Protein Binding Pocket Prediction

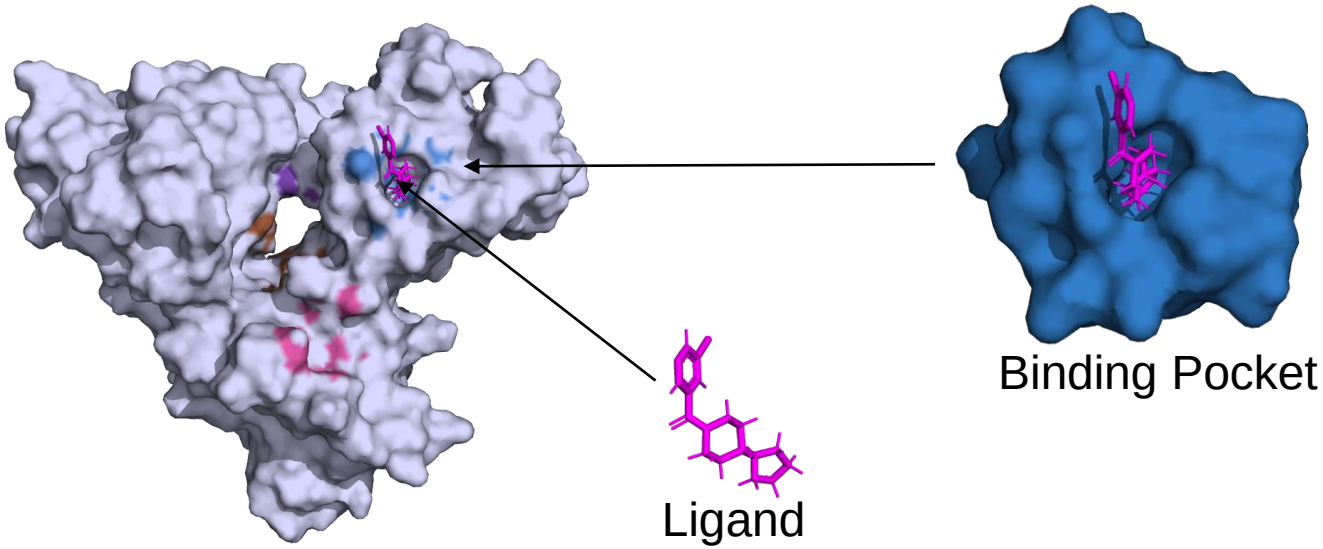
Using Equivariant Graph Neural Networks with Virtual Nodes

Lisa Schneckenreiter, 14-03-2023

Protein Binding Pockets …

- … are regions on a protein to which a **ligand** (e.g. small drug molecule) can bind.
- … usually lie in **cavities** on the protein **surface**.
- … often build **active sites**, i.e. binding triggers chemical modifications or conformational change.

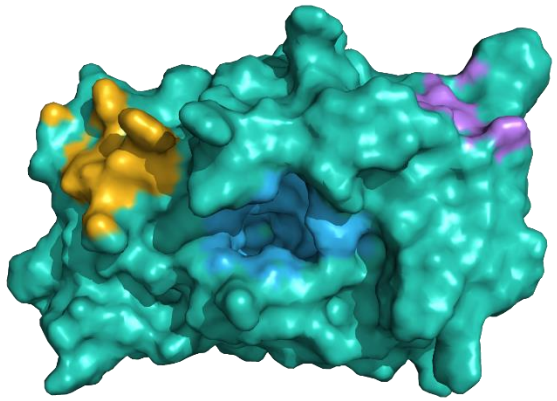
  → **Biological functions** of proteins can be modulated by ligands (drugs) binding to them.



Binding Pocket Prediction vs. Docking

Binding Pocket Prediction:

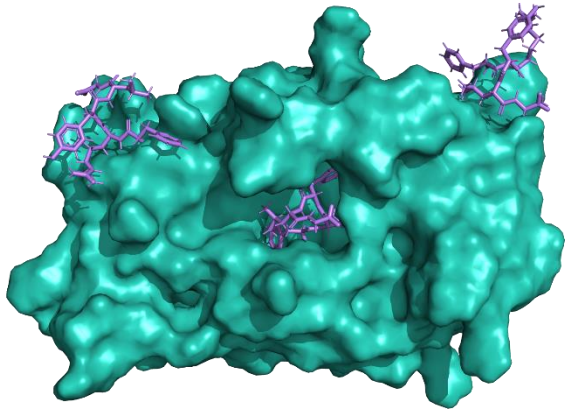
Where are regions on the protein to which a potential (unknown) ligand can bind?



only protein given, no ligand information

Docking:

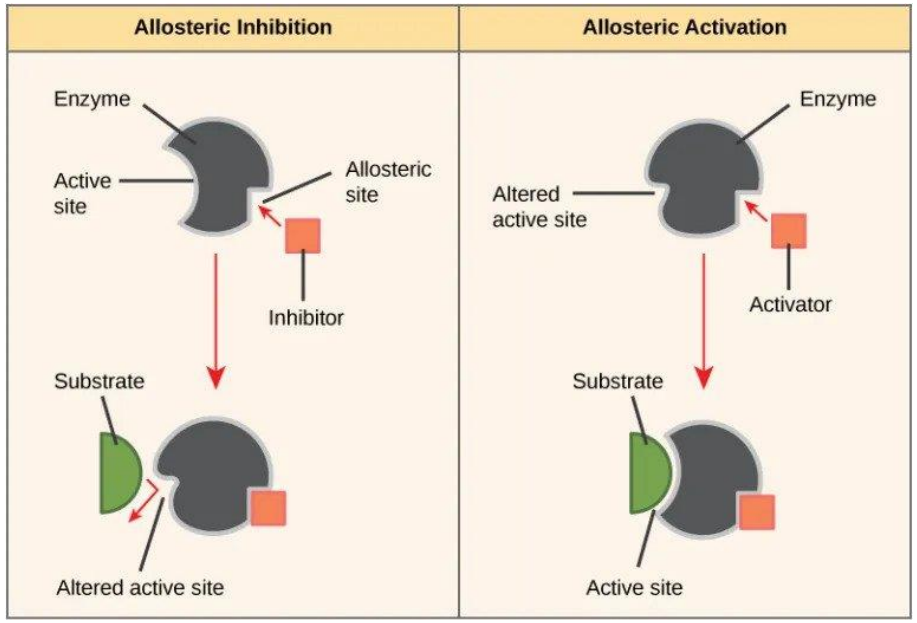
Where does a given ligand fit on a protein?



protein and ligand given as input

Why Binding Pocket Prediction?

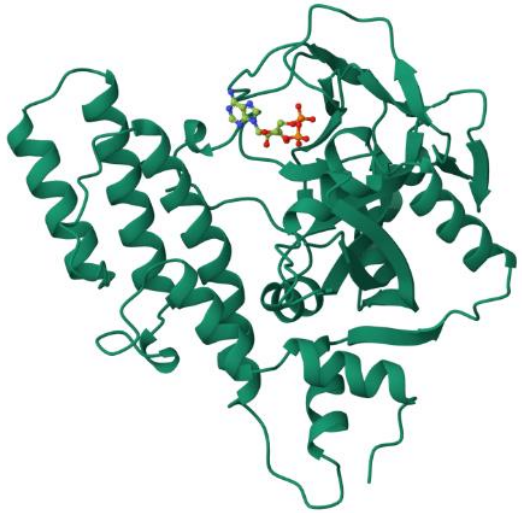
- provides valuable information for understanding **protein function**
- to identify a protein as a potential **drug target**
- to identify **allosteric binding sites**
- to gain information about potential drug ligands, guiding **rational drug design**
- as a **prerequisite** for many docking or generative models



Data for Pocket Prediction

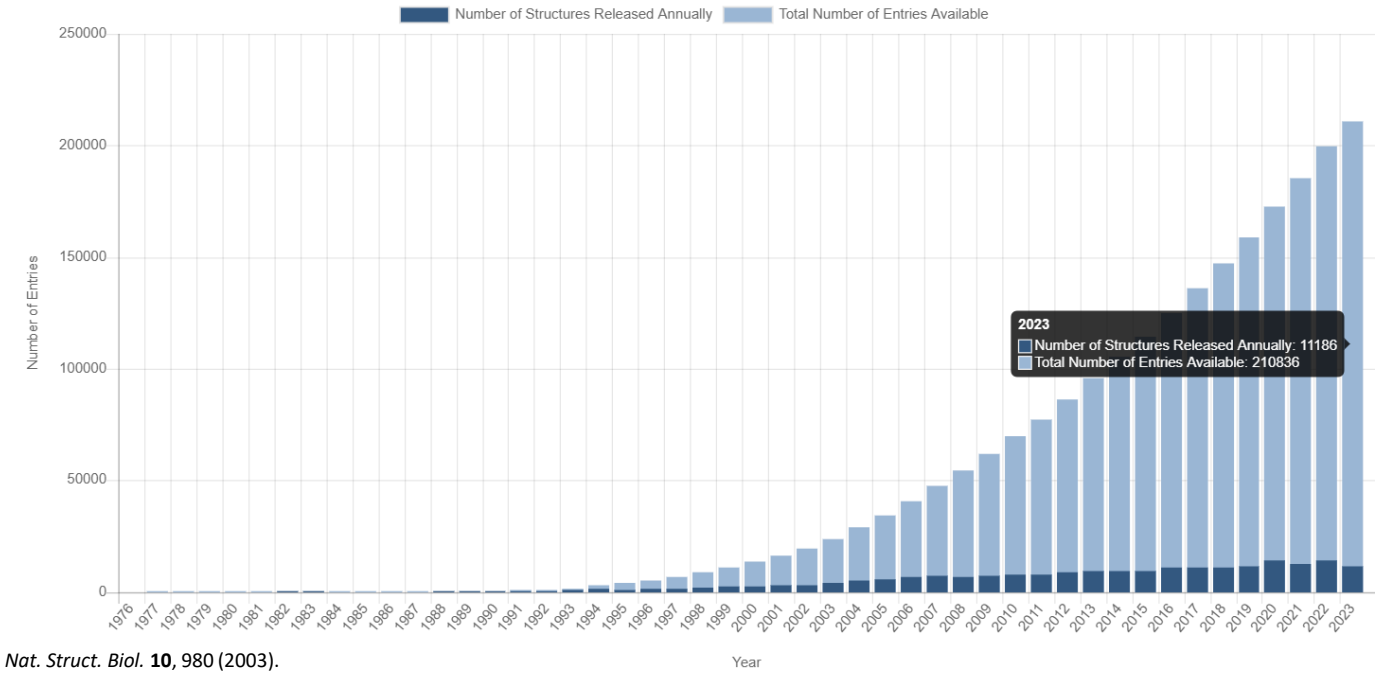
- experimentally measured (X-ray crystallography, NMR) 3D structures of protein-ligand complexes
- atom coordinates saved in PDB files

```
                                  Chain name
      Amino Acid                 /   Sequence Number
                  \             /   /
       Element     \           /   /      -----Coordinates-----
                    \   \     /   /       X       Y       Z
ATOM       1   N    ASP L    1          4.060   7.307   5.186
ATOM       2   CA   ASP L    1          4.042   7.776   6.553
ATOM       3   C    ASP L    1          2.668   8.426   6.644
ATOM       4   O    ASP L    1          1.987   8.438   5.606
ATOM       5   CB   ASP L    1          5.090   8.827   6.797
ATOM       6   CG   ASP L    1          6.338   8.761   5.929
ATOM       7   OD1  ASP L    1          6.576   9.758   5.241
ATOM       8   OD2  ASP L    1          7.065   7.759   5.948
                  \\
                    Element position within amino acid
```



The Worldwide Protein Data Bank (wwPDB)

- international collaboration between PDB in Europe, USA, Japan and UK
- data curated by one member → synchronized with all
- publicly available archive of macro-molecular structures solved by X-ray crystallography or NMR spectroscopy
- 210,836 structures in total



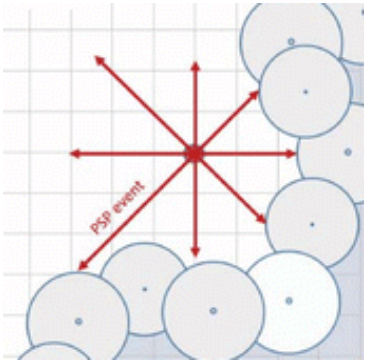
[1] Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).

Definition of a Binding Pocket

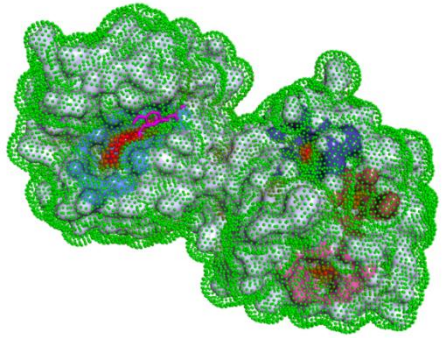
- **Residue-centric definition:**
  - segmentation of protein surface residues or atoms as binding or non-binding
  - typically protein atoms within 4Å of any ligand atom belong to binding pocket

- **Pocket-centric definition:**
  - defined by pocket center and/or as a set of points around the protein surface that characterize the shape of the pocket
  - e.g. spaced grid points (CNN-based methods), points on a solvent accessible surface (P2Rank) or virtual node position (VN-EGNN)

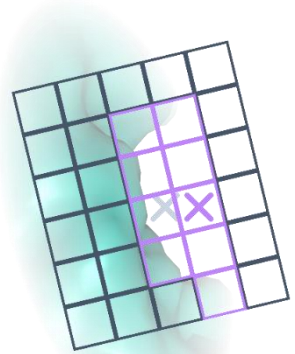
Types of Pocket Prediction Methods



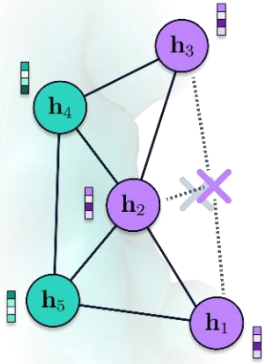
Geometry- and Energy-Based Methods



Random Forest Classification of Protein Surface



Convolutional Neural Networks

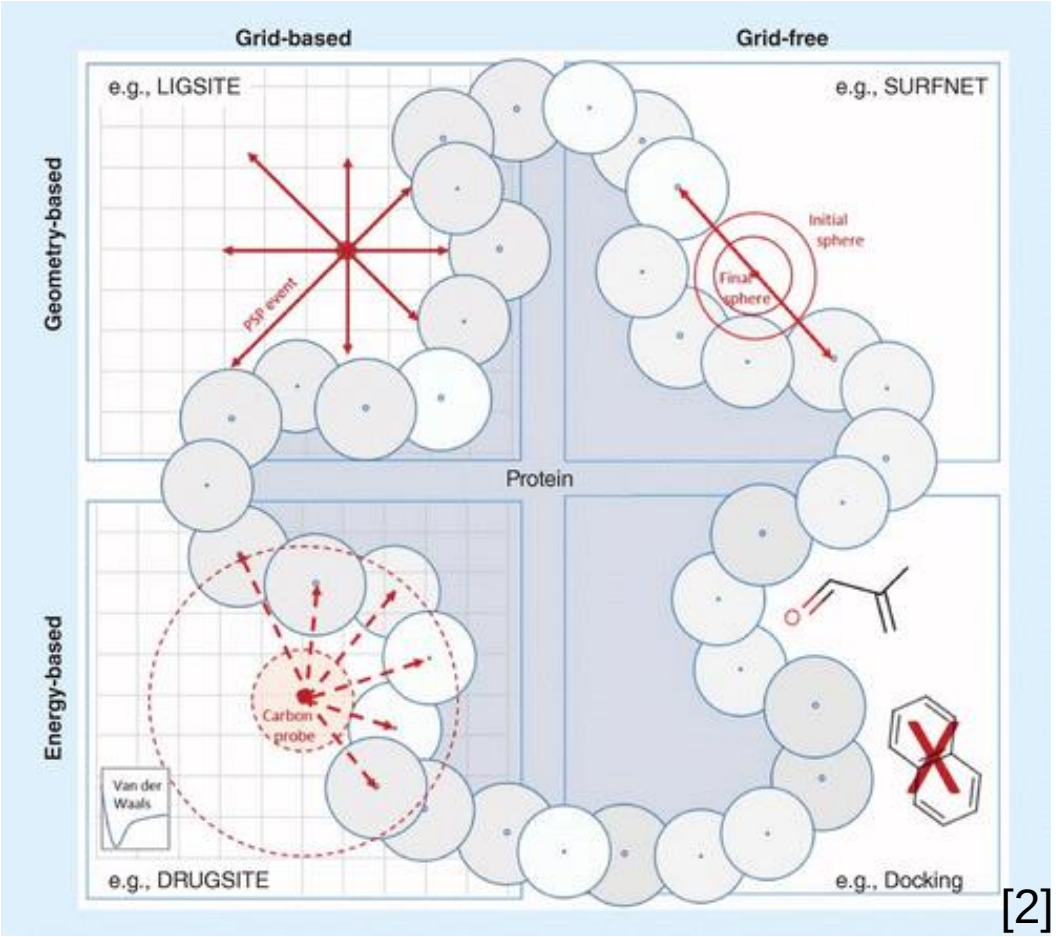


Graph Neural Networks

Early Binding Site Detection Approaches

- Geometry-based approaches: analyze the shape of a molecular surface
- Energy-based approaches: interactions of probes or molecular fragments with the protein

Both strategies can be performed on a Cartesian grid-based representation of the protein or grid-free.

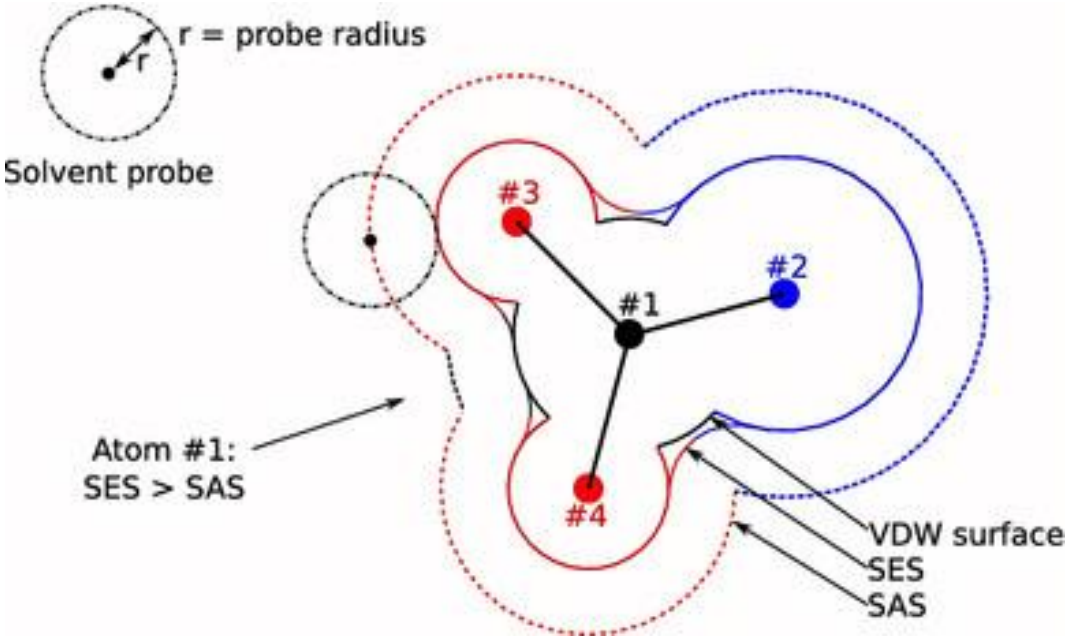


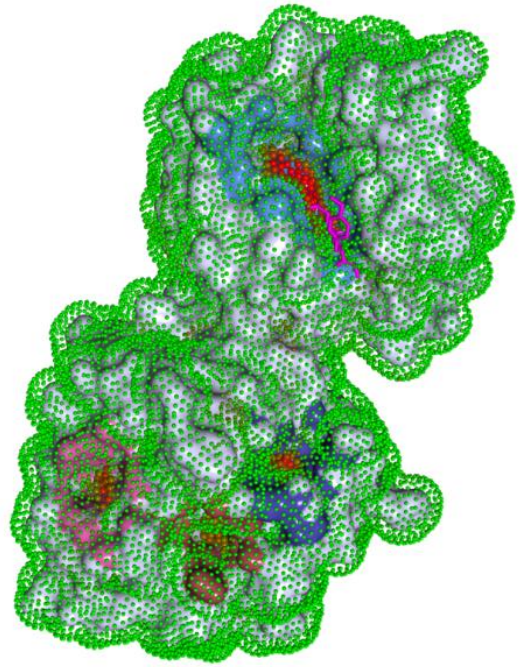
[2]

[2] Volkamer, A. & Rarey, M. Exploiting structural information for drug-target assessment. *Future Med. Chem.* **6**, 319–331 (2014).

P2Rank [3] – a Random Forest Based Approach

- Step 1: generate set of regularly spaced points on solvent accessible surface (SAS)

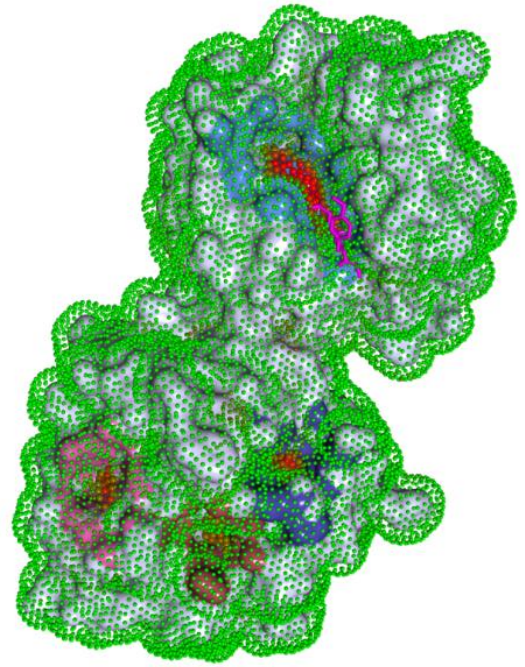




[3] Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminformatics* **10**, 39 (2018).

P2Rank [3] – a Random Forest Based Approach

- Step 1: generate set of regularly spaced points on solvent accessible surface (SAS)
- Step 2: define feature vectors of SAS points based on distance-weighed atomic features of closest atoms
- Step 3: **random forest classifier** for “ligandability”
- Step 4: clustering of ligandable SAS points
- Step 5: ranking by cumulative ligandability score



- used as part of some docking methods (e.g. TankBind [4])

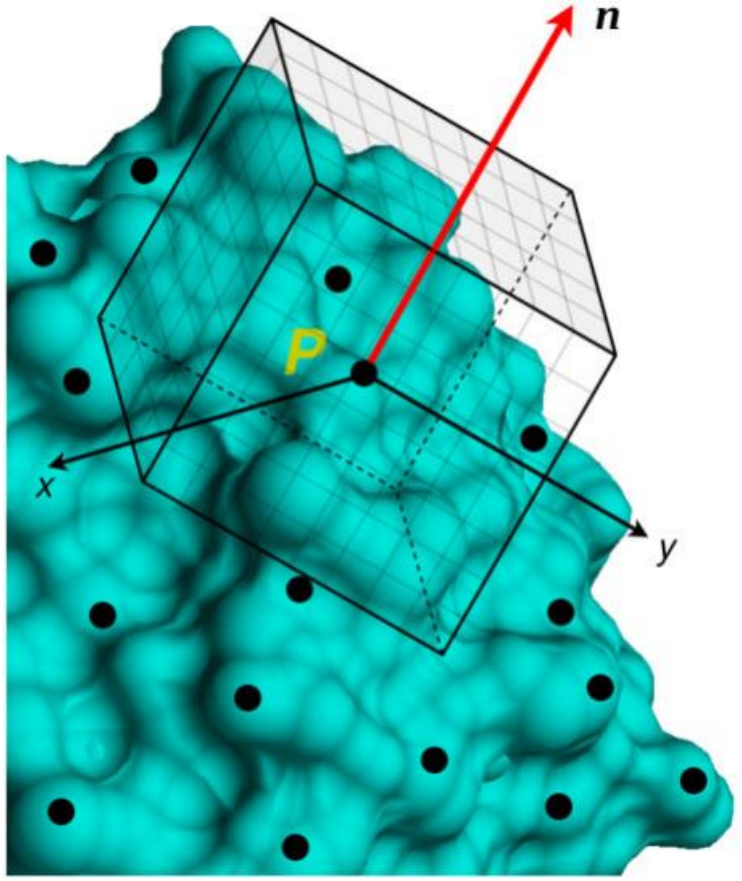
[3] Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminformatics* **10**, 39 (2018).

[4] Lu, W. *et al. TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction*. http://biorxiv.org/lookup/doi/10.1101/2022.06.06.495043 (2022) doi:10.1101/2022.06.06.495043.

DeepSurf [5] – a CNN-Based Approach

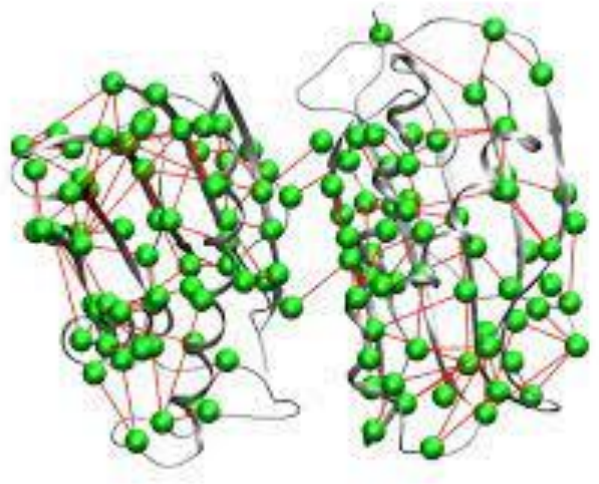
- Step 1: get SAS points
- Step 2: create local grid around normal vector for each point and assign physico-chemical features to voxels
- Step 3: apply **3D-CNN** to grid → ligandability score
- Step 4: discard points with low score
- Step 5: cluster remaining points and assign to closest atoms
- Step 6: rank according to average score

- other CNN-based methods: DeepSite, DeepPocket



[5] Mylonas, S. K., Axenopoulos, A. & Daras, P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinforma. Oxf. Engl.* **37**, 1681–1690 (2021).

Proteins as Graphs

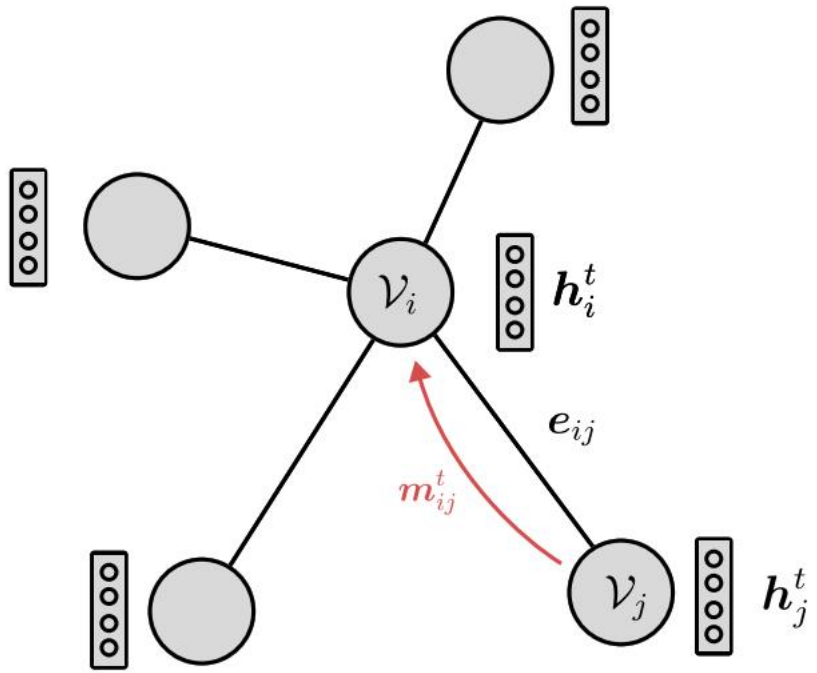


- Nodes:
  - (surface) atoms
  - (surface) amino acid residues
- Node features:
  - atomic features
    - hand-crafted features (e.g. atom type, amino acid type, distance to surface)
    - learned features (e.g. Evolutionary Scale Modeling (ESM) [9] embeddings)
  - coordinates → geometric graphs
- Edges:
  - chemical bonds
  - spatial edges (distance less than a cut-off)
  - nearest neighbours (fixed number)

} → geometric graphs

[9] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Message Passing Neural Networks (MPNN)



- Message and Aggregate

$$\boldsymbol{m}_{ij}^{t+1} = M(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t, \boldsymbol{e}_{i,j}; \boldsymbol{w})$$

$$\boldsymbol{m}_i^{t+1} = \sum_{j:a_{ij}=1} \boldsymbol{m}_{ij}$$

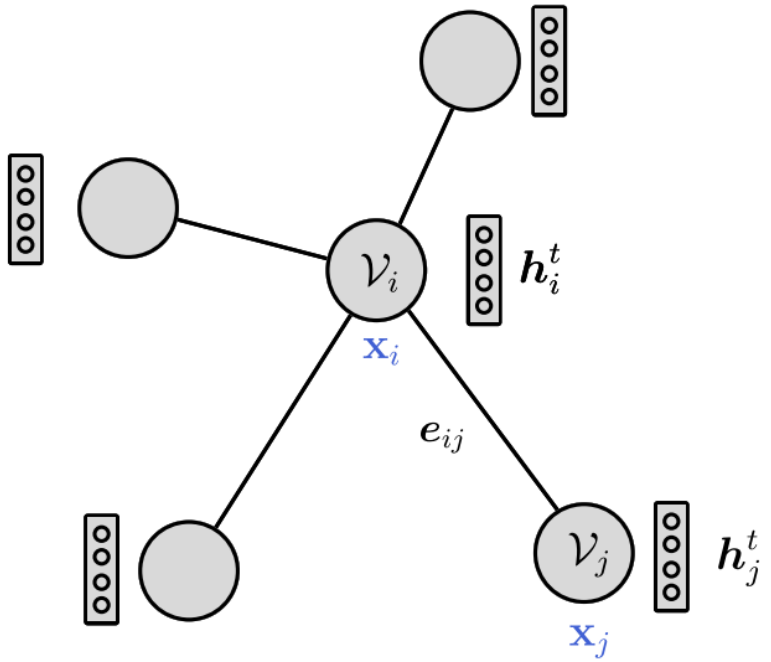
- Update

$$\boldsymbol{h}_i^{t+1} = U(\boldsymbol{h}_i^t, \boldsymbol{m}_i^{t+1}; \boldsymbol{v})$$

[6] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning* 1263–1272 (PMLR, 2017).

E(n) – Equivariant Graph Neural Networks (EGNN)

- equivariant w.r.t. rotations and translations



- Message and Aggregate

$$\boldsymbol{m}_{ij} = M\left(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t, \left\|\mathbf{x}_i^l - \mathbf{x}_j^l\right\|^2, \boldsymbol{e}_{i,j}; \boldsymbol{w}\right)$$

$$\boldsymbol{m}_i^{t+1} = \sum_{j:a_{ij}=1} \boldsymbol{m}_{ij}$$

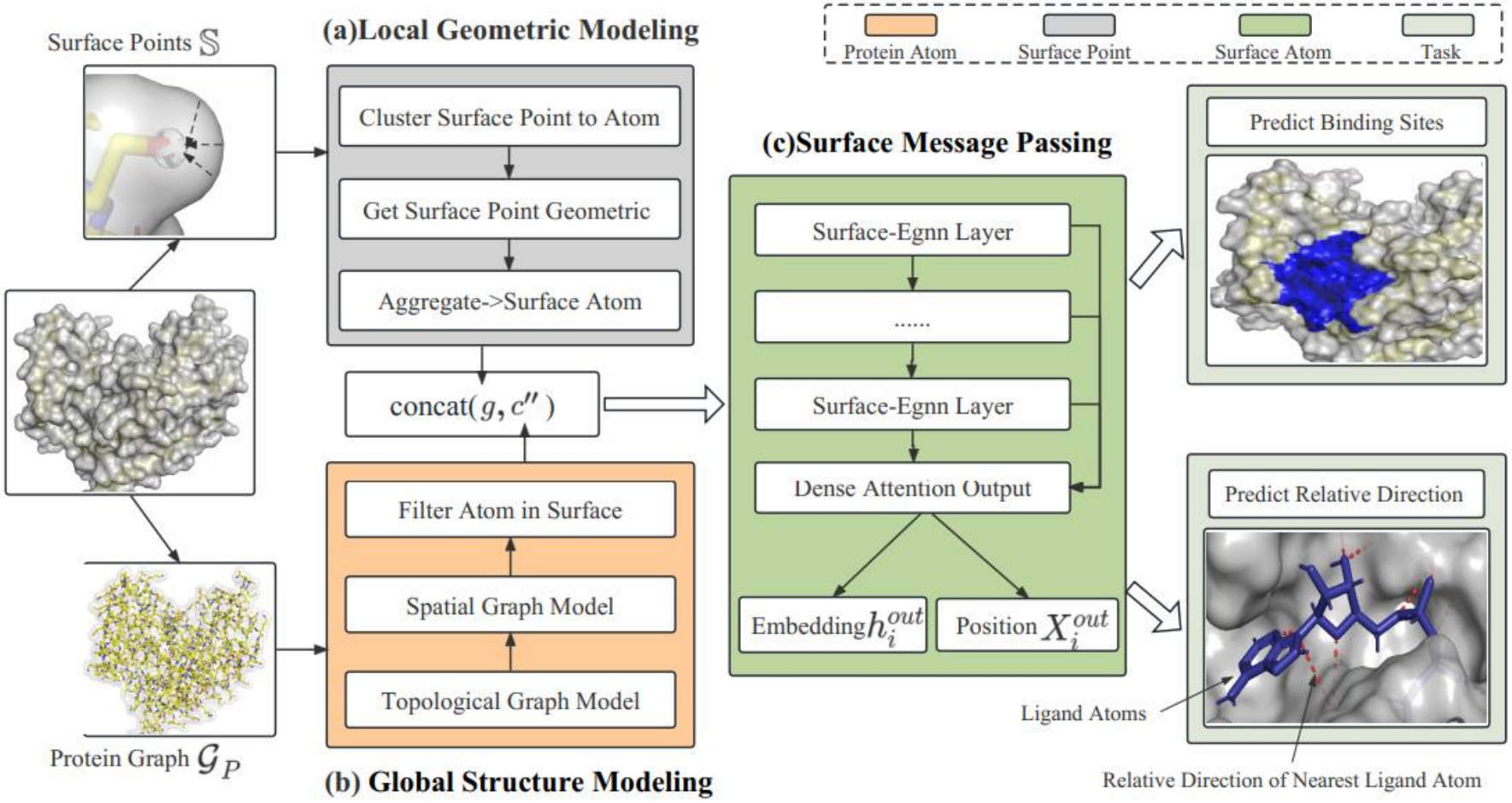
- Update (ϕ_x denotes a NN too)

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + C \sum_{j:a_{ij}=1} (\mathbf{x}_i^l - \mathbf{x}_j^l)\phi_x(\boldsymbol{m}_{ij})$$

$$\boldsymbol{h}_i^{l+1} = U(\boldsymbol{h}_i^l, \boldsymbol{m}_i; \boldsymbol{v})$$

[7] Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. Preprint at http://arxiv.org/abs/2102.09844 (2022).

EquiPocket – an EGNN-Based Approach



[8] Zhang, Y., Huang, W., Wei, Z., Yuan, Y. & Ding, Z. EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction. Preprint at http://arxiv.org/abs/2302.12177 (2023).

EquiPocket – an EGNN-Based Approach

Objectives:

1. Segmentation:

- Prediction: $\hat{y}_i = \mathbf{Sigmoid}(\mathrm{MLP}(\boldsymbol{h}_i^{\mathrm{out}}))$.
- Dice loss: $\mathcal{L}_b = 1 - \frac{2 \cdot \sum(\hat{y}_i \cdot y_i)}{\sum(\hat{y}_i) + \sum(y_i) + \epsilon}$,

2. Relative direction of nearest ligand atom:

- Prediction: $\hat{\boldsymbol{d}}_i = \frac{\boldsymbol{x}_i^{\mathrm{out}} - \boldsymbol{x}_i}{\|\boldsymbol{x}_i^{\mathrm{out}} - \boldsymbol{x}_i\|_2}$.
- Direction loss: $\mathcal{L}_d = \sum(1 - \cos(\hat{\boldsymbol{d}}_i, \boldsymbol{d}_i))$.

$\widehat{y_i}$ … prediction for node i (in [0,1])
y_i … label for node i (in {0,1})

h_i^{out} … output features of atom i
$\mathrm{x_i}$ … initial position of atom i
$\mathrm{x_i}^{\mathrm{out}}$ … output position of atom i

d_i … direction of nearest ligand atom

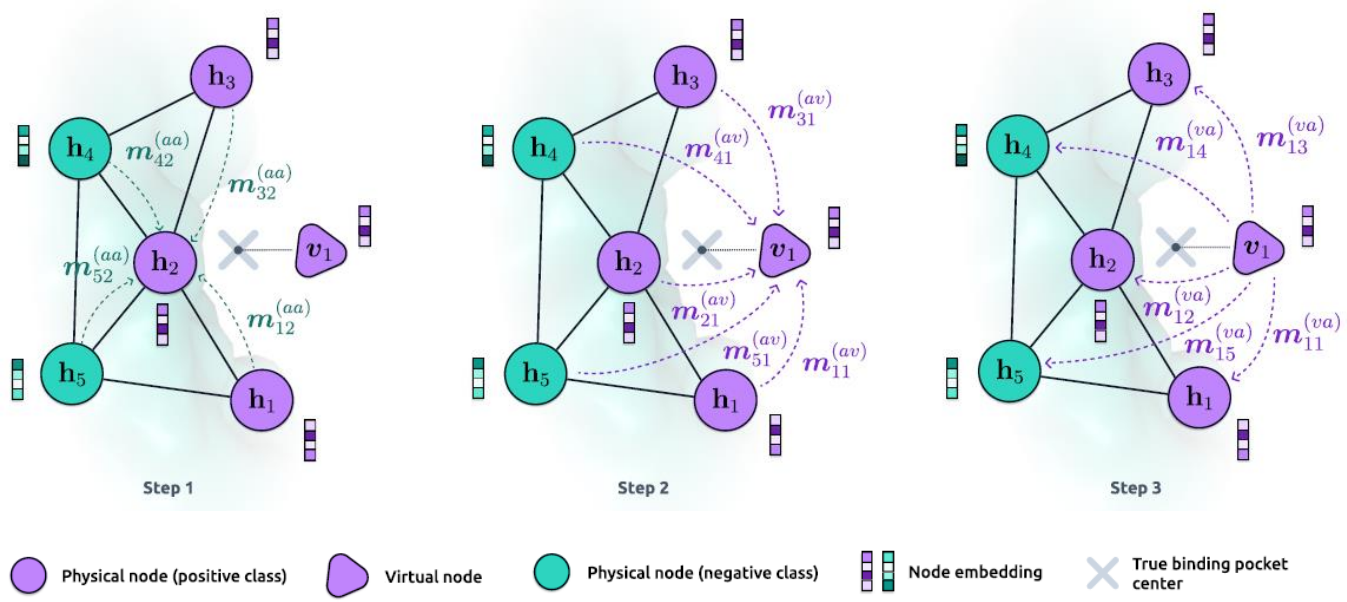
[8] Zhang, Y., Huang, W., Wei, Z., Yuan, Y. & Ding, Z. EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction. Preprint at http://arxiv.org/abs/2302.12177 (2023).

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

Florian Sestak, Lisa Schneckenreiter, Johannes Brandstetter, Sepp Hochreiter, Andreas Mayr, Günter Klambauer

VN-EGNN Overview

- **E(3)-equivariant graph neural network** on protein **residue** graph
- additional **virtual nodes (VN)** connected to all physical nodes
- **3** message passing phases



Protein Representation

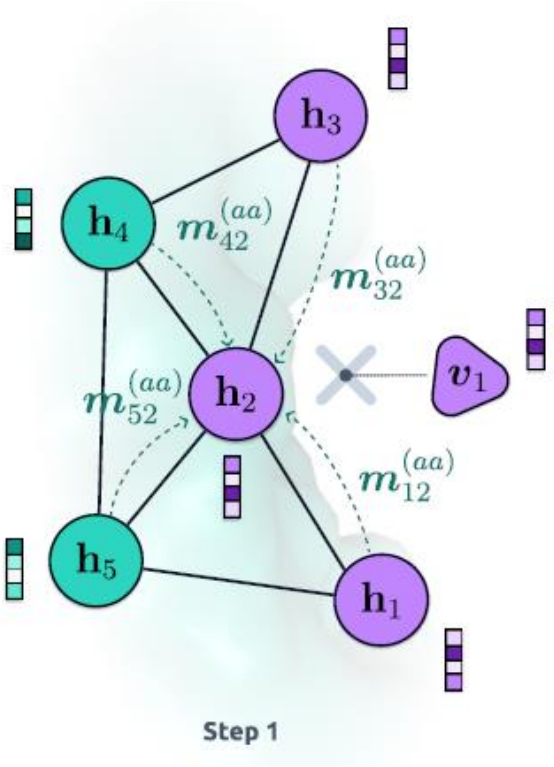
- Protein residue graph
  - Nodes = amino acid residues:
    - coordinates: $\mathbf{x}_n \in \mathbb{R}^3$
    - ESM (Evolutionary Scale Modeling [9]) features: $\boldsymbol{h}_n \in \mathbb{R}^D$
  - Edges:
    - incoming from 10 nearest neighbors (if closer than 30Å)
  - Ground truth labels:
    - binary classification (pocket residue or not): $y_n \in \{0,1\}$
    - binding site centers: $\mathbf{y}_m \in \mathbb{R}^3$

[9] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Virtual Nodes

- Coordinates:
  - $\mathbf{z}_k \in \mathbb{R}^3$
  - evenly distributed on a sphere around the protein (Fibonacci grid)
- Feature vectors:
  - $\boldsymbol{v}_k \in \mathbb{R}^D$
  - initialized with a average feature vector of all residues
- Edges:
  - Connected to all physical (residue) nodes
- Number:
  - variable but 8 per default

Message Passing Phase I



Message from residue i to residue j:

$$\boldsymbol{m}_{ij}^{(aa)} = \phi_{e^{(aa)}}(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, \|\mathbf{x}_i^l - \mathbf{x}_j^l\|, a_{ij})$$

Aggregation of messages:

$$\boldsymbol{m}_j^{(aa)} = \frac{1}{|\mathcal{N}(j)|} \sum_{i \in \mathcal{N}(j)} \boldsymbol{m}_{ij}^{(aa)}$$

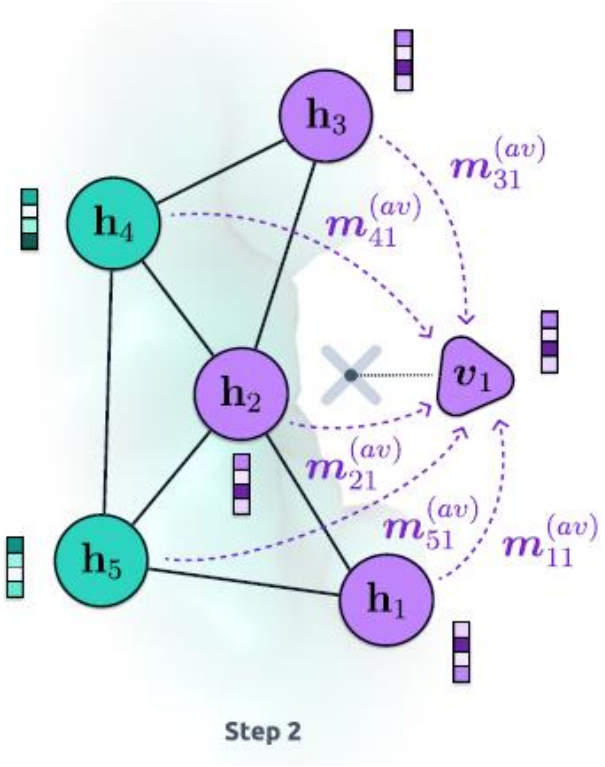
Feature update:

$$\boldsymbol{h}_j^{l+1/2} = \phi_{h^{(aa)}}\left(\boldsymbol{h}_j^l, \boldsymbol{m}_j^{(aa)}\right)$$

Coordinate update:

$$\mathbf{x}_j^{l+1/2} = \mathbf{x}_j^l + \frac{1}{|\mathcal{N}(j)|} \sum_{i \in \mathcal{N}(j)} \frac{\mathbf{x}_i^l - \mathbf{x}_j^l}{\|\mathbf{x}_i^l - \mathbf{x}_j^l\|} \phi_{x^{aa}}(\boldsymbol{m}_{ij}^{(aa)})$$

Message Passing Phase II



Message from atom i to virtual node j:

$$\boldsymbol{m}_{ij}^{(av)} = \phi_{e^{(av)}}(\boldsymbol{h}_i^{l+1/2}, \boldsymbol{v}_j^l, \|\mathbf{x}_i^{l+1/2} - \mathbf{z}_j^l\|, d_{ij})$$

Aggregation of messages:

$$\boldsymbol{m}_j^{(av)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}_{ij}^{(av)}$$

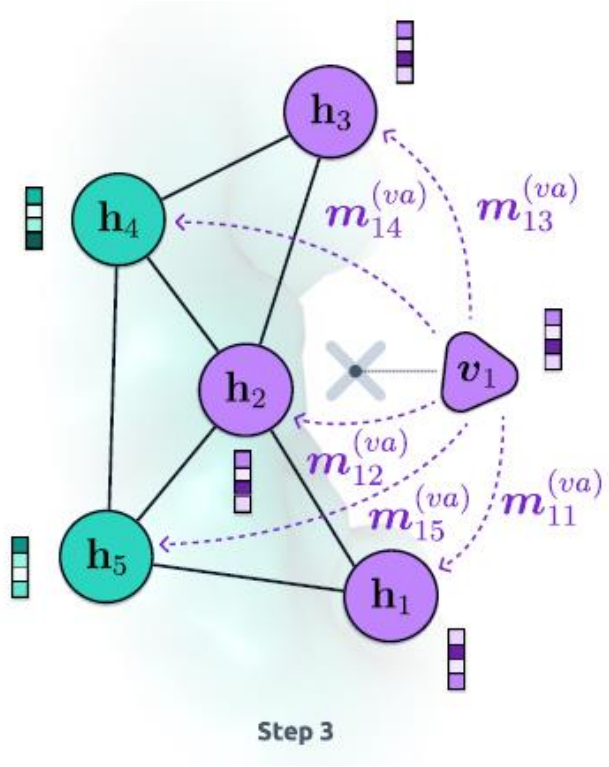
Feature update:

$$\boldsymbol{v}_j^{l+1} = \phi_{h^{(av)}}\left(\boldsymbol{v}_j^l, \boldsymbol{m}_j^{(av)}\right)$$

Coordinate update:

$$\mathbf{z}_j^{l+1} = \mathbf{z}_j^l + \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{x}_i^{l+1/2} - \mathbf{z}_j^l}{\|\mathbf{x}_i^{l+1/2} - \mathbf{z}_j^l\|}\phi_{x^{av}}(\boldsymbol{m}_{ij}^{(av)})$$

Message Passing Phase III



Message from virtual node i to atom j:

$$\boldsymbol{m}_{ij}^{(va)} = \phi_{e^{(va)}}(\boldsymbol{v}_i^{l+1}, \boldsymbol{h}_j^{l+1/2}, \|\mathbf{z}_i^{l+1} - \mathbf{x}_j^{l+1/2}\|, d_{ij})$$

Aggregation of messages:

$$\boldsymbol{m}_j^{(va)} = \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{m}_{ij}^{(va)}$$

Feature update:

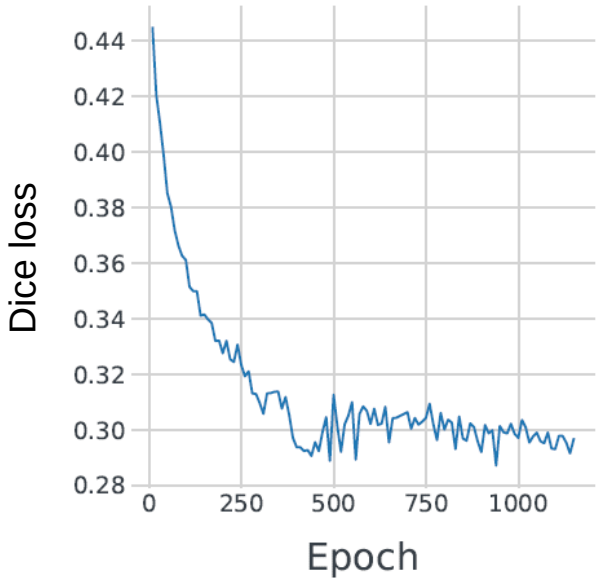
$$\boldsymbol{h}_j^{l+1} = \phi_{h^{(va)}} \left(\boldsymbol{h}_j^{l+1/2}, \boldsymbol{m}_j^{(va)} \right)$$

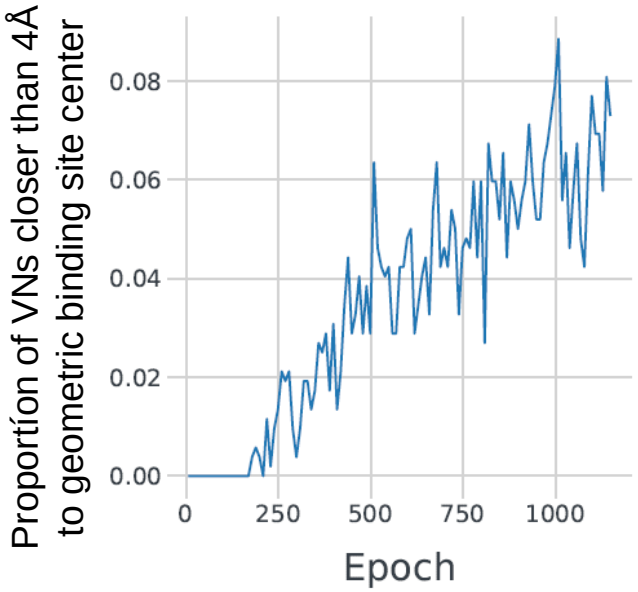
Coordinate update:

$$\mathbf{x}_j^{l+1} = \mathbf{x}_j^{l+1/2} + \frac{1}{K} \sum_{i=1}^{K} \frac{\mathbf{z}_i^{l+1} - \mathbf{x}_j^{l+1/2}}{\|\mathbf{z}_i^{l+1} - \mathbf{x}_j^{l+1/2}\|} \phi_{x^{va}}(\boldsymbol{m}_{ij}^{(va)})$$

Initial Objective: Classification of Residues

- Read-out function: $\hat{y}_n = \sigma(w^\top h_n^L)$
- Loss function (Dice loss): $\mathcal{L}_{\text{segm}} = \text{Dice}\left((y_1, \ldots, y_N), (\hat{y}_1, \ldots, \hat{y}_N)\right) \coloneqq 1 - \frac{2\sum_{n=1}^{N} y_n \hat{y}_n + \epsilon}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} \hat{y}_n + \epsilon}$





$\hat{y}_n$ … prediction for node n (in [0,1])
y_n … label for node n (in {0,1})

h_n^L … output features of atom n
w … classifier parameters

Virtual nodes move towards real binding site centers!

Final Objective

Segmentation loss:

$$\mathcal{L}_{\text{segm}} = \text{Dice}\left((y_1, \ldots, y_N), (\hat{y}_1, \ldots, \hat{y}_N)\right) := 1 - \frac{2\sum_{n=1}^{N} y_n\,\hat{y}_n + \epsilon}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} \hat{y}_n + \epsilon}$$

$\hat{y}_n$… prediction for node n (in [0,1])
y_n… label for node n (in {0,1})

$\hat{\mathbf{y}}_k$ … predicted binding site center/output coordinates of VN k
$\mathbf{y}_m$ … true binding site center

Binding site center loss:

$$\mathcal{L}_{\text{bsc}} = \text{Dist}\left(\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}, \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_K\}\right) := \frac{1}{M}\sum_{m=1}^{M} \min_{k \in 1,\ldots,K} \|\mathbf{y}_m - \hat{\mathbf{y}}_k\|^2$$

→ Combined loss function:

$$\mathcal{L} = \text{Dist}\left(\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}, \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_K\}\right) + \alpha\text{Dice}\left((y_1, \ldots, y_N), (\hat{y}_1, \ldots, \hat{y}_N)\right)$$

Model Details

- 5 layers of VN-EGNN
- feature and message dimension: 100
- outputs:
  - segmentation of residues
  - position of virtual nodes (= binding pocket center)
  - binding pocket representations (output feature vectors of virtual nodes)

Metrics

- **DCC:** minimal distance between ground-truth geometric binding site center and a virtual node/predicted binding site center
- **DCA:** minimal distance between a ligand atom in the binding pocket and a virtual node/predicted binding site center

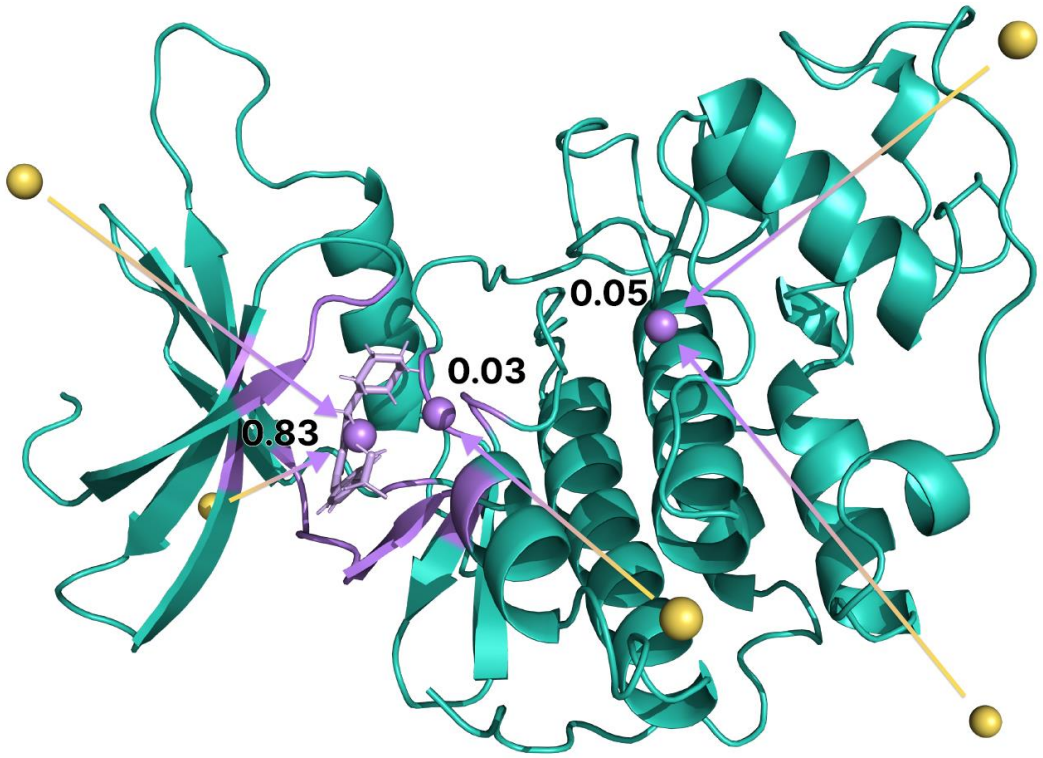
- **DCC/DCA success rate:** proportion of pockets with DCC/DCA$\leq 4\text{Å}$

- **Problem:**
  - The more virtual nodes the better this metric gets
    → clustering of closely located virtual nodes
    → ranking of virtual nodes and only evaluating top M positions (M = number of pockets)

Self-Confidence Module

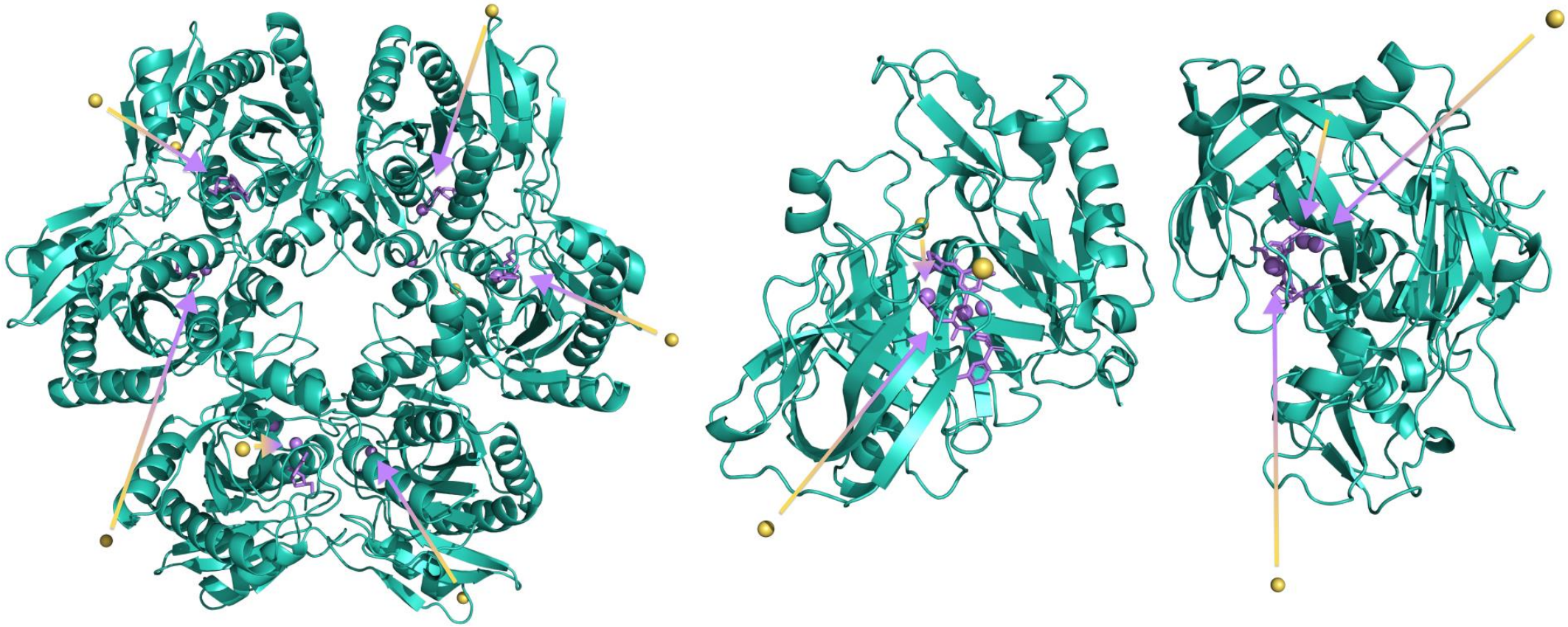
- predicted confidence of position of VN k: $\hat{c}_k = \psi(\boldsymbol{v}_k)$ for a MLP ψ
- confidence labels: $c_k = \begin{cases} 1 - \frac{1}{2\gamma} \cdot \|\mathbf{y}_k - \hat{\mathbf{y}}_k\| & \text{if } \|\mathbf{y}_k - \hat{\mathbf{y}}_k\| \leq \gamma, \\ c_0 & \text{otherwise} \end{cases}$ with $c_0 = 0.001$ and $\gamma = 4$
- confidence loss function: $\mathcal{L}_{\text{confidence}} = \frac{1}{K} \sum_{k=1}^{K} (c_k - \hat{c}_k)^2$

Visualizations



yellow: initial VN positions
purple: final VN positions

Visualizations



Results

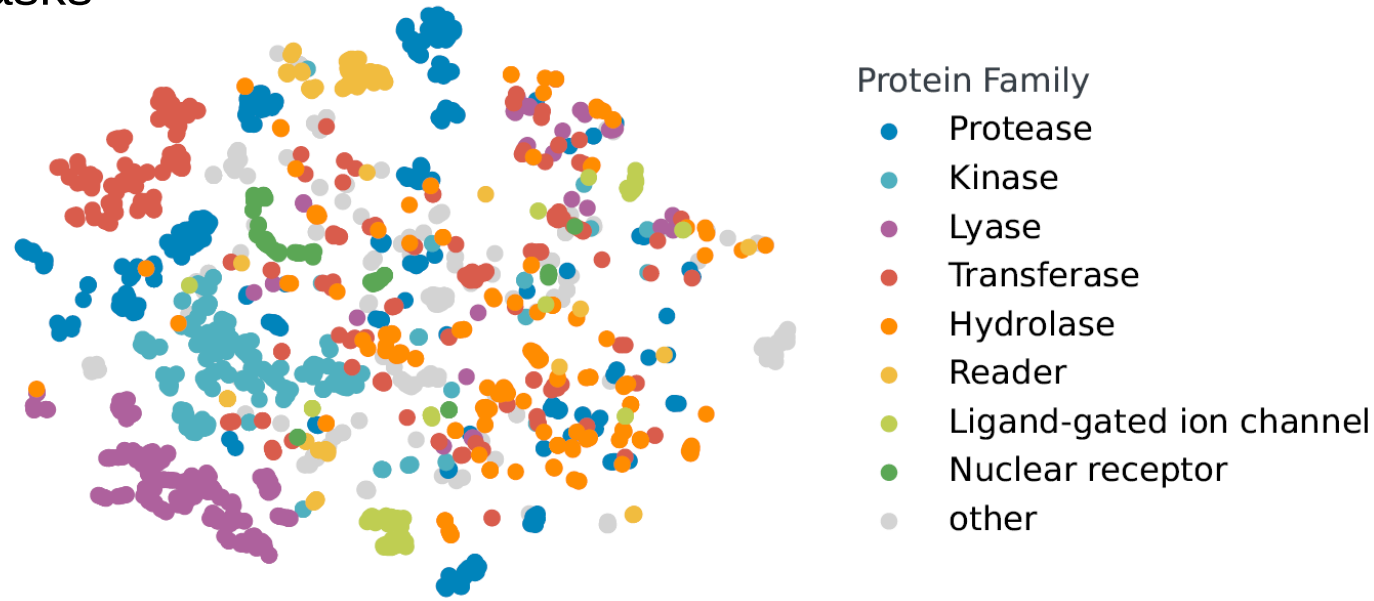
Methods	Param	COACH420		HOLO4K[d]		PDBbind2020	
	(M)	DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
Fpocket (Le Guilloux et al., 2009)[b]	\	0.228	0.444	0.192	0.457	0.253	0.371
P2Rank (Krivák & Hoksza, 2018)[c]	\	*0.464*	*0.728*	*0.474*	**0.787**	*0.653*	**0.826**
DeepSite (Jiménez et al., 2017)[b]	1.00	\	0.564	\	0.456	\	\
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
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)**

Ablation Studies

Methods	VN	heterog. MP	ESM	COACH420		HOLO4K		PDBbind2020	
				DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
EGNN (Satorras et al., 2021)[b]	✗	✗	✗	0.156(0.017)	0.361(0.020)	0.127(0.005)	0.406(0.004)	0.143(0.007)	0.302(0.006)
VN-EGNN (residue emb.)	✓	✓	✗	0.503(0.022)	0.684(0.016)	0.438(0.019)	0.605(0.013)	0.551(0.017)	0.751(0.009)
VN-EGNN (homog.)	✓	✗	✓	0.575(0.008)	0.708(0.009)	0.479(0.012)	0.595(0.010)	0.649(0.010)	0.805(0.006)
VN-EGNN (full)	✓	✓	✓	**0.605(0.009)**	**0.750(0.008)**	**0.532(0.021)**	**0.659(0.026)**	**0.669(0.015)**	**0.820(0.010)**

Representations of Binding Pockets

- feature vectors of VNs represent binding pockets
- used for ranking predictions
- might be useful for down-stream tasks

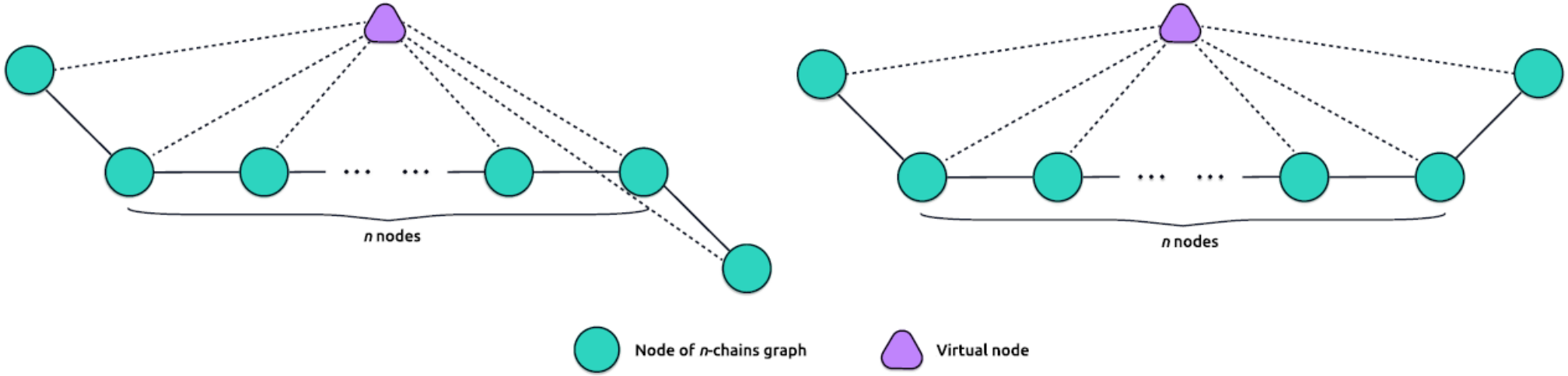


TSNE-embeddings of VN features colored by protein family

Increased Expressivity of VN-EGNN

To distinguish two n-chain graphs:

- need $\left\lfloor \frac{n}{2} \right\rfloor + 1$ layers of EGNN
- one layer of VN-EGNN is sufficient



Increased Expressivity – Empirical Results

Experiments on 4-chain graphs:

	Dim.	1 Layer	2 Layers	3 Layers	4 Layers	5 Layers	6 Layers	7 Layers	8 Layers
EGNN	8	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	98.0 ± 9.8	94.0 ± 16.2	93.0 ± 17.3	99.5 ± 5.0	99.5 ± 5.0
	16	50.0 ± 0.0	50.0 ± 0.0	86.0 ± 22.4	97.5 ± 10.9	99.5 ± 5.0	99.5 ± 5.0	99.5 ± 5.0	100.0 ± 0.0
	32	50.0 ± 0.0	50.0 ± 0.0	56.5 ± 16.8	50.0 ± 0.0	50.0 ± 0.0	96.5 ± 12.8	99.0 ± 7.0	93.5 ± 16.8
	64	50.0 ± 0.0	50.0 ± 0.0	100.0 ± 0.0	99.0 ± 7.0	100.0 ± 0.0	99.0 ± 7.0	100.0 ± 0.0	100.0 ± 0.0
	128	50.0 ± 0.0	50.0 ± 0.0	96.5 ± 12.8	98.5 ± 8.5	95.0 ± 15.0	99.5 ± 5.0	99.5 ± 5.0	99.5 ± 5.0
VN-EGNN	8	65.5 ± 23.1	50.0 ± 0.0	84.5 ± 23.1	92.5 ± 17.9	64.0 ± 22.4	97.0 ± 11.9	86.5 ± 23.3	97.5 ± 10.9
	16	86.0 ± 23.5	95.0 ± 15.0	98.5 ± 8.5	99.5 ± 5.0	99.5 ± 5.0	98.0 ± 9.8	99.5 ± 5.0	100.0 ± 0.0
	32	95.0 ± 15.0	100.0 ± 0.0	99.5 ± 5.0	99.5 ± 5.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	64	97.5 ± 10.9	100.0 ± 0.0	99.5 ± 5.0	99.5 ± 5.0	99.0 ± 7.0	100.0 ± 0.0	100.0 ± 0.0	99.5 ± 5.0
	128	99.0 ± 7.0	99.5 ± 5.0	99.5 ± 5.0	99.0 ± 7.0	99.5 ± 5.0	99.5 ± 5.0	99.0 ± 7.0	99.0 ± 7.0

Summary

- We propose **VN-EGNN**, an equivariant method for binding site identification.
- VN-EGNN uses **virtual nodes** to represent the binding pocket.
- Presumably, this is the **first application** of virtual nodes to geometric graph networks.
- VN-EGNN learns a **feature representation** of the binding pocket which can be beneficial for down-stream tasks.
- VN-EGNN has **increased expressivity** compared to traditional EGNNs.

JOHANNES KEPLER
UNIVERSITY LINZ

References I

[1] Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).

[2] Volkamer, A. & Rarey, M. Exploiting structural information for drug-target assessment. *Future Med. Chem.* **6**, 319–331 (2014).

[3] Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminformatics* **10**, 39 (2018).

[4] Lu, W. *et al. TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction*. http://biorxiv.org/lookup/doi/10.1101/2022.06.06.495043 (2022) doi:10.1101/2022.06.06.495043.

[5] Mylonas, S. K., Axenopoulos, A. & Daras, P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinforma. Oxf. Engl.* **37**, 1681–1690 (2021).

References II

[6] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning* 1263–1272 (PMLR, 2017).

[7] Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. *Preprint at http://arxiv.org/abs/2102.09844* (2022).

[8] Zhang, Y., Huang, W., Wei, Z., Yuan, Y. & Ding, Z. EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction. *Preprint at http://arxiv.org/abs/2302.12177* (2023).

[9] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

JOHANNES KEPLER
UNIVERSITY LINZ