DeepLearning inLifeSciences

# Deep learning models for protein-ligand binding site prediction

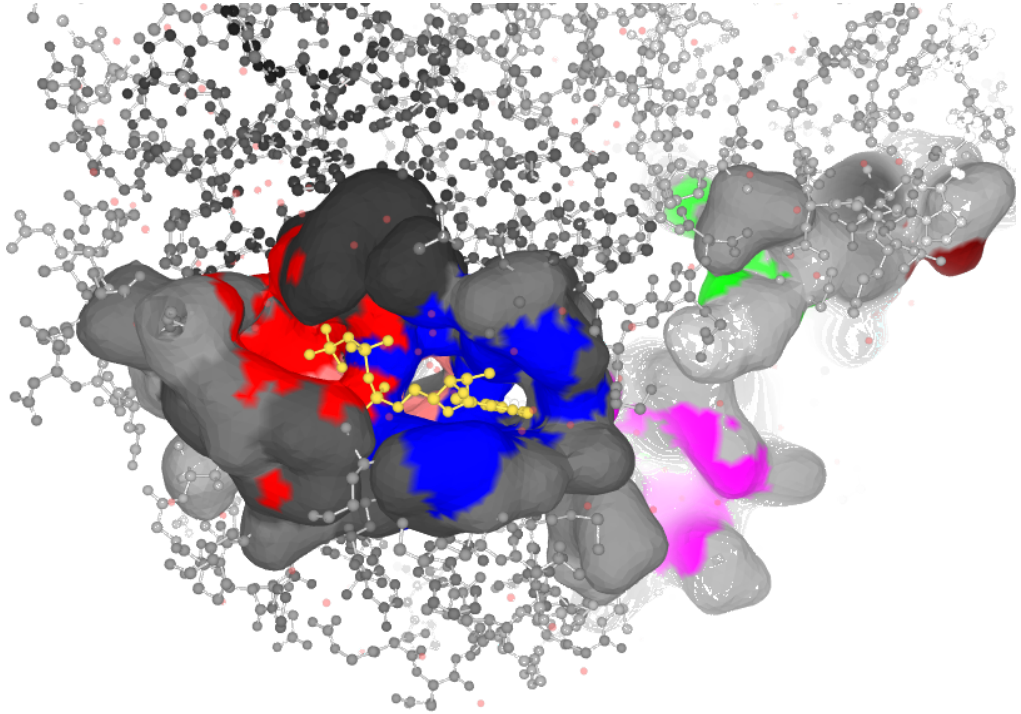

**David Hoksza**
https://bioinformatics.cuni.cz/hoksza
https://ksi.mff.cuni.cz

# Outline

- Binding sites prediction task
  - Sequence vs structure
- **Traditional** (ML-based) sequence- and structure-based approaches
- **Deep learning** structure- and sequence-based approaches
- Limitations, considerations

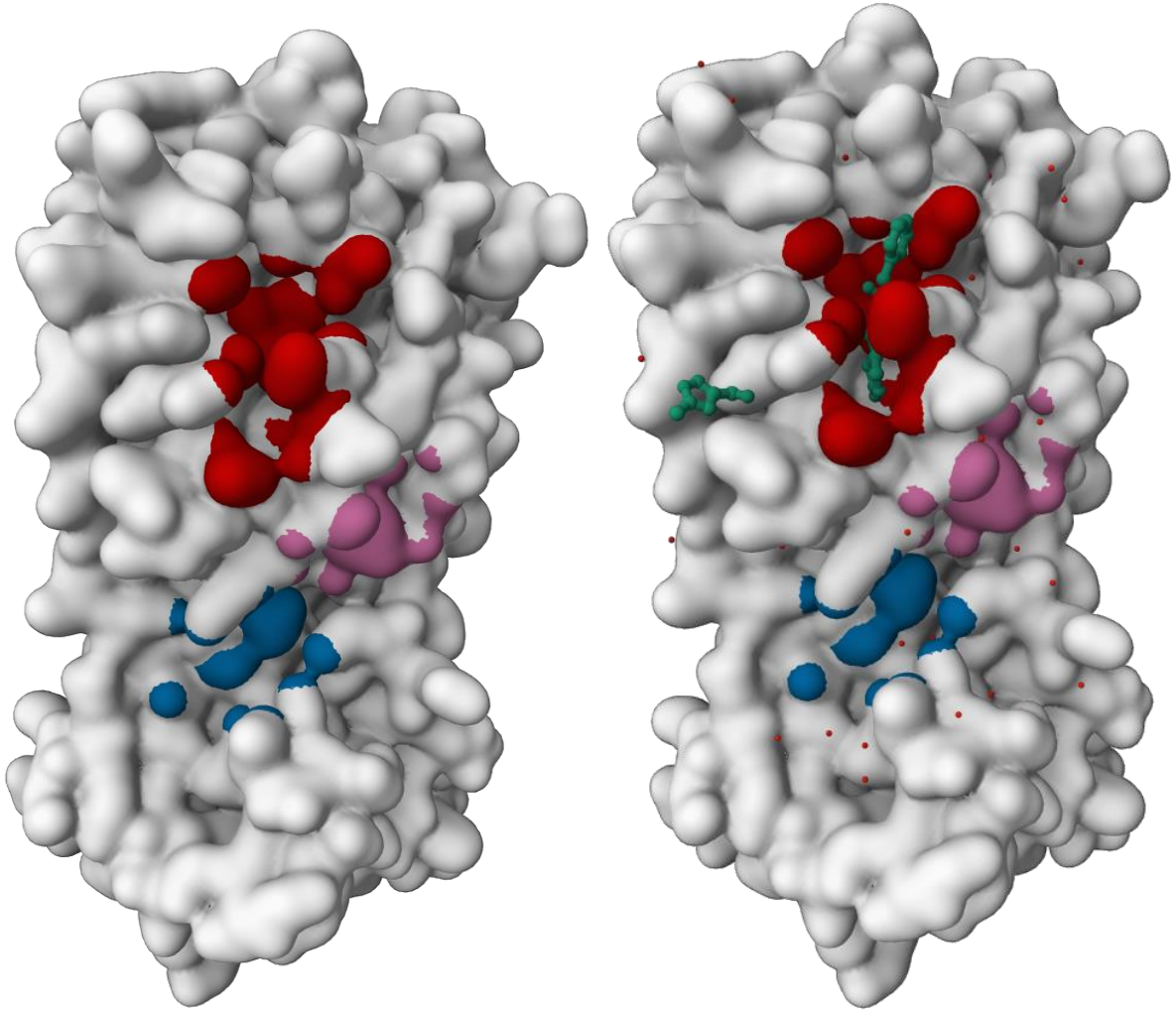
# Protein binding sites

- Where a protein forms **contact with other molecules**
  - Triggering events such as ligand modification or conformational change
  - Focus on **small molecules** in this lecture



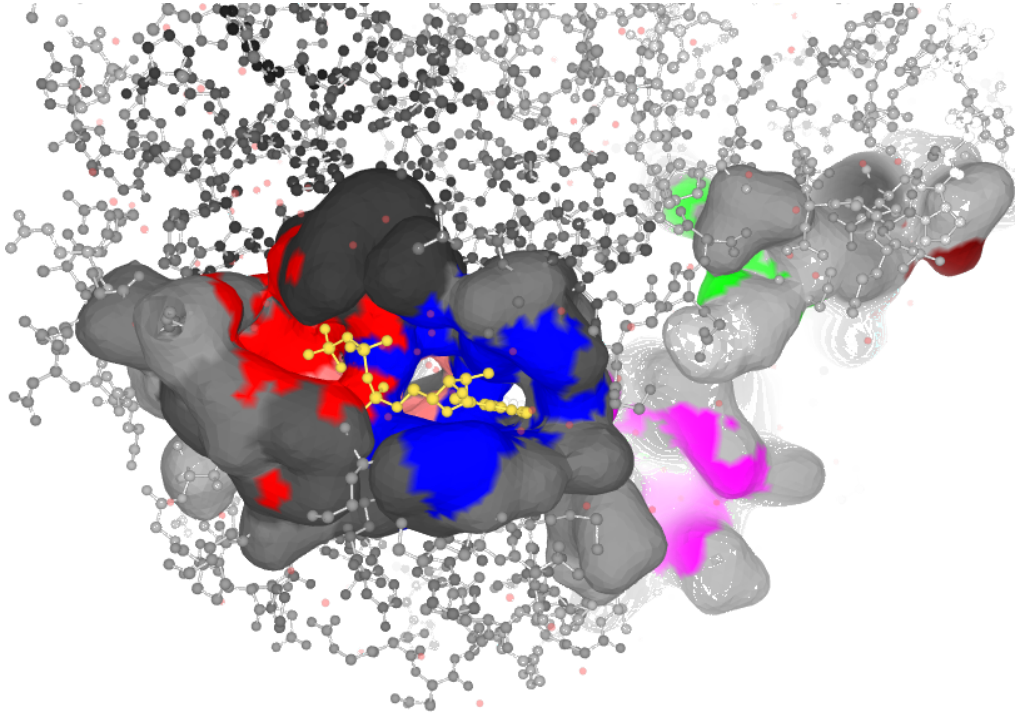
# Motivation

- Understanding protein function
- Structure-based virtual screening/drug design
  - Macromolecular target identification
  - **Pocket detection**
  - Small molecules identification via docking
  - Optimization



# Binding site/residue detection task

**Structure**-based →
**pocket**-level prediction



**Sequence**-based →
**residue**-level prediction

....FPWFGMDIGGTLVKLSYFEPIDITAEEEQEEVES....

# Sequence prediction task

- Goal: **identify residues** that **are part of a binding site**
- **Input**: a protein sequence

....FPWFGMDIGGTLVKLSYFEPIDITAEEEQEEVES....

- **Output**: a list of residues being part of **a** ligand binding site

....FPWFGMDIGGTLVKLSYFEPIDITAEEEQEEVES....

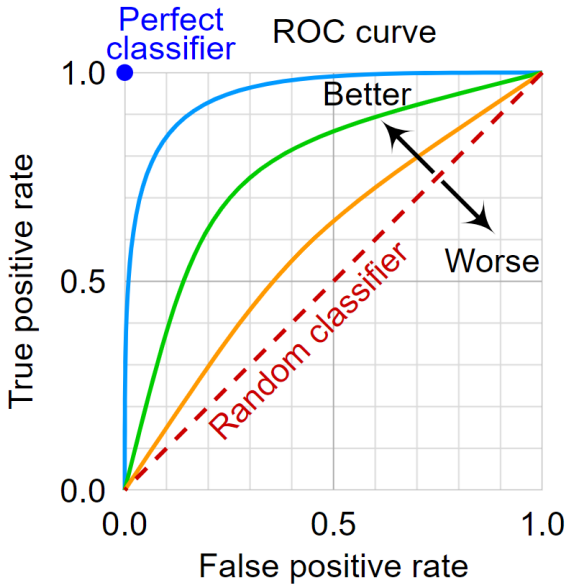
# Evaluation metrics (sequence)

| | | Predicted condition | |
|---|---|---|---|
| | Total population = P + N | Positive (PP) | Negative (PN) |
| Actual condition | Positive (P) | True positive (TP) | False negative (FN) |
| | Negative (N) | False positive (FP) | True negative (TN) |

- A **residue** can be either predicted as **binding or non-binding** → traditional binary classification metrics based on **confusion matrix** such as **F1** score, **MCC** (Matthews correlation coefficient)
  - MCC preferred in the domain
  - Highly imbalanced data → accuracy not suitable

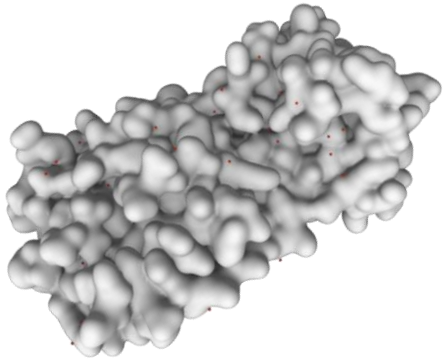
Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21 (2020): 1-13.

- When **thresholding** possible → receiver operating characteristics **(ROC)/AUC**

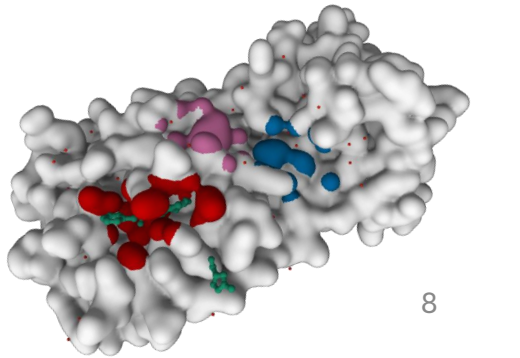


# Structure prediction task

- Goal: **identify surface regions** that are **capable of binding** an unspecified small molecule



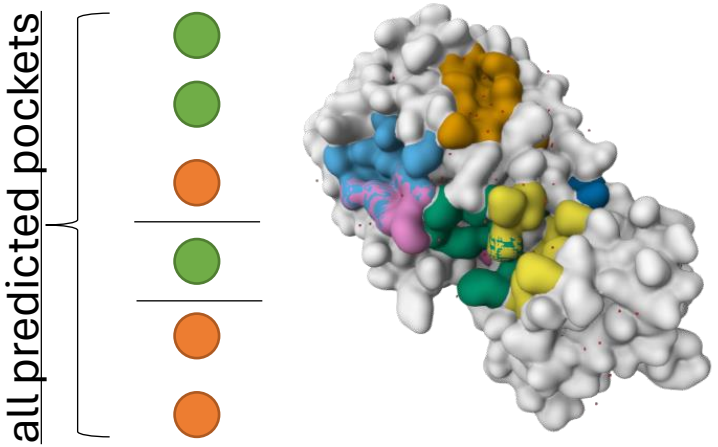
- **Input**: a protein structure
- **Output**: a list of (surface) regions probably capable of ligand binding



# Evaluation metrics (structure)

- Typical binary classification problem metrics MCC, F1, or ROC/AUC not suitable as there are no well-defined negatives
  - Can be solved by projecting surface regions onto residues

- **Success rate** with respect to **Top-$n$ + $k$** pockets
  - $n$ - pockets per protein (e.g. protein has 3 true pockets)
  - $k$ - room for error (k = 1 ⇒ check first 4 positions)

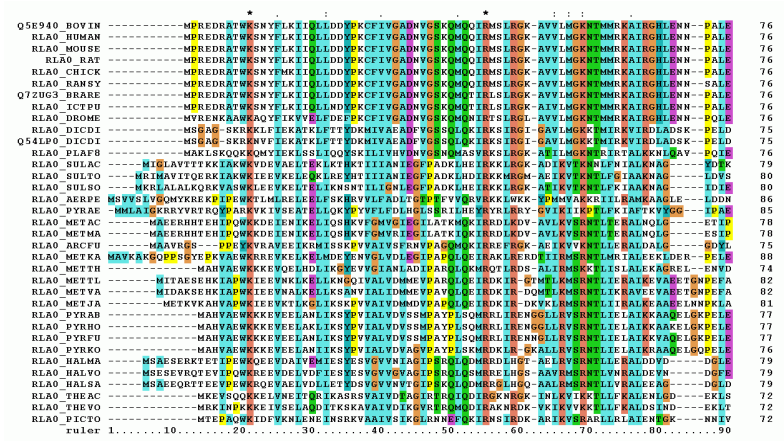


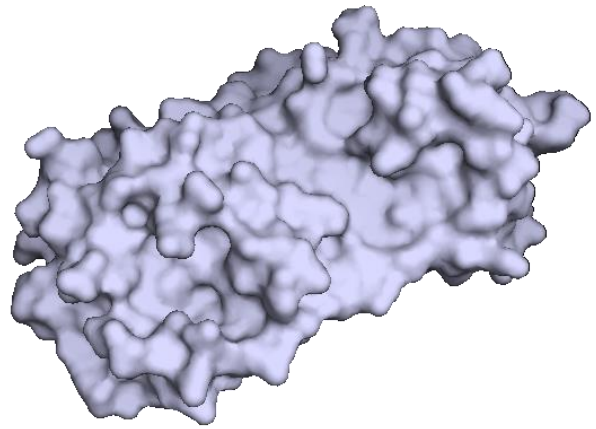
# Pocket detection criteria (structure)

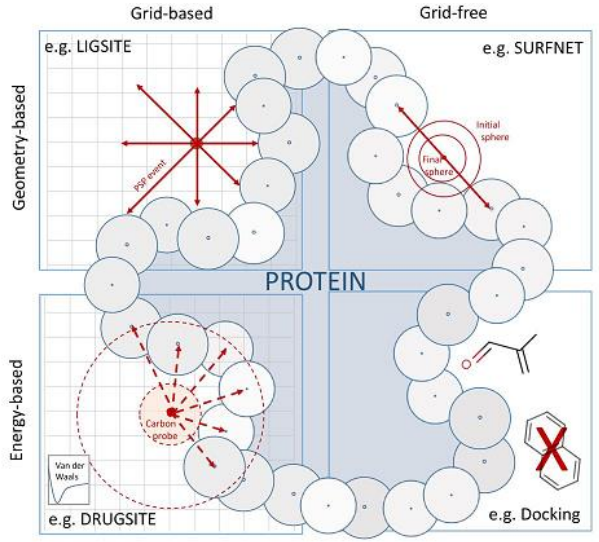
- How do you decide that a pocket was correctly predicted?
  - **Distance-based**
    - **DCA**
      - the minimum distance between the center of the predicted pocket and any atom of the ligand
    - **DCC**
      - the distance between the centers of the predicted and true binding sites
  - Volume-based
    - **DVA** - Discretized volumetric overlap
      - Overlap between the predicted and real binding sites
      - Jaccard Index on the discretized convex hulls of the predicted and true binding sites
    - **MOC** - Mutual overlap criterion (Fpocket)

# Detection approaches

- Template-based
- Template-free
  - Geometry
  - Energy
  - Machine learning










source: Volkamer, Andrea, et al. "Prediction, analysis, and comparison of active sites." Applied Chemoinformatics: Achievements and Future Opportunities (2018): 283-311.

Published 3D structure-based LBS prediction methods.

| Method | Type | Feature | Year |
|---|---|---|---|
| A computational procedure (with no specific name) [39] | Probe Energy-based | Contour surfaces at appropriate energy levels are calculated for each probe and displayed with the protein structure | 1985 |
| POCKET [27] | Spatial Geometry Measurement | Place spheres between atoms and surfaces of pockets are modeled using marching cubes algorithm | 1992 |
| SURFNET [40] | Spatial Geometry Measurement | Place spheres at the gap between any two protein atoms | 1995 |
| LIGSITE [26] | Spatial Geometry Measurement | Set up some regular 3D meshes to cover the target protein | 1997 |
| CAST [41] | Spatial Geometry Measurement | Calculate by using alpha shape and discrete flow theory | 1998 |
| CASTp [42,43] | Spatial Geometry Measurement | Use alpha shape and the pocket algorithm [44] developed in computational geometry | 2003 |
| QSiteFinder [45] | Probe Energy-based | Use the interaction energy between the protein and a simple van der Waals probe | 2005 |
| LIGSITE$^{CSC}$ [46] | Spatial Geometry Measurement | An extension and implementation of the LIGSITE algorithm by using the Connolly surface | 2006 |
| VISCANA [47] | Probe Energy-based | A total energy of the molecule is evaluated by summation of fragment energies and interfragment interaction energies | 2006 |
| Fpocket [48] | Spatial Geometry Measurement | Voronoi tessellation and alpha spheres are used to detect pockets | 2009 |
| SITEHOUND [28,49] | Probe Energy-based | The carbon probe and phosphate probe used to detect interaction force between the probe and the protein | 2009 |
| MSPocket [50] | Spatial Geometry Measurement | Identify surface pocket regions according to the normal vector directions at the vertices on the surface | 2010 |
| FTSite [51] | Probe Energy-based | Use 16 different probes on these grids to detect free energy | 2011 |
| SiteComp [52] | Probe Energy-based | Discovery of subsites with different interaction properties and for fast calculations of residue contribution to binding sites | 2012 |
| LISE [53] | Spatial Geometry Measurement | Compute a score by counting geometric motifs extracted from substructures of interaction networks connecting protein and ligand atoms | 2013 |
| Patch-Surfer2. 0 [54] | Spatial Geometry Measurement | Represent and compare pockets at the level of small local surface patches that characterize physicochemical properties of the local regions | 2014 |
| CurPocket [55] | Spatial Geometry Measurement | Compute the curvature distribution of protein surface and identify the clusters of concave regions | 2019 |

Published template similarity-based LBS prediction methods.

| Method | Type | Feature | Year |
|---|---|---|---|
| ConSurf [56] | Sequence Template-based | Phylogenetic relationships among the sequences and the similarity between the amino acids are taken into account | 2003 |
| A Sequence template-based approach with no specific name [57] | Sequence Template-based | An information-theoretic approach for estimating sequence conservation based on Jensen–Shannon divergence | 2007 |
| FINDSITE [58] | Structure Template-based | PROSPECTOR 3 threading algorithm and TMalign tool are used | 2008 |
| A two-stage template-based LBS prediction method [59] | Structure Template-based | Construct protein's 3D model and use structural clustering of ligand-containing templates on the predicted 3D model | 2009 |
| 3DLigandSite [29] | Structure Template-based | MAMMOTH is used | 2010 |
| FunFOLD [60] | Structure Template-based | Use an automatic approach for cluster identification and residue selection | 2011 |
| COFACTOR [61] | Structure and Sequence Template-based | Use global-to-local sequence and structural comparison algorithm | 2012 |
| webPDBinder [62] | Structure Template-based | Search a protein structure against a library of known binding sites and a collection of control nonbinding pockets. | 2013 |
| S-SITE [31] | Sequence Template-based | Needleman–Wunsch algorithms are used | 2013 |
| TM-SITE [31] | Structure and Sequence Template-based | Mix Structure Template-based and Sequence Template-based method | 2013 |

Traditional machine learning-based LBS prediction and binding affinity research methods.

| Method | Machine Learning Algorithm | Year |
|---|---|---|
| Knowledge-based QSAR approach [69] | Kernel-Partial Least Squares (K-PLS) [70] | 2004 |
| Multi-RELIEF [71] | RELIEF algorithm [72] | 2007 |
| SFCscore [73] | multiple linear regression partial least squares analysis | 2008 |
| ATPint [74] | Support Vector Machine | 2009 |
| ConCavity [75] | K-Means algorithm | 2009 |
| MetaPocket [76] | hierarchical clustering algorithm [77] | 2009 |
| RF-Score [4] | The Random Forest algorithm | 2010 |
| MetaDBSite [78] | Support Vector Machine | 2011 |
| NsitePred [79] | Support Vector Machine | 2011 |
| NNSCORE [80,81] | Artificial Neural Network (shallow neural network [82]) | 2011 |
| L1pred [30] | L1-Logreg Regression classifier | 2012 |
| TargetS [83] | Support Vector Machine | 2013 |
| eFindSite [84] | Support Vector Machine | 2013 |
| VitaPred [85] | Support Vector Machine | 2013 |
| COACH [31] | Support Vector Machine | 2013 |
| LigandRFs [86] | The Random Forest algorithm | 2014 |
| OSML [87] | Support Vector Machine | 2015 |
| LigandDSES [88] | The Random Forest algorithm | 2015 |
| PRANK [89] | The Random Forest algorithm | 2015 |
| A method for protein-ligand binding affinity prediction [90] | Gradient Boosting Regressor [91] | 2018 |
| SAnDReS [92] | Regression Analysis | 2016 |
| P2Rank [93] | The Random Forest algorithm | 2018 |
| COACH-D [94] | Support Vector Machine | 2018 |
| Taba [95] | Regression Analysis | 2019 |

Deep learning-based LBS prediction and binding affinity research methods.

| Method | Main Goal | Network Type | Year |
|---|---|---|---|
| A deep learning framework for modeling structural features of RNA-binding protein targets [118] | Binding references modeling of RNA-binding proteins | DBN | 2015 |
| DeepBind [119] | Sequence specificities prediction of DNA- and RNA-binding proteins | CNN | 2015 |
| DeepDTA [3] | Drug-target interaction identification | CNN | 2018 |
| $K_{DEEP}$ [120] | Protein-ligand binding affinity prediction | CNN | 2018 |
| DEEPSite [36] | LBS Prediction | CNN | 2017 |
| DeepCSeqSite [121] | LBS Prediction | CNN | 2019 |
| DeepConv-DTI [122] | Drug-target interaction identification | CNN | 2019 |
| DeepDrug3D [35] | Binding pockets characterization and classification | CNN | 2019 |
| Onionnet [123] | Protein-ligand binding affinity prediction | CNN | 2019 |

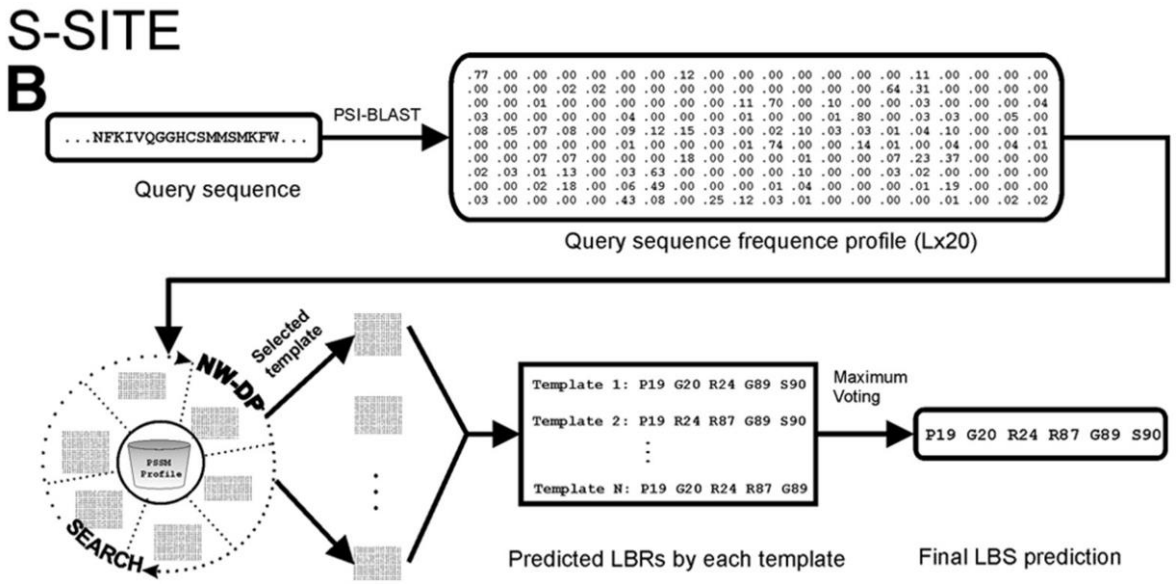


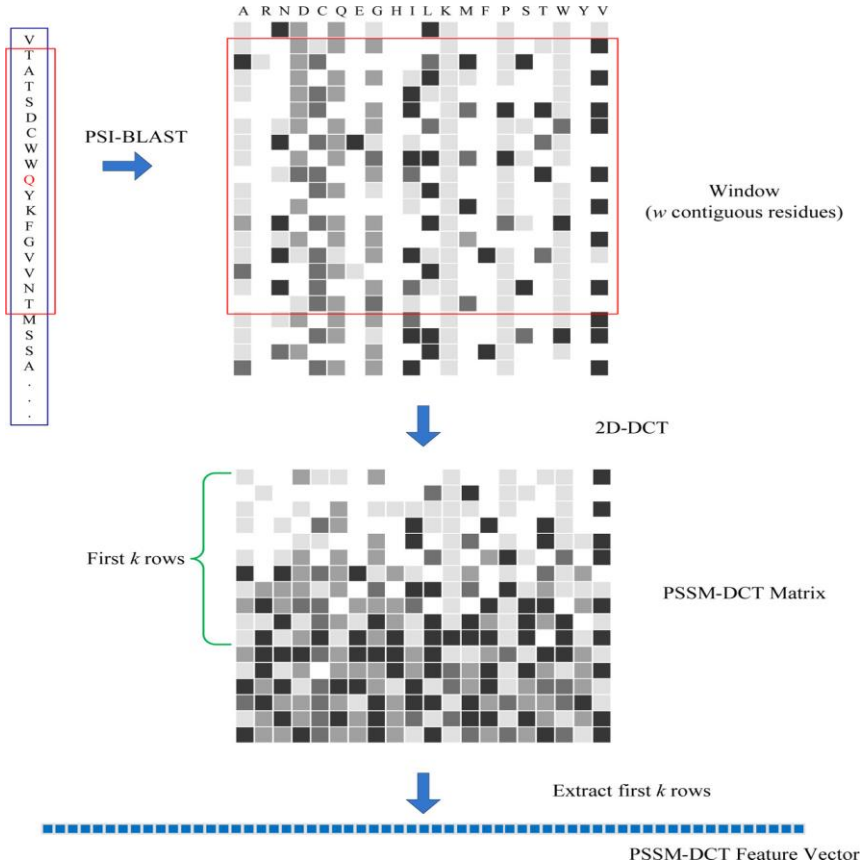
source: Zhao, Jingtian, Yang Cao, and Le Zhang. "Exploring the computational methods for protein-ligand binding site prediction." Computational and structural biotechnology journal 18 (2020): 417-426.

# Traditional (ML-based) approaches

Sequence

# Utilization of evolutionary information






Yang, Jianyi, Ambrish Roy, and Yang Zhang. "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment." Bioinformatics 29.20 (2013): 2588-2595.

Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of protein–ligand binding sites by sequence information and ensemble classifier." Journal of Chemical Information and Modeling 57.12 (2017): 3149-3161.

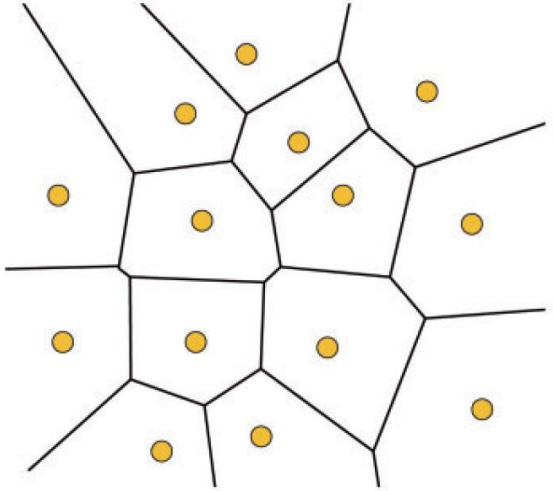
# Traditional (ML-based) approaches

Structure

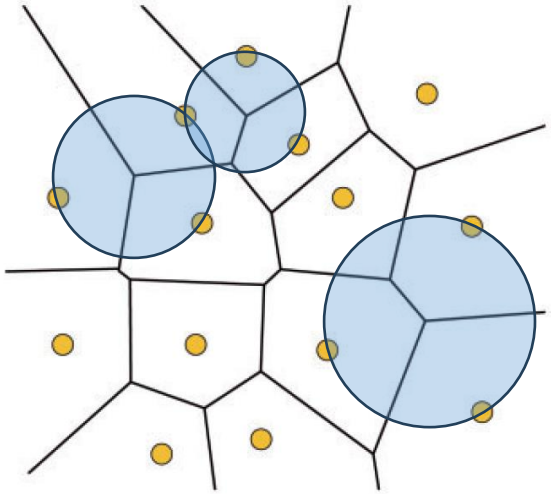
# Fpocket

- Based on the concept of **alpha spheres** obtained from **Voronoi tessellation** given protein atoms

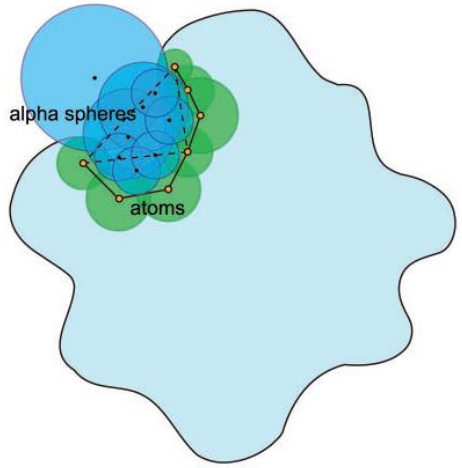
- **Cavities** → alpha spheres of **intermediate radii**



Voronoi diagram



Voronoi diagram




Fpocket

source: Zhou, Weiqiang, and Hong Yan. "Alpha shape and Delaunay triangulation in studies of protein-related interactions." Briefings in bioinformatics 15.1 (2014): 54-64.



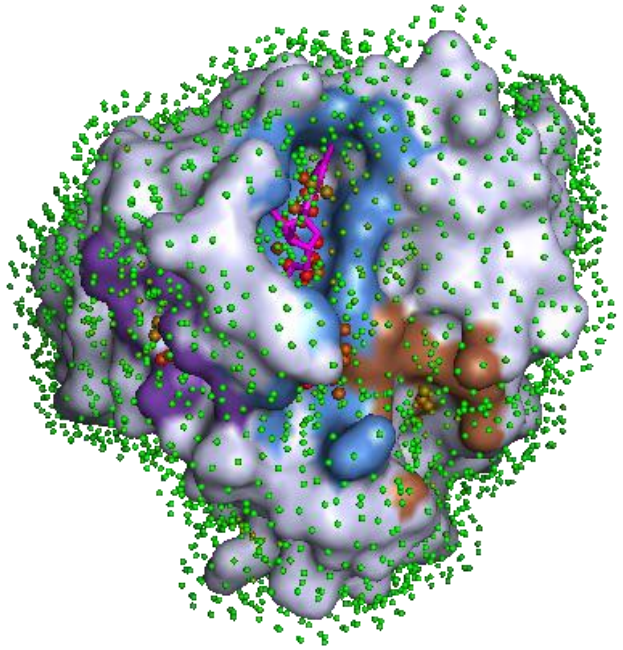
Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery. "Fpocket: an open source platform for ligand pocket detection." BMC bioinformatics 10 (2009): 1-11.

# Fpocket algorithm

- **Voronoi tessellation** and $\alpha$-sphere detection
  - Pruning based on max and min $\alpha$-sphere size → elimination of solvent inaccessible alpha spheres and too exposed $\alpha$-spheres
  - $\alpha$-sphere labeling – polar/apolar
- **Clustering** of $\alpha$-spheres
  - **Segmentation** – neighboring Voronoi vertices close enough form clusters → removal of singletons (large spheres on protein surface) + center of mass (CoM) of clusters computation
  - **Aggregation** of clusters having proximate CoMs
  - Multiple-linkage **clustering** → final clusters/**pockets**
  - **Drop small** and **hydrophobic** clusters
- Characterization and **ranking** of the pocket
  - Number of alpha spheres, alpha sphere density, local hydrophobicity, portion of apolar $\alpha$-spheres, polarity score

# P2RANK

- Based on the concept of ML-based prediction of **ligandability** of **surface points** and clustering





Krivák, Radoslav, and David Hoksza. 2018. “P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure.” Journal of Cheminformatics 10 (1): 39.

# Model construction

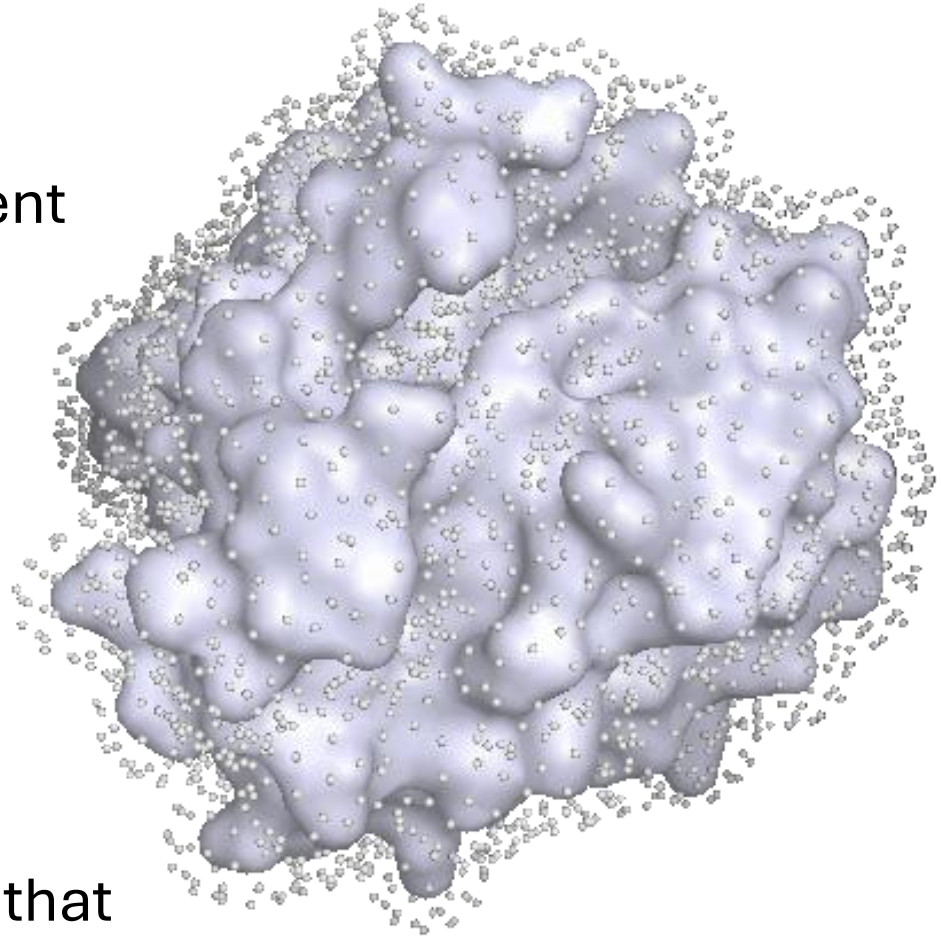
1. Obtaining known protein-ligand complexes

2. Cover the surface with a **mesh of points** (solvent accessible surface – SAS points)

3. **Extract a vector of physico-chemical and structural features** for each SAS point of each protein

4. **Label points** as binding/nonbinding

5. **Build a model** (RF) that is able for a given point (feature vector) to decide with what probability that point is part of a pocket

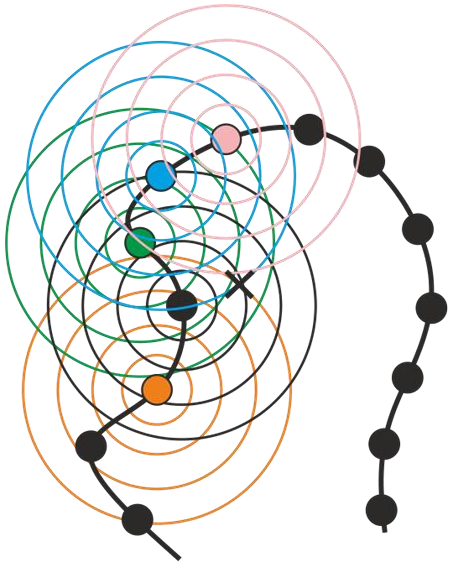


# Feature extraction

- **35 attributes** describing the local neighborhood of a given point based on structural properties and physical-chemical properties of neighboring amino acids

$$\mathrm{IFV}(P) = \sum_{\mathrm{A}_i \in\ \mathrm{A}(P)} \mathrm{AFV}\,(\mathrm{A}_i) \cdot w(\mathrm{dist}(P, \mathrm{A}_i)) \quad || \quad \mathrm{FV}(P)$$

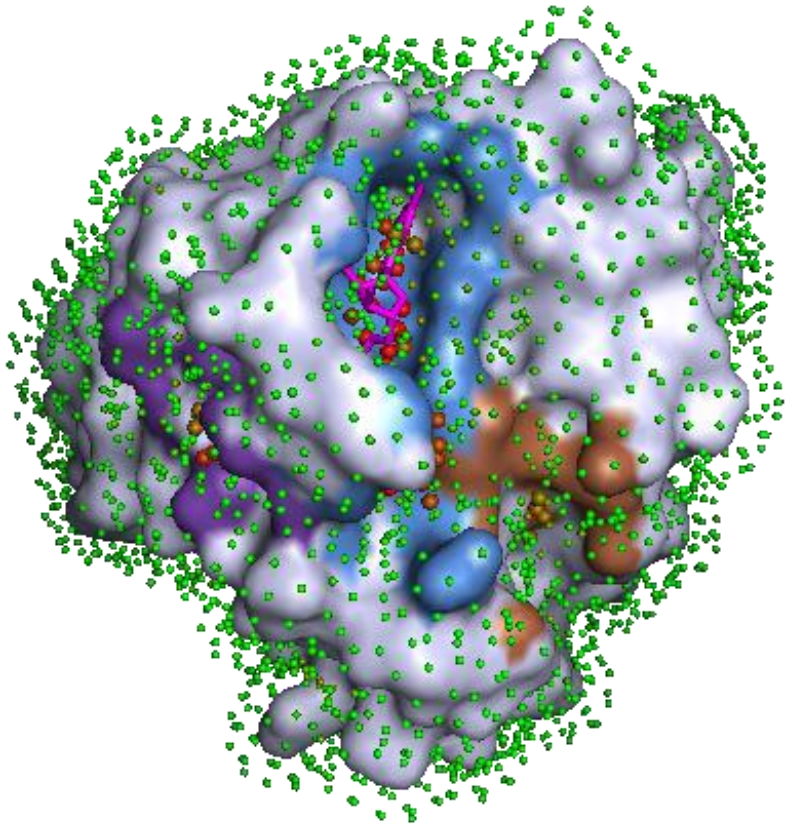
$$w(d) = 1 - d/8$$



| feature | importance |
|---|---|
| protrusion | 0.084528 |
| bfactor | 0.013888 |
| apRawInvalids | 0.011785 |
| vsAromatic | 0.010165 |
| apRawValids | 0.009403 |
| atomO | 0.009275 |
| hydrophobic | 0.008630 |
| hydrophilic | 0.007643 |
| vsAcceptor | 0.006244 |
| vsHydrophobic | 0.005273 |
| atoms | 0.005188 |
| aromatic | 0.004433 |
| atomN | 0.004236 |
| hydrophatyIndex | 0.004232 |
| atomC | 0.003687 |
| vsDonor | 0.003451 |
| aliphatic | 0.003350 |
| atomicHydrophobicity | 0.002663 |
| hBondDonorAcceptor | 0.002650 |
| hDonorAtoms | 0.002626 |
| atomDensity | 0.002549 |
| polar | 0.002402 |
| ionizable | 0.002142 |
| hAcceptorAtoms | 0.001904 |
| hBondAcceptor | 0.001705 |
| sulfur | 0.001621 |
| negCharge | 0.001538 |
| acidic | 0.001504 |
| basic | 0.001467 |
| hydroxyl | 0.001328 |
| vsAnion | 0.001072 |
| hBondDonor | 0.001059 |
| posCharge | 0.001021 |
| vsCation | 0.000832 |
| amide | 0.000831 |

# Inference

1. Cover the surface with a **mesh of SAS points**
2. **Apply the model to every point of the mesh** → ligandability score
3. **Filter** out points with low ligandability score
4. **Cluster** the remaining **points** → **binding pocket**
5. **Score the pockets** – cumulative ligandability score → raw score → confidence score
6. **Map** pocket SAS points onto atoms



| | COACH420 | | HOLO4K | |
|---|---|---|---|---|
| | Top-$n$ | Top-($n$+2) | Top-$n$ | Top-($n$+2) |
| Fpocket 1.0 | 56.4 | 68.9 | 52.4 | 63.1 |
| Fpocket 3.1 | 42.9 | 56.9 | 54.9 | 64.3 |
| SiteHound[a] | 53.0 | 69.3 | 50.1 | 62.1 |
| MetaPocket 2.0[a] | 63.4 | 74.6 | 57.9 | 68.6 |
| DeepSite[a] | 56.4 | 63.4 | 45.6 | 48.2 |
| P2Rank | 72.0 | **78.3** | 68.6 | 74.0 |
| P2Rank+Cons.[b] | **73.2** | 77.9 | **72.1** | **76.7** |

**Table 2.** Number of predicted binding sites and dataset statistics.

| | COACH420 | HOLO4K |
|---|---|---|
| Proteins | 420 | 4009 |
| Avg. protein atoms | 2179 | 3908 |
| Avg. ligands | 1.2 | 2.4 |
| Fpocket 1.0 | 14.6 | 27.0 |
| Fpocket 3.1 | 13.9 | 16.0 |
| SiteHound | 66.2 | 99.5 |
| MetaPocket 2.0 | 6.3 | 6.4 |
| DeepSite | 3.2 | 2.8 |
| P2Rank | 6.3 | 12.6 |
| P2Rank+Conservation | 3.4 | 7.7 |

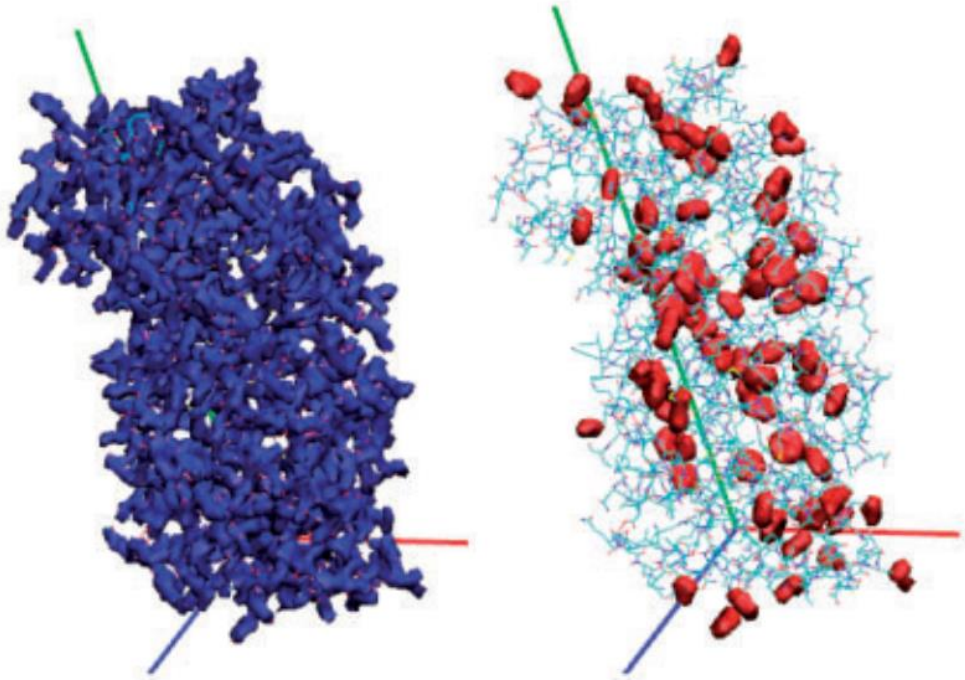
Displayed is the average total number of binding sites predicted per protein by each method on a given dataset.

# Deep-learning approaches

Structure

# DeepSite

- Main idea
  - **Voxelization** of the 3D space
  - Assigning **properties** to **voxels**
  - Application of **CNN**





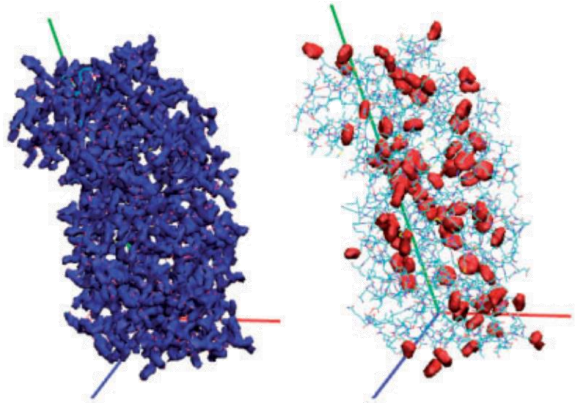
Jiménez, J., S. Doerr, G. Martínez-Rosell, A. S. Rose, and G. de Fabritiis. 2017. “DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks.” Bioinformatics 33 (19): 3036–42. ttps://doi.org/10.1093/bioinformatics/btx350.

# DeepSite – structure conversion

- Discretize 3D image into $1 \times 1 \times 1\ \dot{A}^3$ grid → **voxels**
- Assign **pharmacophoric properties** to atoms → channels
- Voxel **channel** values are function of distance of neighboring atom **properties**

| Element | Description |
|---|---|
| C | Non H-bonding aliphatic carbon |
| A | Non H-bonding aromatic carbon |
| NA | Acceptor 1 H-bond nitrogen |
| NS | Acceptor S Spherical nitrogen |
| OA | Acceptor 2 H-bonds oxygen |
| OS | Acceptor S Spherical oxygen |
| SA | Acceptor 2 H-bonds sulfur |
| HD | Donor 1 H-bond hydrogen |
| HS | Donor S Spherical hydrogen |
| MG | Non H-bonding magnesium |
| ZN | Non H-bonding zinc |
| MN | Non H-bonding manganese |
| CA | Non H-bonding calcium |
| FE | Non H-bonding iron |

| Property | Rule |
|---|---|
| Hydrophobic | atom type C or A |
| Aromatic | atom type A |
| Hydrogen bond acceptor | atom type NA or NS or OA or OS or SA |
| Hydrogen bond donor | atom type HD or HS with O or N partner |
| Positive ionizable | atom with positive charge |
| Negative ionizable | atom with negative charge |
| Metal | atom type MG or ZN or MN or CA or FE |
| Excluded volume | all atom types |

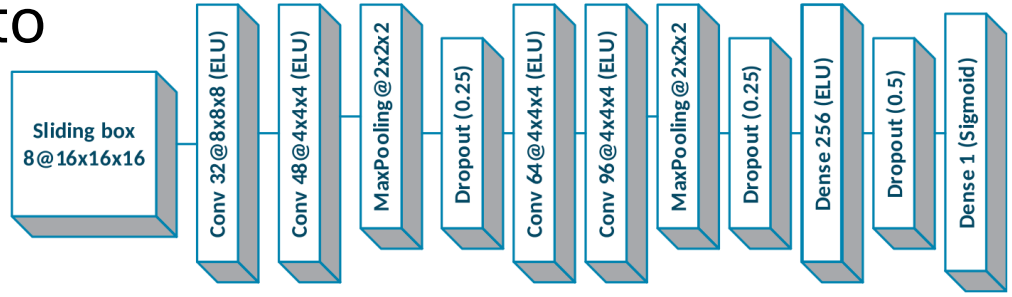


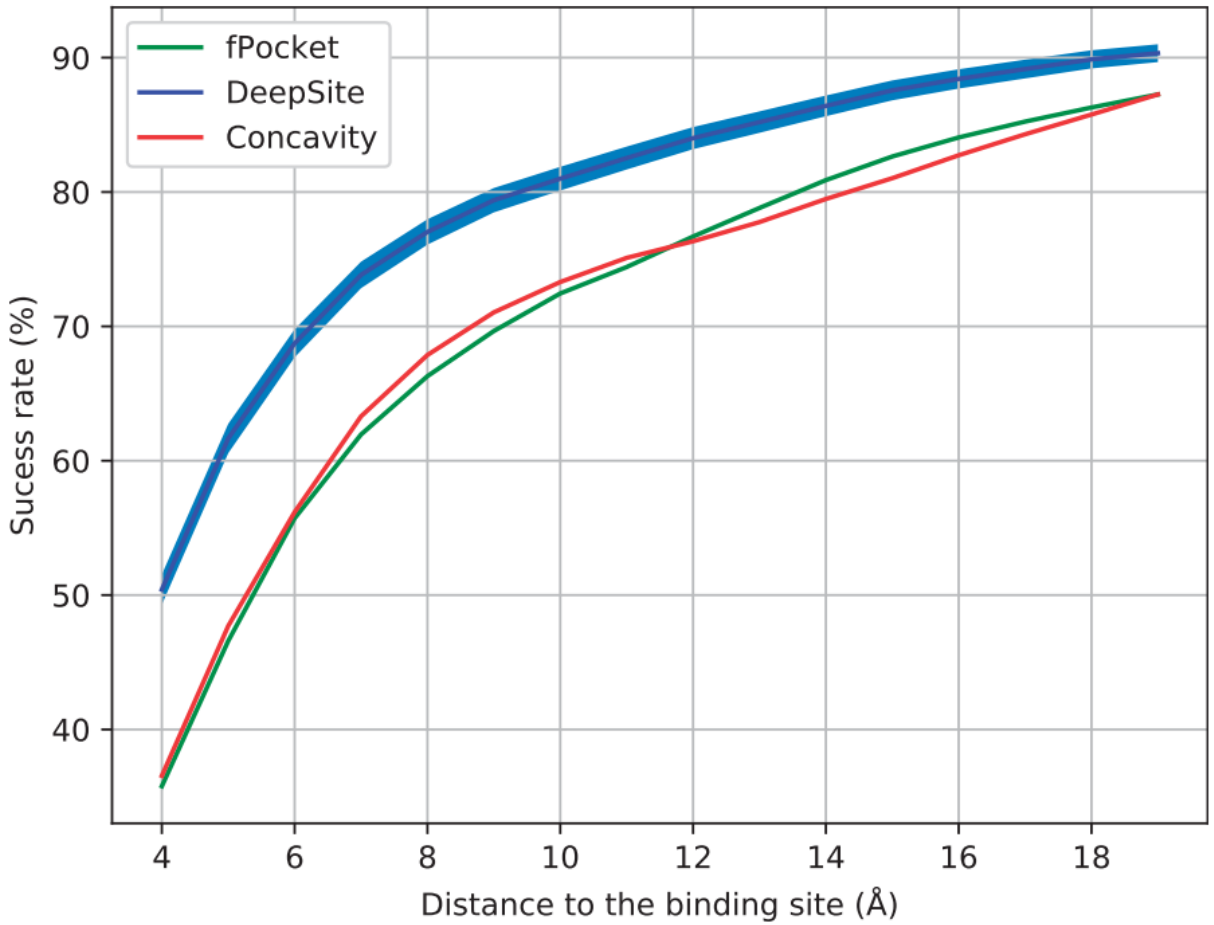
hydrophobic channel

aromatic channel

# DeepSite - classification

- Sliding **cuboid** window → **subgrids**
- Application of CCN to every voxel → a volumetric map
- **Low-probability** voxels **removed** and Mean-Shift **clustering** algorithm applied → resulting binding sites
- **Convolutional neural network** of 4 layers (smaller amount of training data compared to computer vision) that outputs subgrids **binding site label probability**
  - Class imbalance 1:100 of positive to negative → undersampling the majority class

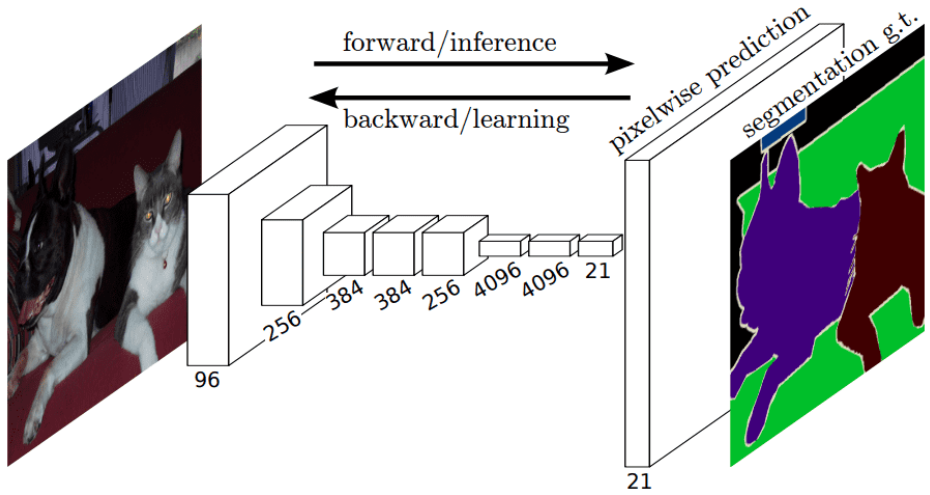





| | Average DVO | SD | *P*-value |
|---|---|---|---|
| DeepSite | **0.652** | 0.129 | — |
| fPocket | 0.619 | 0.169 | 0 |
| Concavity | 0.489 | 0.172 | 0 |

# Kalasanty

- Idea
  - Application of 3D **image segmentation**
    - 3D image ↔ protein
    - channels ↔ protein properties
    - objects ↔ pockets




source: https://theaisummer.com/unet-architectures/



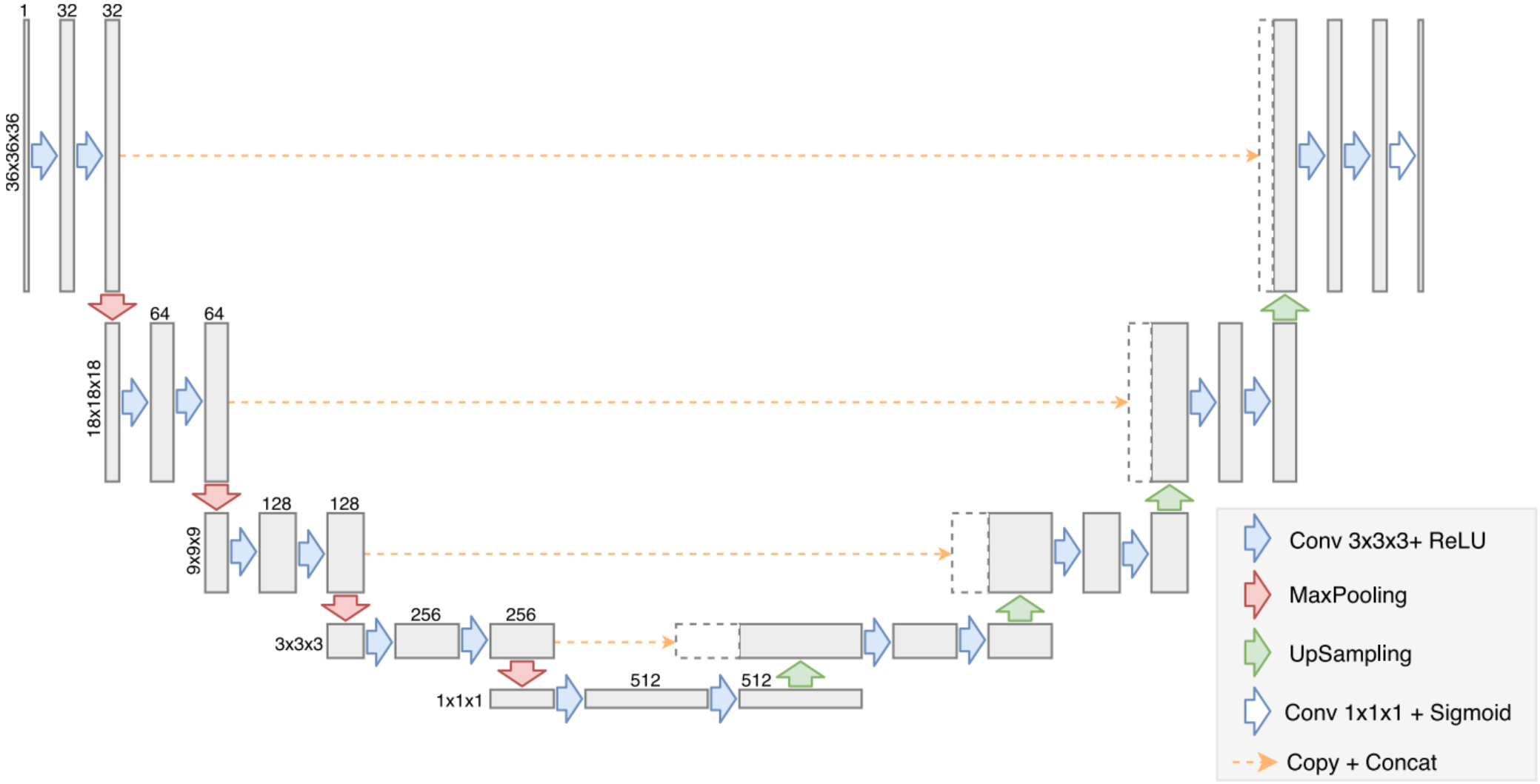
Stepniewska-Dziubinska, Marta M., Piotr Zielenkiewicz, and Pawel Siedlecki. 2020. "Improving Detection of Protein-Ligand Binding Sites with 3D Segmentation." Scientific Reports 10 (1): 1–9. https://doi.org/10.1038/s41598-020-61860-z.

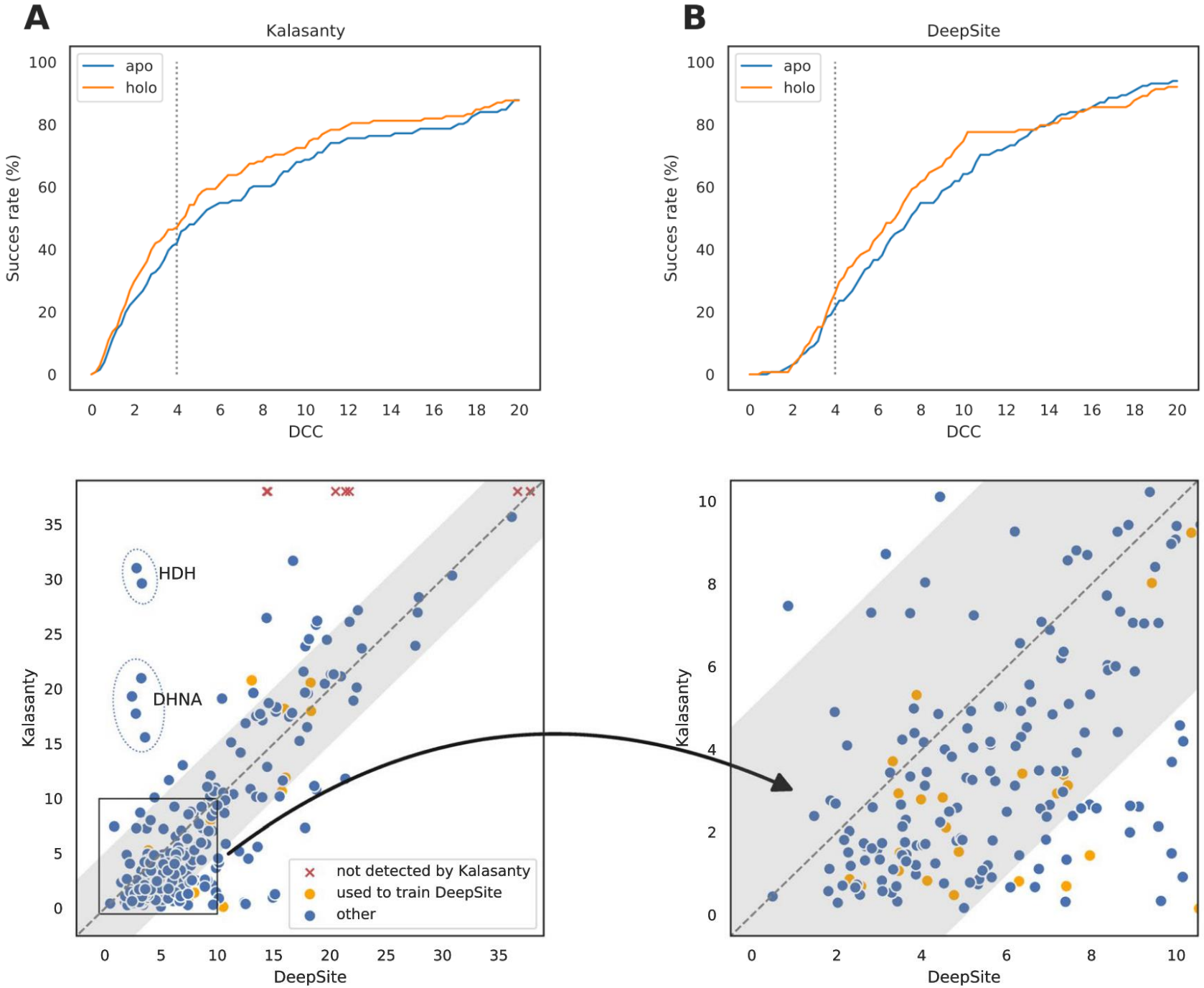
# Kalasanty

- Input: 3D grid
  - 2Å resolution centered on a protein center, 70Å in each direction
  - 18 channels – features such as atom type, partial charge, number of bonds with other atoms, ...
- Output: 3D grid
  - dimensions as input and a single channel representing pocket membership probability

- **U-Net** architecture (encoder-decoder CNN – see the following slide)
  - Training
    - 3D binary grid
    - Augmentation - random translations and rotations

# Kalasanty U-Net-like architecture

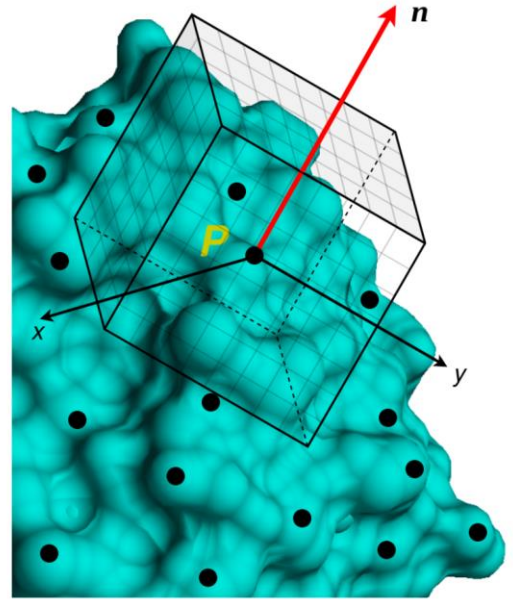


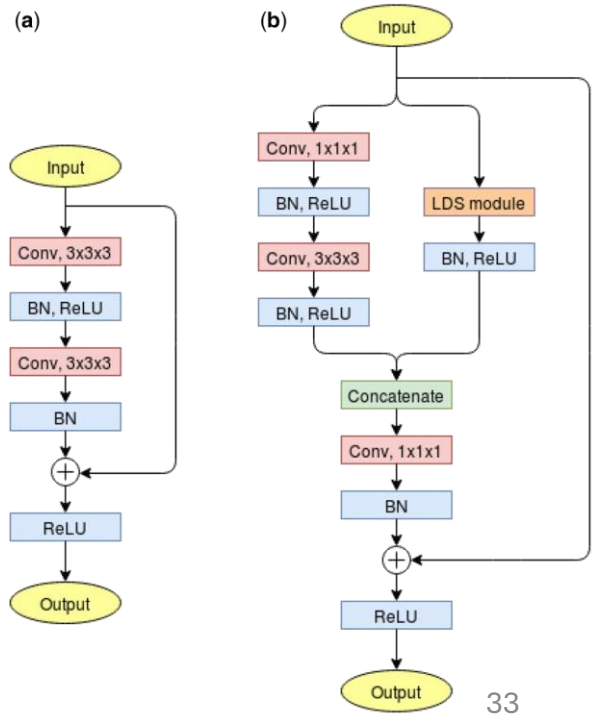



Comparison between DCC values achieved by Kalasanty and DeepSite

# DeepSurf

- Generate a reduced set of **SAS points** (k-means clustering)
- Create a **local grid** for each point and assign features to every voxel (Kalasanty)
- Use the **grid as input** to **CNN** (18-layer 3D-LDS-ResNet) → **ligandability** score
- **Filter** low-scoring point → **clustering** (Mean-Shift)
- **Map** clusters **onto atoms**






Mylonas, Stelios K., Apostolos Axenopoulos, and Petros Daras. "DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins." Bioinformatics 37.12 (2021): 1681-1690.

# Comparison of DL and non-DL approaches

| | COACH420 | | HOLO4K | |
|---|---|---|---|---|
| | **Top-n** | **Top-(n+2)** | **Top-n** | **Top-(n+2)** |
| Fpocket | 56.4 | 68.9 | 52.4 | 63.1 |
| Fpocket+PRANK[a] | 63.6 | 76.5 | 62.0 | 71.0 |
| SiteHound[†] | 53.0 | 69.3 | 50.1 | 62.1 |
| MetaPocket 2.0[†] | 63.4 | 74.6 | 57.9 | 68.6 |
| DeepSite[†] | 56.4 | 63.4 | 45.6 | 48.2 |
| P2Rank[protrusion][b] | 64.2 | 73.0 | 59.3 | 67.7 |
| P2Rank | *72.0* | *78.3* | *68.6* | *74.0* |

| | COACH420 | | HOLO4K | |
|---|---|---|---|---|
| | Top-n | Top-(n+2) | Top-n | Top-(n+2) |
| DeepSite[21] | 57.5 | 65.1 | 45.6 | 48.2 |
| Jiang *et al.*[22] | 55 | 58.7 | 38.2 | 41.5 |
| Kalasanty[23] | 68 | 70.4 | 32.1 | 32.3 |
| DeepSurf (ResNet-18) | 71.9 | 72.3 | 50.7 | 51.1 |
| DeepSurf (Bot-LDS-ResNet-18) | 71.3 | 72.9 | 50.4 | 50.9 |

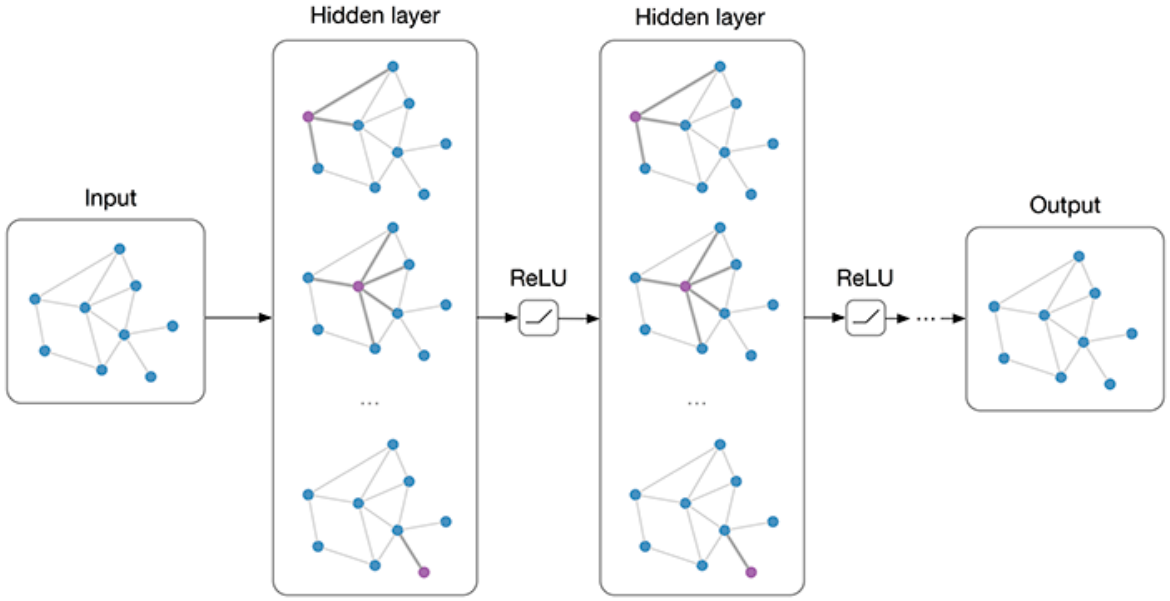
# Runtime

| Method | Time† |
|---|---|
| COACH (web server) | 15 h (self reported estimate) |
| eFindSite (web server) | 6.9 ± 0 h |
| COACH (stand-alone) | 6.4 ± 2 h |
| GalaxySite (web server) | 2 h (self reported estimate) |
| 3DLigandSite (web server) | 1–3 h (self reported estimate) |
| ISMBLab-LIG (web server) | 71 ± 2 min |
| FTSite (web server) | 39 ± 3 min |
| LISE (web server) | 39 ± 0.1 min |
| MetaPocket 2.0 (web server) | 2.8 ± 0.4 min |
| DeepSite (web server) | 38 ± 0.03 s |
| SiteHound (stand-alone) | 12 ± 0.5 s |
| P2Rank (stand-alone) | 6.8 ± 0.2 s (cold start*) |
| | 0.9 s (in larger dataset*) |
| Fpocket (stand-alone) | 0.2 ± 0.01 s |

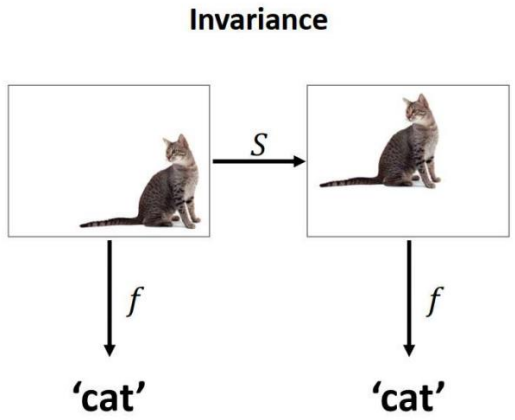
† Average time required for LBS prediction on a single protein. Displayed is self reported estimate or a result of our test on a small dataset of 5 proteins á ~ 2500 atoms. Stand-alone tools were tested on a single 3.7 GHz CPU core. For web servers the wall time from submitting a job to receiving the result was measured.

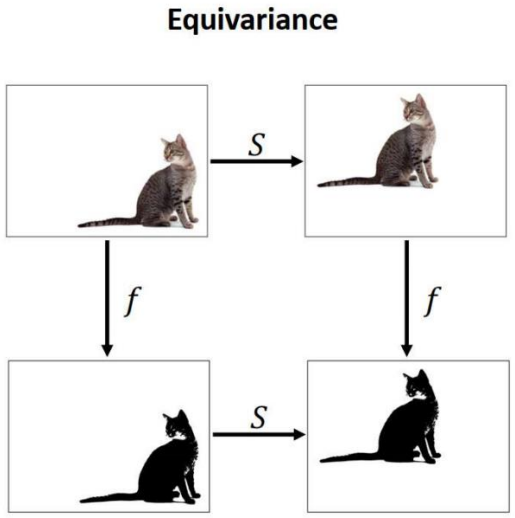
*Difference is due to JVM initialization and model loading cost

# Equivariant graph neural networks



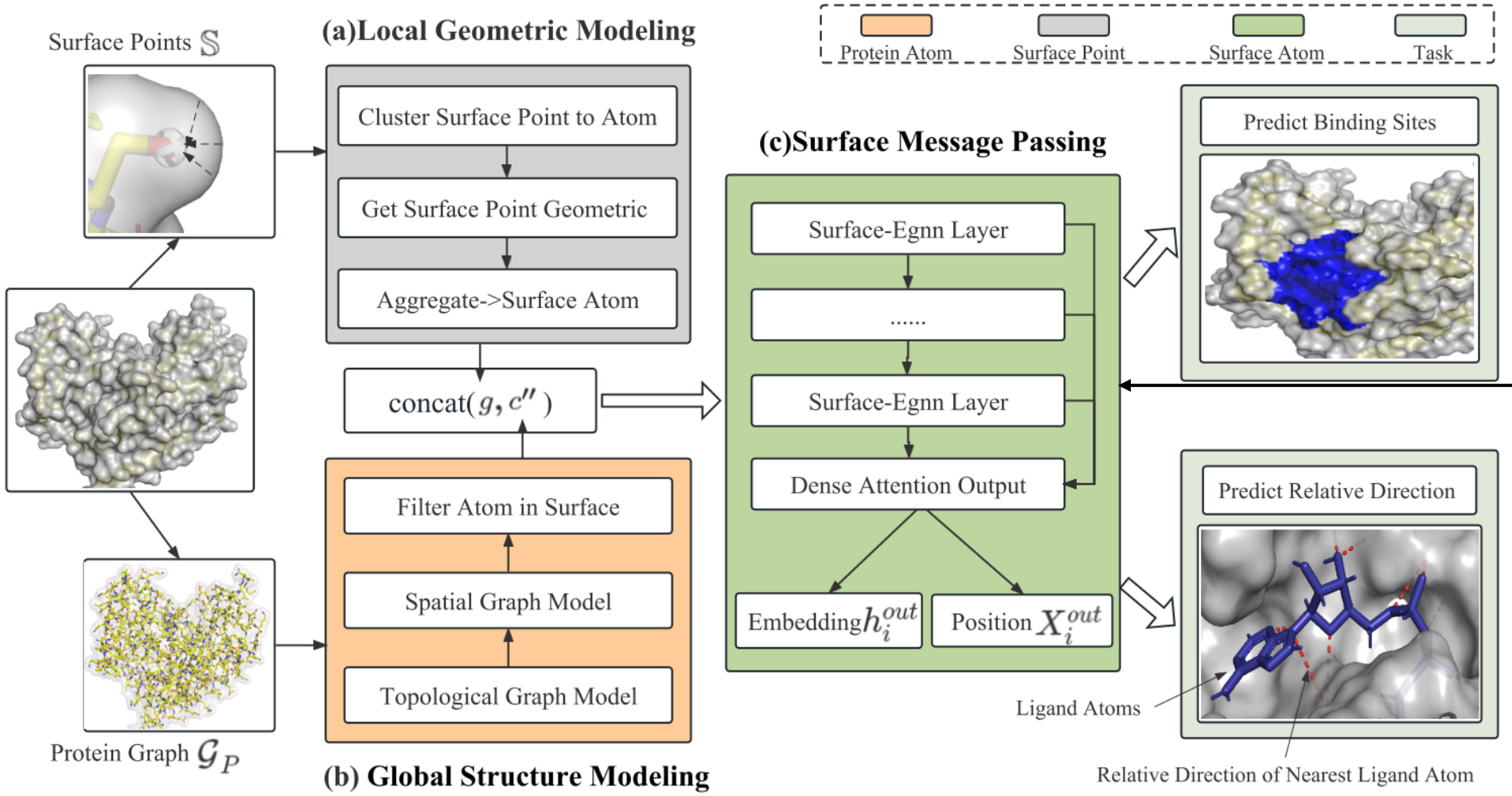

source: https://tkipf.github.io/graph-convolutional-networks/




source: https://www.doc.ic.ac.uk/~bkainz/teaching/DL/notes/equivariance.pdf

# EquiPocket




$$h_i^{(0)} = [c_i'', g_i]$$

$$X_i^{(0)} = [x_i, \bar{x}_i]$$

$$m_{ij} = \phi_m\left(h_i^{(l)}, h_j^{(l)}, f_x(X_i^{(l)}, X_j^{(l)}), e_{ij}\right),$$

$$h_i^{(l+1)} = \phi_h\left(h_i^{(l)}, \sum_{j\in\mathbb{N}(i)} m_{ij}\right),$$

$$X_i^{(l+1)} = X_i^{(l)} + \frac{1}{|\mathbb{N}(i)|}\sum_{j\in\mathcal{N}(i)}(x_i^{(l)} - x_j^{(l)})\phi_x(m_{ij}),$$

$$f_x(X_i, X_j) := [\|x_{ij}\|_2, \|x_{ci}\|_2, \|x_{cj}\|_2, \angle_{ci,ij}, \angle_{cj,ij}, \angle_{ci,cj}]$$

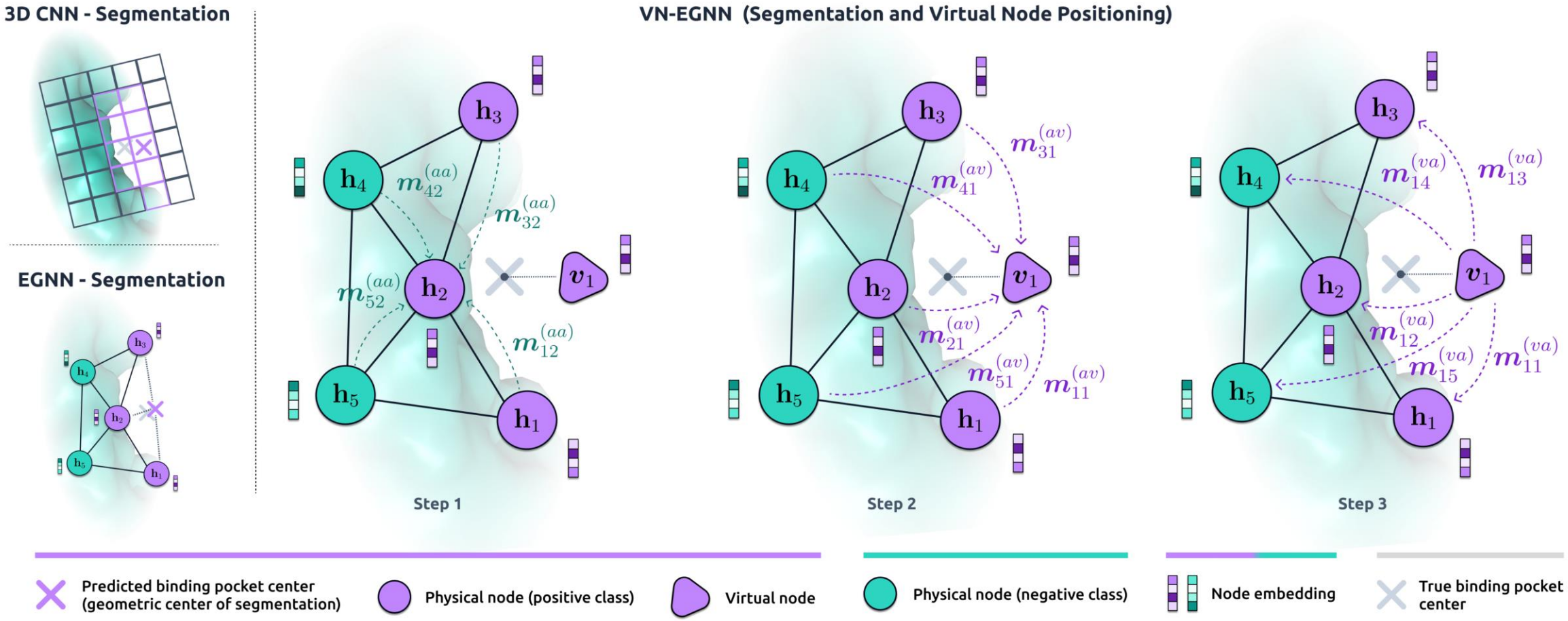
$\phi_m, \phi_h, \phi_x$ are MLPs

E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction



Zhang, Yang, et al. "Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction." arXiv preprint arXiv:2302.12177 (2023).

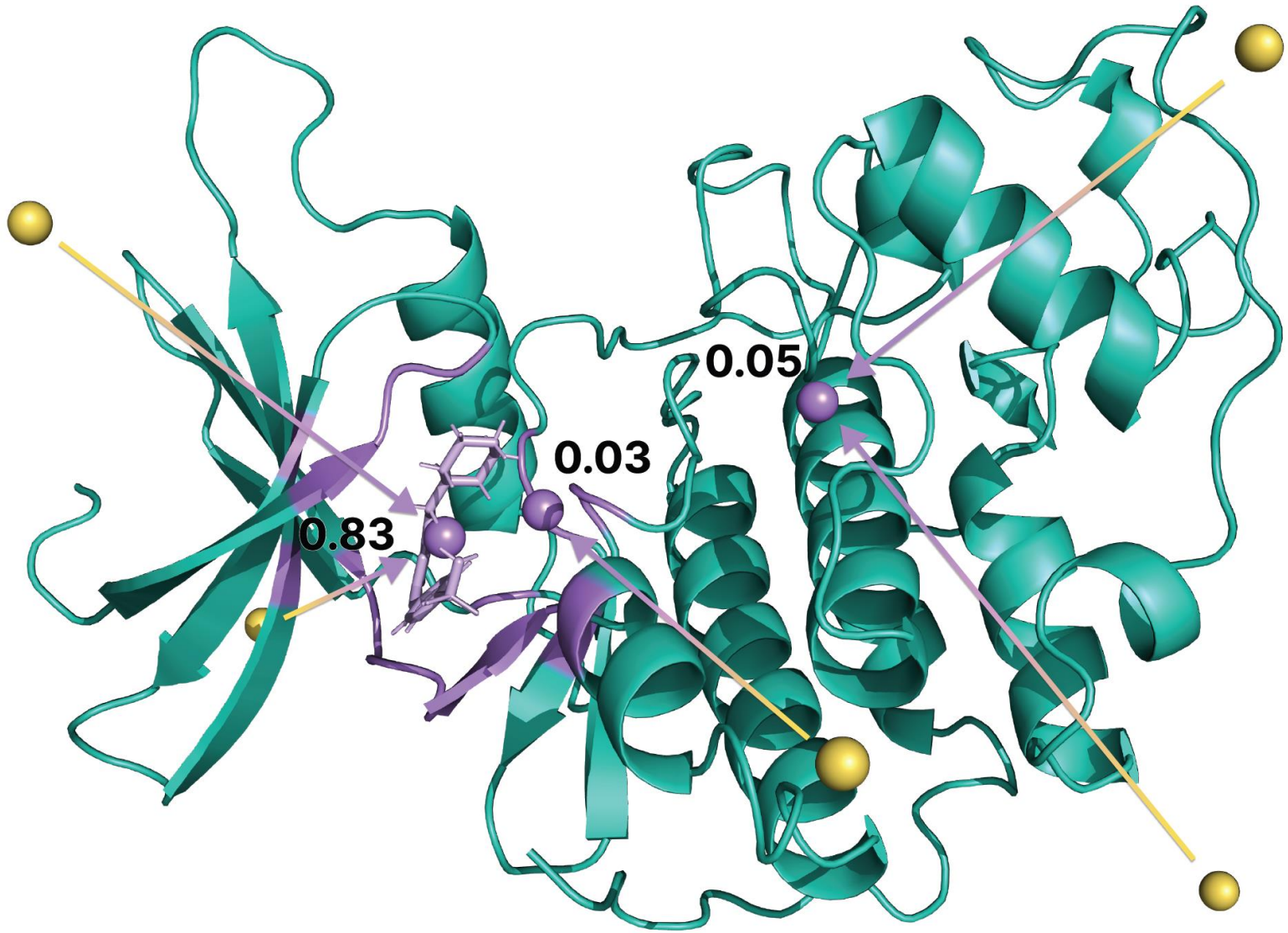
# VN-EGNN




E(3)-Equivariant Graph Neural Networks with Virtual Nodes Enhance Protein Binding Site Identification



Sestak, Florian, et al. "VN-EGNN: E (3)-Equivariant Graph Neural Networks with Virtual Nodes Enhance Protein Binding Site Identification." arXiv preprint arXiv:2404.07194 (2024).


0.05
0.03
0.83

| Methods | Param (M) | COACH420 | | HOLO4K[d] | | PDBbind2020 | |
|---|---|---|---|---|---|---|---|
| | | DCC↑ | DCA↑ | DCC↑ | DCA↑ | DCC↑ | DCA↑ |
| Fpocket (Le Guilloux et al., 2009)[b] | \ | 0.228 | 0.444 | 0.192 | 0.457 | 0.253 | 0.371 |
| P2Rank (Krivák & Hoksza, 2018)[c] | \ | *0.464* | *0.728* | *0.474* | **0.787** | *0.653* | **0.826** |
| DeepSite (Jiménez et al., 2017)[b] | 1.00 | \ | 0.564 | \ | 0.456 | \ | \ |
| Kalasanty (Stepniewska-Dziubinska et al., 2020)[b] | 70.64 | 0.335 | 0.636 | 0.244 | 0.515 | 0.416 | 0.625 |
| DeepSurf (Mylonas et al., 2021)[b] | 33.06 | 0.386 | 0.658 | 0.289 | 0.635 | 0.510 | 0.708 |
| DeepPocket (Aggarwal et al., 2022b)[c] | \ | 0.399 | 0.645 | 0.456 | 0.734 | 0.644 | **0.813** |
| GAT (Veličković et al., 2018)[b] | **0.03** | 0.039(0.005) | 0.130(0.009) | 0.036(0.003) | 0.110(0.010) | 0.032(0.001) | 0.088(0.011) |
| GCN (Kipf & Welling, 2017)[b] | 0.06 | 0.049(0.001) | 0.139(0.010) | 0.044(0.003) | 0.174(0.003) | 0.018(0.001) | 0.070(0.002) |
| GAT + GCN[b] | 0.08 | 0.036(0.009) | 0.131(0.021) | 0.042(0.003) | 0.152(0.020) | 0.022(0.008) | 0.074(0.007) |
| GCN2 (Chen et al., 2020)[b] | 0.11 | 0.042(0.098) | 0.131(0.017) | 0.051(0.004) | 0.163(0.008) | 0.023(0.007) | 0.089(0.013) |
| SchNet (Schütt et al., 2017)[b] | 0.49 | 0.168(0.019) | 0.444(0.020) | 0.192(0.005) | 0.501(0.004) | 0.263(0.003) | 0.457(0.004) |
| EGNN (Satorras et al., 2021)[b] | 0.41 | 0.156(0.017) | 0.361(0.020) | 0.127(0.005) | 0.406(0.004) | 0.143(0.007) | 0.302(0.006) |
| EquiPocket (Zhang et al., 2023b)[b] | 1.70 | 0.423(0.014) | 0.656(0.007) | 0.337(0.006) | *0.662*(0.007) | 0.545(0.010) | 0.721(0.004) |
| VN-EGNN (ours) | 1.20 | **0.605(0.009)** | **0.750(0.008)** | **0.532(0.021)** | 0.659(0.026) | **0.669(0.015)** | **0.820(0.010)** |

[a] The standard deviation across training re-runs is indicated in parentheses. [b] Results from Zhang et al. (2023b). [c] Uses different training set and, thus, limited comparability. [d] This dataset represents a strong domain shift from the training data for all methods (except for P2Rank). Details on the domain shift in Section I.

# Deep-learning approaches

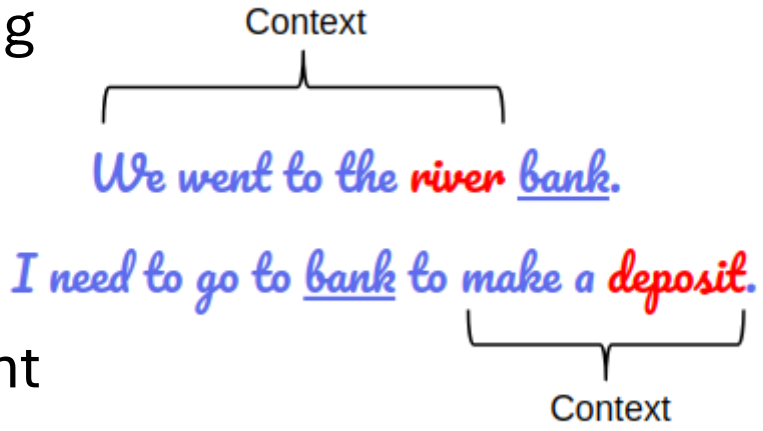
Sequence

# Language models recap

- Vector-based  word (token) representations/embeddings - similar words similar representations

- Context independent (Word2Vec, …)
  - No need for a “model” (dictionary of word-embedding pairs)

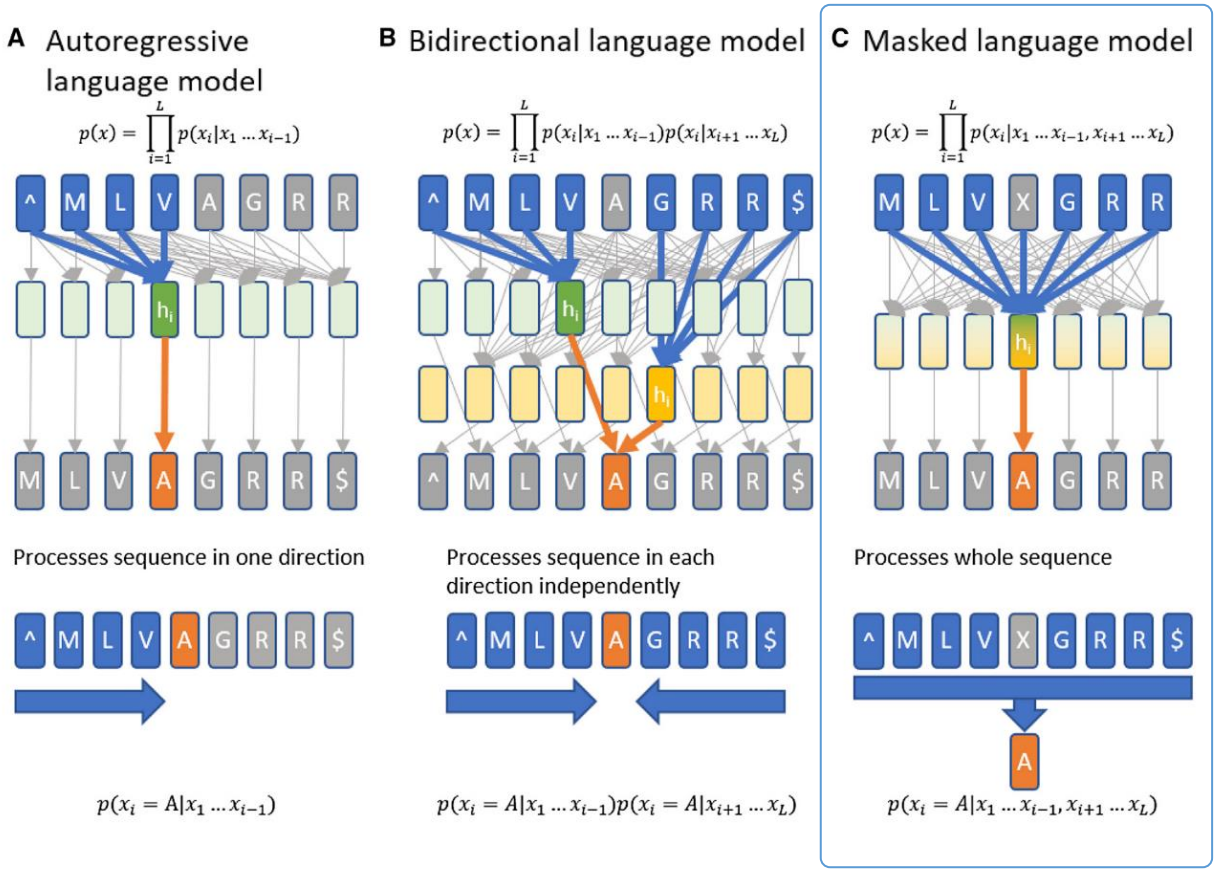
- **Context dependent** (BERT, …)
  - Different representation for the same word in different contexts – model accepting sentence as input



# Protein Language Models

token ↔ AA

sentence ↔ protein



# ProtTrans

- Pretrained language models (BERT, Albert, Electra T5, ...) for proteins

| *Data LM* | *UniRef50* | *UniRef100* | *BFD* |
|---|---|---|---|
| *Number proteins* [in m] | 45 | 216 | 2,122 |
| *Number of amino acids* [in b] | 14 | 88 | 393 |
| *Disk space* [in GB] | 26 | 150 | 572 |



Elnaggar, Ahmed, et al. "Prottrans: Toward understanding the language of life through self-supervised learning." IEEE transactions on pattern analysis and machine intelligence 44.10 (2021): 7112-7127.

# ProtTrans - models

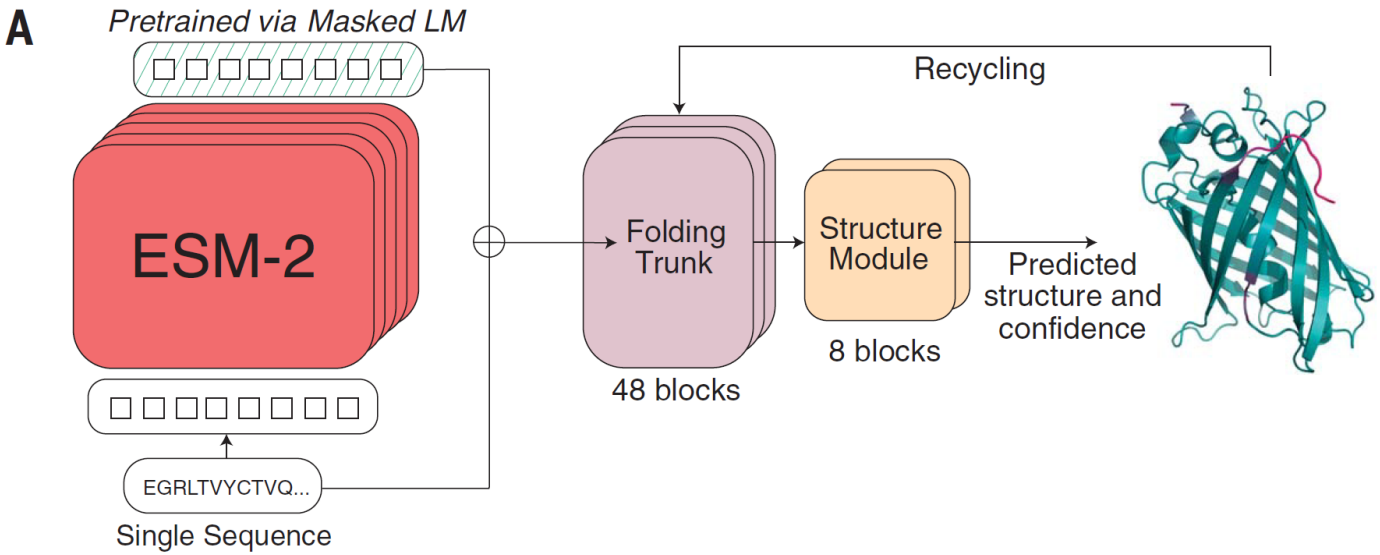
https://github.com/agemagician/ProtTrans

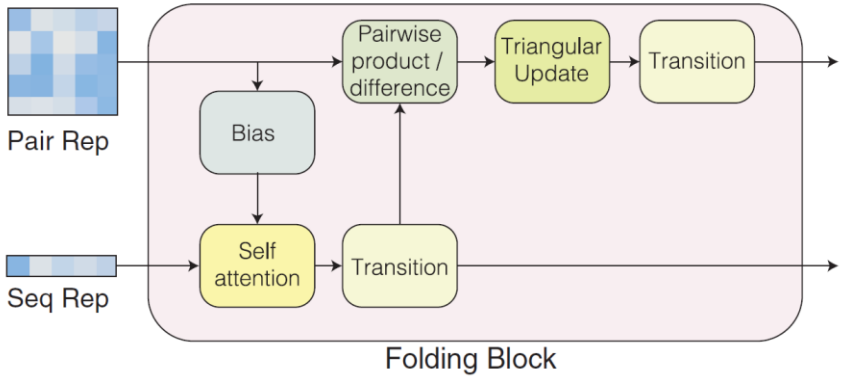
https://github.com/sacdallago/bio_embeddings

| Hyperparameter | ProtTXL | | ProtBert | | ProtXLNet | ProtAlbert | ProtElectra | ProtT5-XL | | ProtT5-XXL | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | BFD100 | UniRef100 | BFD100 | UniRef100 | UniRef100 | UniRef100 | UniRef100 | UniRef50 | BFD100 | UniRef50 | BFD100 |
| Number of Layers | 32 | 30 | 30 | | 30 | 12 | 30 | 24 | | 24 | |
| Hidden Layers Size | 1024 | | 1024 | | 1024 | 4096 | 1024 | 1024 | | 1024 | |
| Hidden Layers Intermediate Size | 4096 | | 4096 | | 4096 | 16384 | 4096 | 16384 | | 65536 | |
| Number of Heads | 14 | 16 | 16 | | 16 | 64 | 16 | 32 | | 128 | |
| Positional Encoding Limits | - | | 40K | | - | 40K | 40K | - | | - | |
| Dropout | 0.15 | | 0.0 | | 0.1 | 0.0 | 0.0 | 0.1 | | 0.1 | 0.0 |
| Target Length | 512 | | 512/2048 | | 512 | 512/2048 | 512/1024 | 512 | | 512 | |
| Memory Length | 512 | | - | | 384 | - | - | - | | - | |
| Masking Probability | - | | 15% | | - | 15% | 25% | 15% | | 15% | |
| Local Batch Size | 8 | 5 | 32/6 | 30/5 | 2 | 21/2 | 18/7 | 8 | 4 | 8 | 4 |
| Global Batch Size | 44928 | 22464 | 32768/6144 | 15360/2560 | 1024 | 10752/1024 | 9216/3584 | 2048 | 4096 | 2048 | 4096 |
| Optimizer | Lamb | | Lamb | | Adam | Lamb | Lamb | AdaFactor | | AdaFactor | |
| Learning Rate | 0.0005 | 0.002 | 0.002 | | 0.00001 | 0.002 | 0.002 | 0.01 | | 0.01 | |
| Weight Decay | 0.0 | 0.01 | 0.01 | | 0.01 | 0.01 | 0.01 | 0.0 | | 0.0 | |
| Training Steps | 40.7K | 31.3K | 800K/200K | 300K/100K | 847K | 150K/150K | 400K/400K | 991K | 1.2M | 343K | 920K |
| Warm-up Steps | 13.6K | 5.5K | 140K/20K | 40K/0K | 20K | 40K/5K | 40K/40K | 10K | | 10K | |
| Mixed Precision | FP16 Model Weight<br>Fp32 Master Weight | | None | | None | None | None | None | | None | |
| Number of Parameters | 562M | 409M | 420M | | 409M | 224M | 420M | 3B | | 11B | |
| System | Summit | Summit | TPU Pod | | TPU Pod | TPU Pod | TPU Pod | TPU Pod | | TPU Pod | |
| Number of Nodes | 936 | | 128 | 64 | 64 | 64 | 64 | 32 | 128 | 32 | 128 |
| Number of GPUs/TPUs | 5616 | | 1024 | 512 | 512 | 512 | 512 | 256 | 1024 | 256 | 1024 |

# ESM-2 pLM

https://github.com/facebookresearch/esm

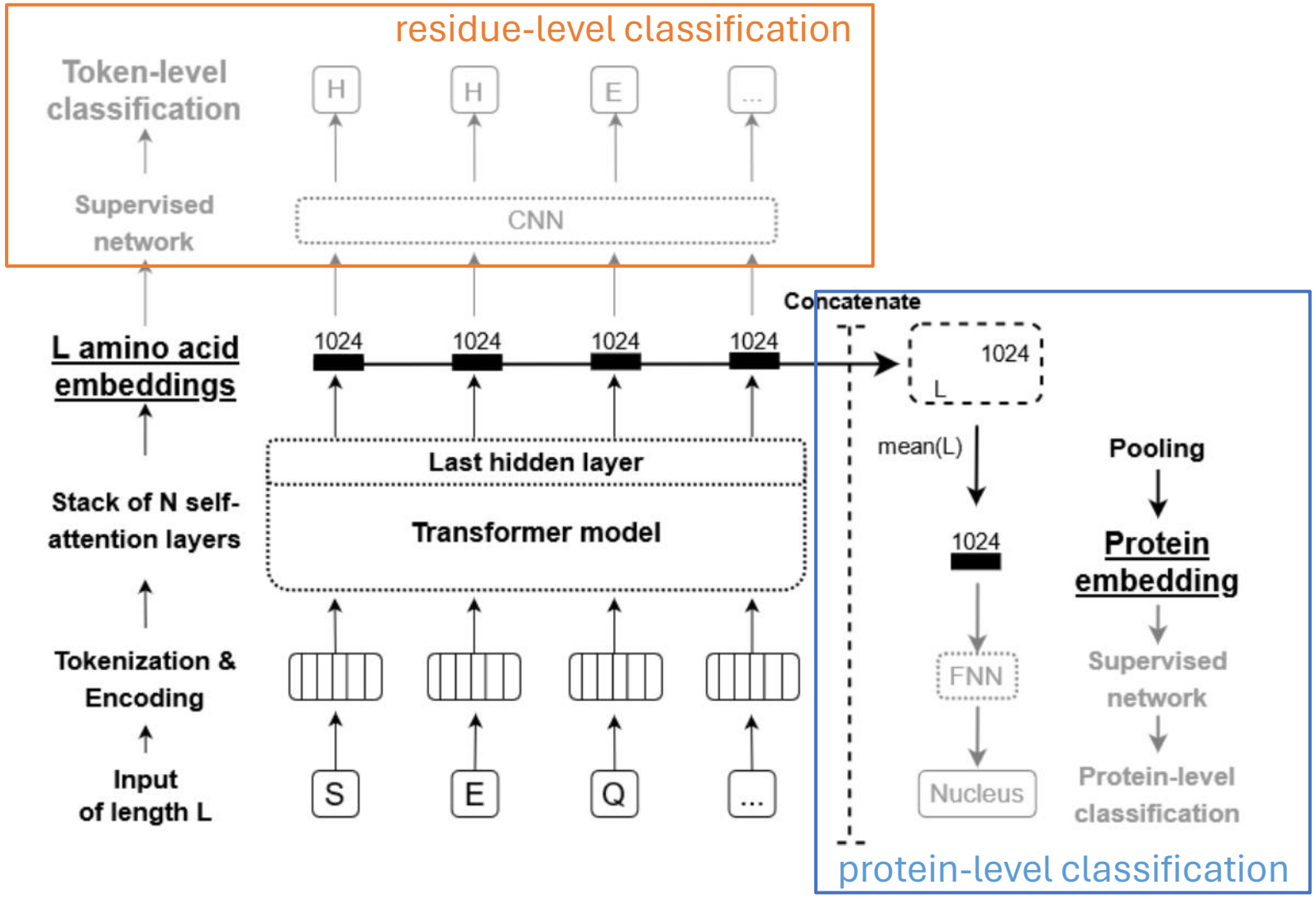




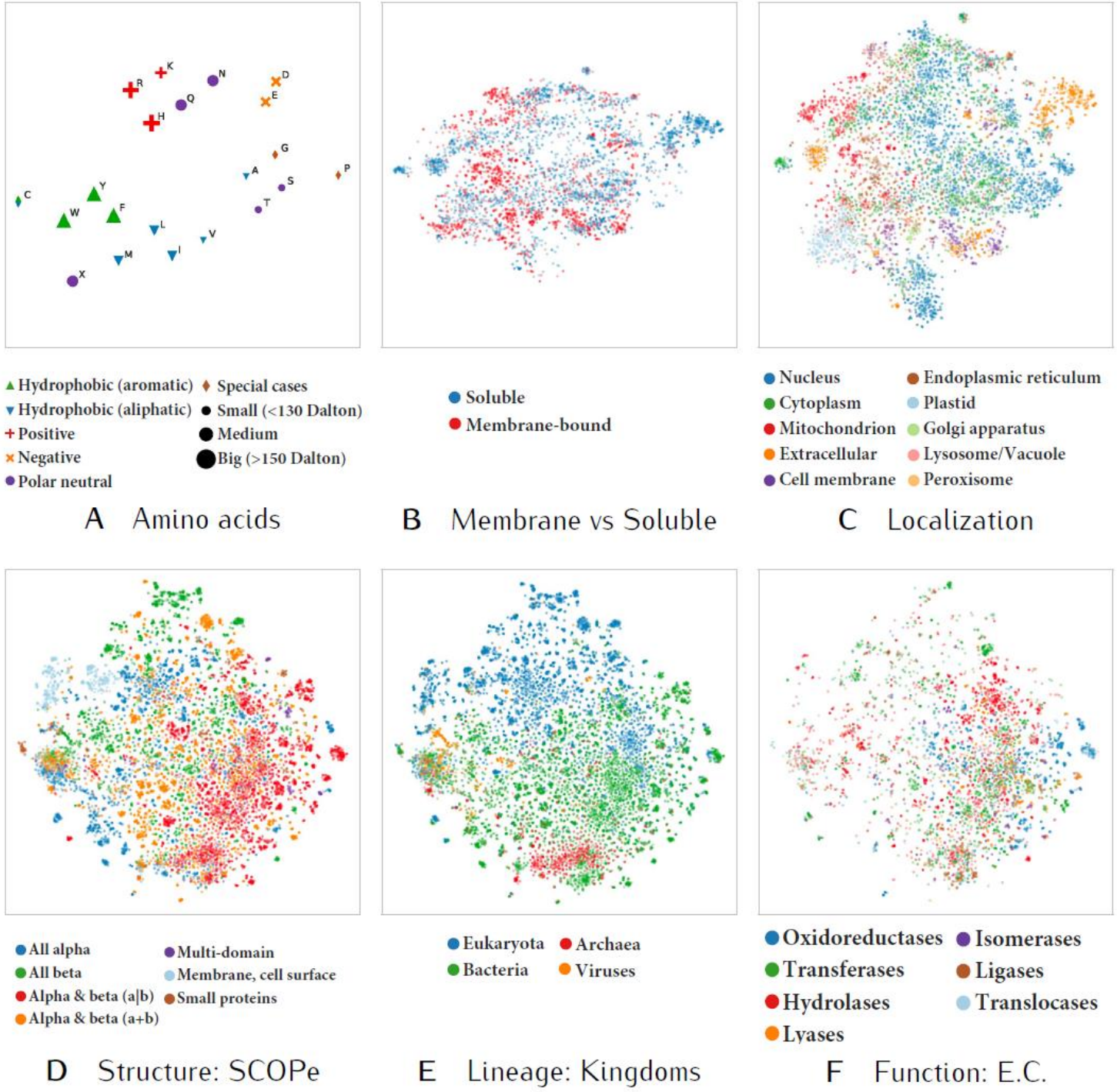

Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379.6637 (2023): 1123-1130.

# Transfer learning with pLMs

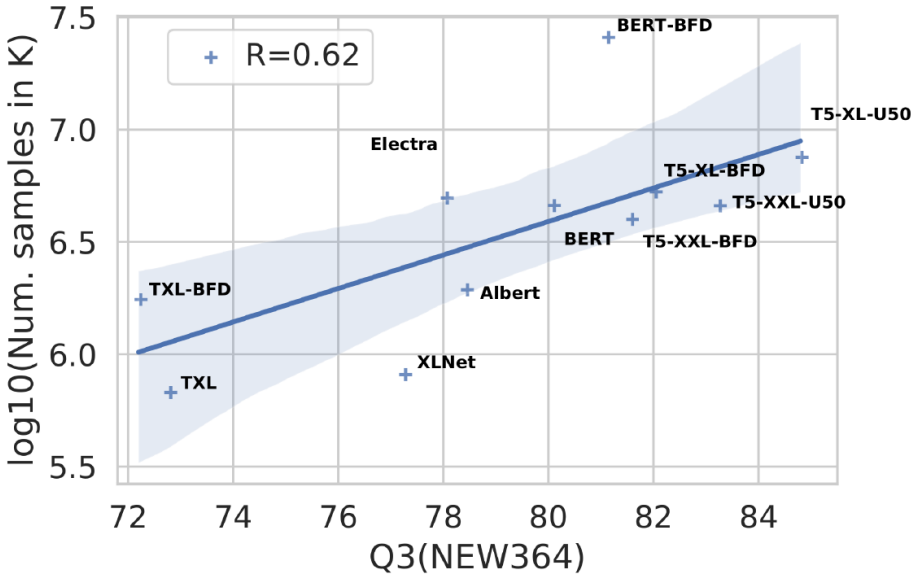


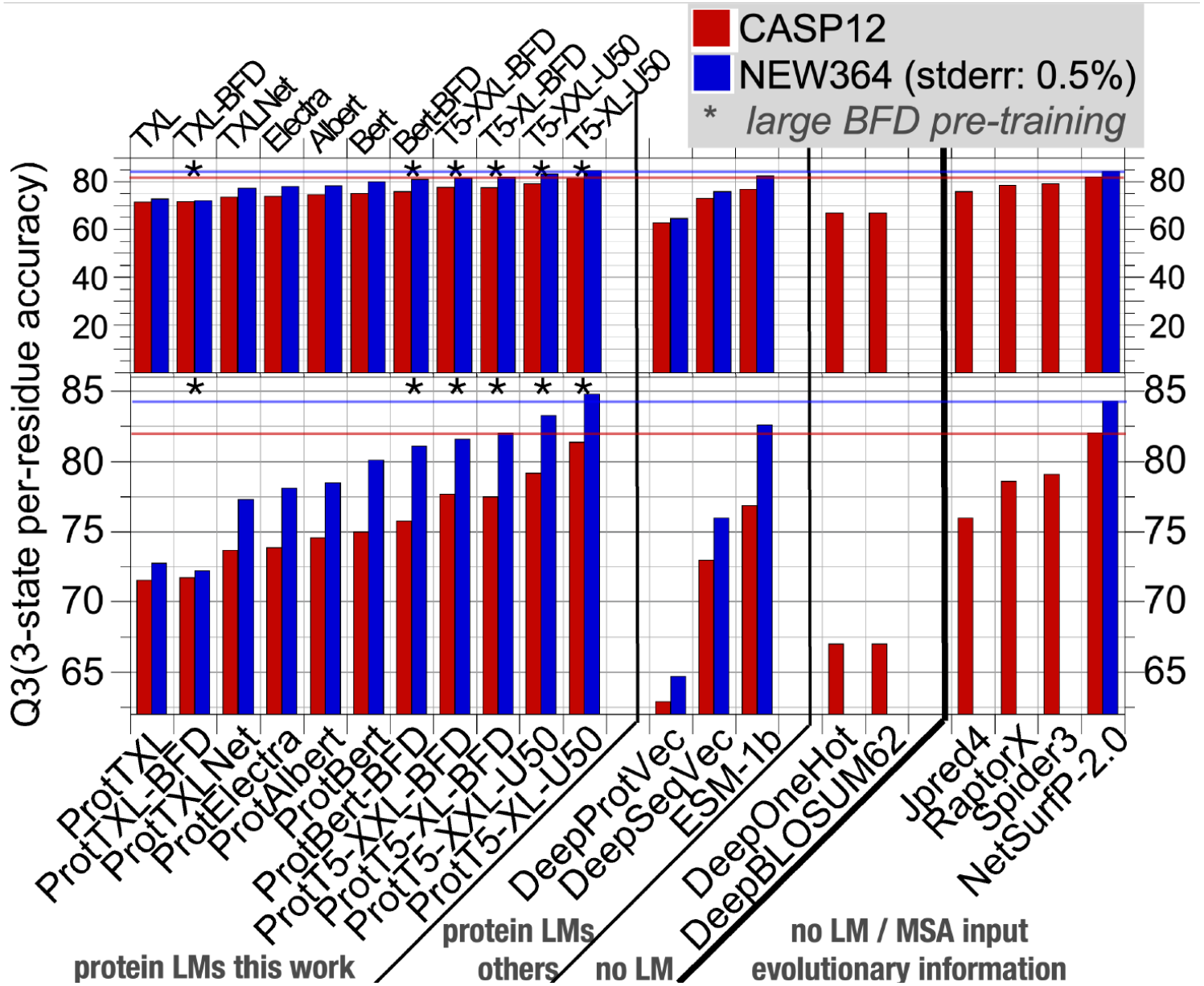

source: Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, et al. 2021. "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning." IEEE TRANS PATTERN ANALYSIS & MACHINE INTELLIGENCE. Vol. 14..

# Embeddings properties – unsupervised

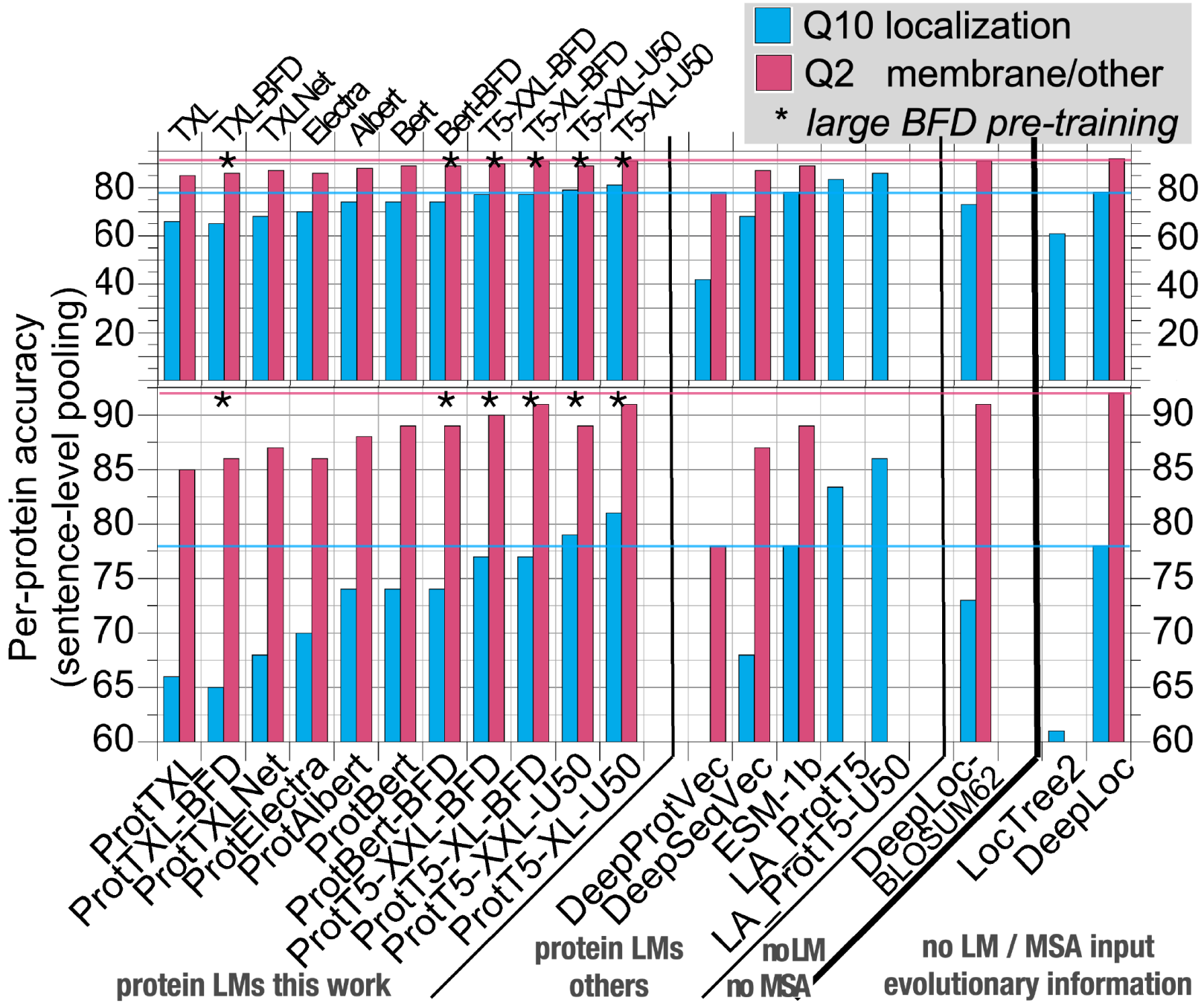


# Supervised residue-level classification

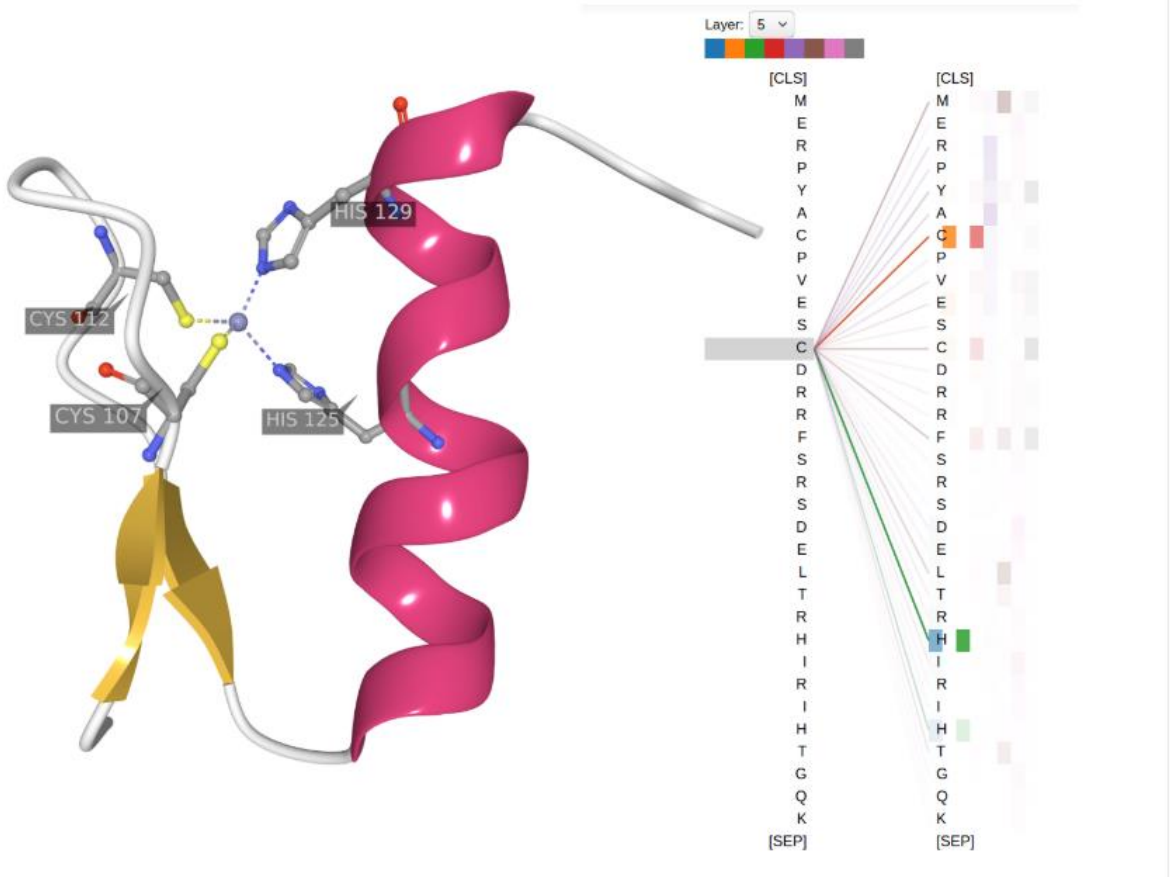


# Supervised protein-level classification

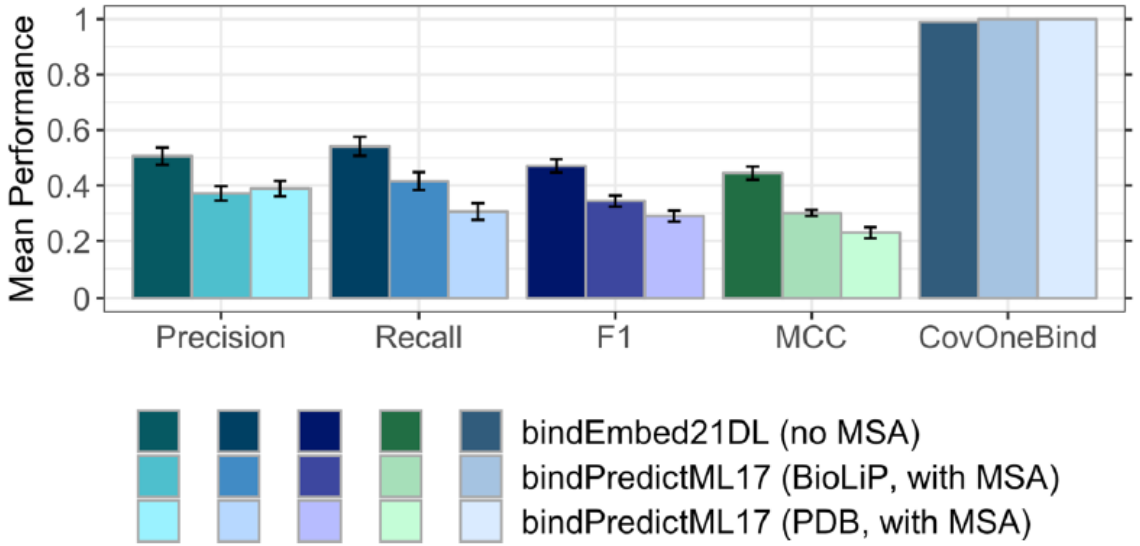


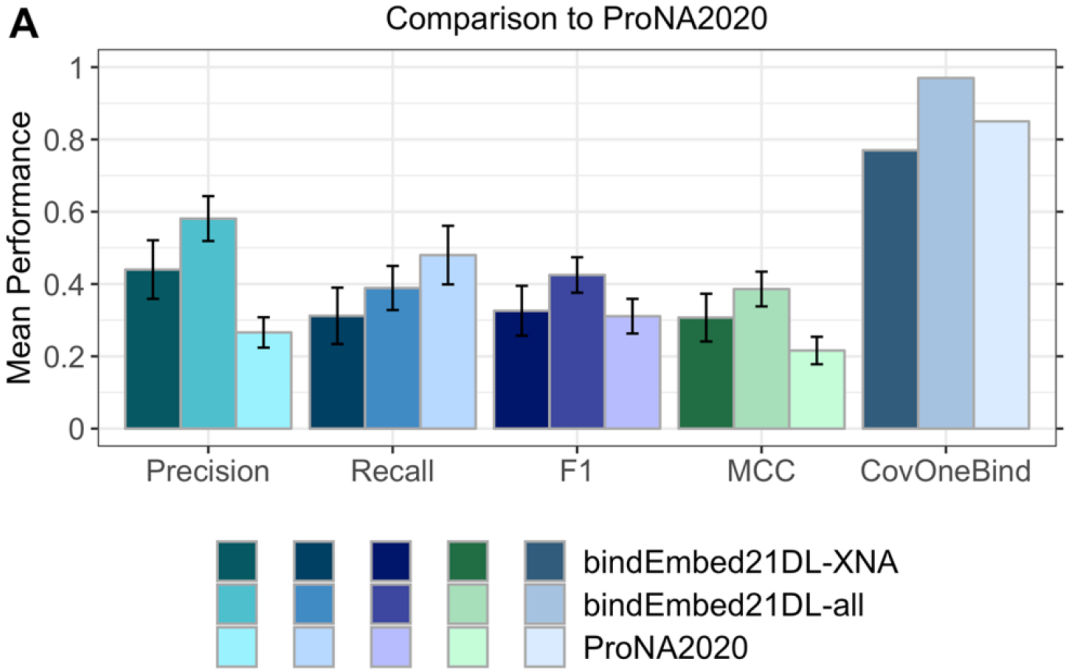
# Interpretation through attention visualization

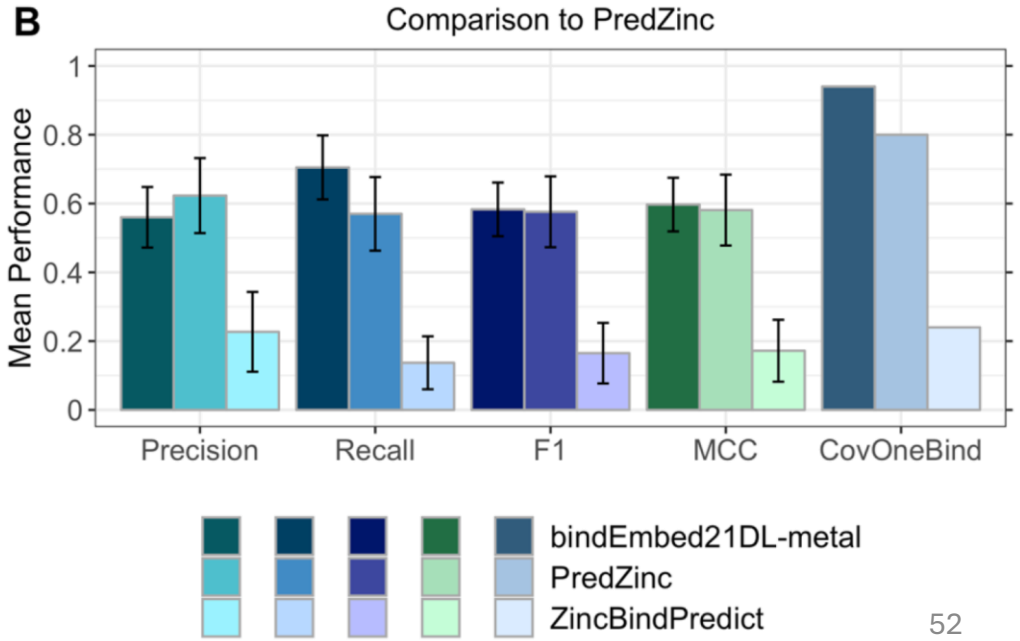


# bindEmbed21DL

- ProtT5 + shallow two-layer CNN



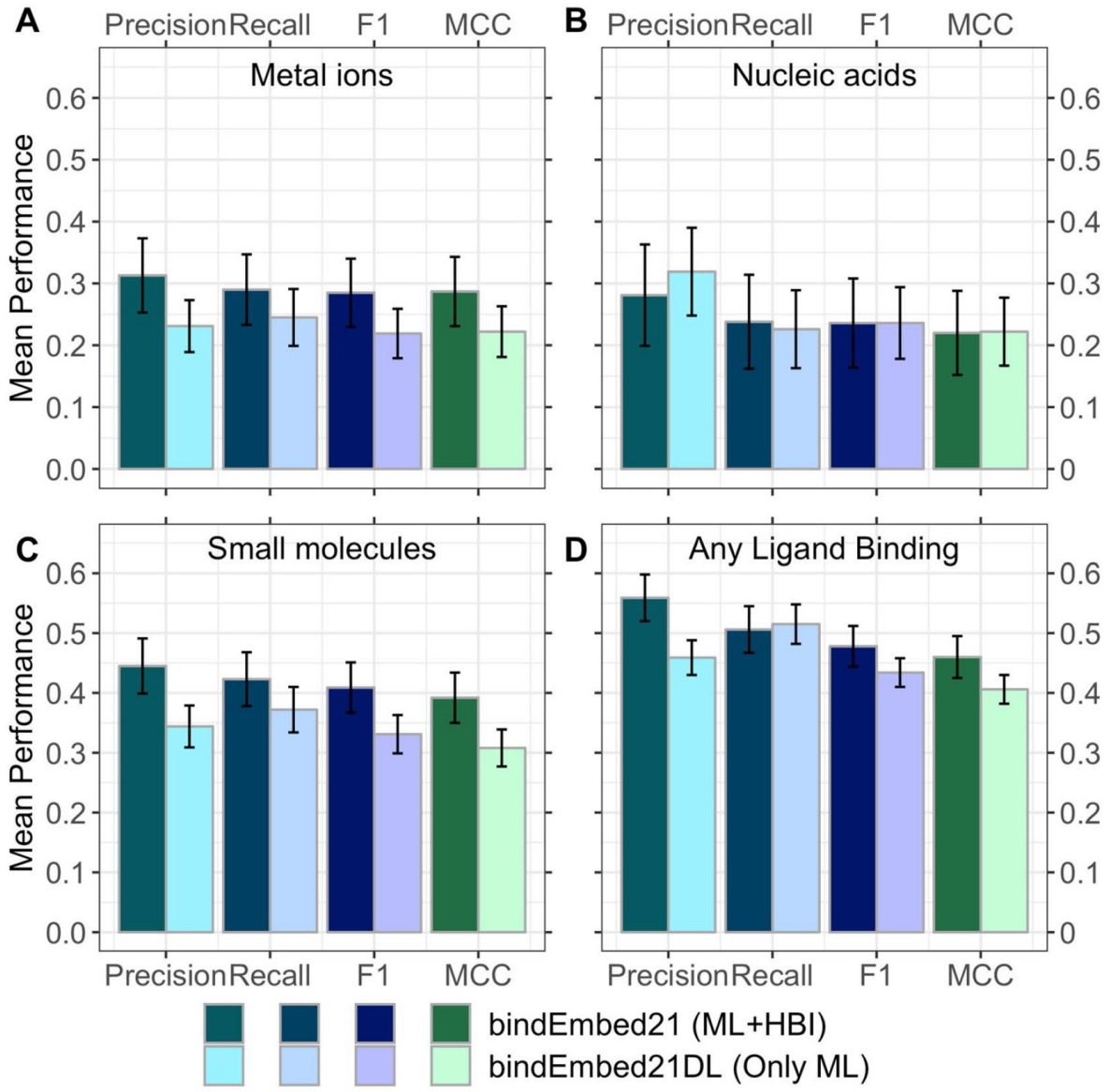





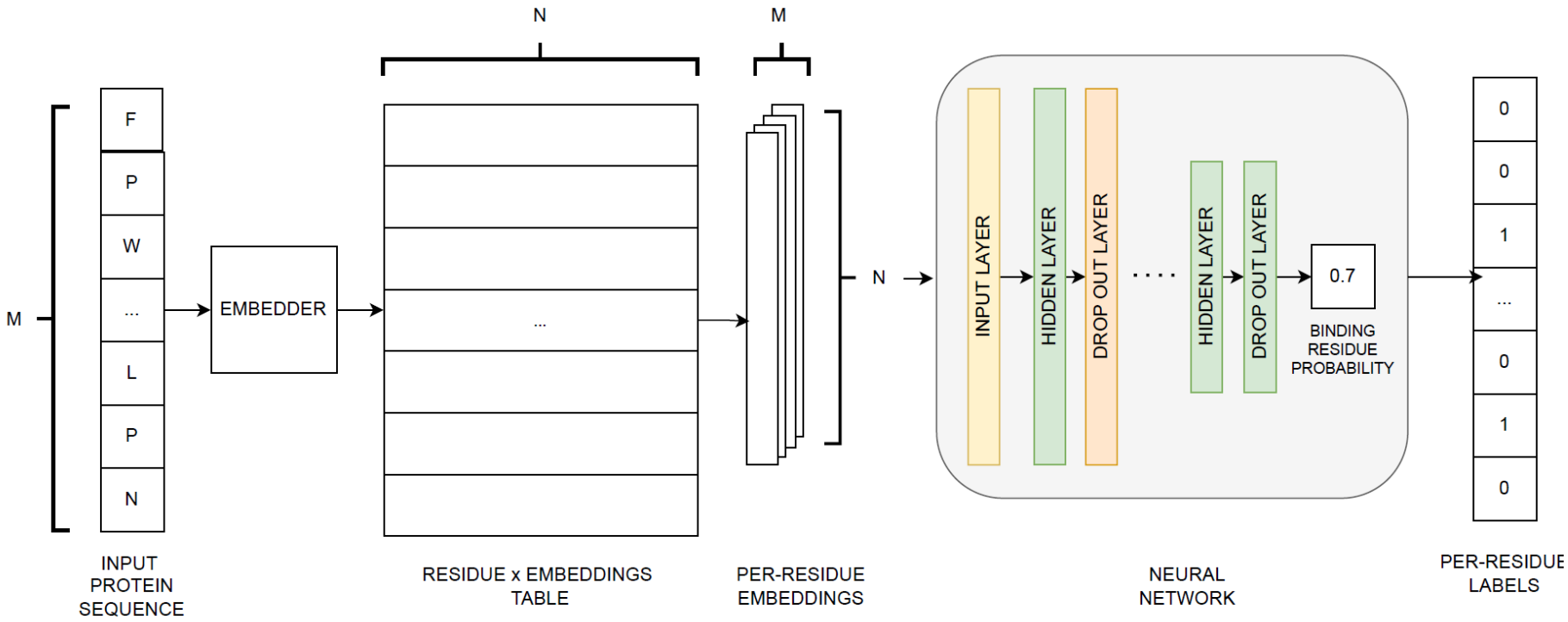

Littmann, Maria, et al. "Protein embeddings and deep learning predict binding residues for various ligand classes." Scientific Reports 11.1 (2021): 23916.

# bindEmbed21

- binEmebed21DL (ML) + **homology-based inference (HBI)**
  - predict binding residues through HBI if an experimentally annotated sequence-similar protein is available, use ML otherwise



# pLM + MLP




| Parameter | Value |
|---|---|
| No. of hidden layers | 1-4 |
| No. neurons in 1. layer | 128, 256, 512 |
| No. neurons in 2.- 4. layer | 32-512 (step 32) |
| Drop out value | 0.3 - 0.7 (step 0.1) |
| Learning rate | 1e-2, 1e-3, 1e-4 |



Hoksza, David, and Hamza Gamouh. "Exploration of protein sequence embeddings for protein-ligand binding site detection." 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2022.

# Comparison with traditional ML sequence-based methods - dataset

**Table 1. Detailed Compositions of 12 Different Data Sets**

| | | Training Set | | Independent Test Set | | |
|---|---|---|---|---|---|---|
| Ligand Category | Ligand Type | No. of Sequences | numP, numN[a] | No. of Sequences | numP, numN[a] | Total No. of Sequences |
| Nucleotides | ATP | 221 | 3021, 72334 | 50 | 647, 16639 | 271 |
| | ADP | 296 | 3833, 98740 | 47 | 686, 20327 | 343 |
| | AMP | 145 | 1603, 44401 | 33 | 392, 10355 | 178 |
| | GDP | 82 | 1101, 26244 | 14 | 194, 4180 | 96 |
| | GTP | 54 | 745, 21205 | 7 | 89,1868 | 61 |
| Metal Ions | $Ca^{2+}$ | 965 | 4914, 287801 | 165 | 785, 53779 | 1130 |
| | $Zn^{2+}$ | 1168 | 4705, 315235 | 176 | 744, 47851 | 1344 |
| | $Mg^{2+}$ | 1138 | 3860, 350716 | 217 | 852, 72002 | 1355 |
| | $Mn^{2+}$ | 335 | 1496, 112312 | 58 | 237, 17484 | 393 |
| | $Fe^{3+}$ | 173 | 818, 50453 | 26 | 120, 9092 | 199 |
| DNA | | 335 | 6461, 71320 | 52 | 973, 16225 | 387 |
| HEME | | 206 | 4380, 49768 | 27 | 580, 8630 | 233 |

[a]numP and numN represent the numbers of positive (binding residues) samples and negative (nonbinding residues) samples, respectively.

# Comparison with traditional ML sequence-based methods

| | Method | ACC | AUC | MCC |
|---|---|---|---|---|
| ATP | Embeddings-T5 | 95.9 | 0.844 | **0.561** |
| | TargetS | 96.5 | 0.898 | 0.502 |
| | EC-RUS | 96.8 | 0.871 | 0.506 |
| | SXGBsite | 96.4 | 0.886 | 0.448 |
| ADP | Embeddings-T5 | 96.0 | 0.819 | 0.515 |
| | TargetS | 97.2 | 0.896 | 0.507 |
| | EC-RUS | 97.6 | 0.872 | 0.511 |
| | SXGBsite | 97.0 | 0.907 | **0.521** |
| AMP | Embeddings-T5 | 91.3 | 0.803 | 0.377 |
| | TargetS | 95.9 | 0.83 | 0.359 |
| | EC-RUS | 97.0 | 0.815 | **0.393** |
| | SXGBsite | 96.4 | 0.851 | 0.366 |
| GDP | Embeddings-T5 | 96.3 | 0.863 | 0.633 |
| | TargetS | 96.2 | 0.896 | 0.55 |
| | EC-RUS | 97.1 | 0.872 | 0.579 |
| | SXGBsite | 97.9 | 0.93 | **0.678** |
| GTP | Embeddings-T5 | 93.6 | 0.811 | 0.477 |
| | TargetS | 96.9 | 0.855 | 0.617 |
| | EC-RUS | 97.0 | 0.861 | **0.641** |
| | SXGBsite | 97.8 | 0.883 | 0.572 |
| CA | Embeddings-T5 | 97.1 | 0.699 | 0.293 |
| | TargetS | 98.8 | 0.767 | **0.243** |
| | EC-RUS | 98.7 | 0.77 | 0.225 |
| | SXGBsite | 98.1 | 0.757 | 0.167 |
| MG | Embeddings-T5 | 97.3 | 0.673 | 0.246 |
| | TargetS | 98.8 | 0.706 | 0.294 |
| | EC-RUS | 99.1 | 0.78 | 0.317 |
| | SXGBsite | 99.0 | 0.819 | **0.326** |
| MN | Embeddings-T5 | 97.8 | 0.854 | **0.494** |
| | TargetS | 98.7 | 0.888 | 0.449 |
| | EC-RUS | 97.3 | 0.891 | 0.31 |
| | SXGBsite | 98.5 | 0.888 | 0.329 |
| FE | Embeddings-T5 | 98.1 | 0.945 | **0.593** |
| | TargetS | 98.7 | 0.945 | 0.479 |
| | EC-RUS | 99.0 | 0.936 | 0.49 |
| | SXGBsite | 98.5 | 0.913 | 0.454 |
| ZN | Embeddings-T5 | 96.7 | 0.89 | 0.473 |
| | TargetS | 98.7 | 0.936 | **0.527** |
| | EC-RUS | 98.6 | 0.958 | 0.437 |
| | SXGBsite | 98.5 | 0.892 | 0.363 |
| DNA | Embeddings-T5 | 92.6 | 0.814 | **0.494** |
| | TargetS | 93.3 | 0.836 | 0.377 |
| | EC-RUS | 95.2 | 0.814 | 0.319 |
| | SXGBsite | 87.2 | 0.827 | 0.27 |
| HEME | Embeddings-T5 | 94.9 | 0.909 | **0.672** |
| | TargetS | 95.9 | 0.907 | 0.598 |
| | EC-RUS | 96.4 | 0.935 | 0.64 |
| | SXGBsite | 95.4 | 0.9 | 0.555 |

# Effect of model complexity

| | Embedding | ACC | AUC | MCC |
|---|---|---|---|---|
| ATP | BB | 87.8 (±1.9) | 0.82 (±0.009) | 0.361 (±0.018) |
| | ProtBert | 90.6 (±1.6) | 0.784 (±0.015) | 0.365 (±0.014) |
| | ProtT5 | 95.3 (±0.2) | 0.843 (±0.011) | 0.548 (±0.009) |
| ADP | BB | 79.1 (±2.9) | 0.765 (±0.007) | 0.242 (±0.016) |
| | ProtBert | 88.0 (±1.9) | 0.8 (±0.009) | 0.336 (±0.02) |
| | ProtT5 | 95.7 (±0.3) | 0.883 (±0.008) | 0.587 (±0.01) |
| AMP | BB | 85.8 (±3.5) | 0.814 (±0.018) | 0.319 (±0.023) |
| | ProtBert | 88.0 (±0.8) | 0.76 (±0.014) | 0.285 (±0.02) |
| | ProtT5 | 94.7 (±0.3) | 0.799 (±0.013) | 0.453 (±0.016) |
| GDP | BB | 91.7 (±1.7) | 0.877 (±0.004) | 0.479 (±0.032) |
| | ProtBert | 94.7 (±0.3) | 0.817 (±0.011) | 0.5 (±0.014) |
| | ProtT5 | 96.2 (±0.5) | 0.884 (±0.011) | 0.626 (±0.03) |
| GTP | BB | 79.7 (±6.6) | 0.795 (±0.014) | 0.272 (±0.035) |
| | ProtBert | 90.7 (±2.6) | 0.796 (±0.022) | 0.366 (±0.024) |
| | ProtT5 | 95.4 (±0.5) | 0.847 (±0.02) | 0.523 (±0.023) |
| CA | BB | 72.8 (±7.5) | 0.712 (±0.005) | 0.126 (±0.016) |
| | ProtBert | 84.0 (±1.2) | 0.766 (±0.007) | 0.184 (±0.007) |
| | ProtT5 | 97.1 (±0.1) | 0.76 (±0.004) | 0.389 (±0.01) |
| MG | BB | 78.4 (±1.5) | 0.767 (±0.007) | 0.134 (±0.003) |
| | ProtBert | 95.5 (±0.3) | 0.748 (±0.006) | 0.248 (±0.008) |
| | ProtT5 | 97.8 (±0.1) | 0.755 (±0.007) | 0.357 (±0.006) |
| MN | BB | 86.3 (±1.5) | 0.843 (±0.01) | 0.223 (±0.009) |
| | ProtBert | 96.3 (±0.4) | 0.86 (±0.013) | 0.403 (±0.02) |
| | ProtT5 | 97.7 (±0.4) | 0.863 (±0.017) | 0.493 (±0.03) |
| FE | BB | 92.4 (±1.3) | 0.875 (±0.01) | 0.337 (±0.018) |
| | ProtBert | 95.1 (±1.7) | 0.9 (±0.014) | 0.437 (±0.06) |
| | ProtT5 | 98.0 (±0.1) | 0.911 (±0.01) | 0.597 (±0.018) |
| ZN | BB | 89.5 (±1.5) | 0.892 (±0.005) | 0.297 (±0.016) |
| | ProtBert | 91.1 (±2.0) | 0.902 (±0.007) | 0.327 (±0.024) |
| | ProtT5 | 96.7 (±0.2) | 0.907 (±0.002) | 0.481 (±0.012) |
| DNA | BB | 72.0 (±3.5) | 0.73 (±0.008) | 0.275 (±0.002) |
| | ProtBert | 82.3 (±2.5) | 0.804 (±0.009) | 0.408 (±0.013) |
| | ProtT5 | 91.1 (±0.2) | 0.833 (±0.009) | 0.548 (±0.011) |
| HEME | BB | 86.0 (±1.4) | 0.848 (±0.006) | 0.482 (±0.013) |
| | ProtBert | 87.7 (±1.6) | 0.836 (±0.009) | 0.491 (±0.014) |
| | ProtT5 | 93.0 (±0.5) | 0.878 (±0.003) | 0.631 (±0.011) |

| **Embedding** | **BeplerBerger** | **ProtBert** | **ProtT5** |
|---|---|---|---|
| **Dataset** | SCOP | BFD | BFD, UniRef50 |
| **Model GPU size (GB)** | 1.4 | 2.8 | 5.9 |
| **Number of parameters** | 32M | 420M | 3B |
| **Embedding dimension** | 121 | 1024 | 1024 |

# Considerations

# Noisy data

- Unlabeled active sites
  - **Uncertain** status of **negatives**
  - Ligands transfer from homologous proteins
- Different properties of different **ligand types**
- Binding sites with **multiple** possible **binders**

# Data leakage

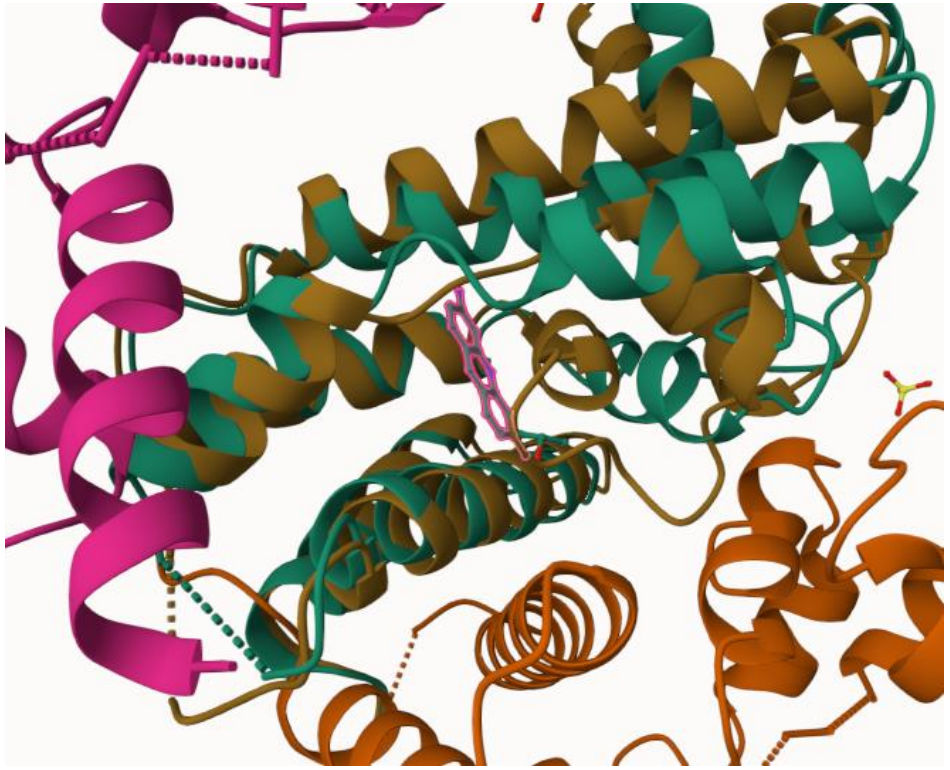
- Difficult to **separate** instances in **train and test** set

**Mitigation** Strategies

- Uniprot ID
- Sequence identity
- Sequence similarity
- Structure similarity
- Pocket similarity

# Conformational changes

- Allosteric effects
- Apo (ligand-free) vs holo (ligand-bound) versions



# Questions