Computational protein design using Geometric Deep Learning

Michael Bronstein

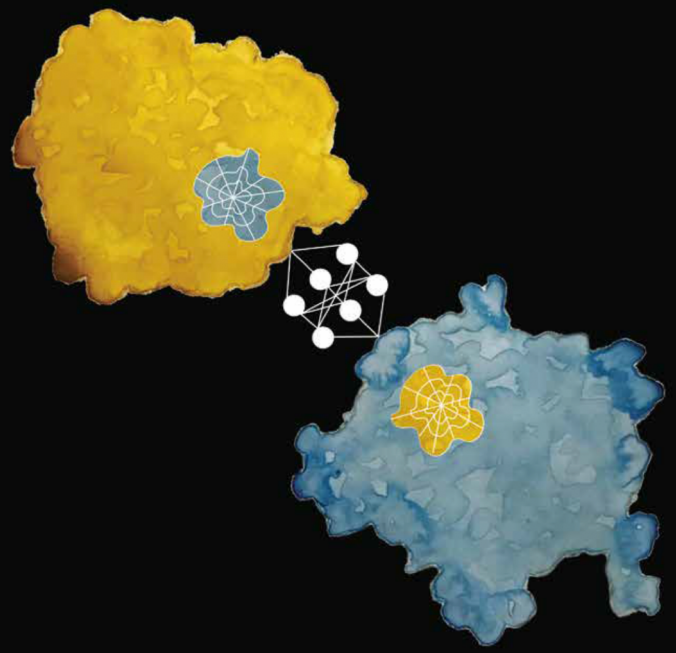
Imperial College London
United Kingdom

Twitter
United Kingdom

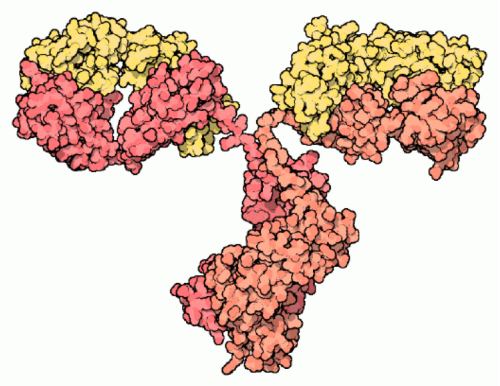
Joint work with B. Correia, P. Gainza, F. Sverisson (EPFL), F. Monti (USI/Twitter), E. Rodolà (Sapienza)

nature.com/nmeth February 2020 Vol. 17 No. 2

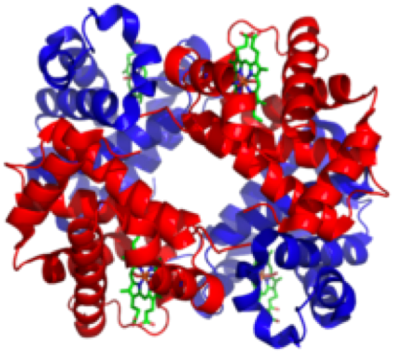
nature methods



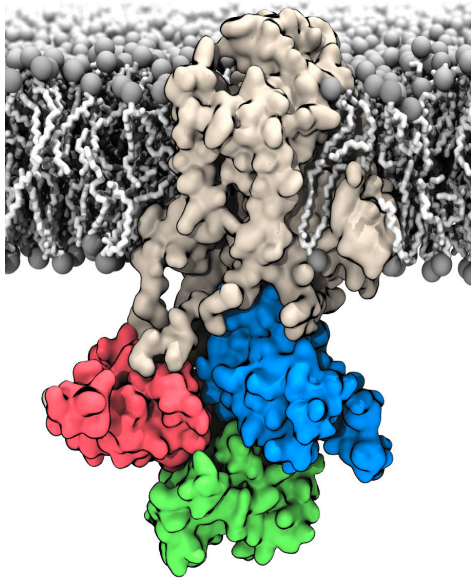
Main cover line here
Localization microscopy twice as precise
A cryo-EM-based structural proteomics approach
Time-resolved crystallography at the European XFEL
Magnetic resonance at high speed



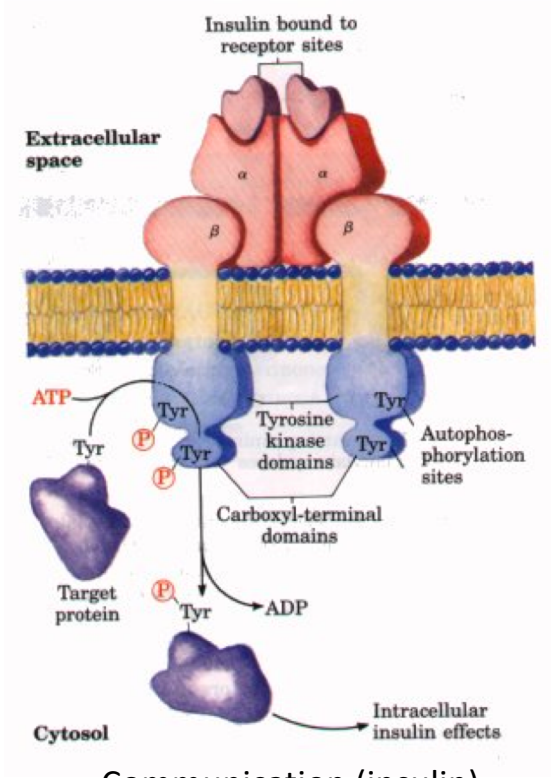
Defense (antibody)



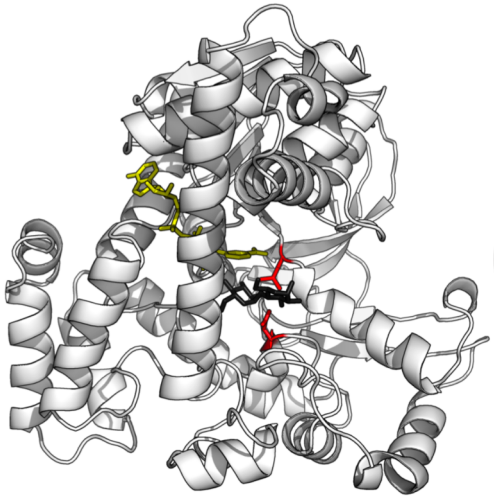
Storage (haemoglobin)

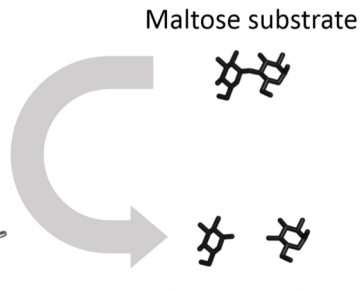


Transport (calcium pump)



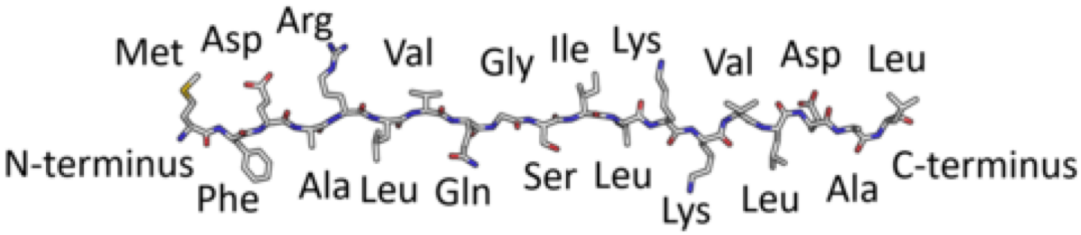
Communication (insulin)

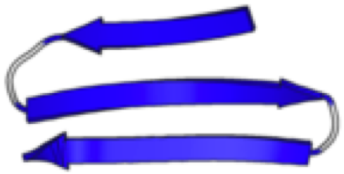




Catalysis (enzyme)

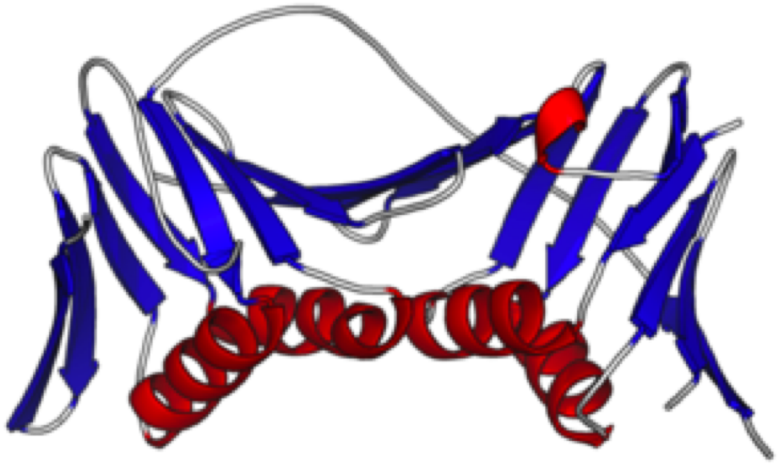
Structure (collagen)

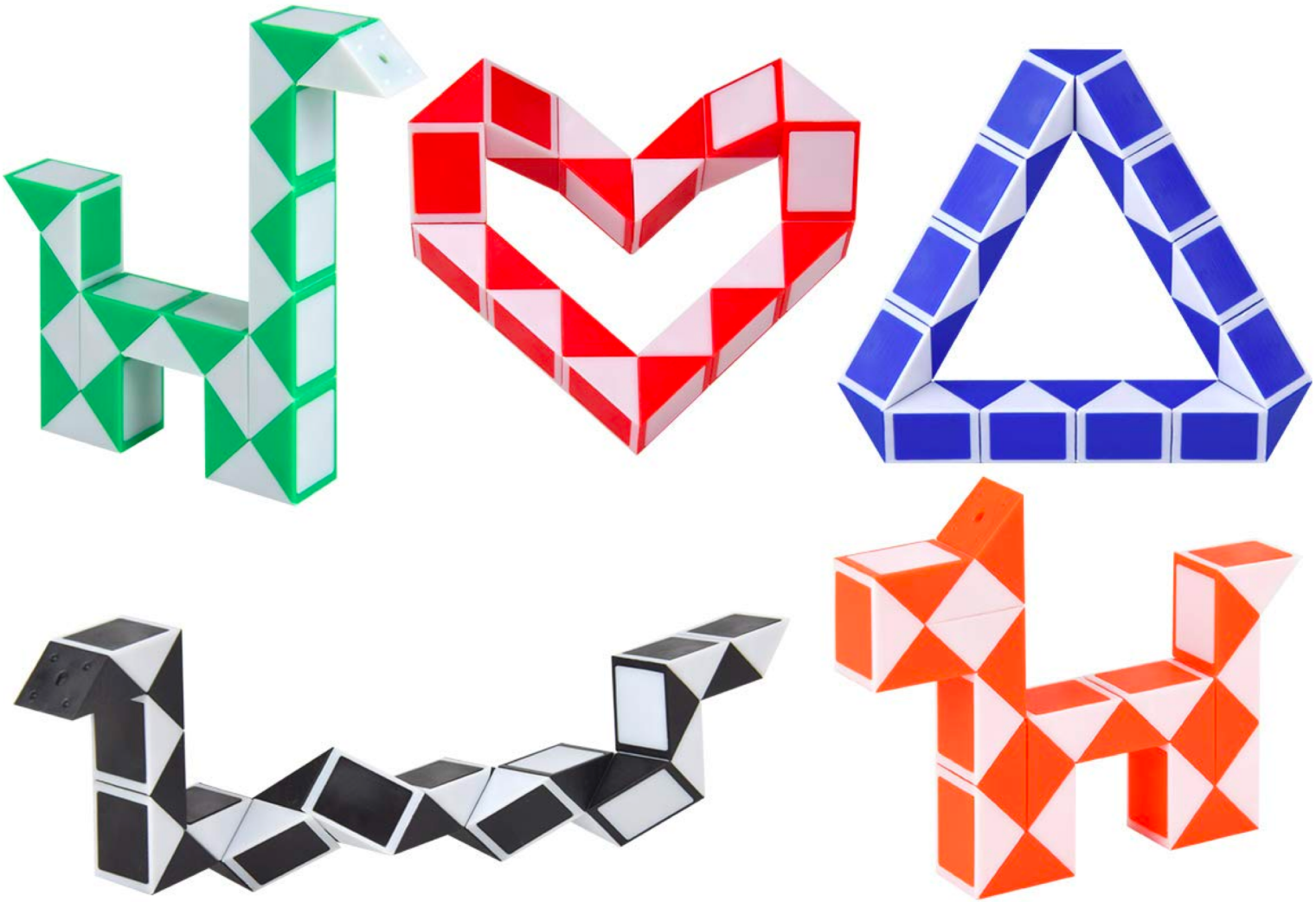




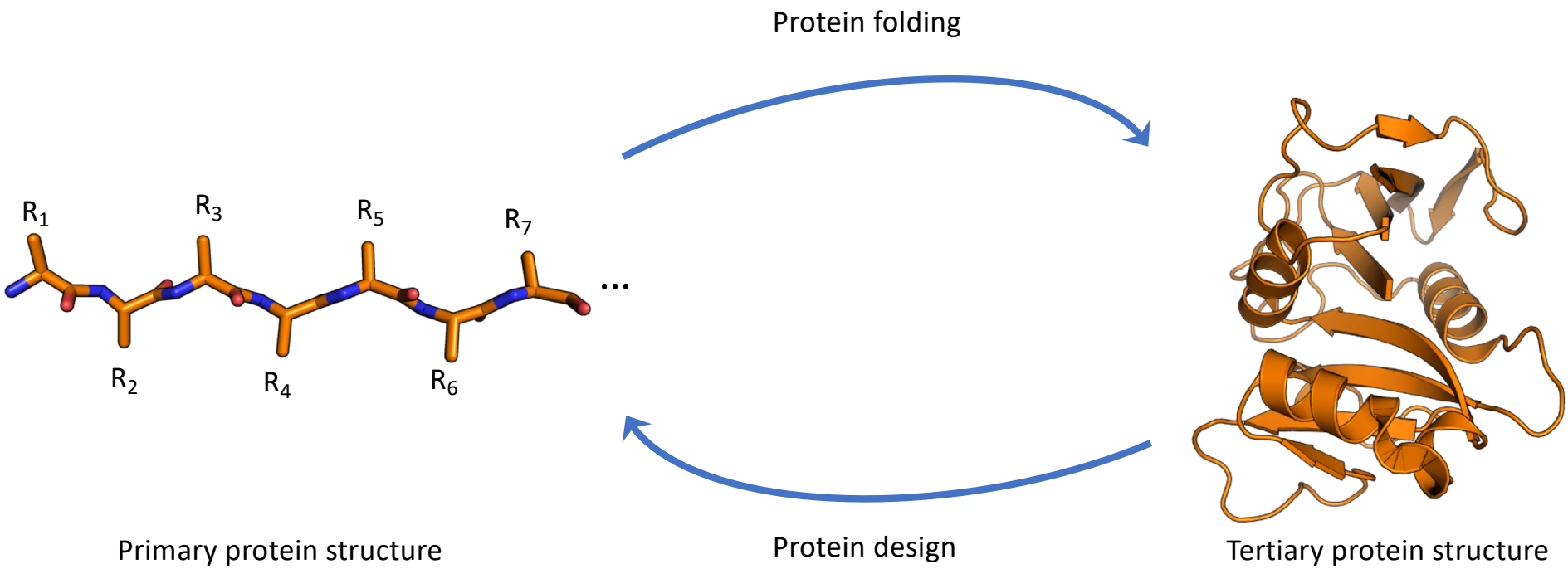
β-Sheet (3 strands)

α-helix

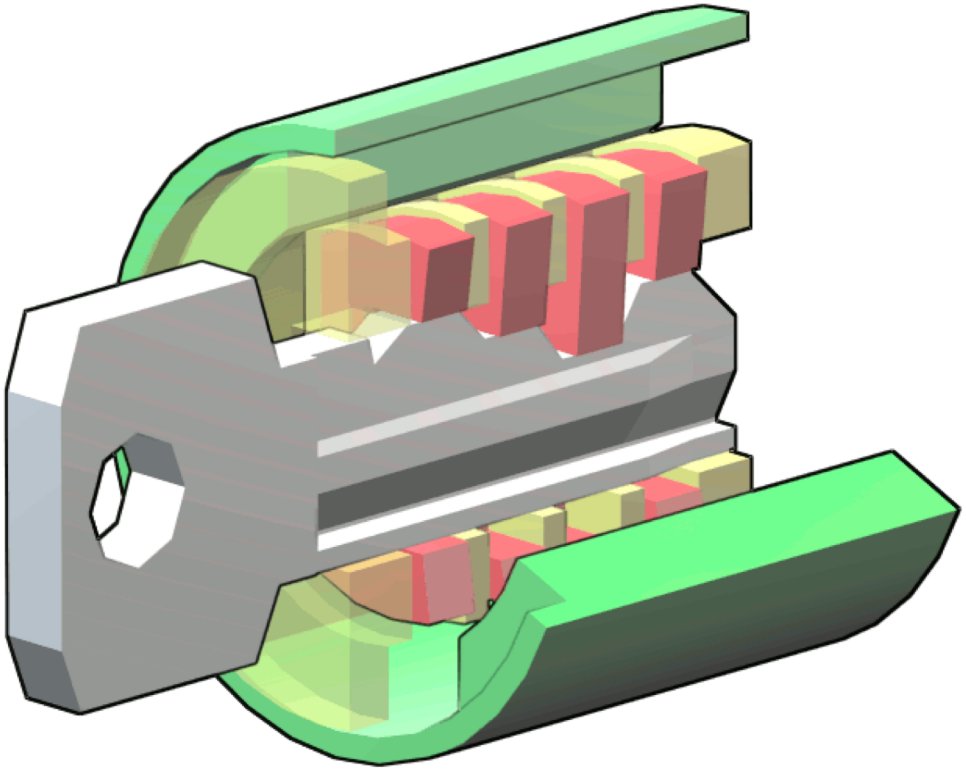




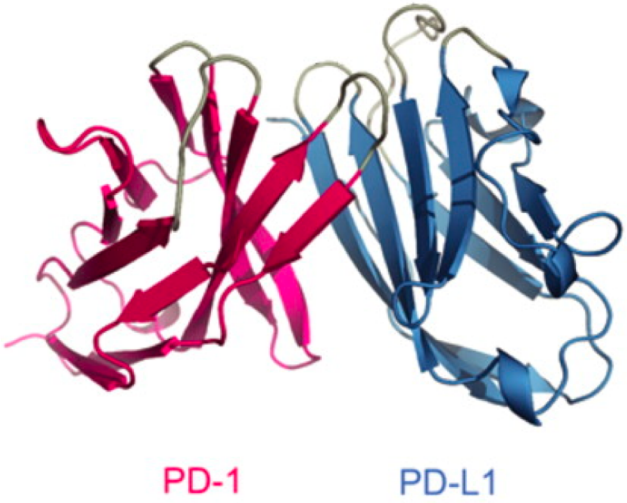
Protein design = inverse folding



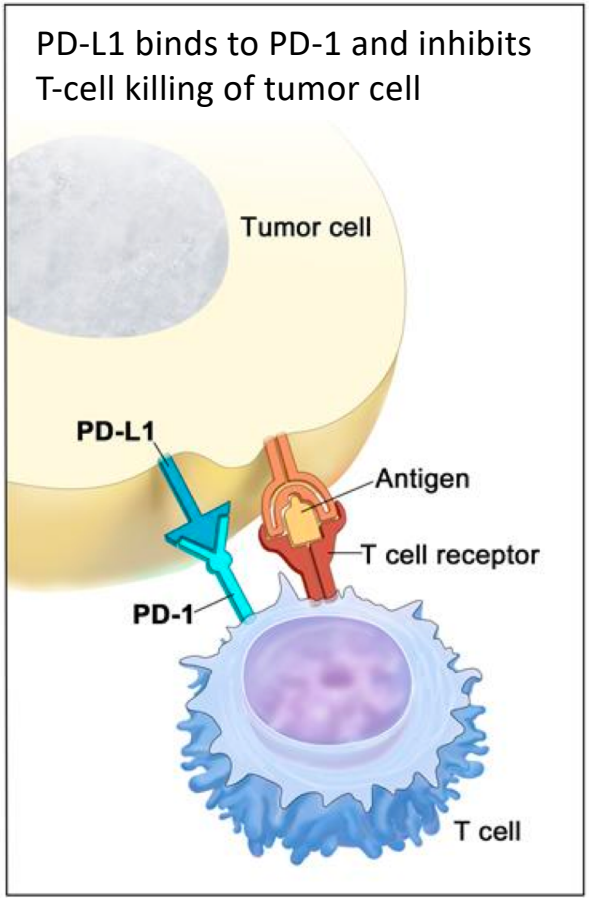
Lock-key model

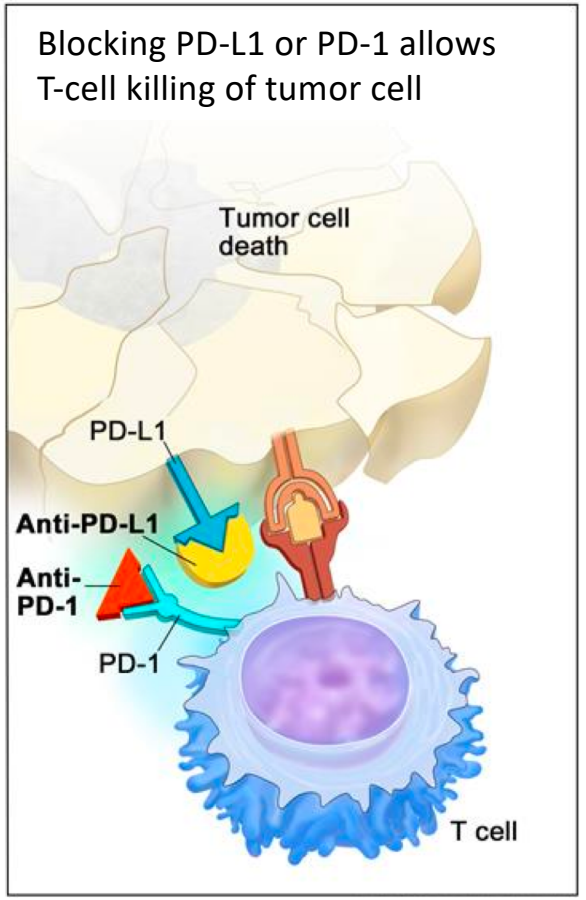


Applications: cancer immunotherapy

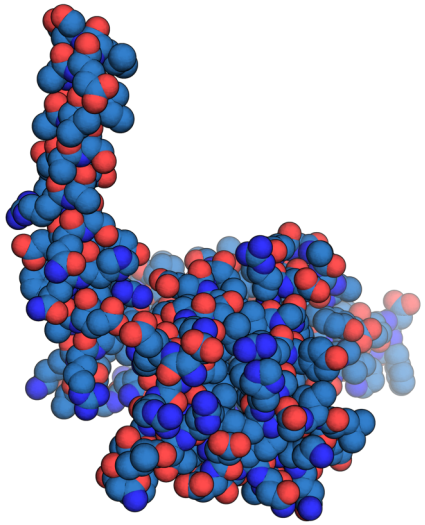


2018 Nobel Prize
PD-proteins role in
immunotherapy

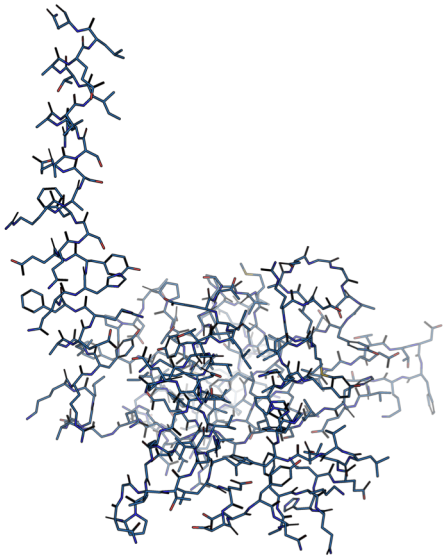




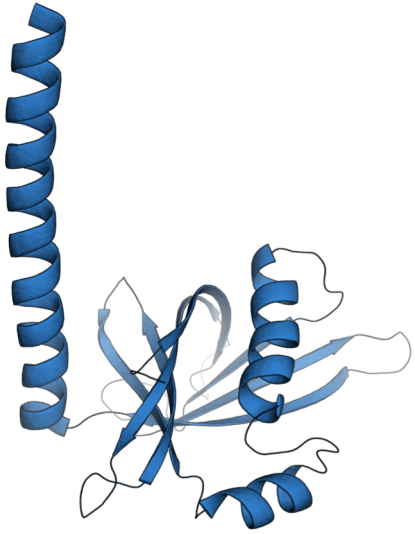
Representation



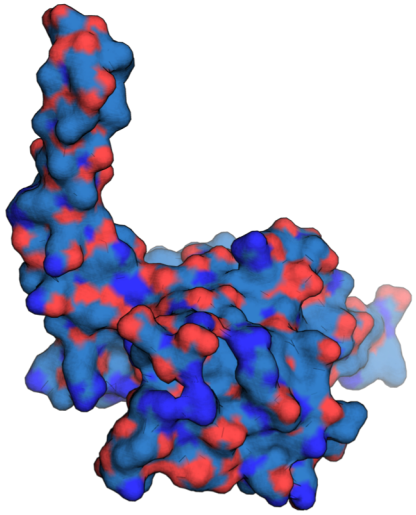
Atom point cloud



Graph



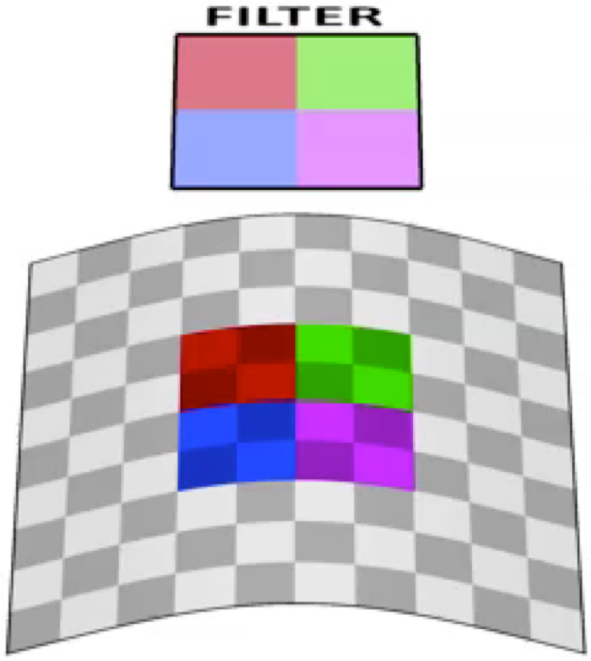
Secondary structures



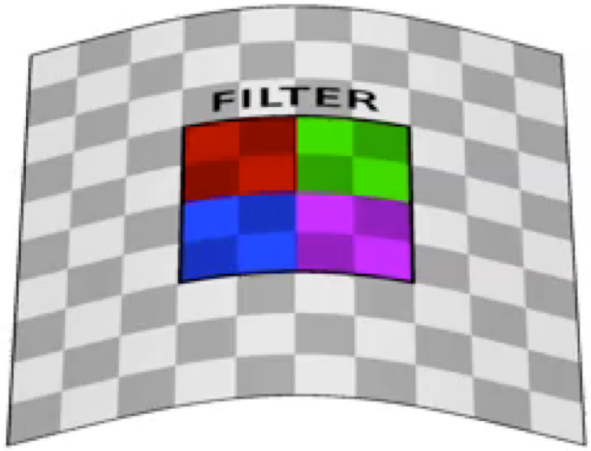
Molecular surface

Geometric deep learning on surfaces

Intrinsic vs extrinsic convolution



Extrinsic



Intrinsic

Masci et al. 2015

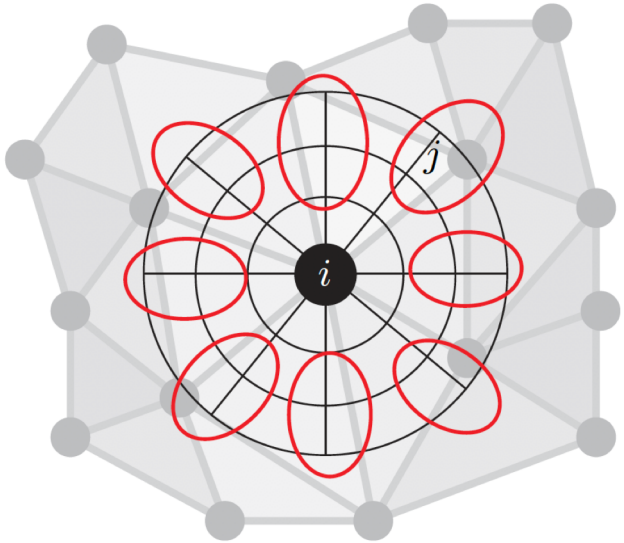
MoNet architecture

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

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

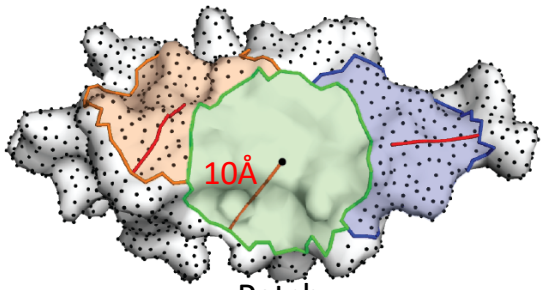
- **Spatial convolution** with filter g:

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



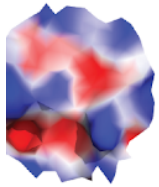
Monti, Boscaini, Masci, Rodolà, Svoboda, B 2017

Molecular surface interaction fingerprinting (MaSIF)

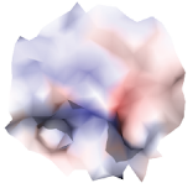


Patch

Geometric features



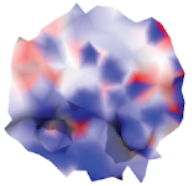
Shape index



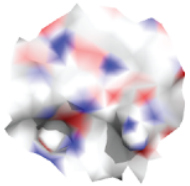
Distance-dependent curvature

Chemical features

Hydropathy



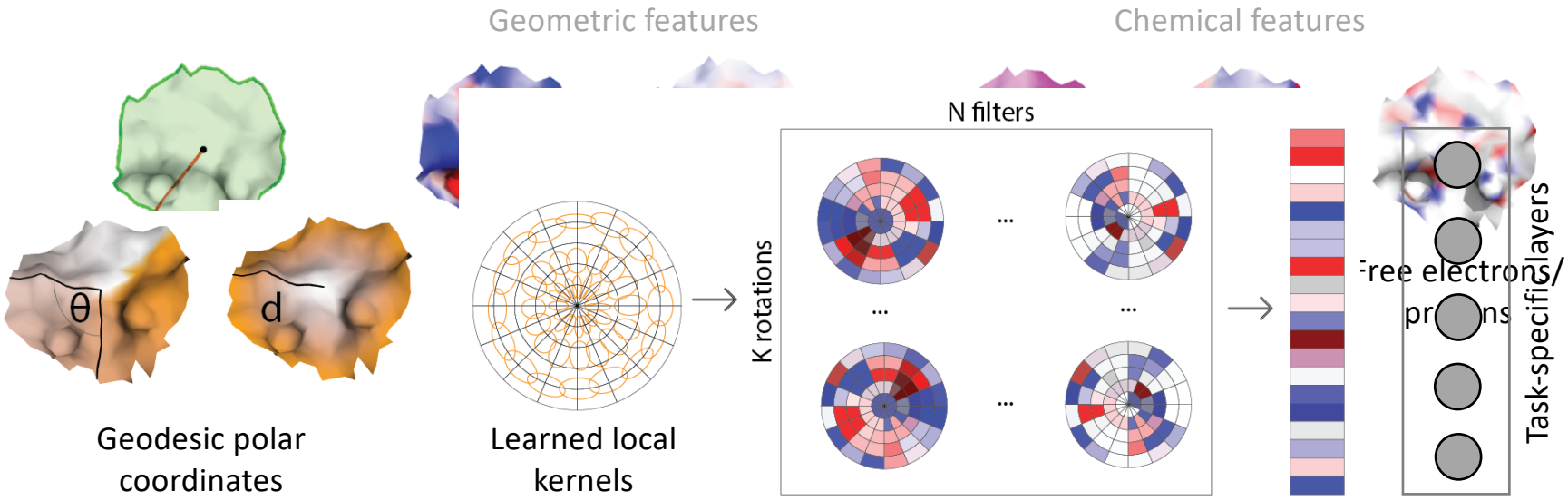
Continuum electrostatics



Free electrons/ protons

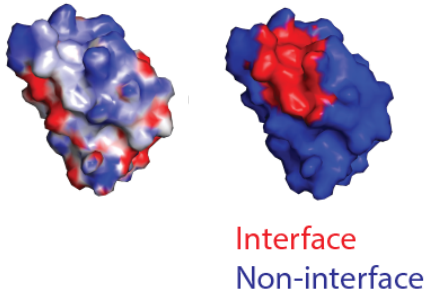
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF architecture



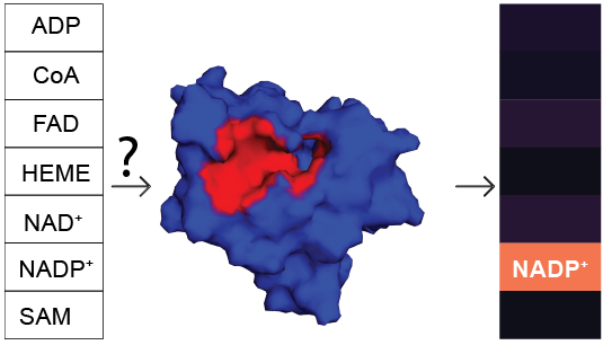
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF applications



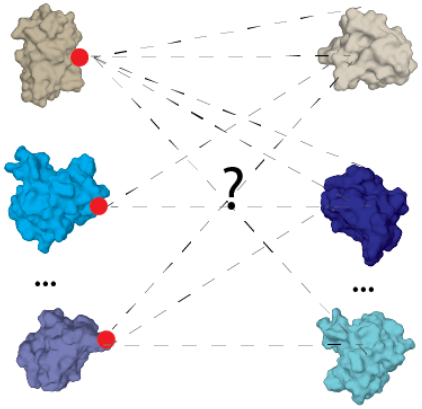
Interface site prediction

MaSIF-site



Pocket classification

MaSIF-ligand

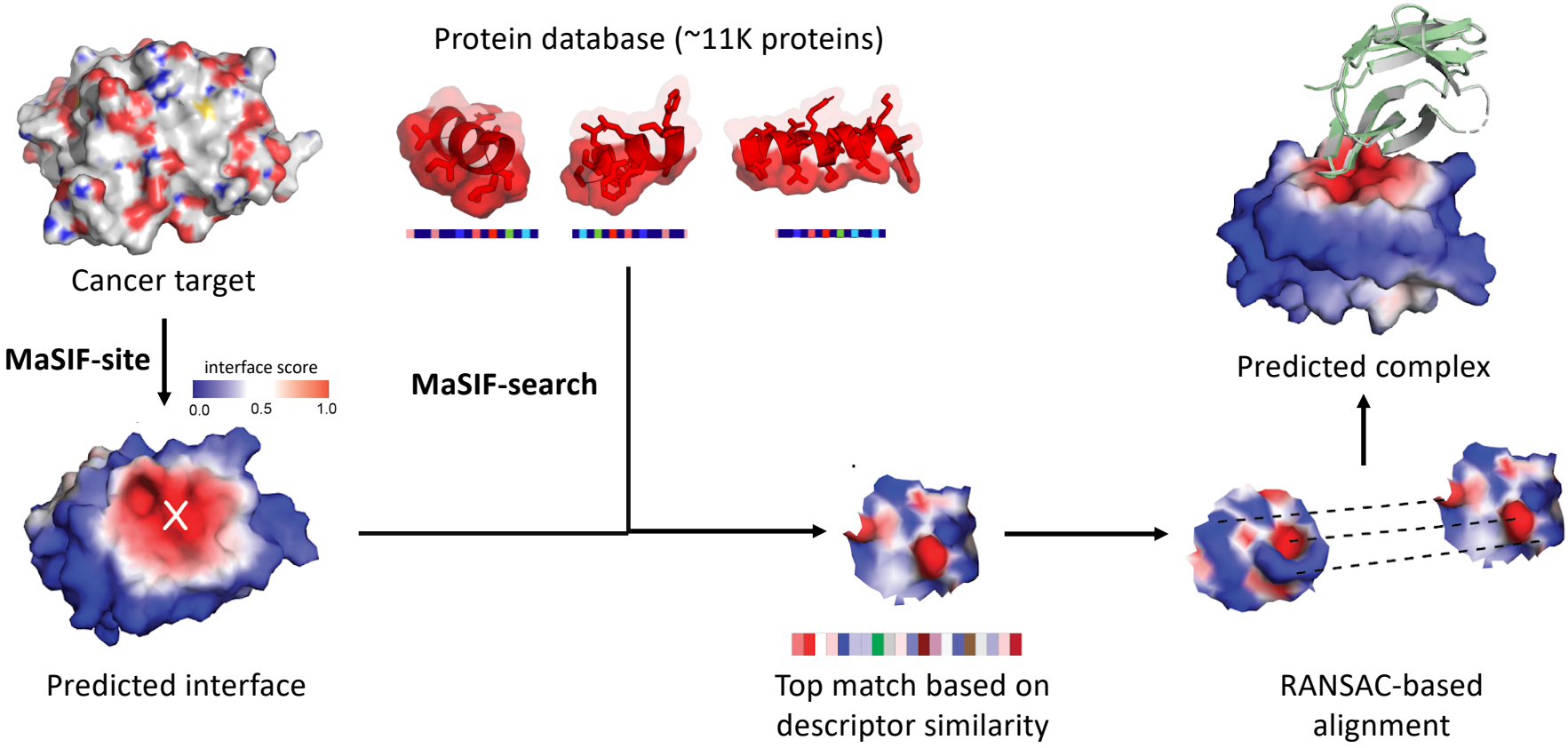


Fast PPI search

MaSIF-search

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

Example: binder design for cancer immunotherapy target



Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF-site

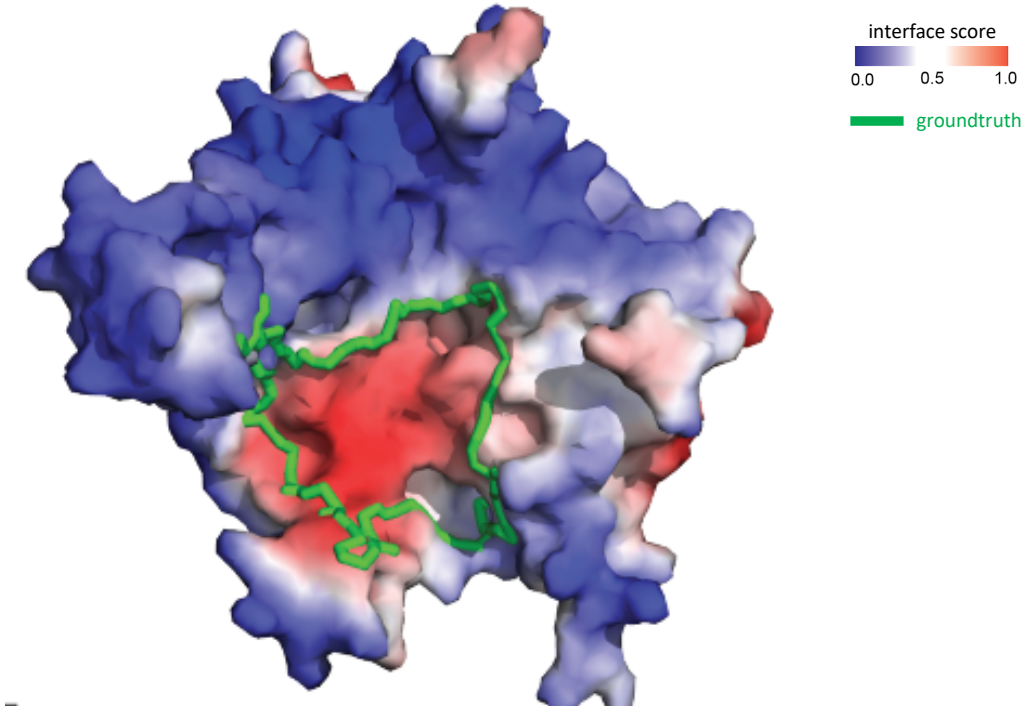
Interface site prediction

MaSIF-site: Prediction of PPI sites

- Point-wise classification problem
- Training set: interface and non-interface points
- Performance criterion: ROC AUC
- Multiple datasets (PRISM, PDBBind, SAbDab antibody:antigen, ZDock)
- Total 3362 crystallized proteins (90% training / 10% testing)

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

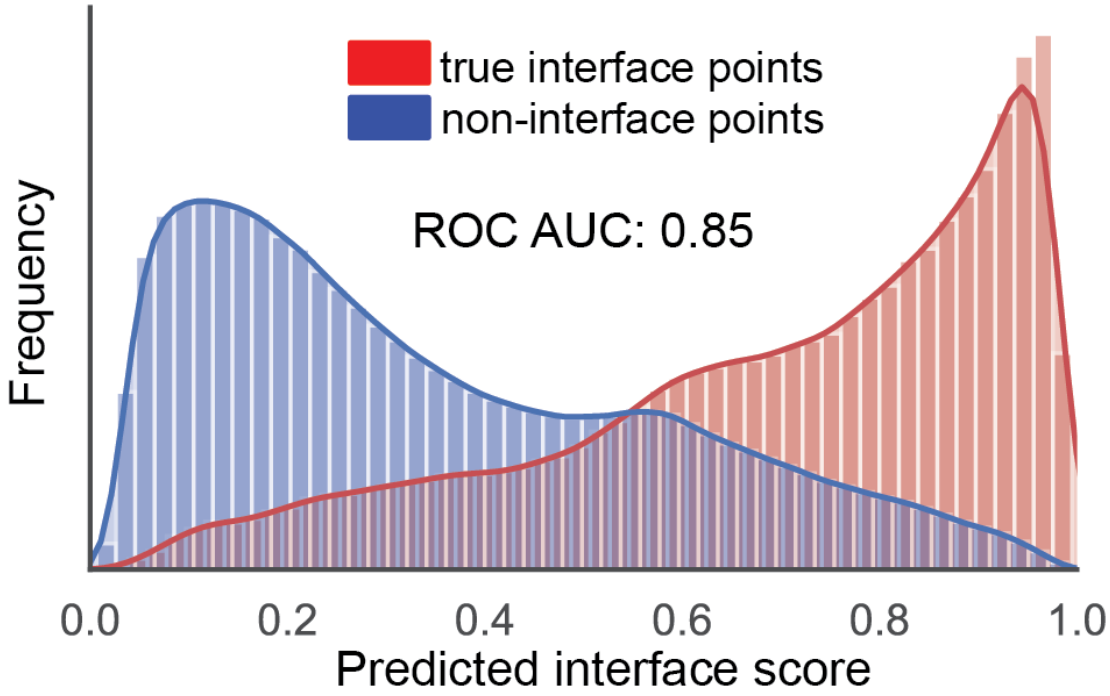
MaSIF-site: Prediction of PPI sites (ubiquitine hydrolase)



Typical example (ROC AUC 0.84)

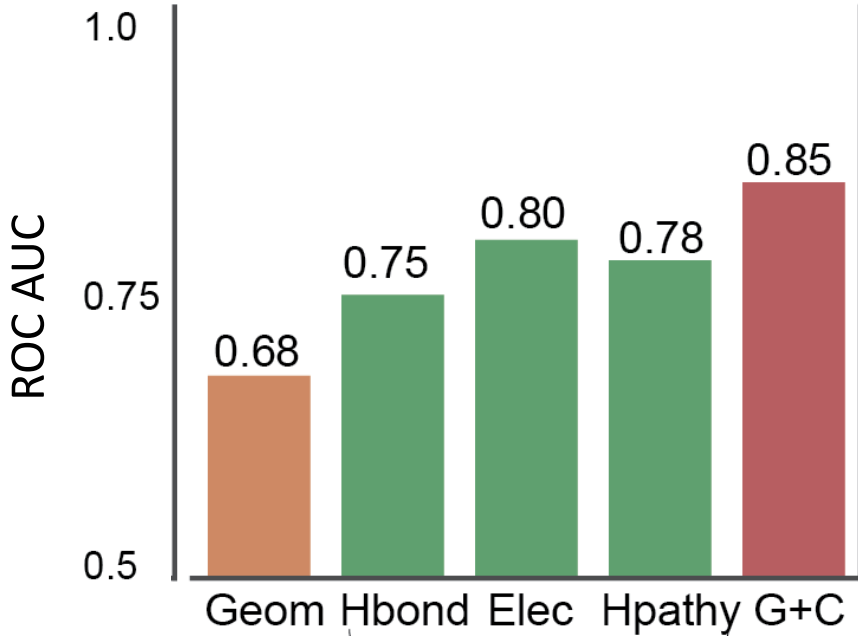
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF-site performance



Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF-site: ablation study

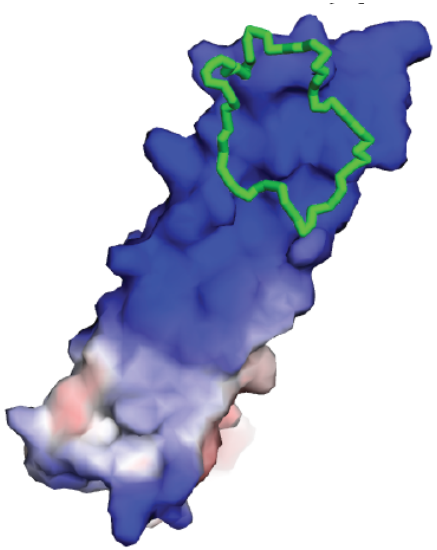


ROC AUC of networks trained with different subsets of features: only geometric (Geom), location of free electrons/proton donors (hbond), Poisson-Boltzmann electrostatics (elec), hydropathy index (hpathy), and all features (G+C).

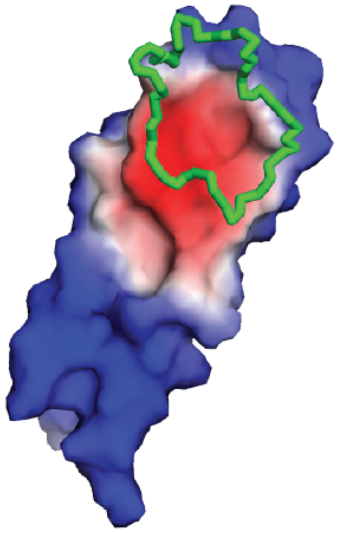
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF-site: Prediction of PPI sites (FFL001 epitope scaffold)

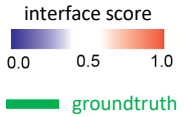
Groundtruth complex



Wildtype



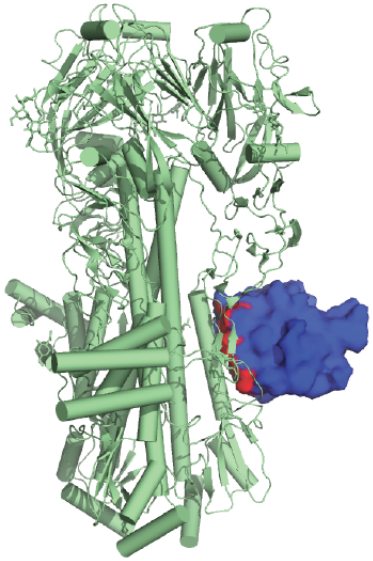
Design



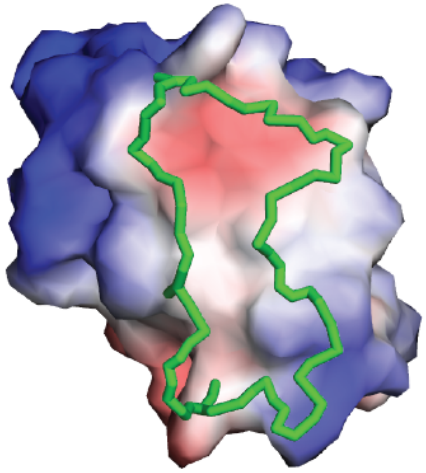
Designed respiratory syncytial virus epitope-scaffold (PDB id: 4JLR) vs. wild type scaffold (PDB id: 1ISE)

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

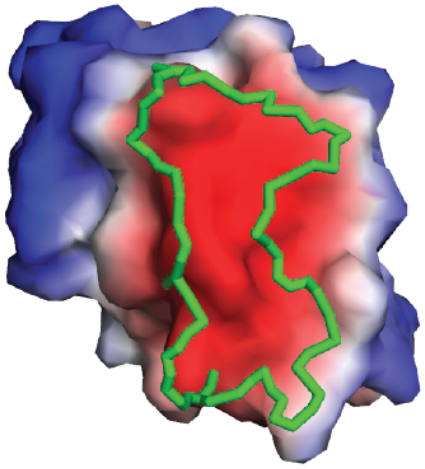
MaSIF-site: Prediction of PPI sites (HB36 influenza inhibitors)



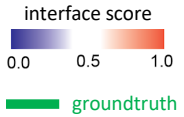
Groundtruth complex



Wildtype



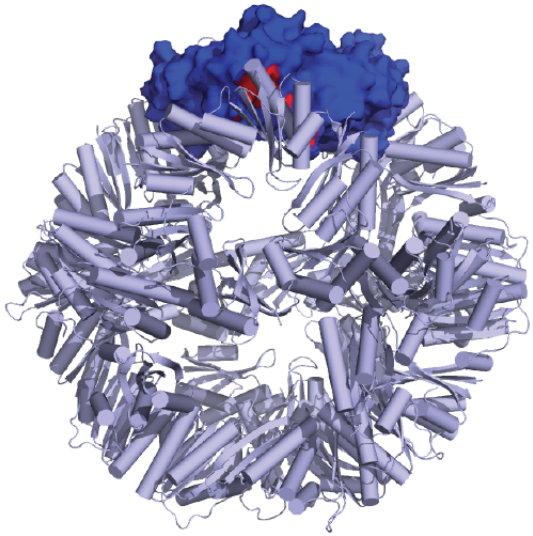
Design



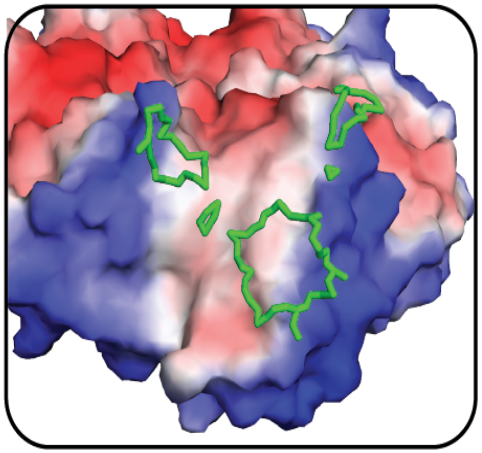
Designed HB36 influenza inhibitors (PDB id: 3R2X) vs. the wild type scaffold protein (PDB id: 1U84)

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

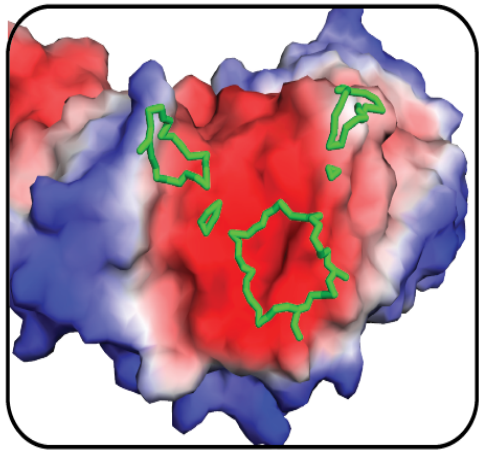
MaSIF-site: Prediction of PPI sites (self-assembling cage proteins)



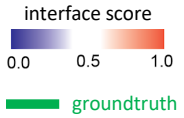
Groundtruth complex



Wildtype



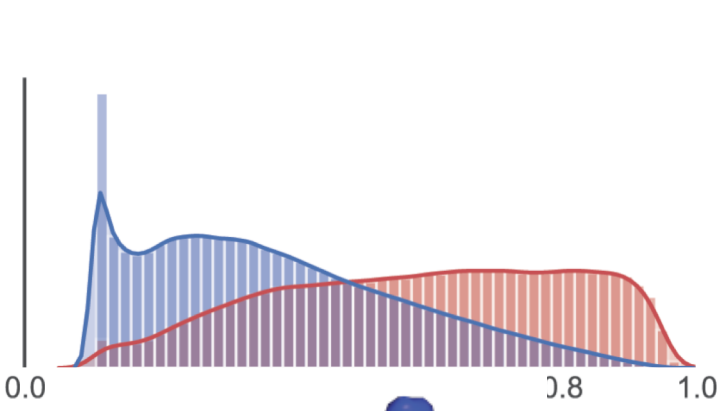
Design

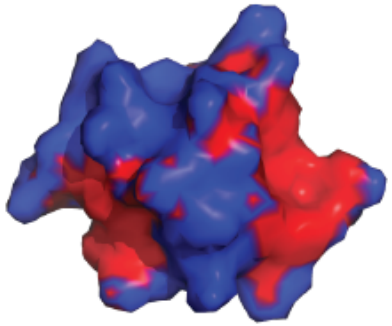


Designed self-assembling nanocage protein (PDB id: 3VCD) vs. the wild type scaffold (PDB id: 3N79)

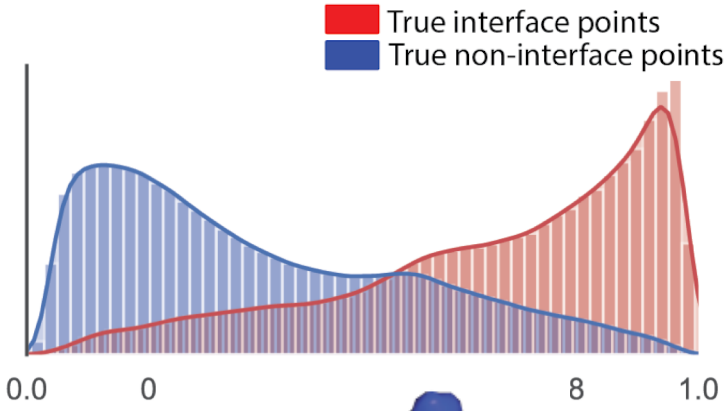
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

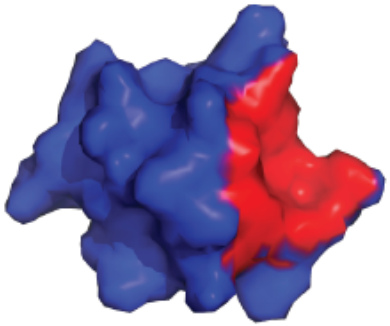
MaSIF-site: going deeper





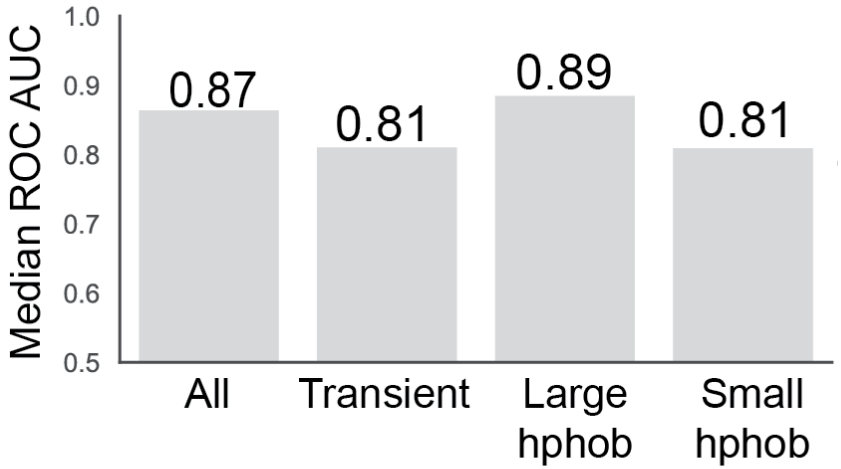
predicted interface points for true and





Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

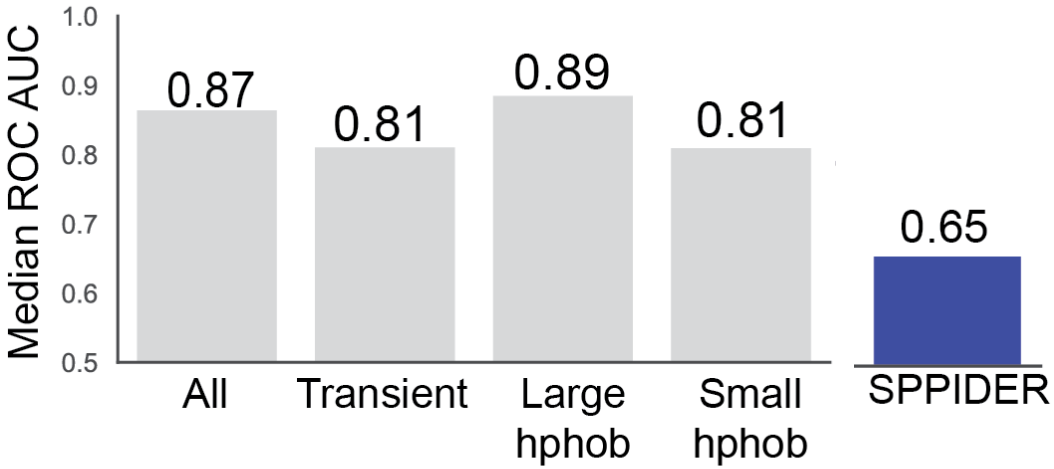
MaSIF-site performance



Performance comparison on different subsets of proteins

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

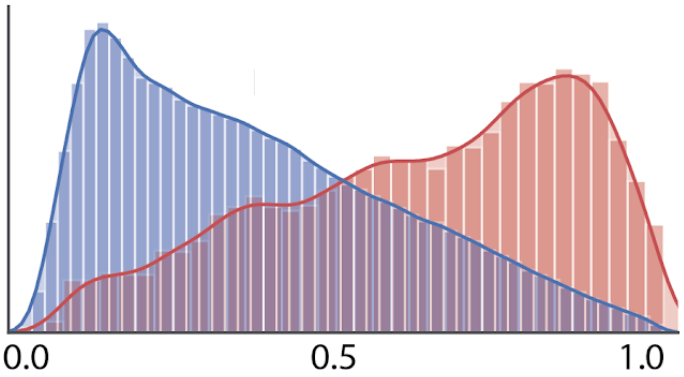
MaSIF-site vs. SPIDDER



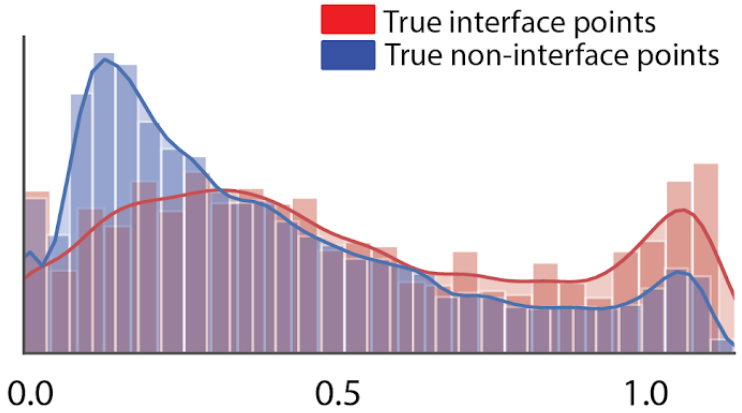
Performance comparison on different subsets of proteins

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Porollo, Meller 2007 (SPIDDER)

MaSIF-site vs. SPIDDER



MaSIF-site
ROC AUC 0.78

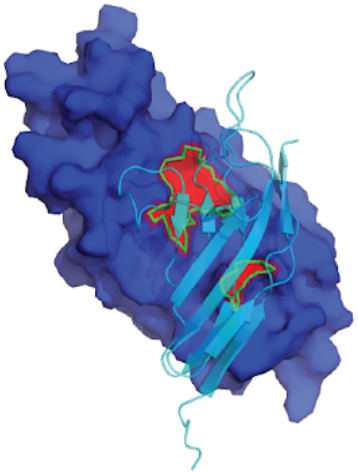


SPIDDER
ROC AUC 0.61

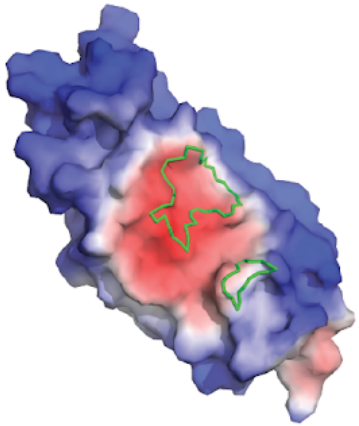
Comparison between MaSIF-site and SPPIDER on 59 transient interactions on a point-by-point basis
(distribution of predicted interface points for true and false interface points)

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Porollo, Meller 2007 (SPIDDER)

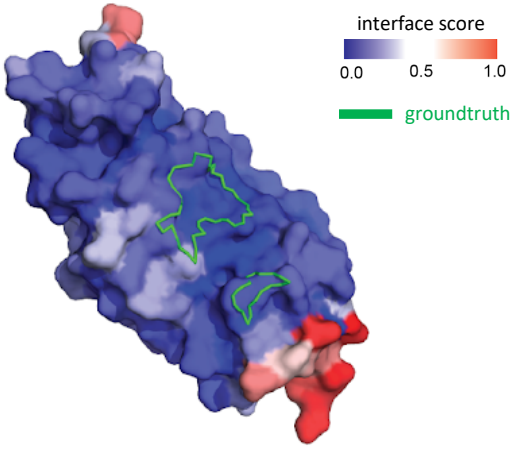
MaSIF-site vs. SPIDDER



Groundtruth



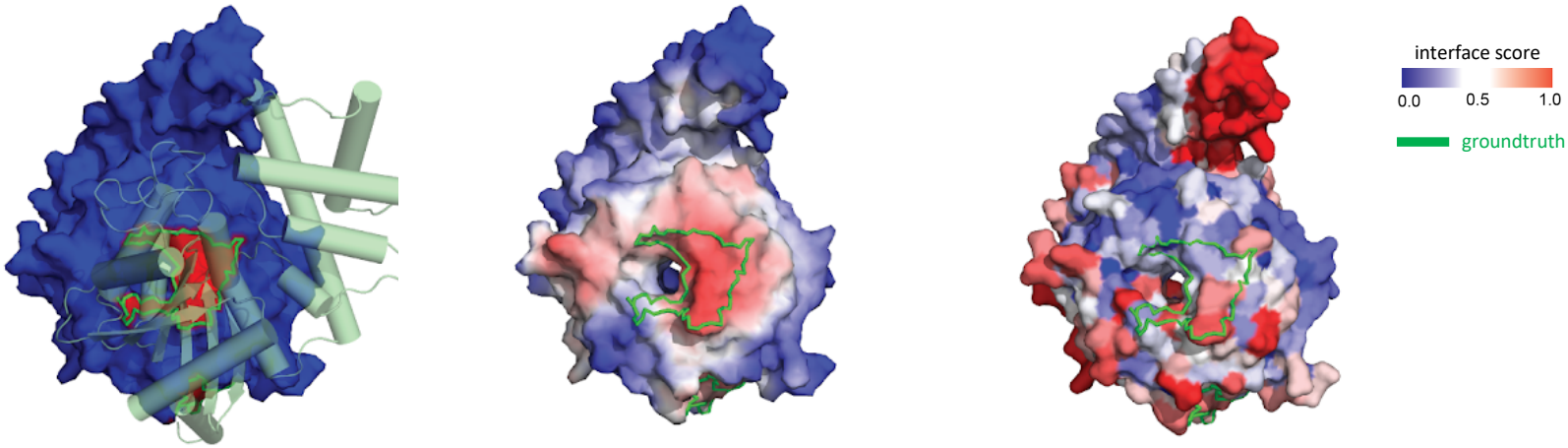
MaSIF-site



SPIDDER

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Porollo, Meller 2007 (SPIDDER)

MaSIF-site vs. SPIDDER



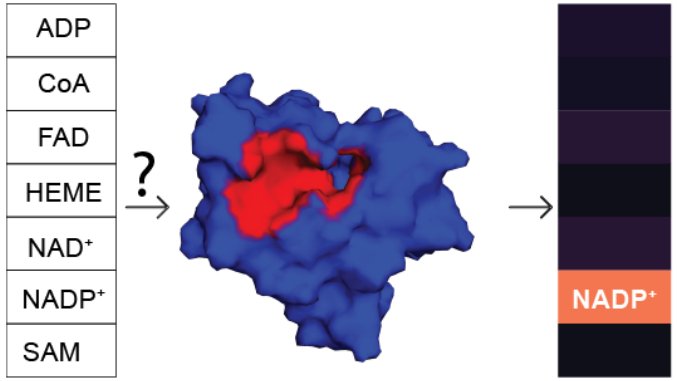
Groundtruth MaSIF-site SPIDDER

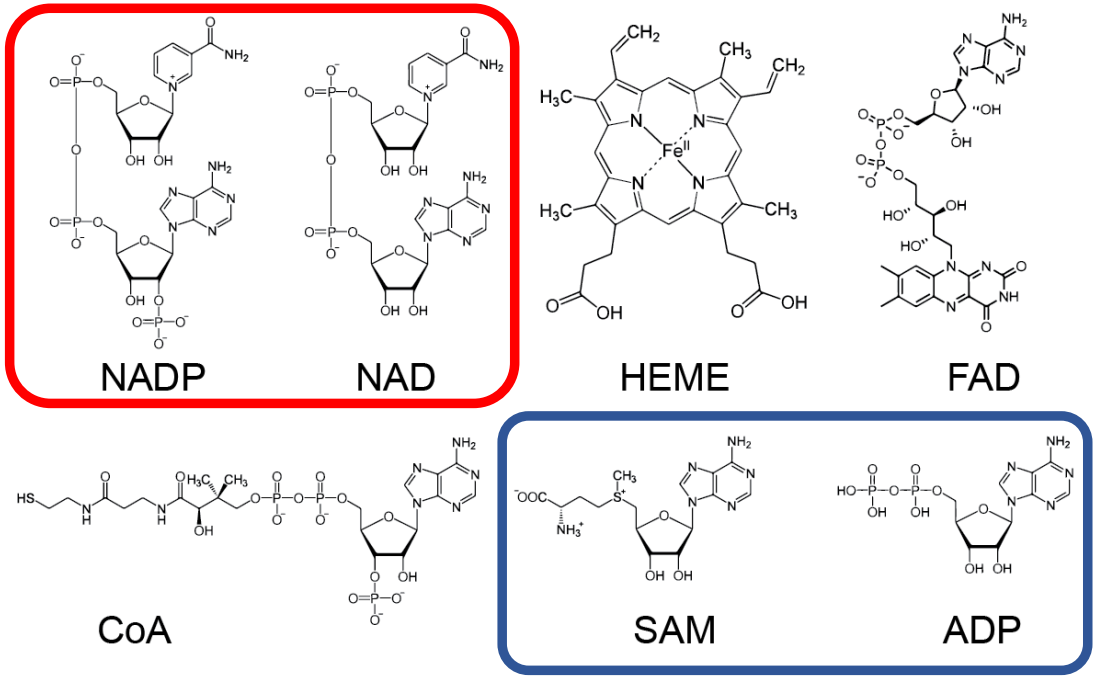
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Porollo, Meller 2007 (SPIDDER)

MaSIF-ligand

Pocket classification

Pocket classification





Structures of the seven cofactors that bind proteins considered for the prediction task

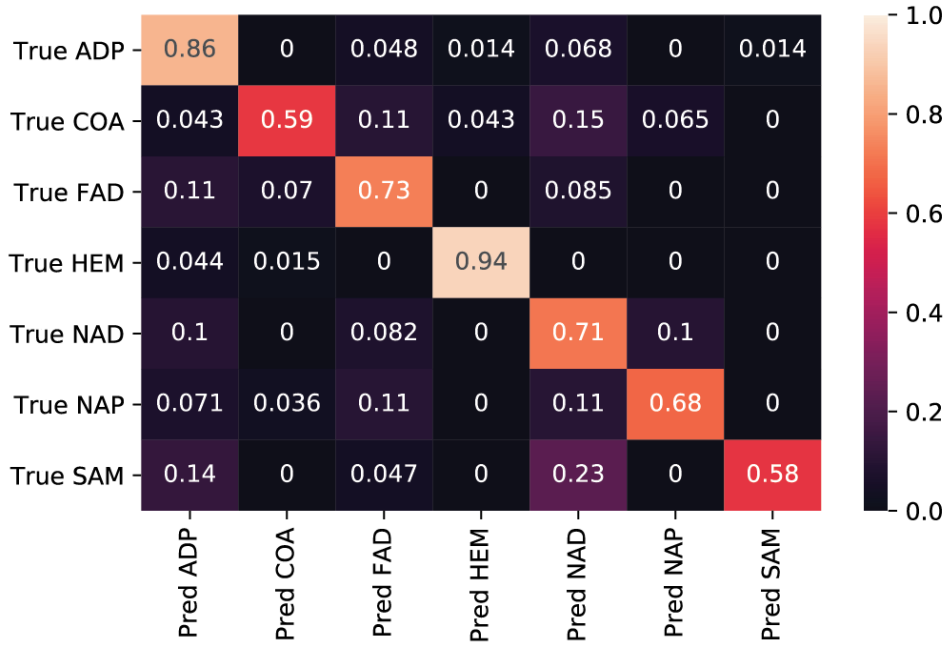
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

MaSIF-ligand: pocket classification

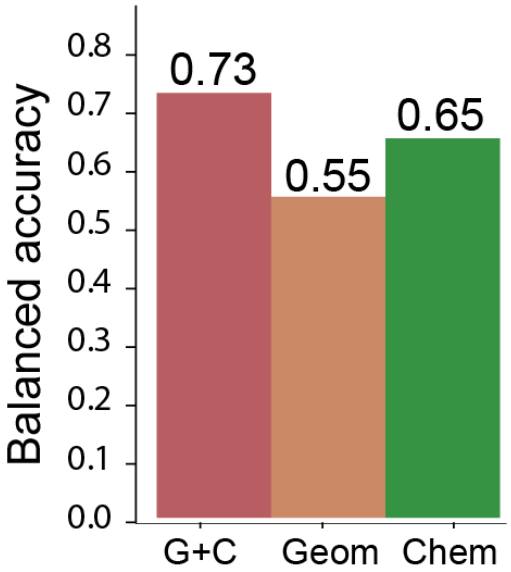
- 7-class point-wise labelling problem
- Training set: proteins interacting with different small molecules
- Total 1459 structures (72% training / 8% validation / 20% testing)
- Careful design of the training and testing sets based on sequence homology

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

Classification of ligand binding sites



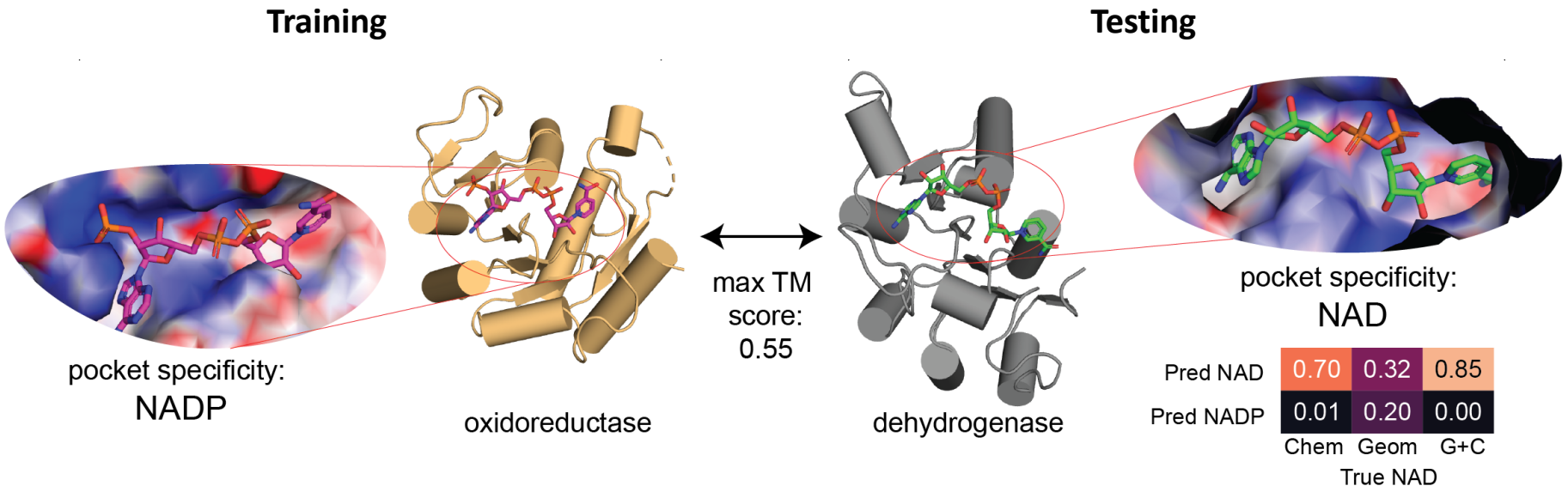
Confusion matrix of ligand specificity on a MaSIF-ligand trained with all features



Balanced accuracy of the prediction of the specificity of binding sites using Geometric, Chemical, and Geometric+Chemical features

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

Classification of ligand binding

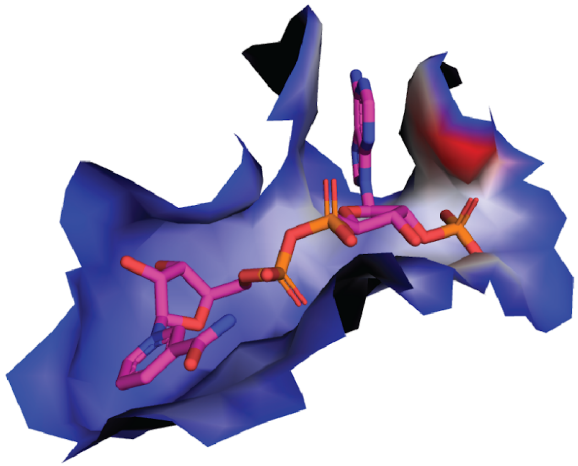


Example of a protein fold that recognizes two similar ligands and yet is correctly predicted. A bacterial dehydrogenase in the test set binds to NAD (PDB id: 2O4C), while its closest structural homologue in the training is a mammalian oxidoreductase (PDB id: 2YJZ), which binds to NADP.

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Yin et al. 2009 (GIF)

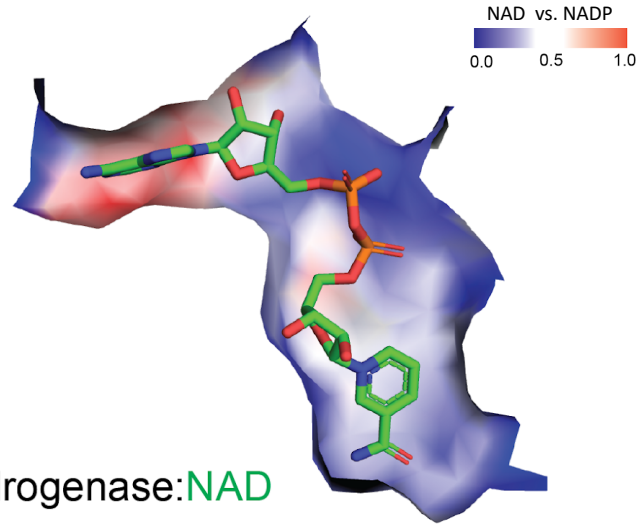
Classification of ligand binding

Training



oxidoreductase:NADP

Testing



dehydrogenase:NAD

Example of a protein fold that recognizes two similar ligands and yet is correctly predicted. A bacterial dehydrogenase in the test set binds to NAD (PDB id: 2O4C), while its closest structural homologue in the training set is a mammalian oxidoreductase (PDB id: 2YJZ), which binds to NADP.

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Yin et al. 2009 (GIF)

MaSIF-search

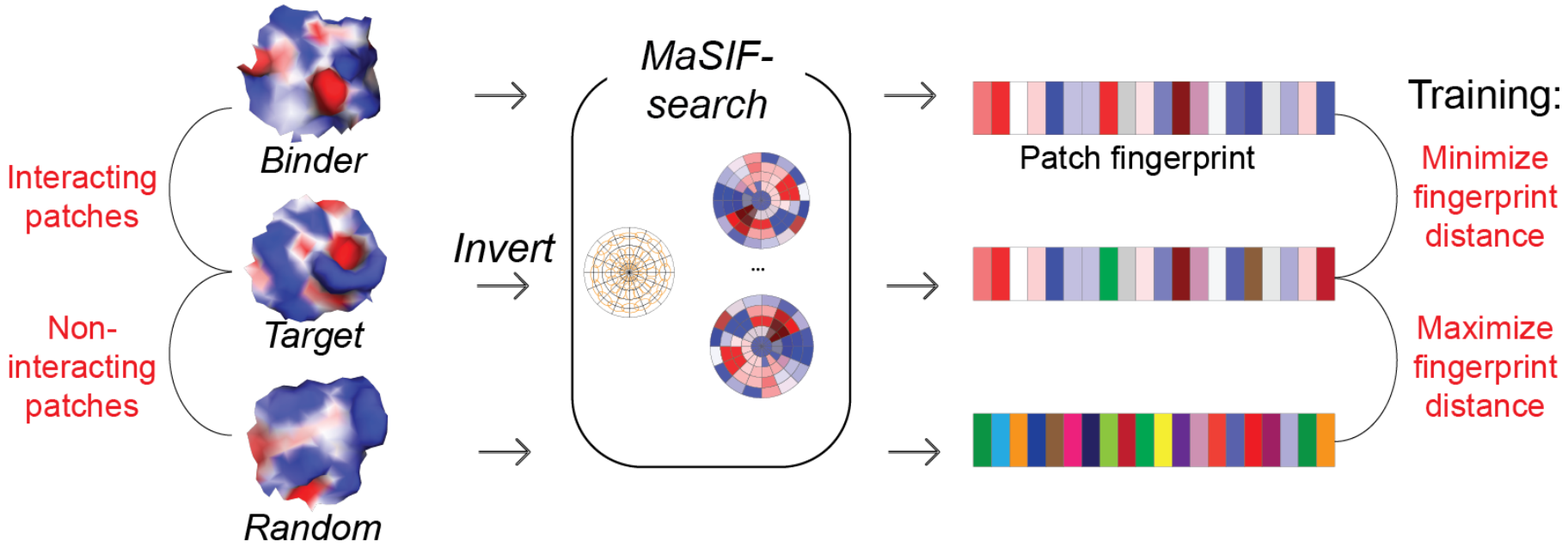
Ultra-fast PPI search

MaSIF-search: PPI prediction

- Local descriptor indicative of interaction (binding)
- Siamese architecture
- Training set: triplets (x,x^+,x^-) where x,x^+ are interacting (positives) and x,x^- are non-interacting (negatives)
- Triplet loss, d-prime loss
- Total 6001 PPIs (80% training / 20% testing)

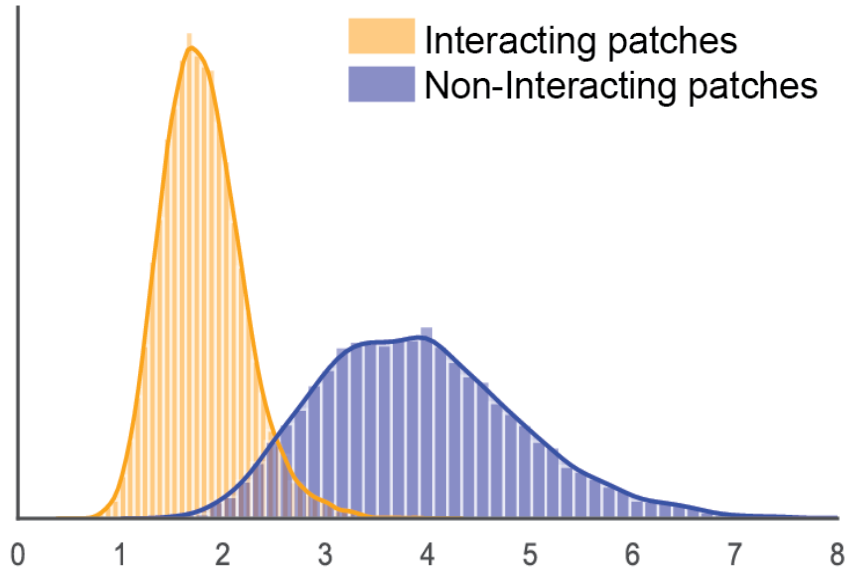
Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

Local descriptors for PPI prediction

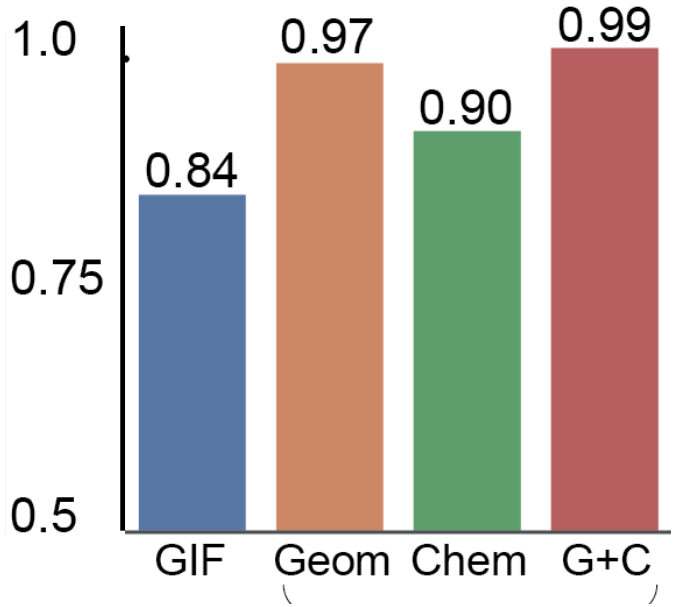


Gainza, Sverisson, Monti, Rodolà, B, Correia 2019

PPI prediction using local surface descriptors



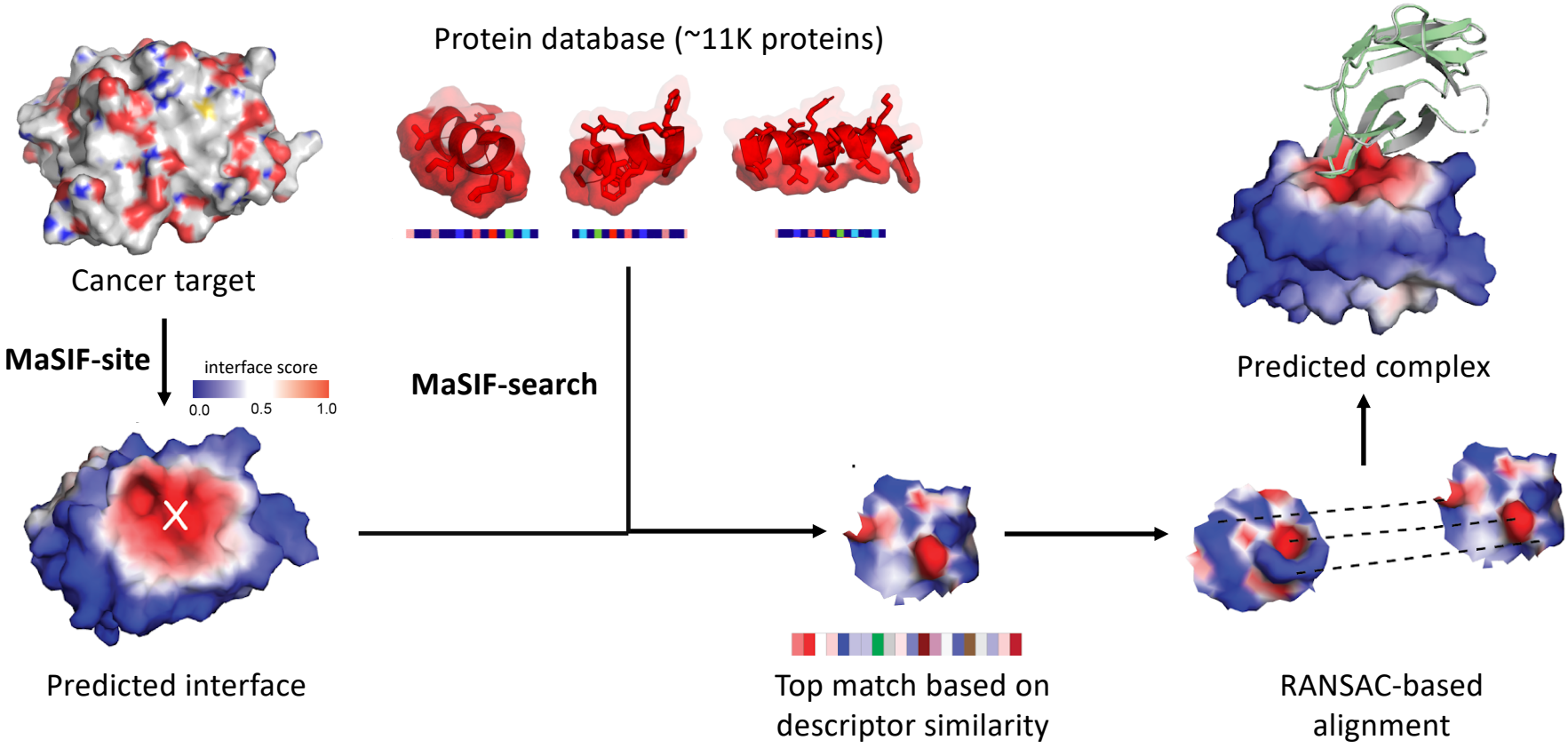
Distribution of descriptor distances between interacting (yellow) and non-interacting (blue) patches in the test set (training/testing with Geometric+Chemical features)



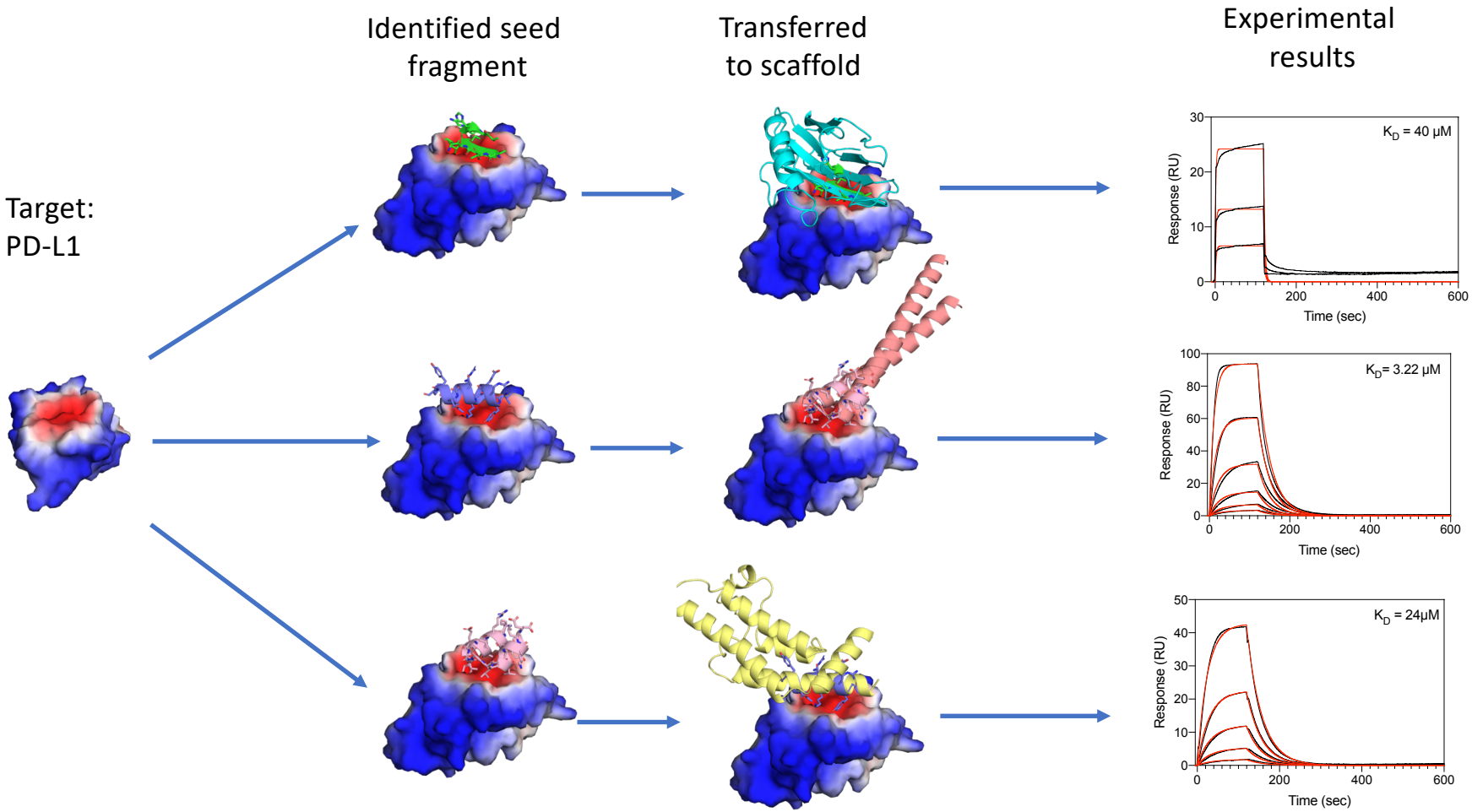
Performance (ROC AUC) using Geometric, Chemical, and Geometric+Chemical features

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Yin et al. 2009 (GIF)

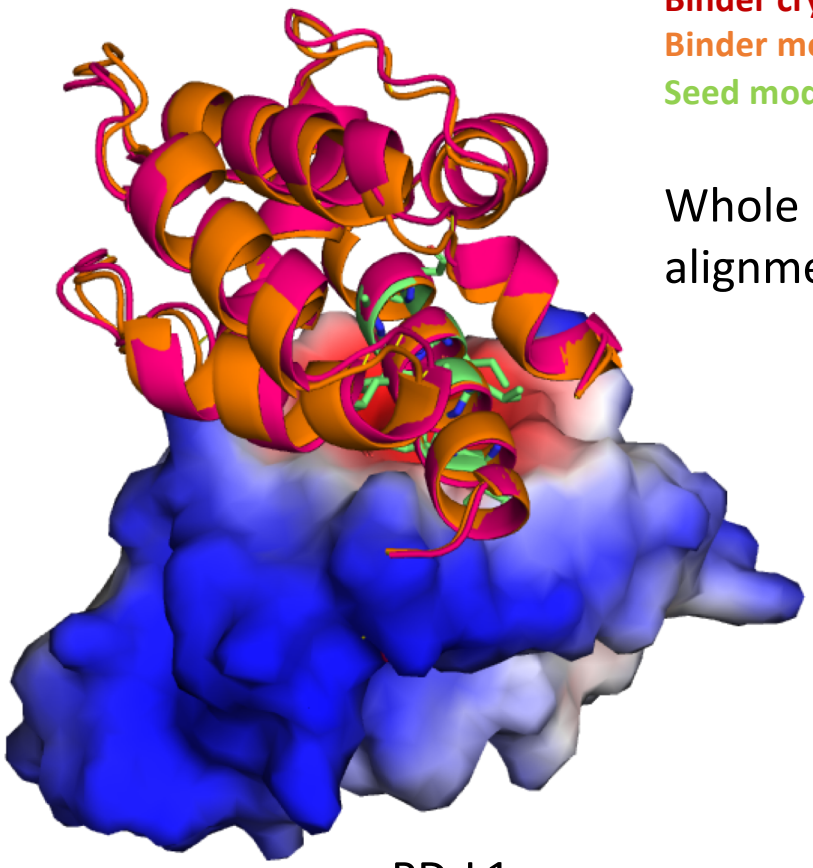
Example: binder design for cancer immunotherapy target



Gainza, Sverisson, Monti, Rodolà, B, Correia 2019



Experimental results: work in progress



Binder crystal structure
Binder model
Seed model

Whole complex alignment: 0.77 Å

PD-L1

Experimental results: work in progress

Large-scale docking

Table 1 | Results for large-scale docking benchmark benchmark for PatchDock, MaSIF-search (with multiple numbers of decoys), ZDock and ZDock+ZRank2 on bound (holo) complexes

Method	Number of solved complexes in the top			Time (min)
	100	10	1	
MaSIF-search decoys = 100	37	36	30	4
MaSIF-search decoys = 2,000	67	56	43	39
PatchDock	43	32	21	2,743
ZDock	58	36	18	134,934
ZDock+ZRank2 decoys = 200,000	77	63	45	159,902

No. of solved complexes in the top, number of target–binder complexes within 5 Å iRMSD found in the top 100, top ten or top one (for holo cases) or top 1,000, top 100 and top ten (for apo cases). Time (min), CPU time in minutes for each program, which excludes precomputation time for MaSIF-search.

Table 2 | Results for large-scale docking benchmark benchmark for PatchDock, MaSIF-search (with multiple numbers of decoys), ZDock and ZDock+ZRank2 on unbound (apo) complexes

Method	Number of solved complexes in the top			Time (min)
	1,000	100	10	
MaSIF-search decoys = 2,000	17	7	2	16
PatchDock	11	4	1	560
ZDOCK	17	13	5	13,174
ZDock+ZRank2 decoys = 80,000	23	12	5	16,866

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Duhovny et al. 2002 (PatchDock); Pierce et al. 2011 (ZDOCK); Pierce, Weng 2008 (ZRank2)

Conclusions

- Novel Geometric DL toolset for protein science
- Task-specific data-driven descriptors for protein structure and functionality
- Significantly more accurate and faster than previous methods
- Independent of sequence (“evolutionary history”)
- Challenge: Bound vs. unbound proteins
- Experimental validation (crystal structure, in vitro, in vivo)

Thank you!