GENEOnet: A GENEO based approach to Pocket Detection.

Giovanni Bocchi
PhD Student in Mathematics
University of Milan

Collaborators

Alessandra Micheletti [1]

[1] Department of Environmental Science and Policy
University of Milan, Italy

Patrizio Frosini [2]

[2] Department of Mathematics
University of Bologna, Italy

Alessandro Pedretti [3]

[3] Department of Pharmaceutical Sciences
University of Milan, Italy

Carmine Talarico Filippo Lunghini Andrea R. Beccari [4]

[4] Dompè Farmaceutici S.p.A., Italy.

In This Talk

GENEOs

Group Equivariant Non - Expansive Operators

Mathematical entities that can be used to build efficient and interpretable networks for data analysis.

Definition

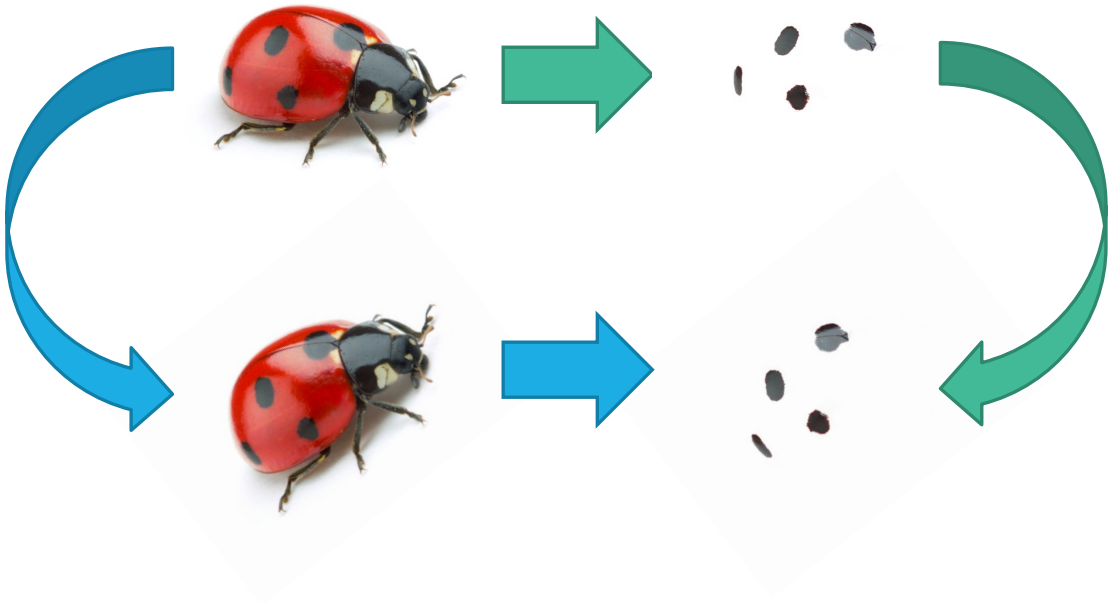
GENEO (Group Equivariant Non-Expansive Operators)

Given two functional spaces $\Phi = \{\varphi: X \to \mathbb{R}\}$ and $\Psi = \{\psi: Y \to \mathbb{R}\}$, two groups G and H of transformations of the functions domains (X and Y) and a fixed homomorphism $T: G \to H$, we define a Group Equivariant Non-Expansive Operator as a function F from Φ to Ψ with the following two properties:

- **Equivariance:** For every $\varphi \in \Phi$ and $g \in G$ it holds that
$$F(\varphi \circ g) = F(\varphi) \circ T(g)$$
- **Non-Expansivity:** For every $\varphi_1, \varphi_2 \in \Phi$ it holds that
$$d(F(\varphi_1), F(\varphi_2)) \leq d(\varphi_1, \varphi_2)$$

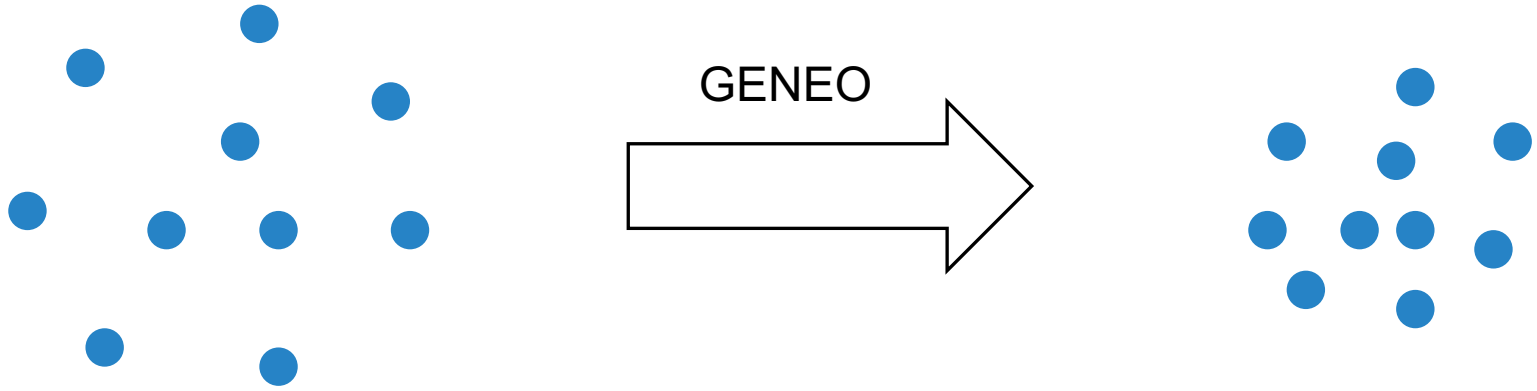
Equivariance

Equivariance means that a GENEO is able to commute with a specified group of geometrical transformations.



Non-Expansivity

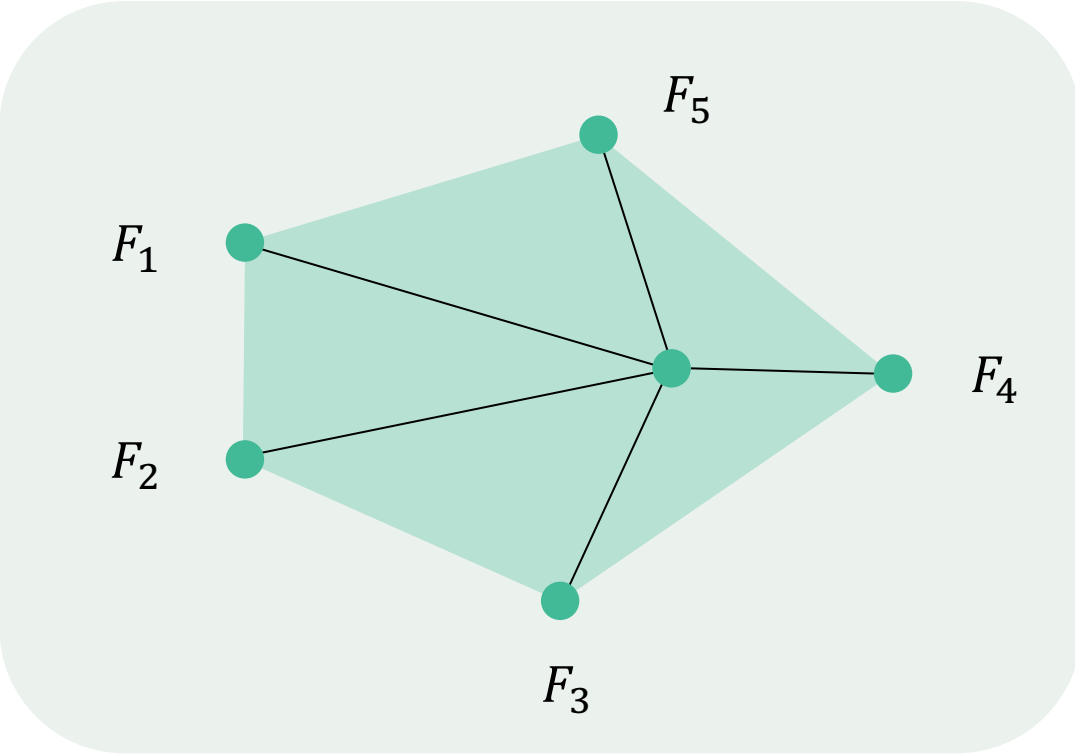
Non-Expansivity means that GENEOs do not increase distances between data functions. In some sense, they give (possibly) simpler representations of the data.



Combining GENEOs

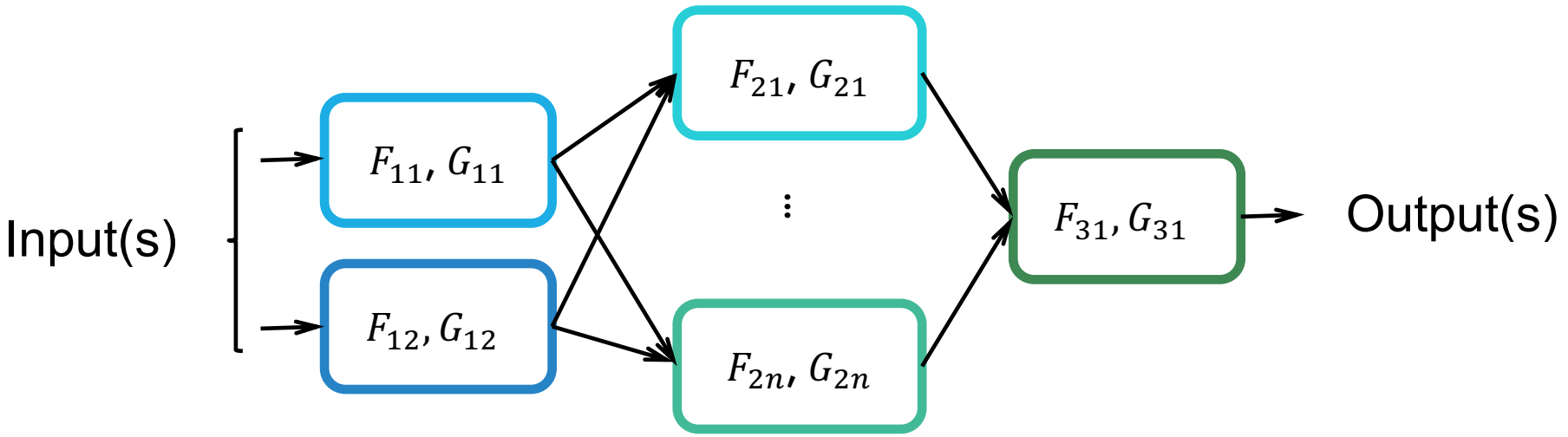
In addition we are allowed to combine GENEOs with some operations:

- Composition
- Minimum and Maximum
- Translation
- Convex combination
- ...



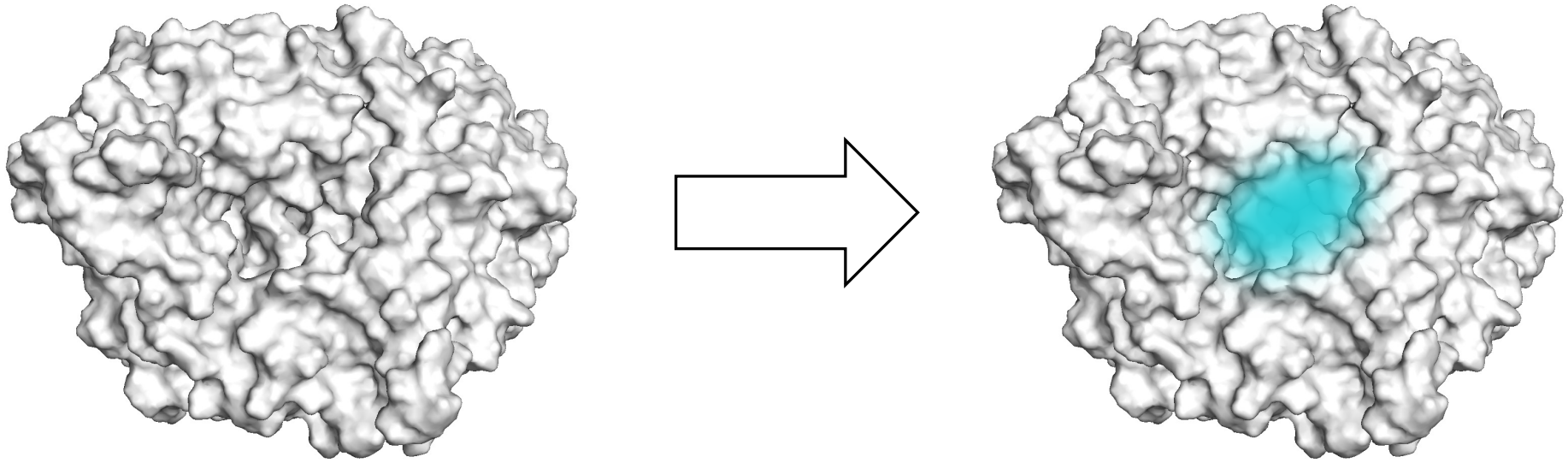
Networking

By combining different families of GENEOs, with possibly different equivariance groups, we can obtain networks of GENEOs and use them to analyze data.



Problem:
Protein Pocket Detection

The first prototype of Network of GENEOs was developed to solve the problem of identifying “druggable” pockets on the surface of proteins.



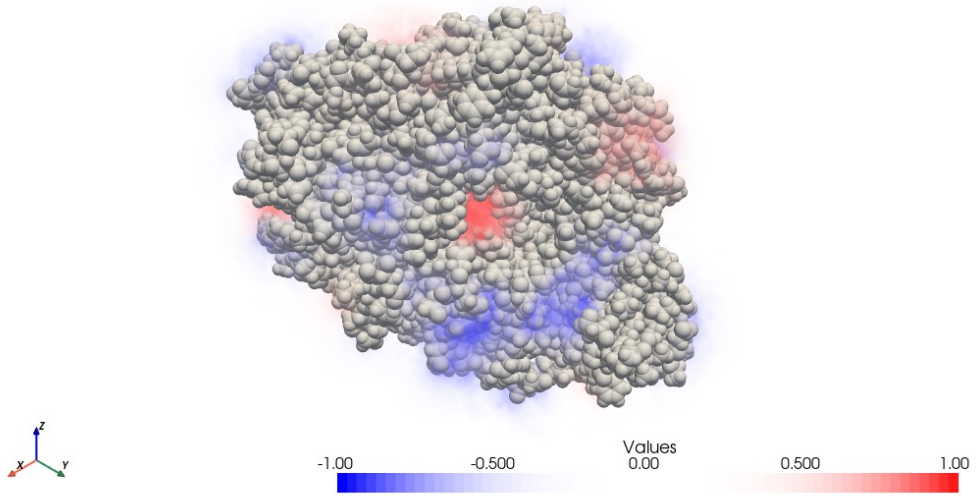
Data

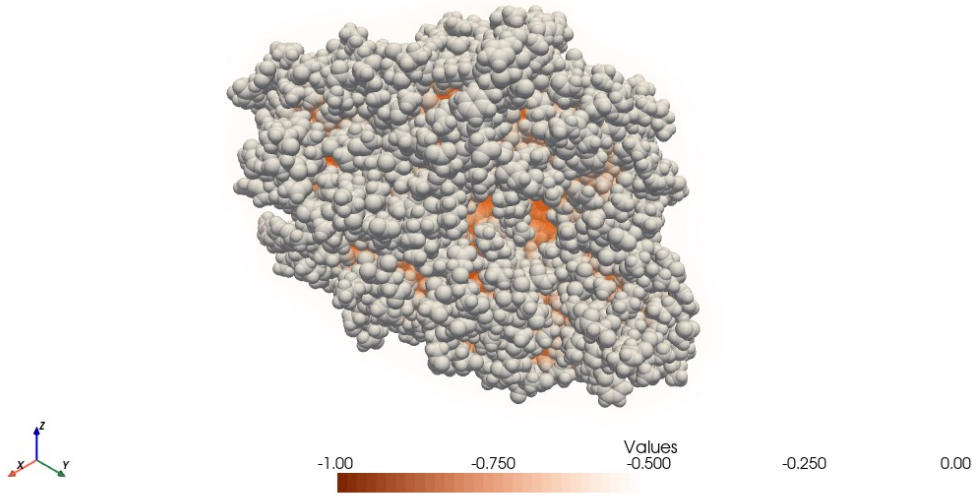
The data used to develop the model are a subset of the PDBbind dataset made of protein/ligand complexes.

These initial data are used to compute 8 functions

$$\varphi_i : B \subseteq \mathbb{R}^3 \to \mathbb{R}$$

that we refer to as “potentials”. They describe the geometrical, physical and chemical properties of a protein and are the actual inputs for GENEOs.



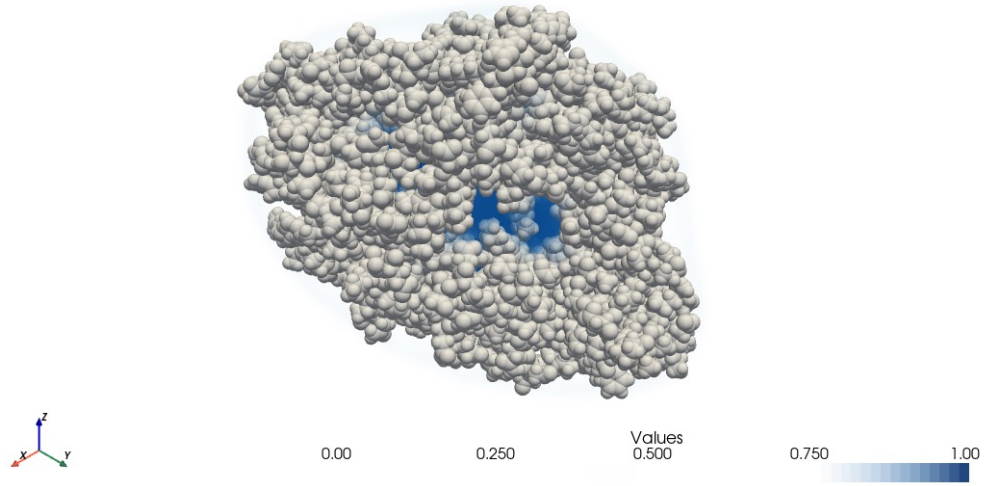


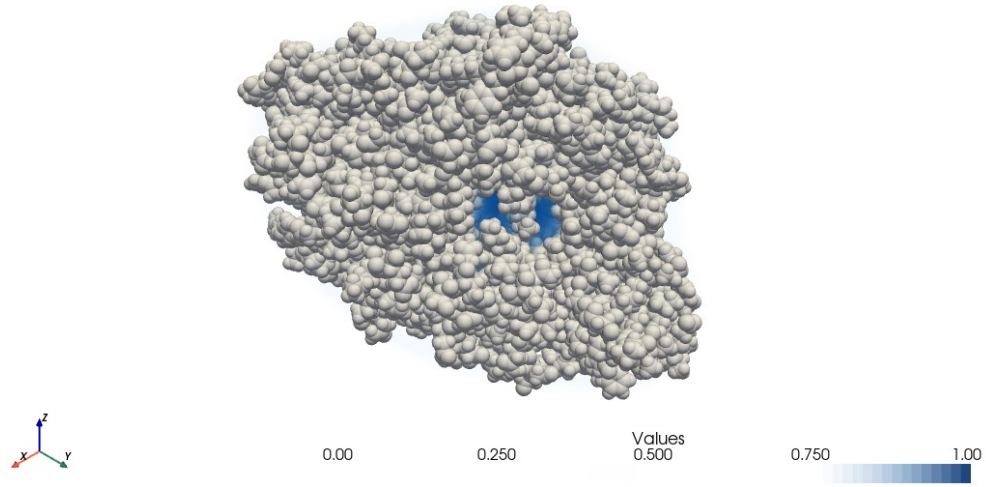
GENEOs

The GENEOs F_i developed to analyze those potentials are all convolutional operators with rotationally invariant kernels.

This guarantees equivariance w.r.t the group of rigid motions of the space.

Each kernel, that was designed to look for a specific property of the corresponding potential, depends on a shape parameter σ_i.





Aggregation

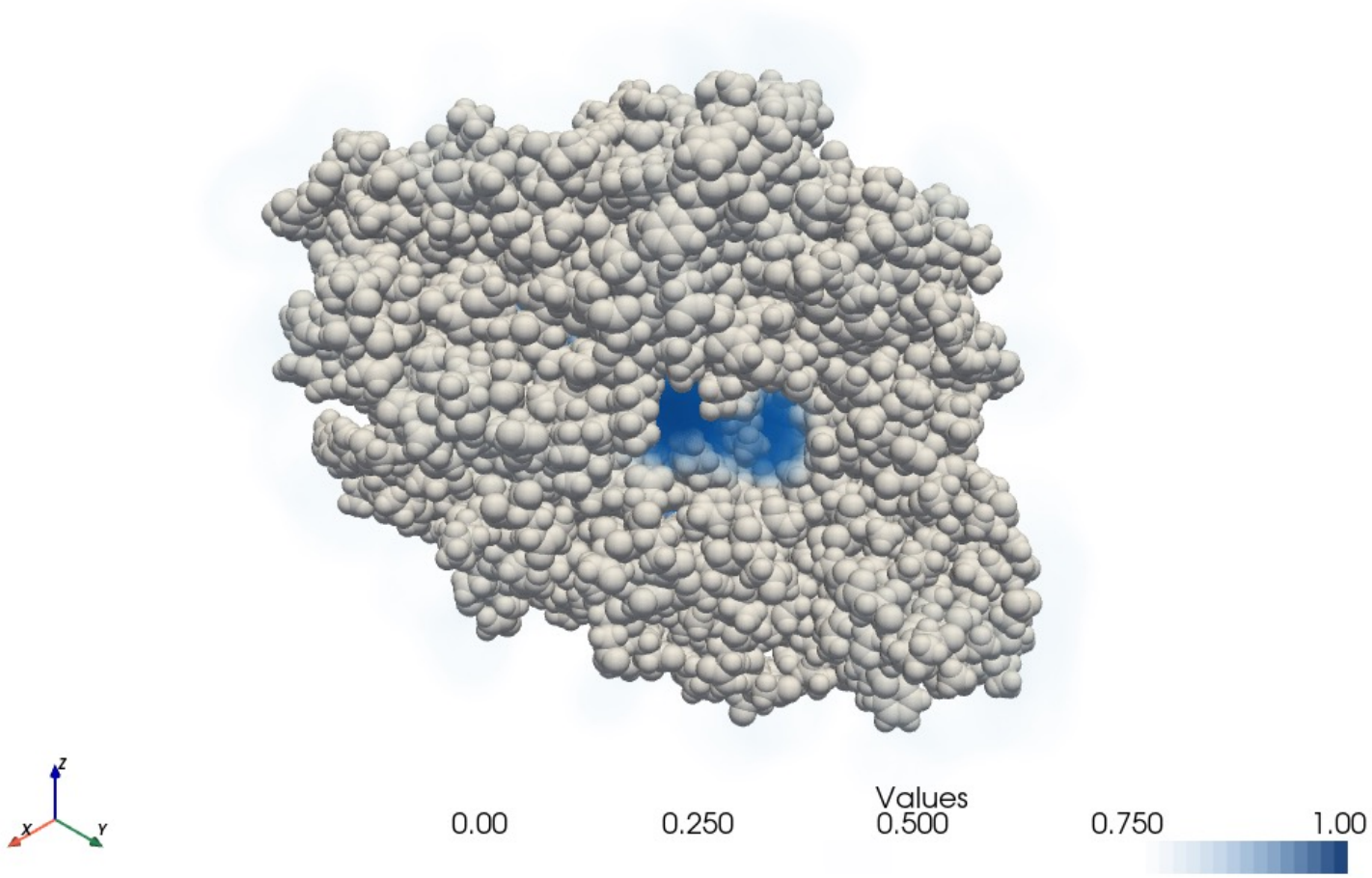
The 8 parametric families of GENEOs are then networked via a convex combination that depends on 8 non negative parameters α_i that add up to 1.

In the end we get an “aggregated” GENEO that blends the information of the various potentials returning a single output function

$$\psi\colon B \subseteq \mathbb{R}^3 \to [0,1]$$

This function can be interpreted as the one that assigns to each point of the space surrounding the protein the probability of belonging to some pocket.

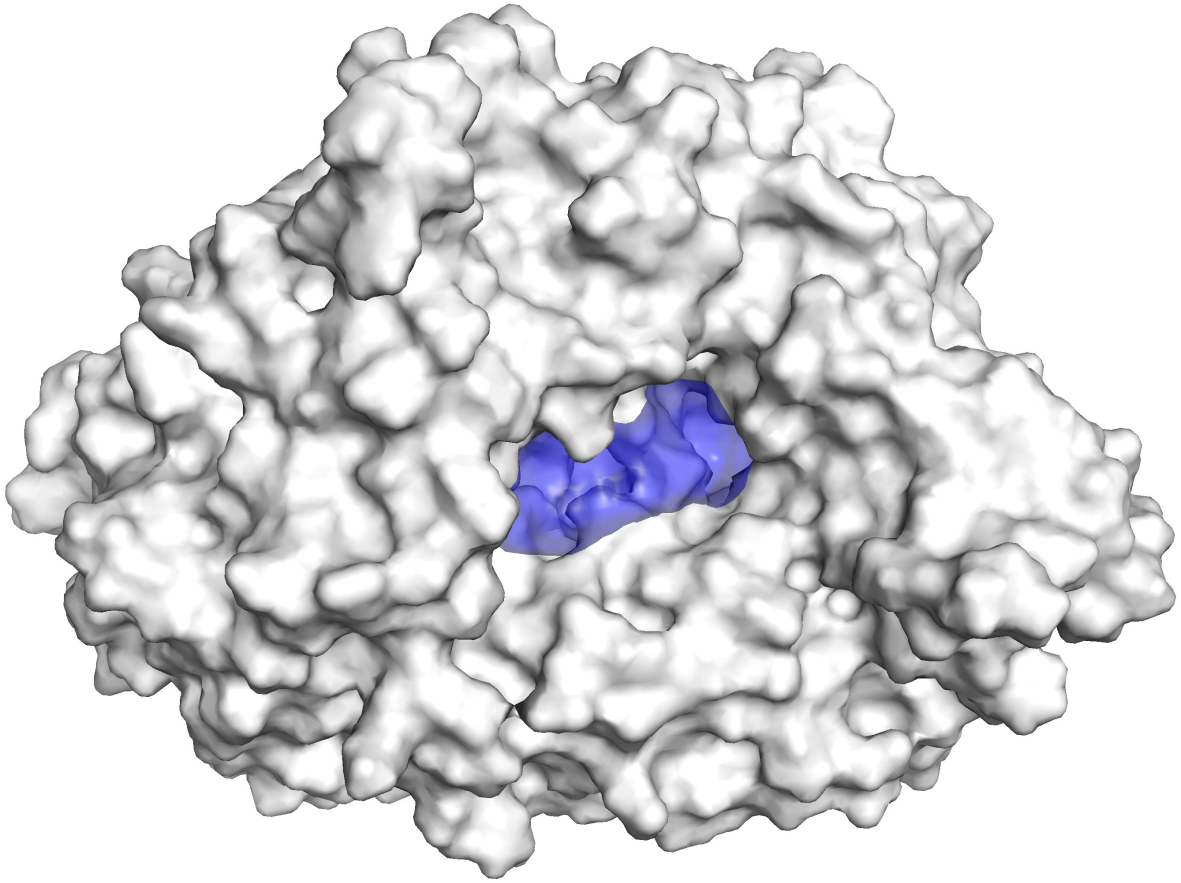
Prediction ψ



Thresholding

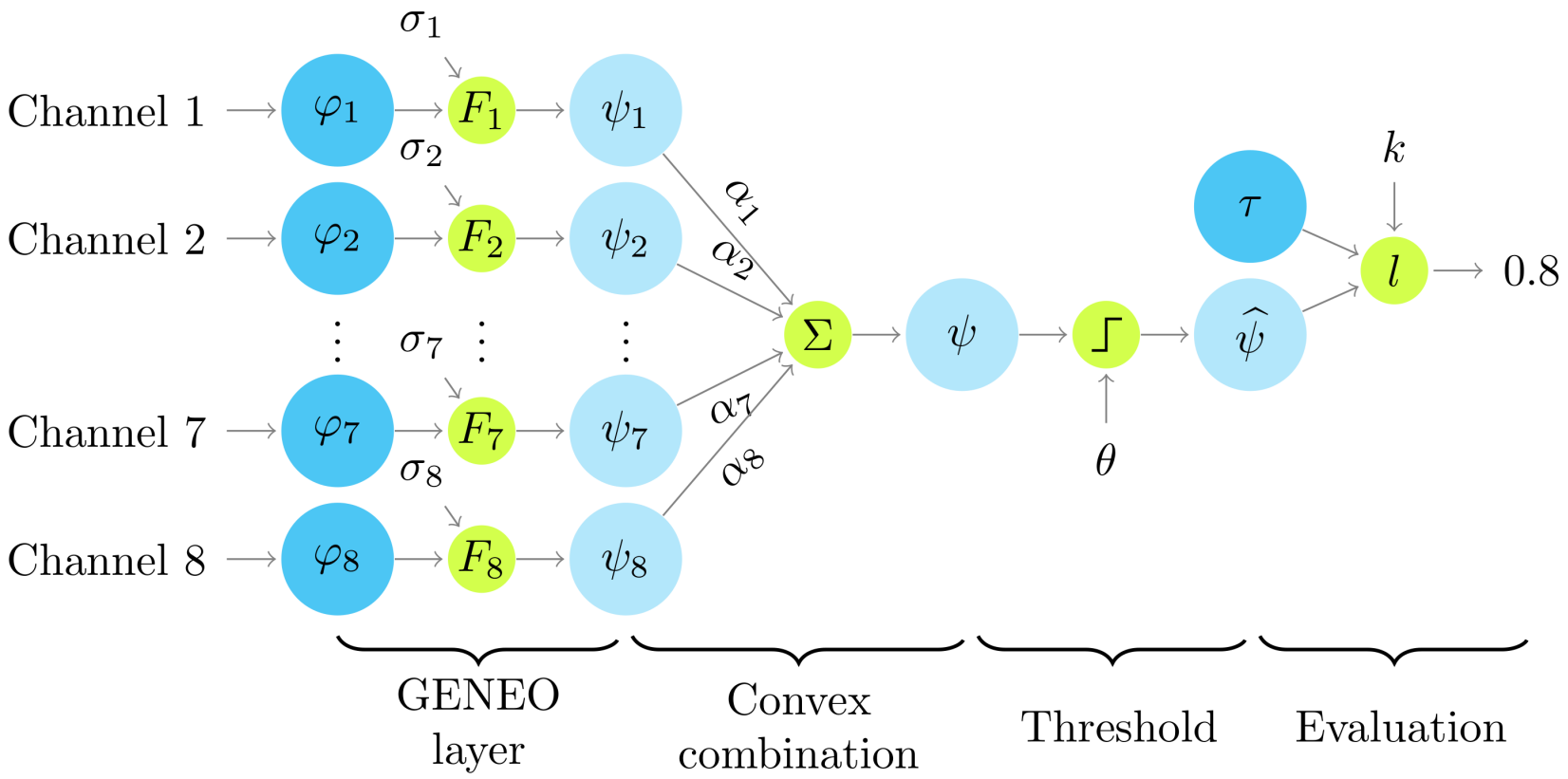
The function ψ encodes all the information useful for the final prediction.

Indeed, we can obtain a finite number of pockets by picking a threshold $\theta \in [0,1]$ and considering the connected components of the superlevel set $\{\psi \geq \theta\}$.



The Model

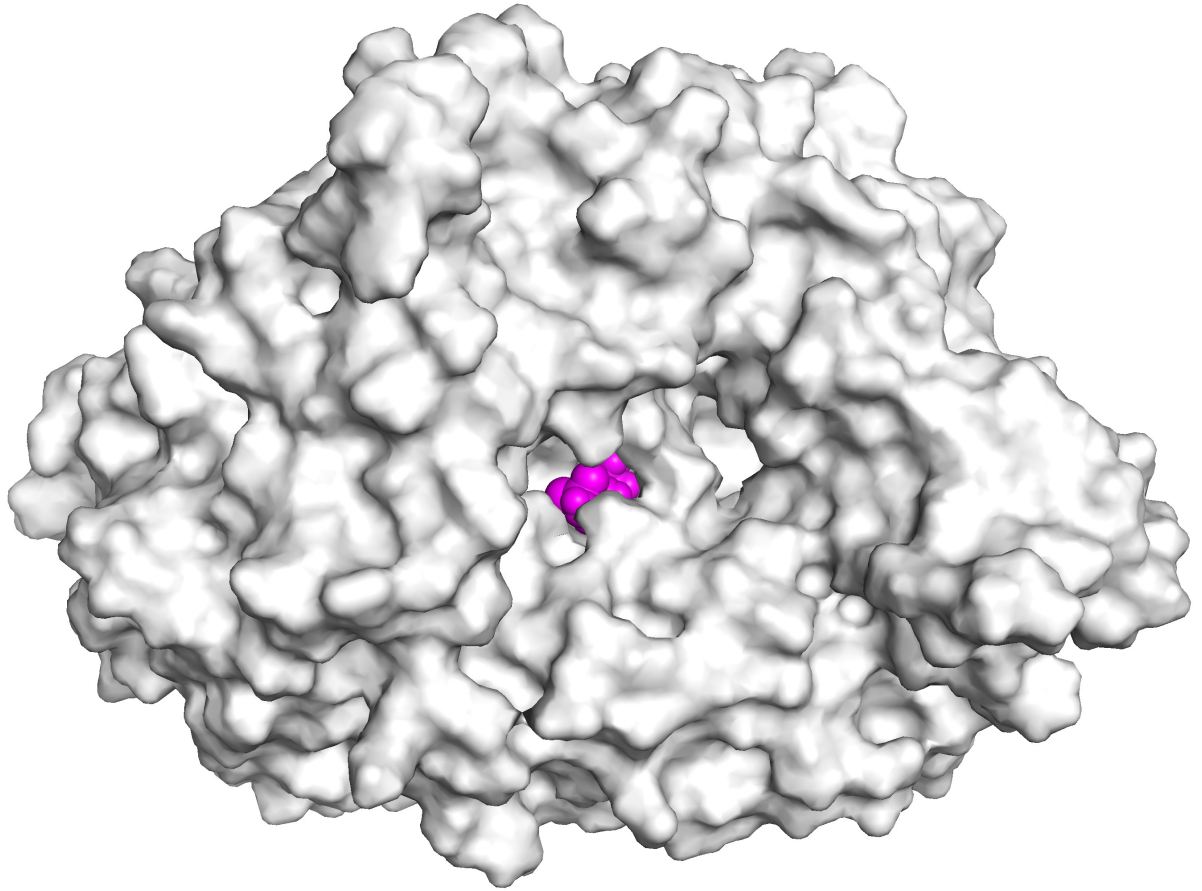
The GENEO-based model that was introduced is called GENEOnet and has the following architecture.



How to Train the Model?

GENEOnet has (just) 17 free parameters that were trained in a neural network fashion using a form of Backpropagation.

But in order to do that we need to specify a ground truth and a loss function.



Loss Function

As ground truth we considered the spatial region occupied by the ligand.

The prediction after thresholding $\hat{\psi}$ is a binary function with values in $\{0,1\}$ that can be compared with the binary ground truth τ through the following accuracy function:

$$l(\hat{\psi},\tau) = \frac{\left|\hat{\psi} \wedge \tau\right| + k\left|(\mathbf{1}-\hat{\psi}) \wedge (\mathbf{1}-\tau)\right|}{|\tau| \; + \; k|\mathbf{1}-\tau|}$$

$k = 1$	Prediction biased towards identifying non-cavities.
$k = 0$	Prediction biased towards including the ligand.
$k \in]0,1[$	Compromise between accuracy and exploration.

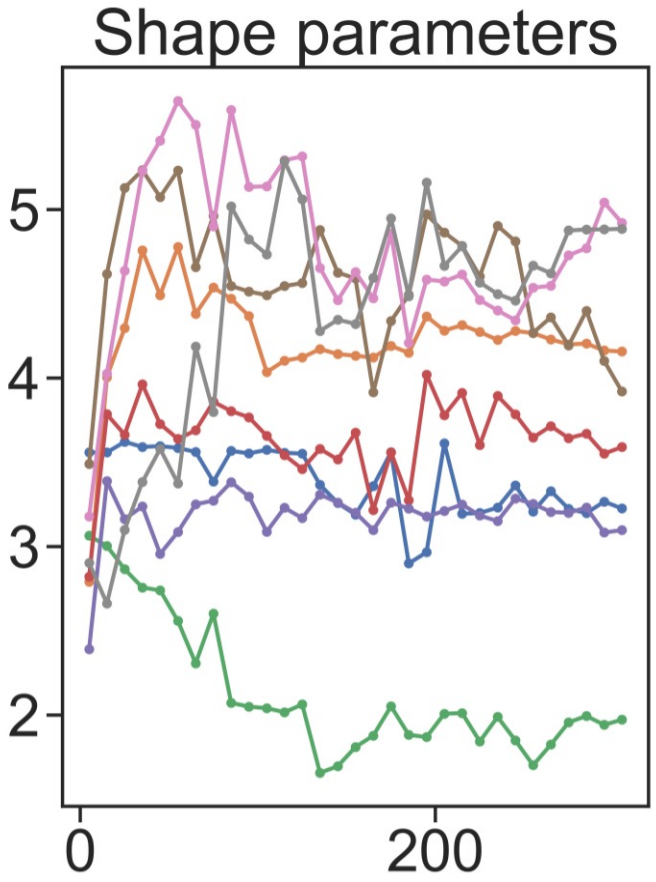
$$\mathrm{k} \approx 0.01 - 0.05$$

Training

The model has been trained to maximize the accuracy function with Adam optimizer on a training set of 200 proteins randomly sampled.

We used such a small training set since:

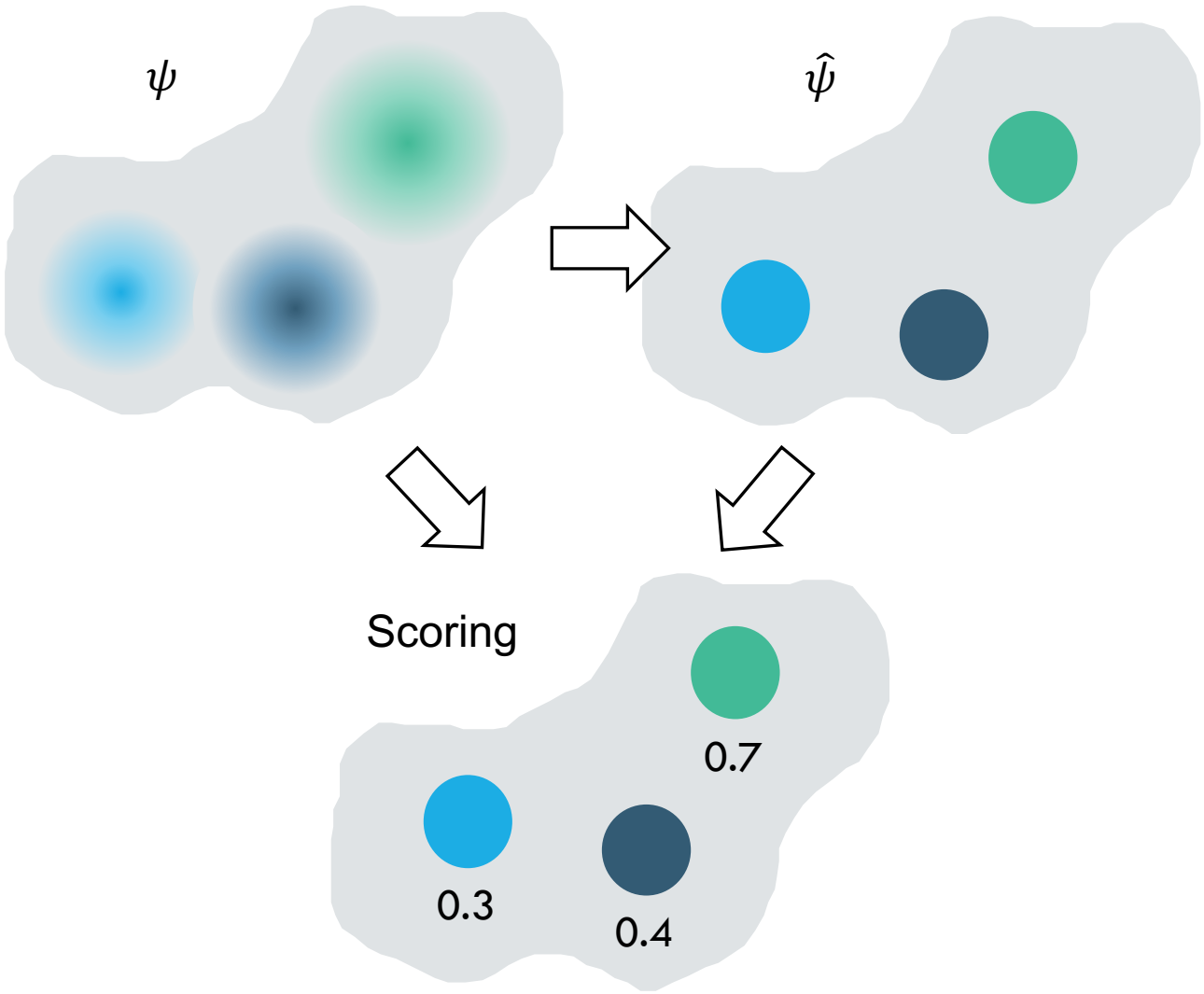
- GENEOnet has a few free parameters.
- Considering larger training sets does not impact significantly the values of the parameters.



Scoring

The final result of GENEOnet is a set of pockets without some sort of ordering.

Thus we derived a scoring function that assigns a pocket a score based on a weighted mean of the values of ψ inside the corresponding connected component.



Comparison and Model Selection

The results of GENEOnet are not easily benchmarkable, since usually different pocket finders have different internal representations of data and pockets. Moreover, when they are ML models, they can be hard to compare due to differences in the loss functions.

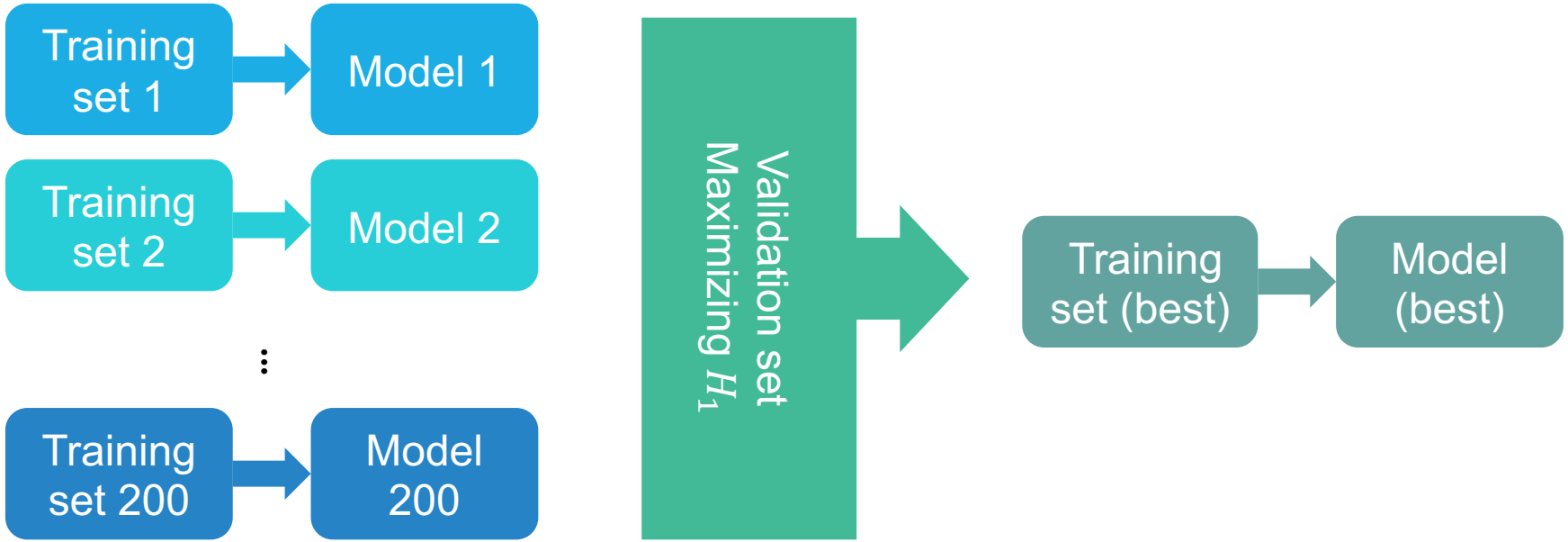
However many models outputs a list of pockets with scores. Thus we chose to compare the models testing how well they can find the right pocket in the top ranked.

$$H_i = \frac{\#\ \text{matchings by the i} - \text{th top ranked}}{\#\ \text{proteins}}$$

$$T_i = \frac{\#\ \text{matchings within the i} - \text{th top ranked}}{\#\ \text{proteins}} = \sum_{j=1}^{i} H_j$$

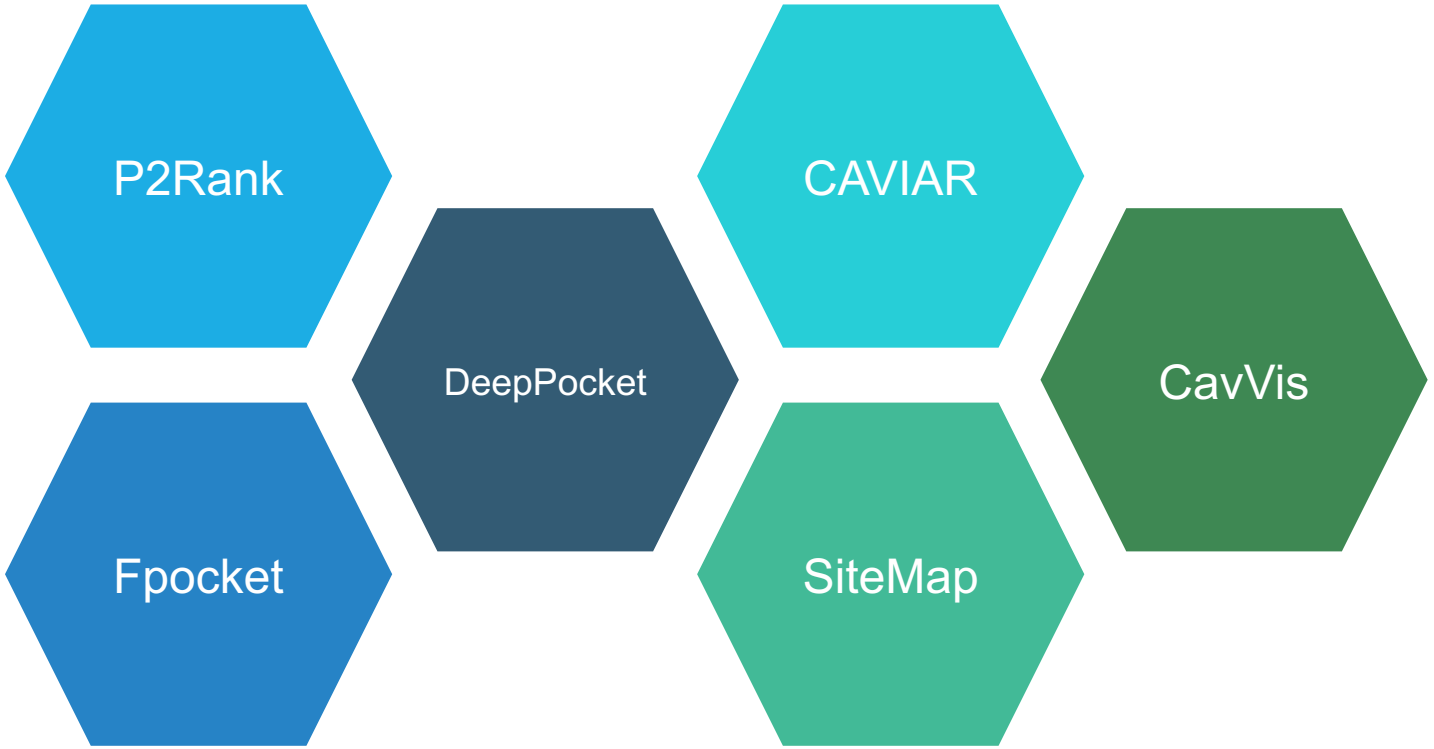
Model Selection

First, H_1 was used to perform model selection on the validation set (almost 3000 proteins).



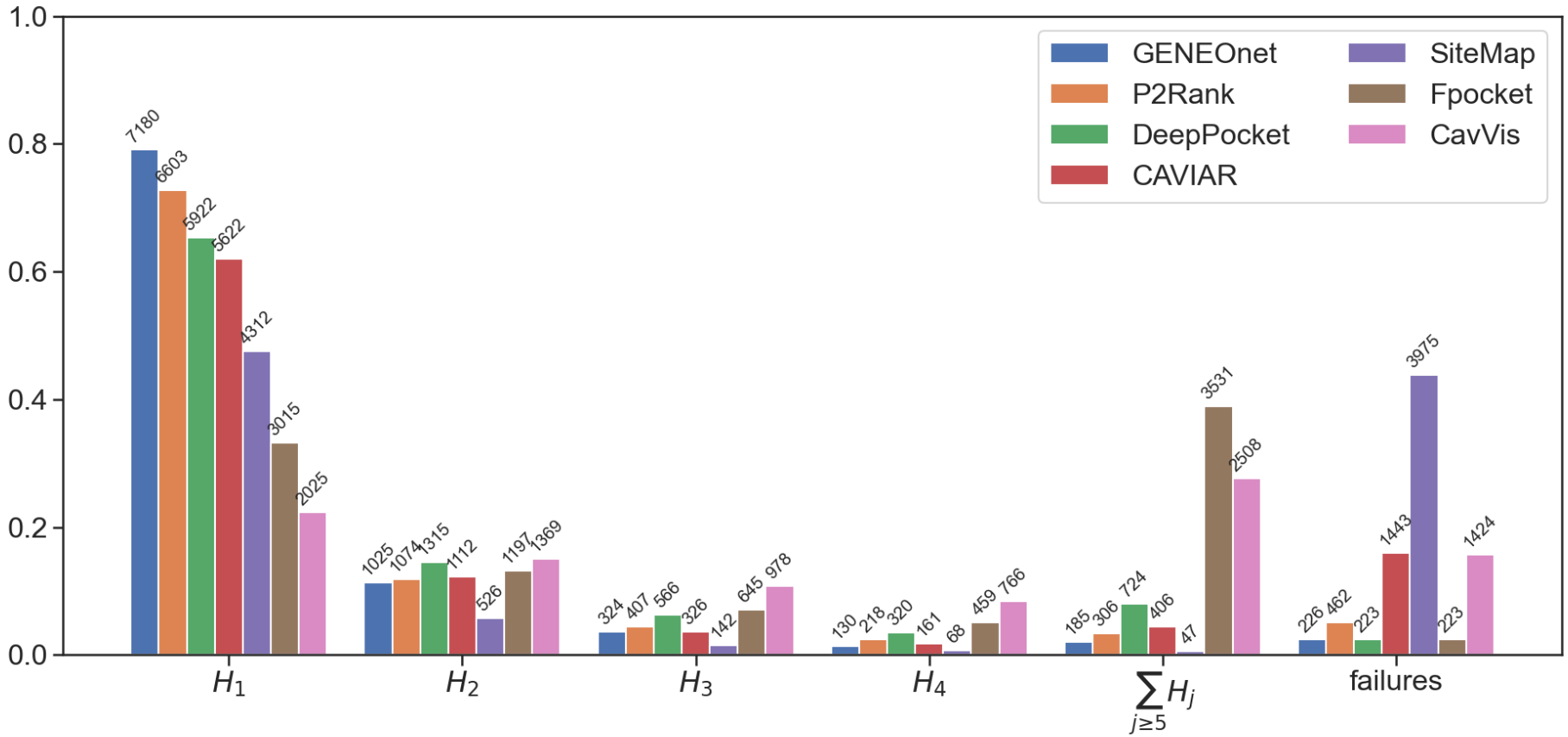
Comparison

The same metrics were used to compare GENEOnet with other state-of-the-art models for protein pocket detection.



Results

The bar chart shows the results of the model comparison on the test set (almost 9000 proteins).



Thank you for your attention!

References

- G. Bocchi, P. Frosini, A. Micheletti, A. Pedretti, C. Gratteri, F. Lunghini, A.R. Beccari and C. Talarico, *“GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection*.” preprint at arXiv 10.48550/arXiv.2202.00451.
- M. G. Bergomi, P. Frosini, D. Giorgi, and N. Quercioli, “*Towards a topological–geometrical theory of group equivariant non-expansive operators for data analysis and machine learning*” Nature Machine Intelligence, pp. 423–433, 2019. [Online]. Available: https://rdcu.be/bP6HV