Computational protein design using Geometric Deep Learning

Michael Bronstein

Imperial College London
United Kingdom

USI Lugano
Switzerland

Joint work with B. Correia, P. Gainza, F. Sverisson (EPFL), F. Monti (USI), E. Rodolà (Sapienza)
β-Sheet (3 strands)

α-helix
Protein design = inverse folding
Protein binding: Lock-key model
Applications: cancer 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

PD-1
PD-L1

2018 Nobel Prize
PD-proteins role in immunotherapy
Applications: biosensors

1. Biosensors are equipped with biomolecules such as antigens, antibodies or enzymes that specifically interact with disease markers.

2. The biomolecules are placed on ultra-thin films, typically measuring only nanometers, forming an active layer.

3. A matrix consisting of various materials such as chitosan and concanavalin, or cystamine and glutaraldehyde, is used to attach the biomolecules to the device.

4. The active layer containing the immobilizing matrix chemically bonds to the electrode, which functions as a signal-transduction pathway.

5. This signal-transduction pathway is able to convert the alterations caused by the interaction between the biomolecules and the markers present in the sample under analysis into an electrical, electrochemical or optical signal.

6. A portable device reads the biosensor, indicating the presence or absence of the biomarkers—and if present, the concentration of the substance in the blood or fluid that served as a sample.
De novo design of high-affinity protein binders
Representation

Atom point cloud
Graph
Secondary structures
Molecular surface
Geometric deep learning

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

- **Local coordinates** $u_{ij}$ around $i$

- **Local weights** $w_1(u), ..., w_L(u)$ w.r.t. $u$, e.g. Gaussians:
  
  $$w_\ell(u) = \exp \left( -(u - \mu_\ell)^T \Sigma_\ell^{-1}(u - \mu_\ell) \right)$$

- **Spatial convolution** with filter $g$:
  
  $$x_i' = \frac{\sum_{\ell=1}^L g_\ell \sum_{j=1}^n w_\ell(u_{ij})x_j}{\sum_{\ell=1}^L g_\ell \sum_{j=1}^n w_\ell(u_{ij})}$$
Molecular surface interaction fingerprinting (MaSIF)

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

Typical example (ROC AUC 0.85)

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019
MaSIF-site performance

ROC AUC: 0.85

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 (HB36 influenza inhibitors)

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 (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 (self-assembling cage proteins)

Groundtruth complex

Wildtype

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 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

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

MaSIF-site
ROC AUC 0.78

SPIDDER
ROC AUC 0.61

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

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

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

- Interface site prediction: **MaSIF-site**
- Pocket classification: **MaSIF-ligand**
- Fast PPI search: **MaSIF-search**
Structures of the seven cofactors that bind proteins considered for the prediction task
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
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
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.
MaSIF applications

- Interface site prediction: MaSIF-site
- Pocket classification: MaSIF-ligand
- Fast PPI search: MaSIF-search
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)
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

Cancer target

Protein database (~11K proteins)

MaSIF-site

MaSIF-search

Predicted interface

Top match based on descriptor similarity

Predicted complex

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019
Predicted binders for PD-L1

Predicted mouse PD1
Groundtruth human PD1

Mouse PD1 (PDB id: 3BIK) ranked #1, RMSD=0.6 Å

Predicted human PD1
Groundtruth human PD1

Human PD1 (PDB id: 4ZQK) ranked #8, RMSD=0.3 Å

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019
Large-scale docking (bound)

<table>
<thead>
<tr>
<th></th>
<th># found (N=100)*</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PatchDock</td>
<td>40</td>
<td>2854</td>
</tr>
<tr>
<td>MaSIF (K=100)</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td>MaSIF (K=500)</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>MaSIF (K=1000)</td>
<td>54</td>
<td>7</td>
</tr>
<tr>
<td>MaSIF (K=3000)</td>
<td>62</td>
<td>17</td>
</tr>
<tr>
<td>GIF (K=3000)</td>
<td>9</td>
<td>15</td>
</tr>
</tbody>
</table>

#solved: number of target:binder complexes within 5Å iRMSD found in the top 100. K: number of patches used for matching. Time: Processing time in minutes for each algorithm - excludes pre-computation time for fingerprint-based algorithms.

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Duhovny et al. 2002 (PatchDock); Yin et al. 2009 (GIF)
Large-scale docking (unbound)

<table>
<thead>
<tr>
<th></th>
<th># found (N=40)**</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PatchDock</td>
<td>11</td>
<td>1045</td>
</tr>
<tr>
<td>MaSIF (K=1000)</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>MaSIF (K=3000)</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>GIF (K=3000)</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

**solved**: number of target:binder complexes within 5Å iRMSD found in the top 1000. **K**: number of patches used for matching. **Time**: Processing time in minutes for each algorithm - excludes pre-computation time for fingerprint-based algorithms.

Gainza, Sverisson, Monti, Rodolà, B, Correia 2019; Duhovny et al. 2002 (PatchDock); Yin et al. 2009 (GIF)
Experimental results
Protein encoding in DNA

DNA

ATGATCTCTCGTAA
TACTAGAGCATT

mRNA

AUGAUCUCGUA

Protein

Met Ile Ser

TRANSCRIPTION

TRANSLATION

STOP
Expressing proteins

DNA sequence encoding
designed protein

yeast cell
*S. cerevisiae*

binding fluorescence
expression fluorescence
target protein
design protein

[Diagram showing the process of expressing proteins in yeast cells with DNA sequence encoding designed proteins, indicating fluorescence and target protein binding.]
Interface evolution (outside seed)
Binding strength

Des1 (500nM Target)

Initial signal

Optimized

Opt. Des1 (5 nM Target)

Opt. Des2 (1nM Target)

Des2 (100 nM Target)

Initial signal

Optimized
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)