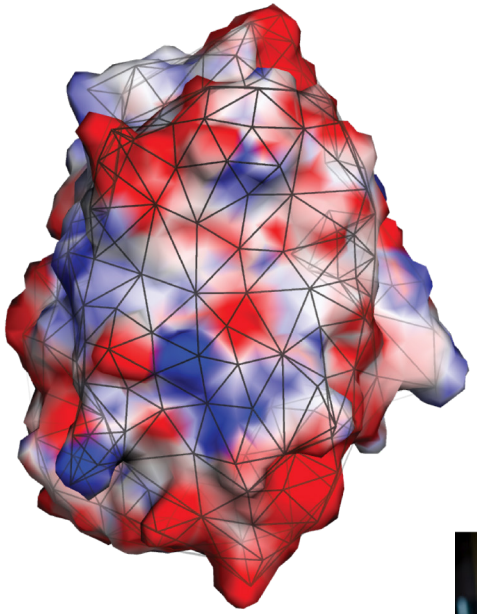


Monti et Bronstein 2017

The many (sur)faces of protein structures

Can we identify surface patterns that reveal functional features of proteins ?





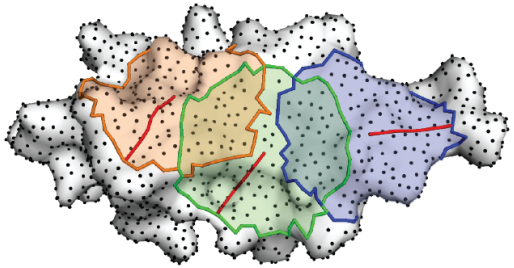
Gainza,…, **Correia**
Nature Methods, 2020

Pablo Gainza & Freyr Sverrisson

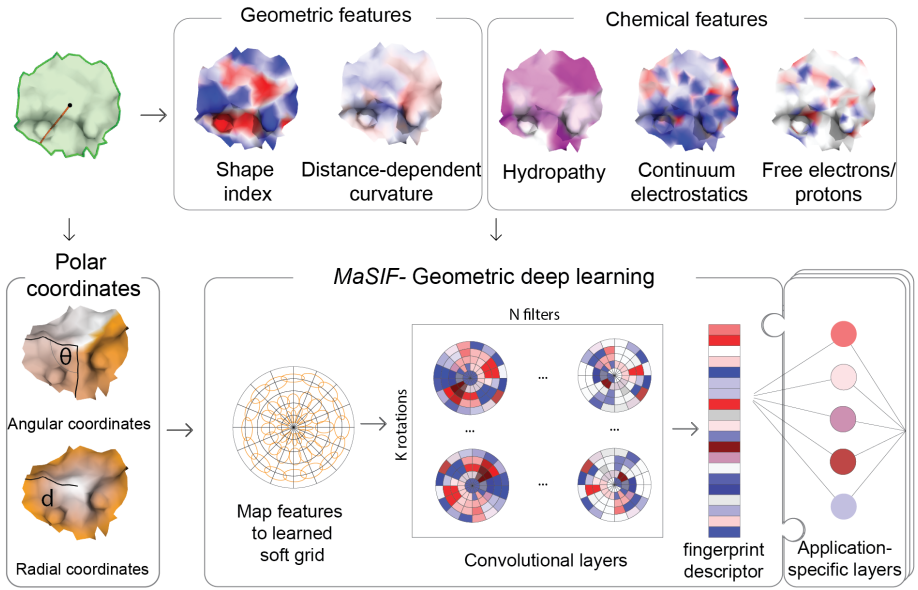
Molecular surface interaction fingerprints (MaSIF)

MaSIF – a framework to generate fingerprint descriptors (vectors) that encode surface features

Approach: systematic extraction of patches



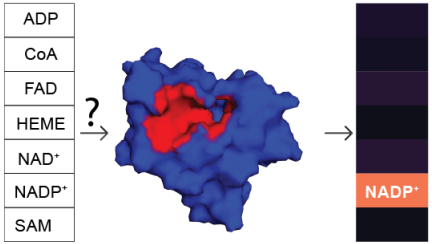
- Patch center points
- Patch radius



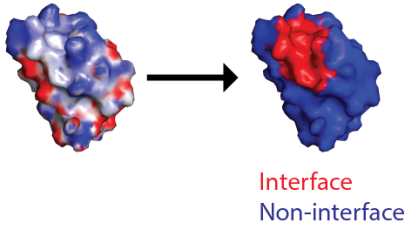
In collaboration with Michael Bronstein – Imperial College

Molecular surface interaction fingerprints (MaSIF)

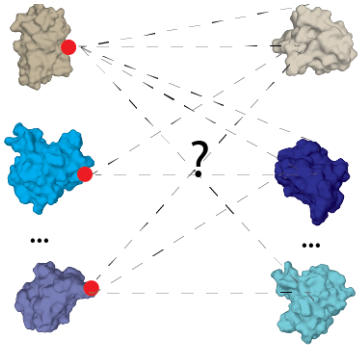
Applications



Pocket classification
MaSIF-ligand

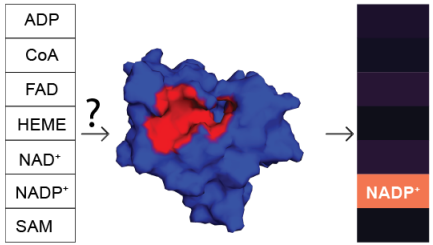


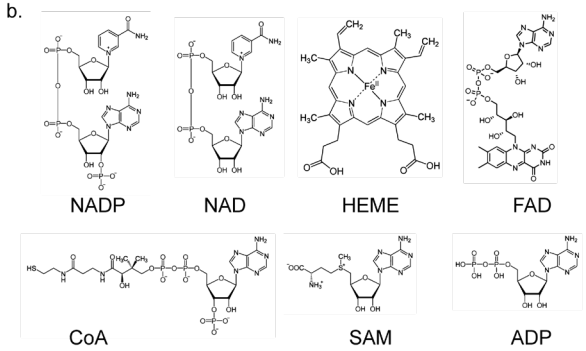
Interface site prediction
MaSIF-site



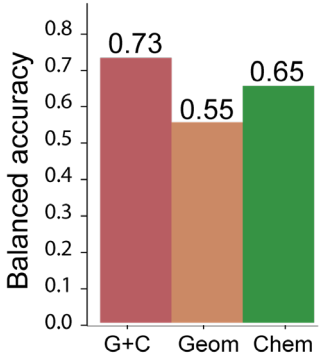
Ultra-fast PPI search
MaSIF-search

Pocket Classification with MaSIF

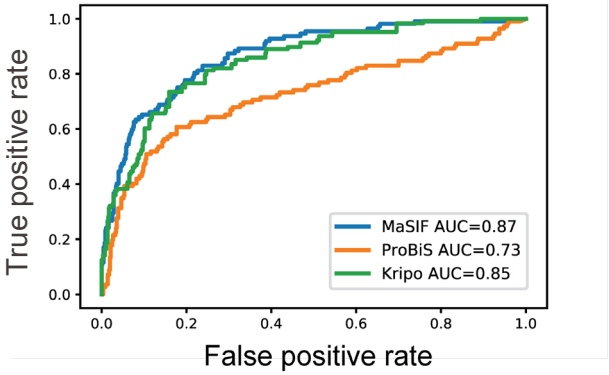




Performance
&
feature contribution

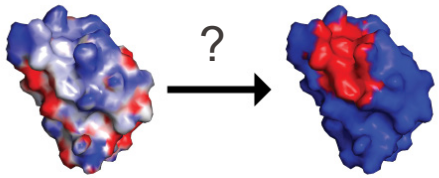


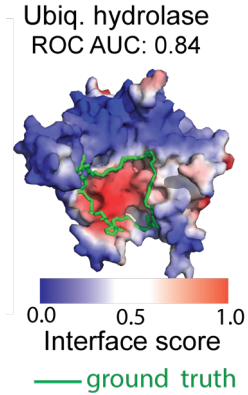
Comparison with other predictors



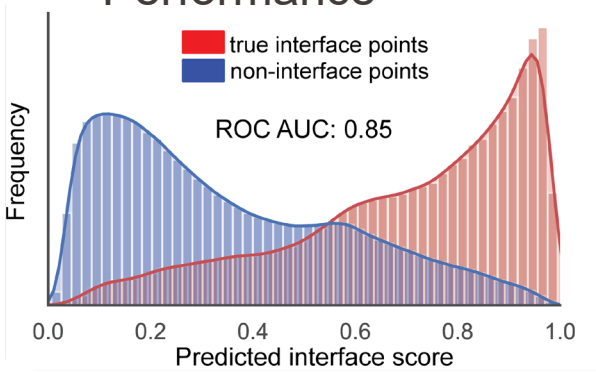
- MaSIF correctly classifies pockets of proteins independently of sequence identity.

Protein-protein interaction site prediction

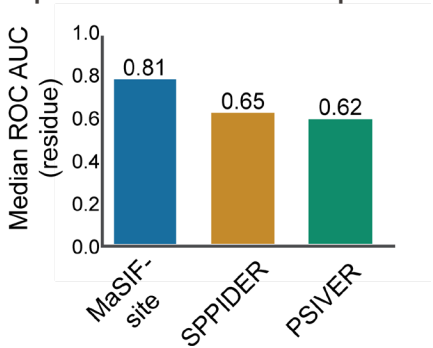




Performance

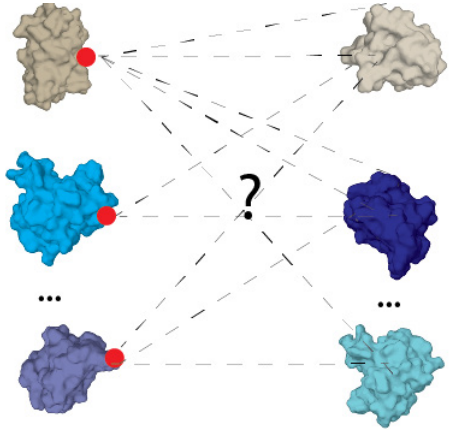


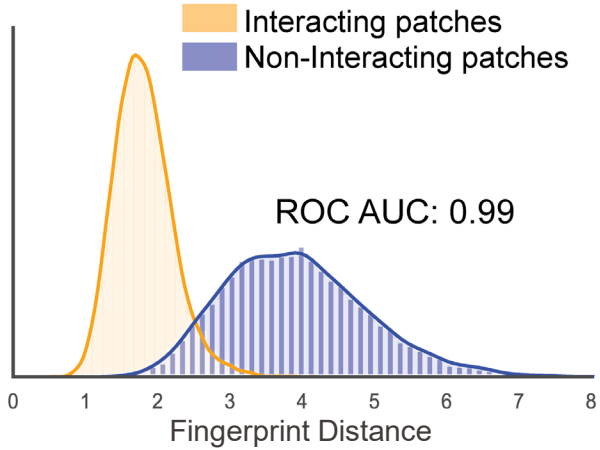
Comparison with other predictors



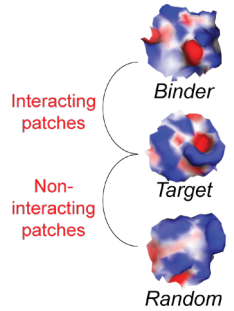
- MaSIF-site predicts PPI sites in the absence of the information of the binding partner.

Super-fast search of protein complexes (docking)



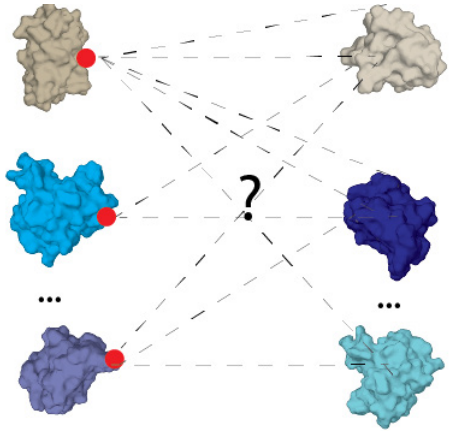


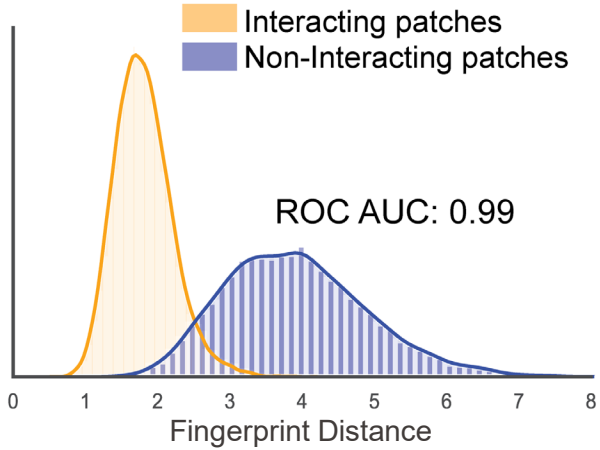
Fingerprint Comparisons



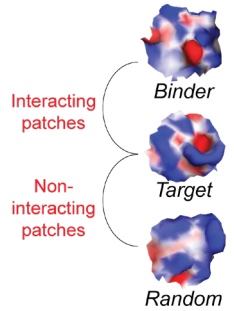
MaSIF-search finds true interacting patches with high accuracy

Super-fast search of protein complexes (docking)





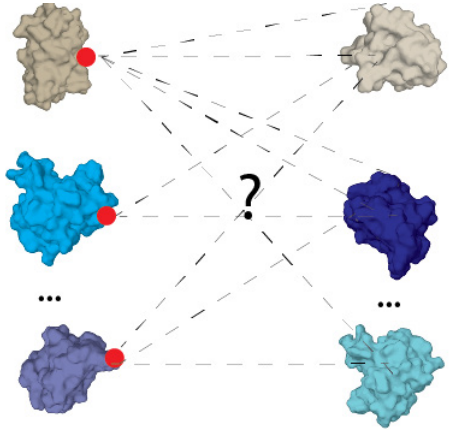
Fingerprint Comparisons



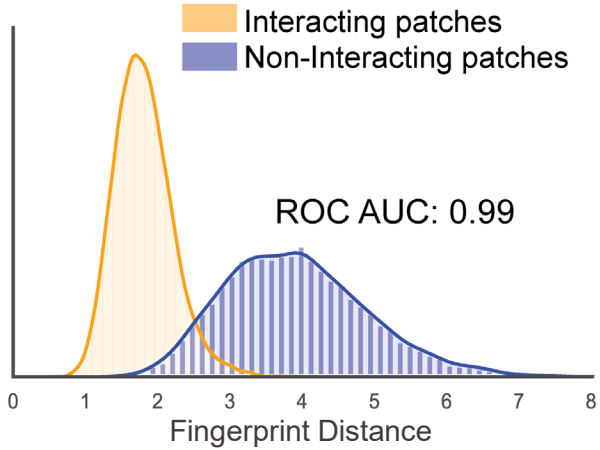
MaSIF-search finds true interacting patches with high accuracy

MaSIF-search workflow

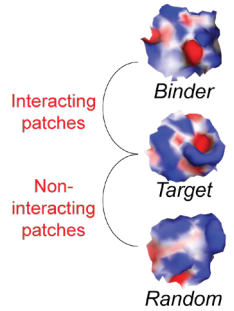
Super-fast search of protein complexes (docking)



Large-scale docking experiment
(100 targets all against all)



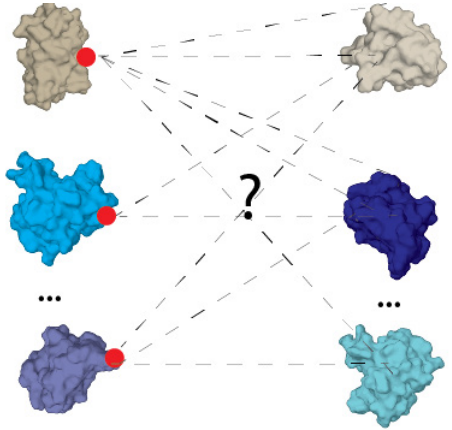
Fingerprint Comparisons



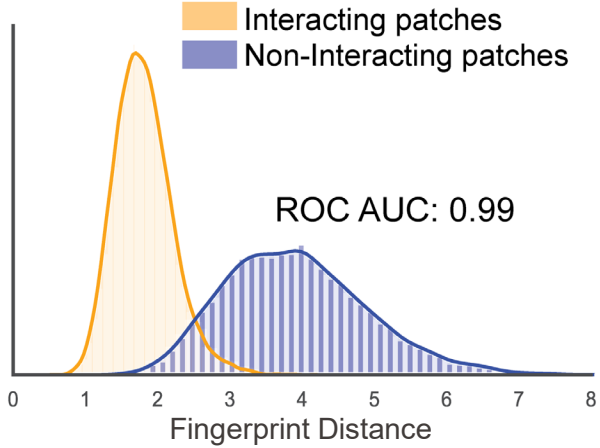
Bound Complexes

Method	# solved complexes in Top			Time (min)
	100	10	1	
PatchDock	40	29	21	2854
MaSIF-search Decoys = 3000	71	63	52	39

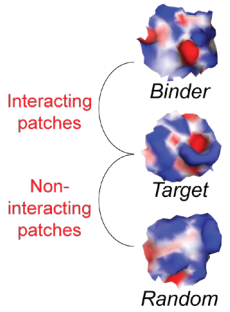
Super-fast search of protein complexes (docking)



Large-scale docking experiment
(100 targets all against all)



Fingerprint Comparisons



Bound Complexes

Method	# solved complexes in Top			Time (min)
	100	10	1	
ZDOCK+ZRANK2 Decoys = 10000	75	63	48	136066
MaSIF-search Decoys = 3000	71	63	52	39

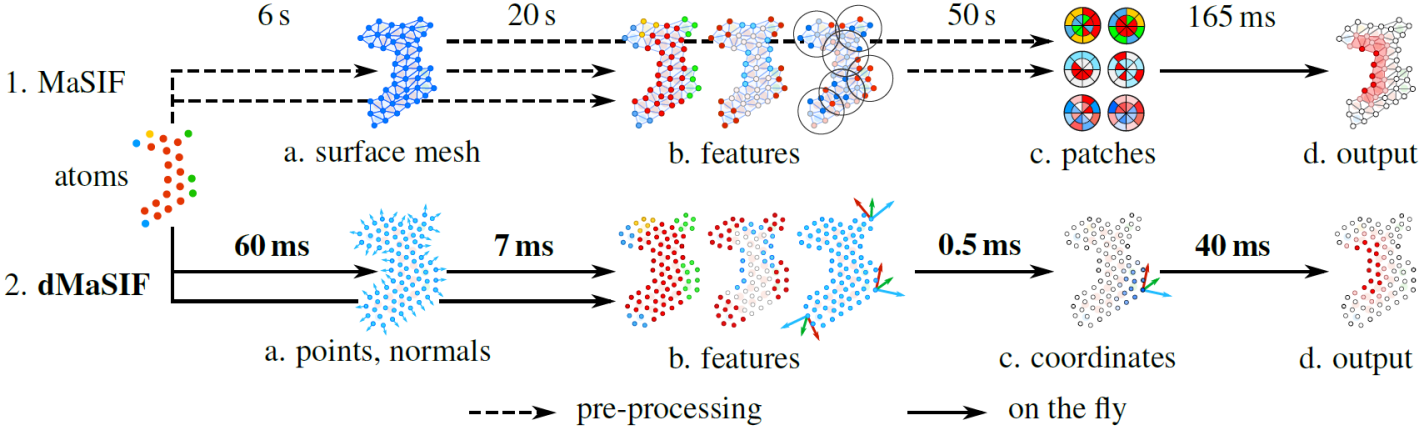
- MaSIF-search performs super-fast docking with similar performances to other programs

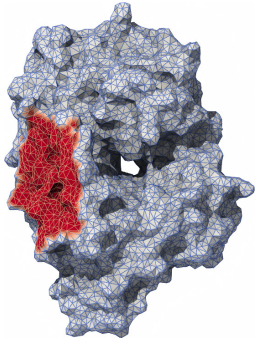
End-to-End MaSIF (dMaSIF)

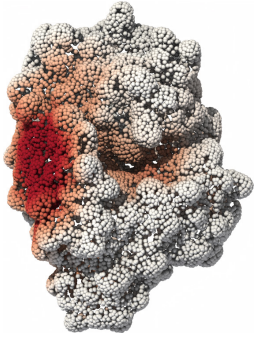
-MaSIF limitations:

Slow and high storage requirements

Pre-computation of handcrafted features







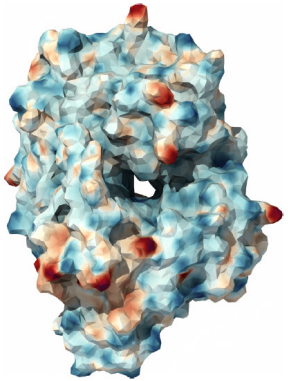
Freyr Sverrisson

Sverrisson, F., Feydy, J., Correia, B. E., & Bronstein, M. M. (2020). Fast end-to-end learning on protein surfaces. *bioRxiv*.

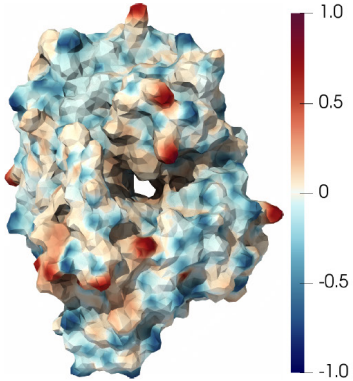
dMaSIF – Prediction of surface chemical features

Electrostatic potentials of the protein surface

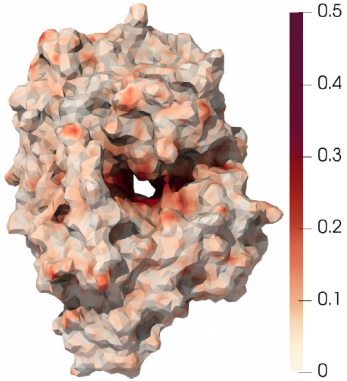
Ground truth



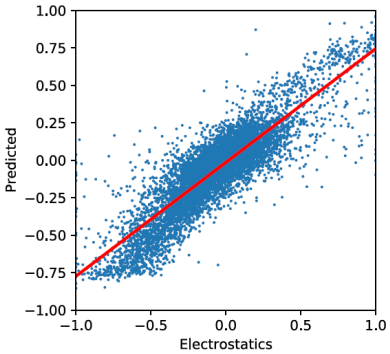
dMaSIF prediciton



Error



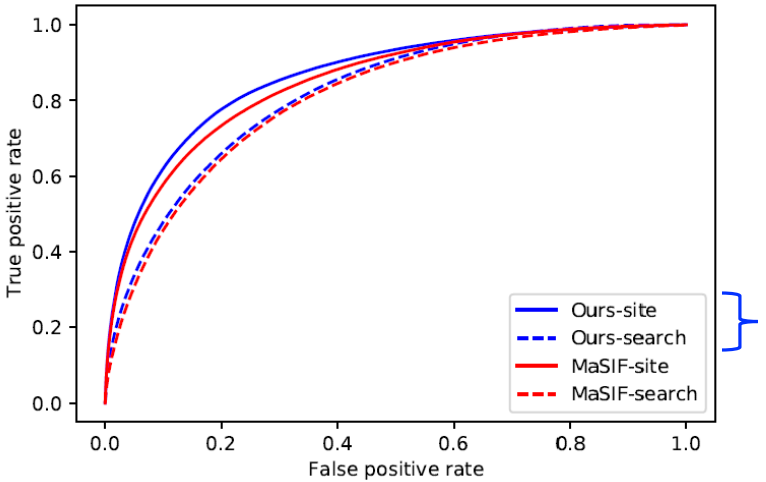
Correlation



Sverrisson, F., Feydy, J., Correia, B. E., & Bronstein, M. M. (2020). Fast end-to-end learning on protein surfaces. *bioRxiv*.

dMaSIF – Performance

-Currently the results are equivalent to the initial MaSIF architecture

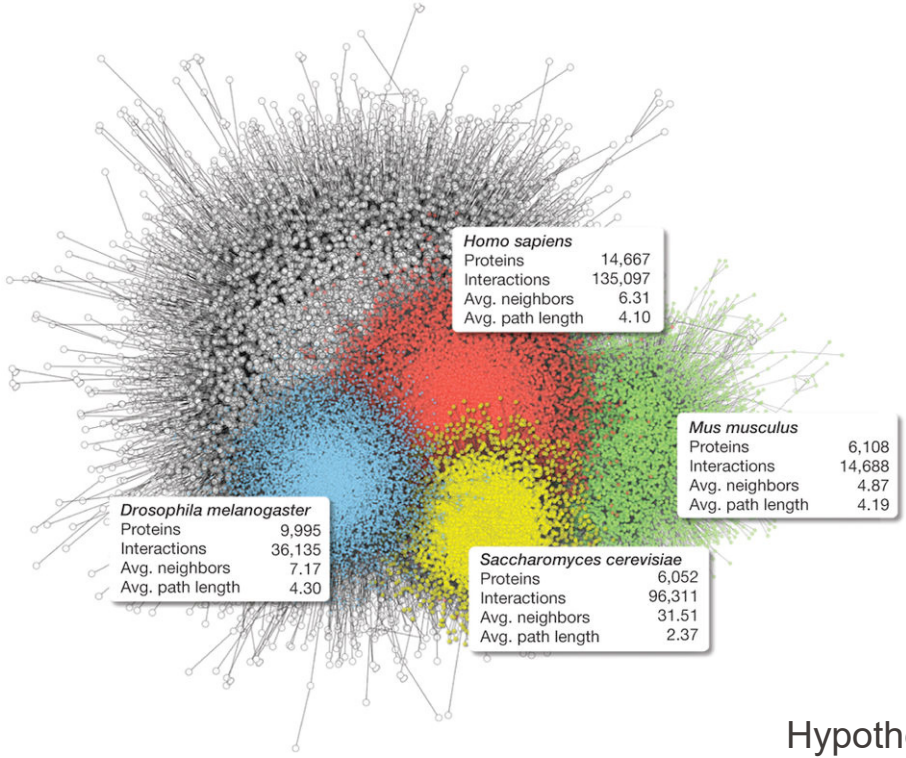


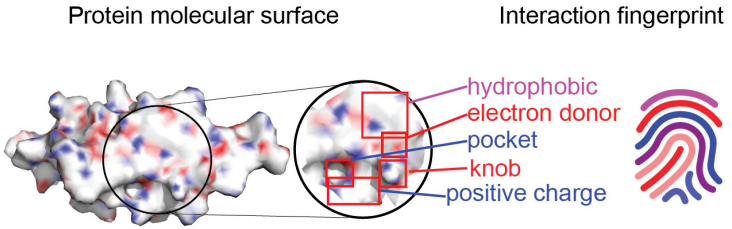
dMaSIF

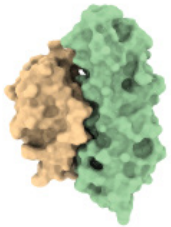
-These technical improvements will be critical for problems related to protein flexibility and design

Sverrisson, F., Feydy, J., Correia, B. E., & Bronstein, M. M. (2020). Fast end-to-end learning on protein surfaces. *bioRxiv*.

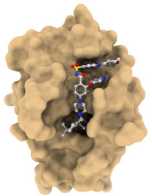
Future perspectives - Deciphering fingerprints on interactomes







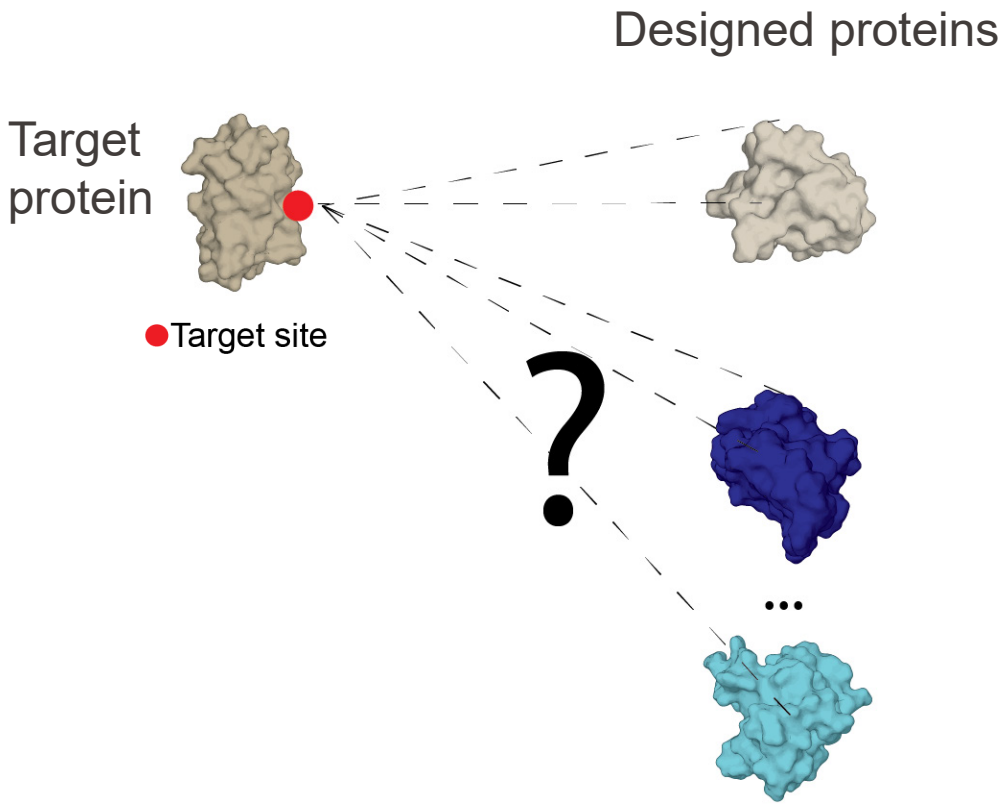
Protein-protein interactions



Protein-metabolite interactions

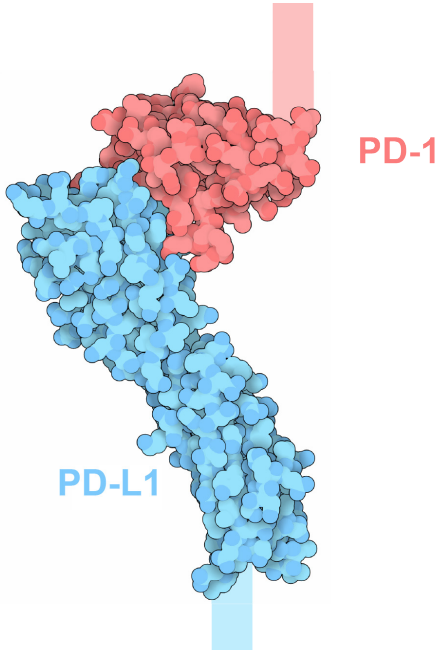
Hypothesis: Proteins that perform similar interactions may display similar '**fingerprints**' *regardless of their evolutionary history*

De novo design of protein interactions – an unsolved problem

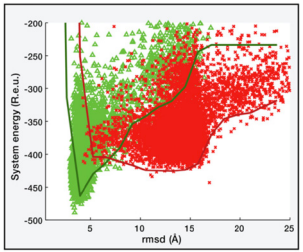


- Aim: One-sided design to bind to a specific site in a protein target

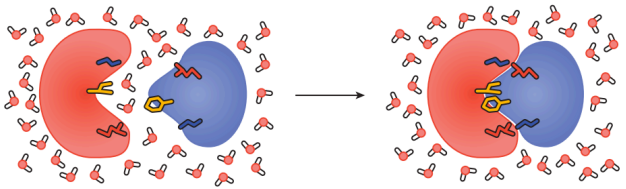
Challenges in designing computationally de novo PPIs



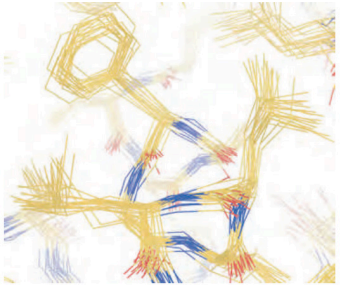
I) Empirical scoring functions lack the accuracy for proper discrimination



II) Solvent absent



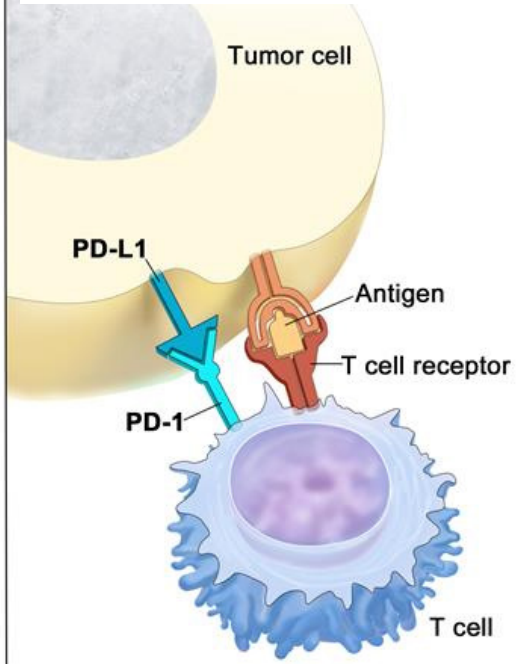
III) Dynamics absent



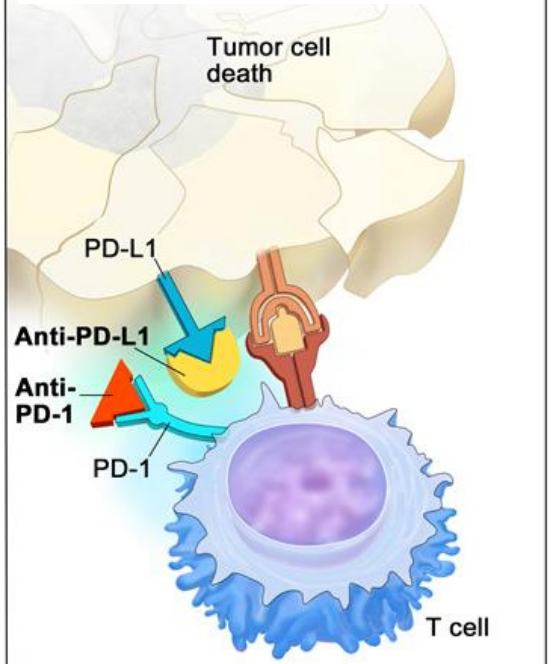
Example: Binder design for cancer immunotherapy target

2018 Nobel Prize
PD-proteins role in immunotherapy

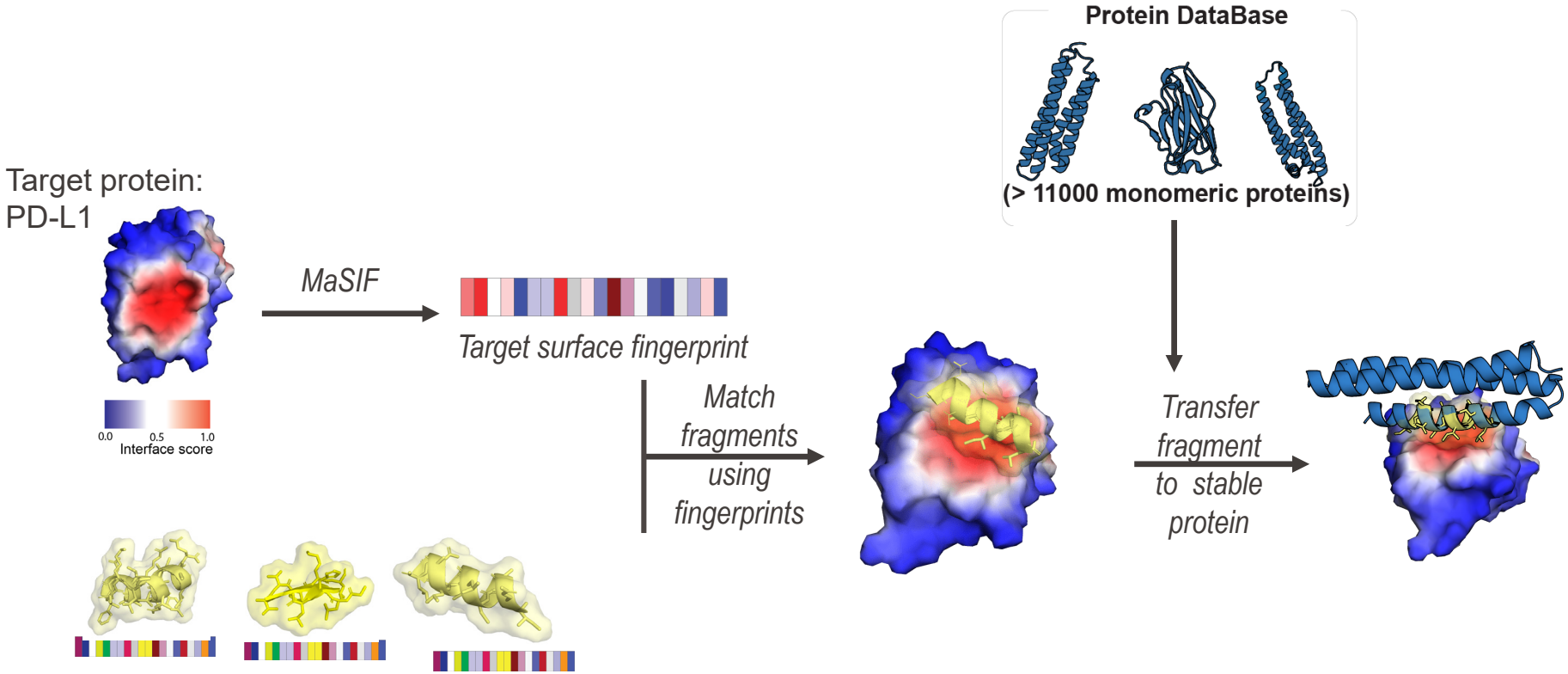
PD-L1 binds to PD-1 and inhibits T-cell killing of tumor cell



Blocking PD-L1 or PD-1 allows T-cell killing of tumor cell



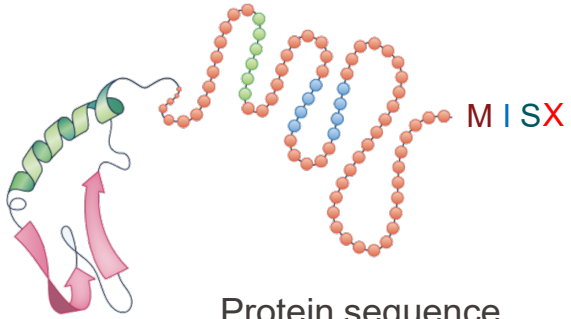
MaSIF – De novo design of PPIs



- Fragment DB : > 100M fragments

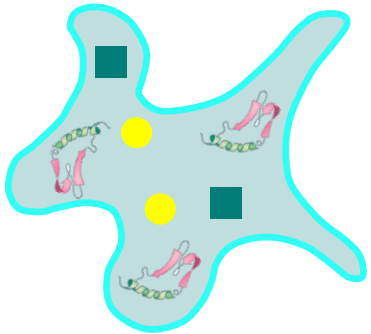
w/ F. Sverrisson

Testing new molecules in the lab



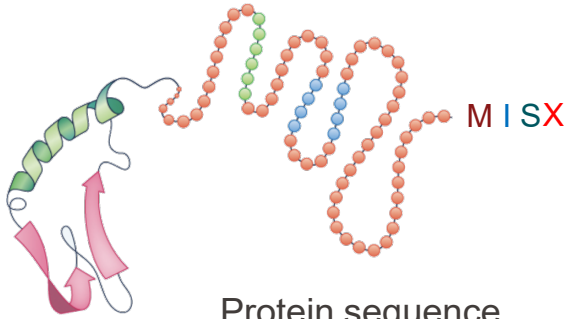
Protein sequence

DNA strand



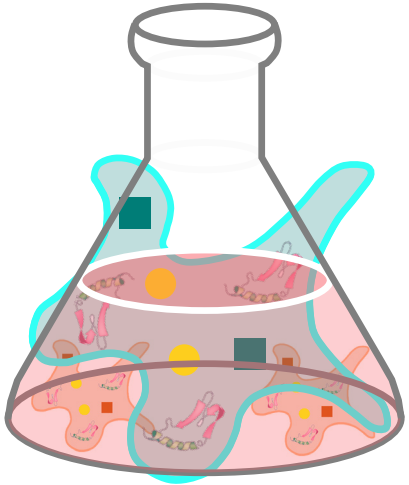
Cell

Testing new molecules in the lab

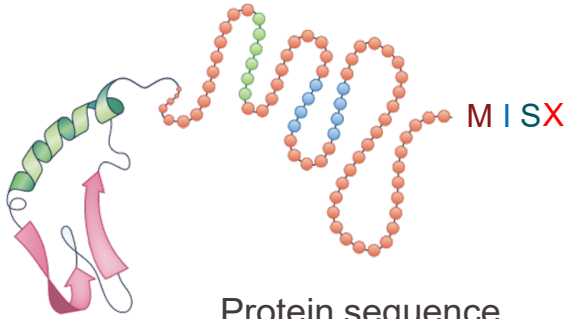


Protein sequence

DNA strand

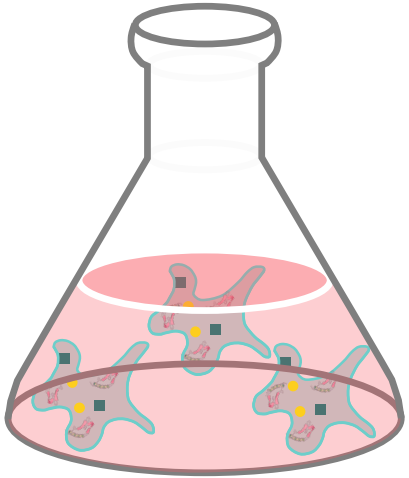


Testing new molecules in the lab



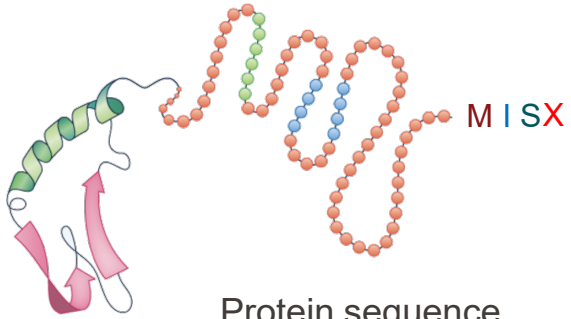
Protein sequence

DNA strand



Cell lysis

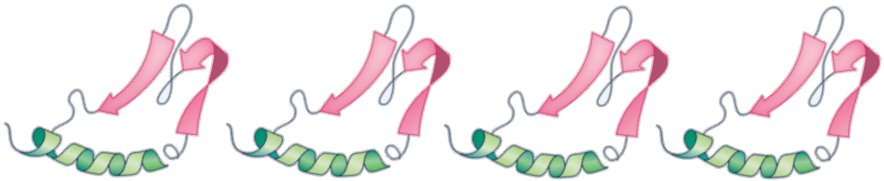
Testing new molecules in the lab

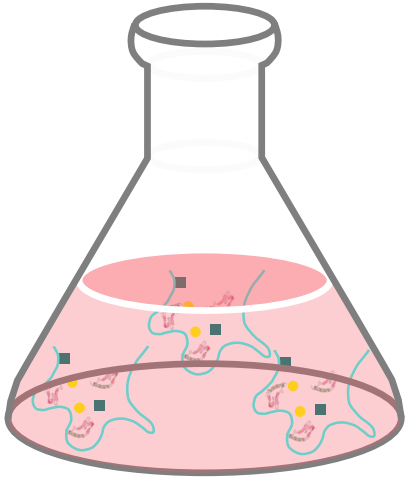


Protein sequence

DNA strand

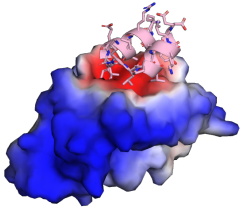
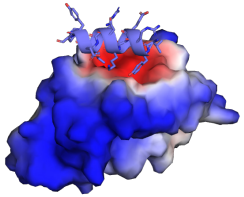
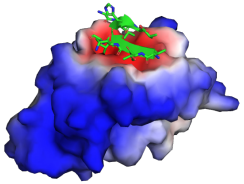
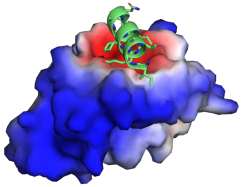
Synthesized protein



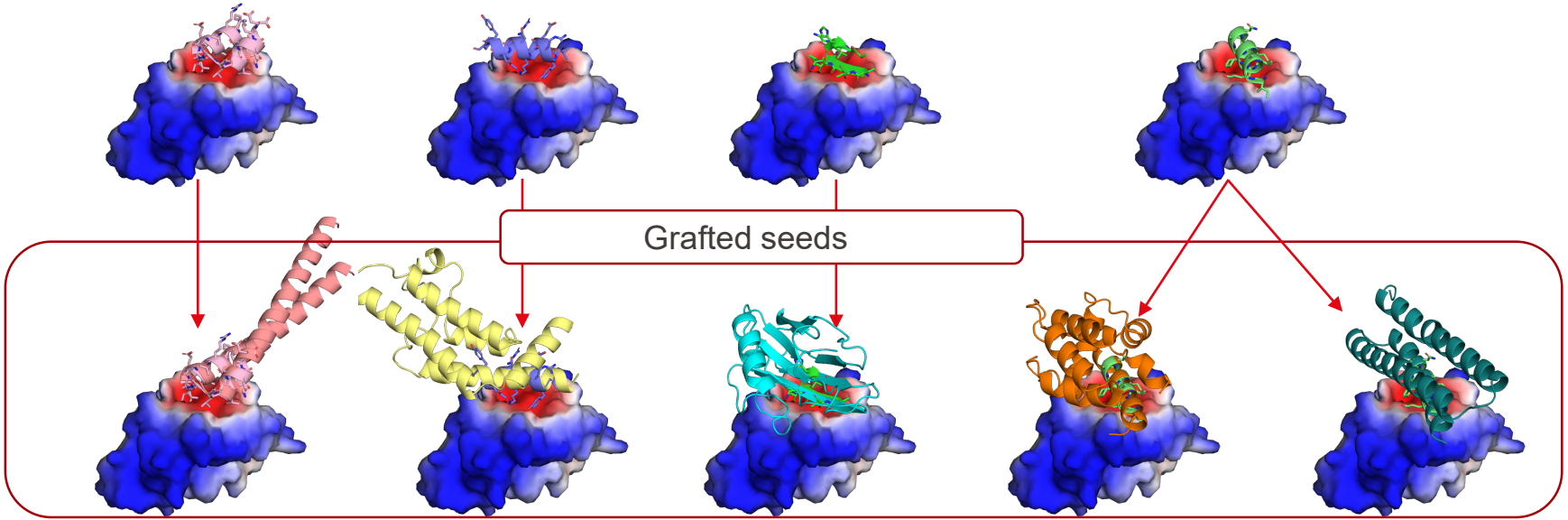


Protein purification

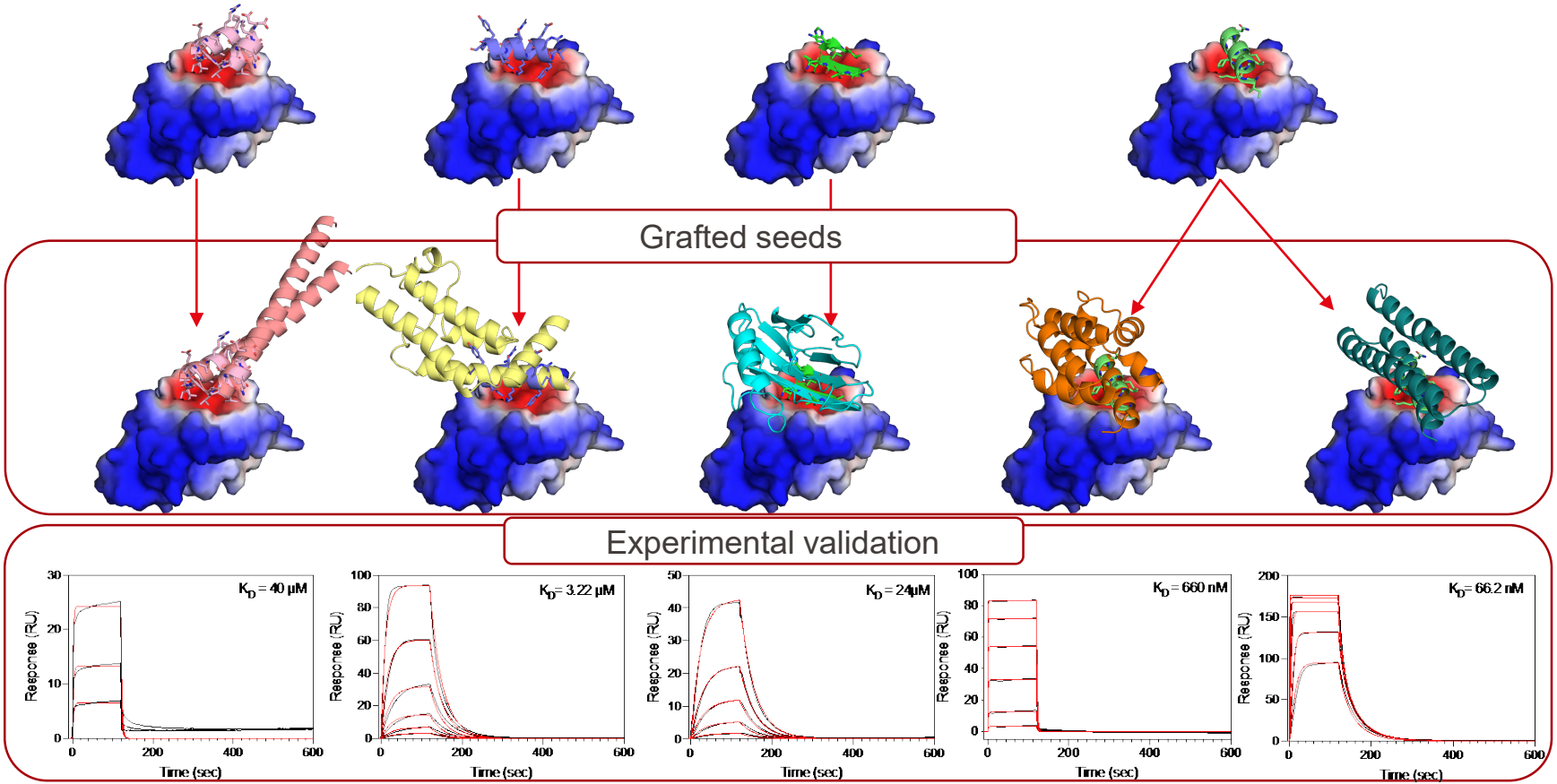
PD-L1 Test Case

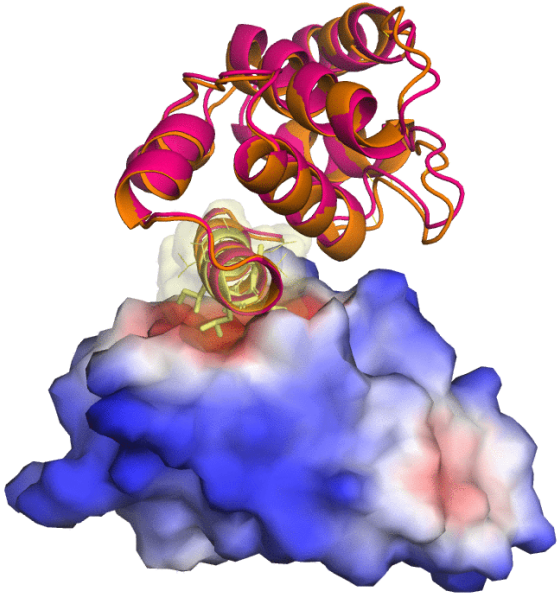
PD-L1 Test Case



PD-L1 Test Case



Structural validation of computationally designed binder



PD-L1

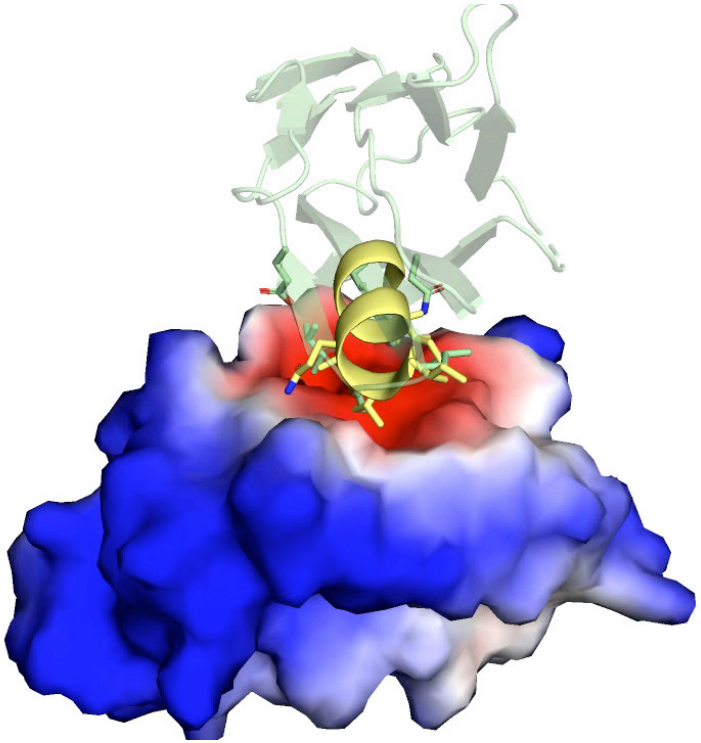
Binder Xtal structure
Binder model
Seed model

Whole complex alignment: 0.77 Å

- Computational model and experimental x-ray structure are in agreement at atomic level

w/ S Wehrle, S Tan, G Gao

MaSIF uncovers binding motifs distinct from native ligands

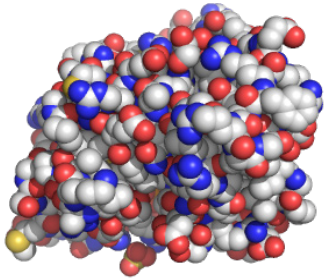
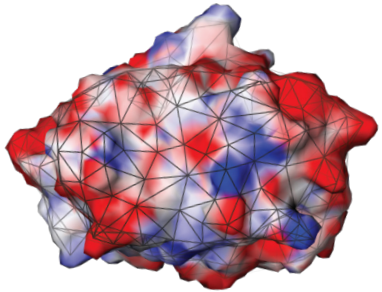
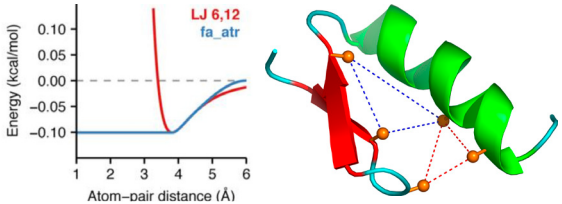
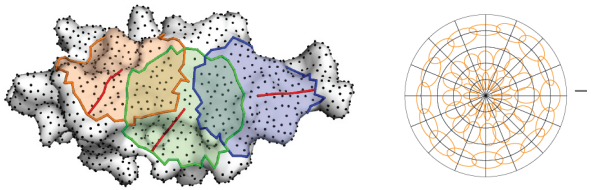


PD-L1

PD1 (Native binder)
Seed model

-Hot-spot residues do not resemble the interactions present in the native ligand

Distinctive points in our modeling framework

	State of the art	MaSIF
Molecular Representation		
Scoring Scheme		
	▪ residue pairwise interactions ▪ pre-defined physical potentials	▪ operates at the patch level ▪ task-specific learned potentials

▪

Conclusions and Future Work

- Vector fingerprints reveal functional signatures from protein structures (independent of sequence evolutionary data)

- Identification of interaction fingerprints for small-molecules and proteins (critical for function)

- Fingerprint-base comparisons enable ultra fast docking simulations (unbound docking largely unsolved)

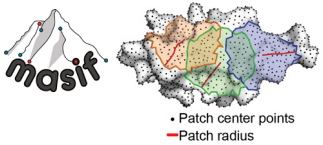
- Generation of protein binders straight of the computational stage (μM range)

- One of the designed binders is in close agreement with the xtal structure

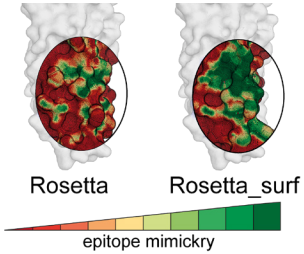
Outlook

Methods

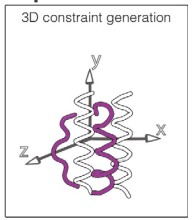
Surface features

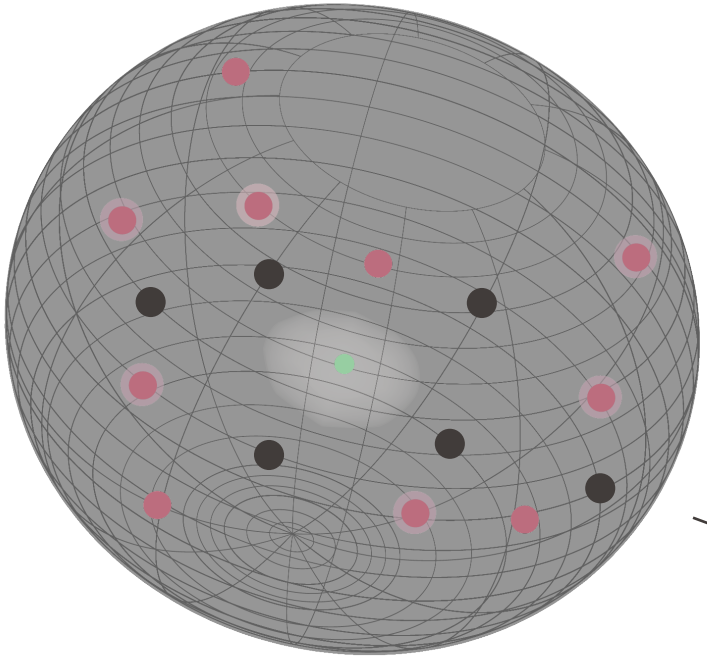


Surface-centric design



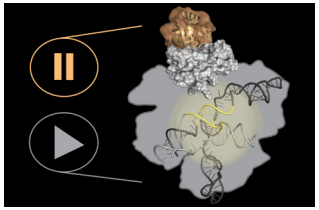
TopoBuilder



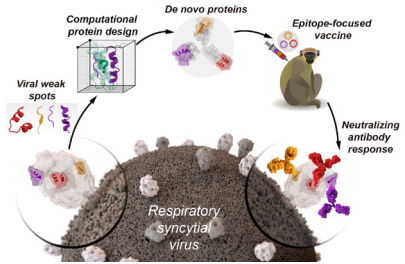


Searching for the functional sequences
in an immense space of possibilities

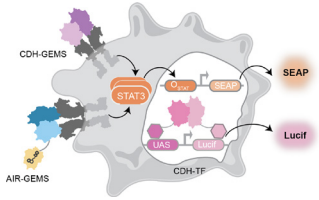
Applications



Genome Engineering

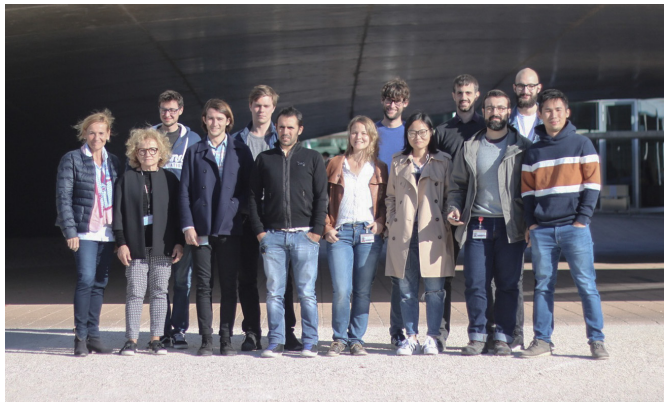


Vaccines



Synthetic Cells

Acknowledgements



FNSNF Swiss National Science Foundation
erc European Research Council Established by the European Commission
Catalyze4Life
SystemsX.ch The Swiss Initiative in Systems Biology
EPFL Fellows
NOVARTIS

LPDI @ EPFL
-Andreas Scheck
-Sarah Wehrle
-Sailan Shui
-Zander Harteveld
-Stéphane Rosset
-Sandrine Georgeon
-Fryer Sverrisson
-Karla Castro
-Leo Scheller
-Anthony Marchand
-Alexandra Beauvais
-Max Jensen

Alumni
-Sabrina Vollers
-Jaume Bonet
-Che Yang
-Fabian Sesterhenn
-Pablo Gainza
-Anastassia Voborieva

EPFL
-Francesco Stellacci
-Li Tang
-Elisa Oriccio
-Beat Fierz
-Pierre Vandergheynst
-Joerg Hulsken
-Hilal Lashuel

Switzerland
-George Coukos (UNIL)
-Martin Fussenegger (ETHZ)
-Tom Ward (UniBas)

Worldwide
-Michael Bronstein (Imperial, UK)
-Marteen Merkx (TUE, NL)
-Sabine Riffault (INRA,FR)
-Barney Graham (NIH, USA)
-Ted Jardetzky (Stanford, USA)
-JP Julien (UToronto, CN)
-Thomas Key (Lubeck, DE)
-Yuxing Li (Maryland, USA)