A new paradigm for Artificial Intelligence based on Group Equivariant Non-Expansive Operators (GENEOs) applied to protein 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 **E**quivariant **N**on - **E**xpansive **O**perator**s**

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

Definition

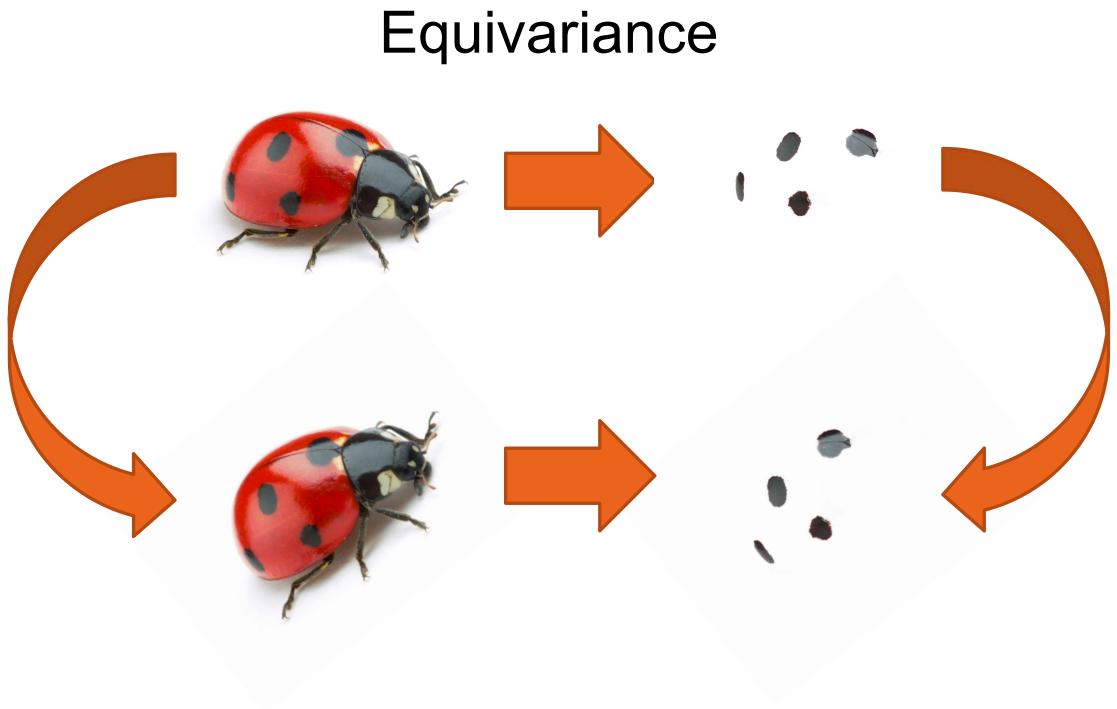
GENEO (Group Equivariant Non-Expansive Operators)

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
$$d(F(\varphi_1), F(\varphi_2)) \leq d(\varphi_1, \varphi_2)$$

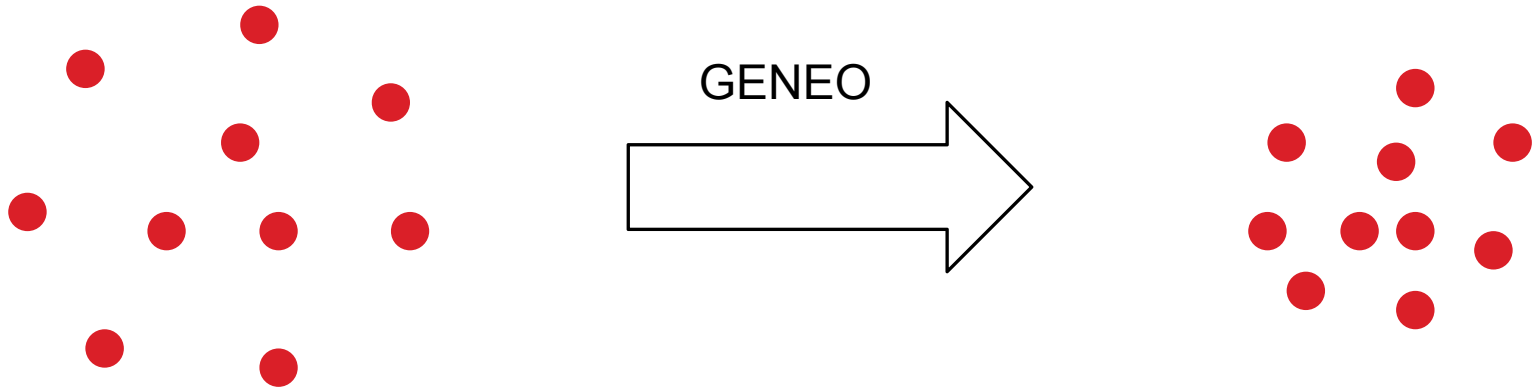
Equivariance

Equivariance implies that a GENEO commutes with a specified group of geometrical transformations.



Non-Expansivity

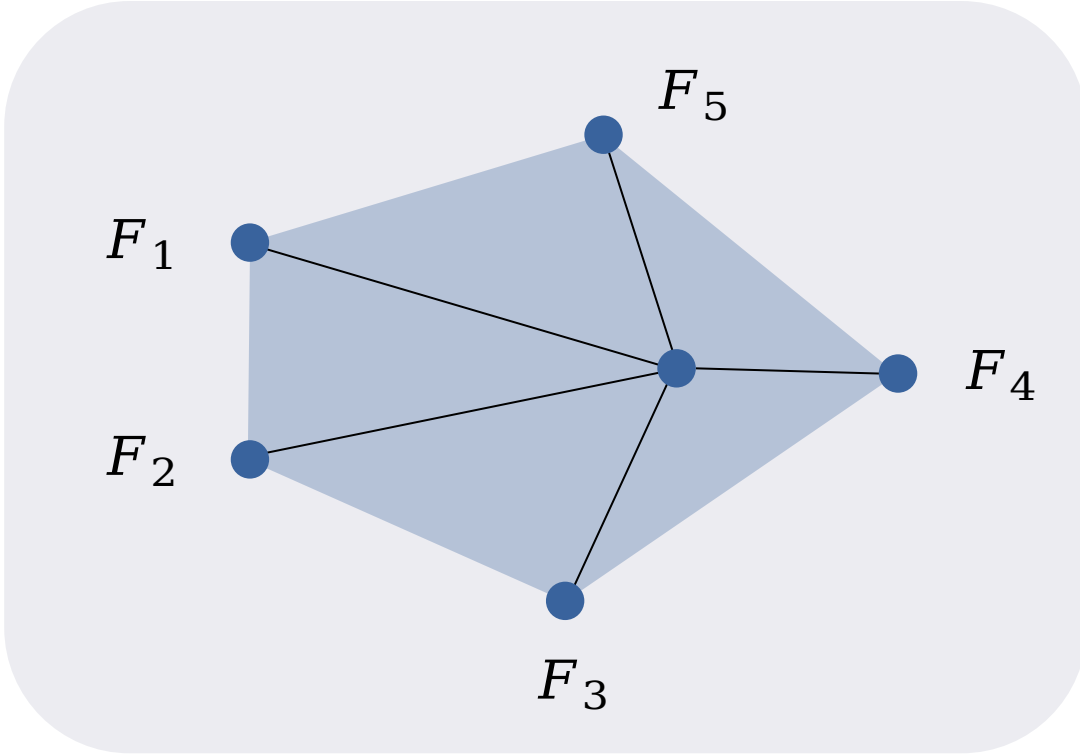
Non-Expansivity implies that GENEOs do not increase distances between data functions. In some sense, they give (possibly) simpler representations of the data.



Combining GENEOs

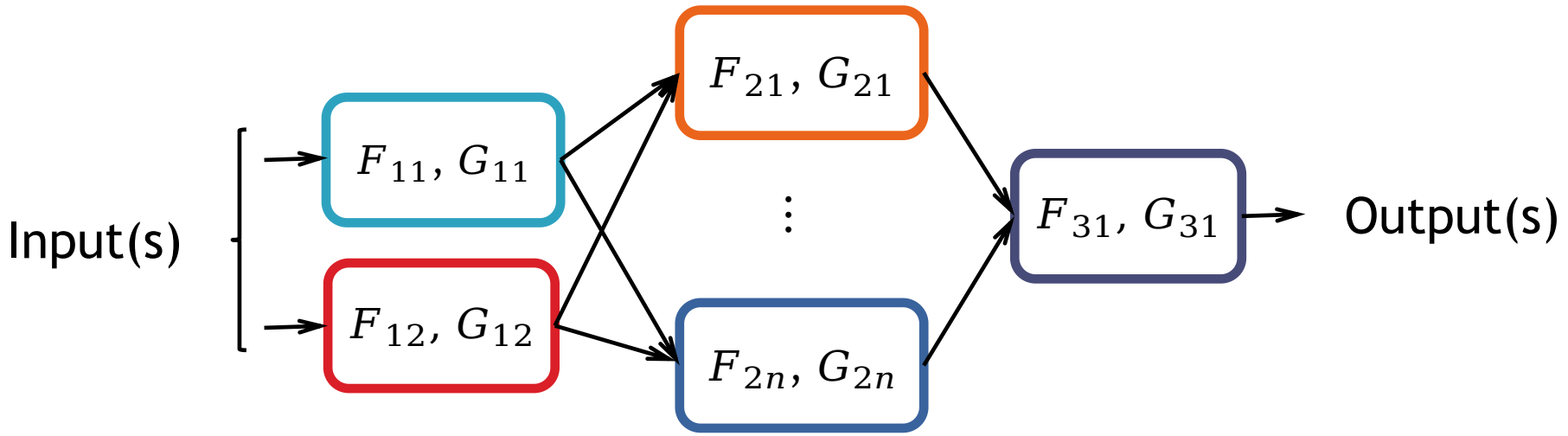
In addition we are allowed to combine GENEOs with some operations:

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



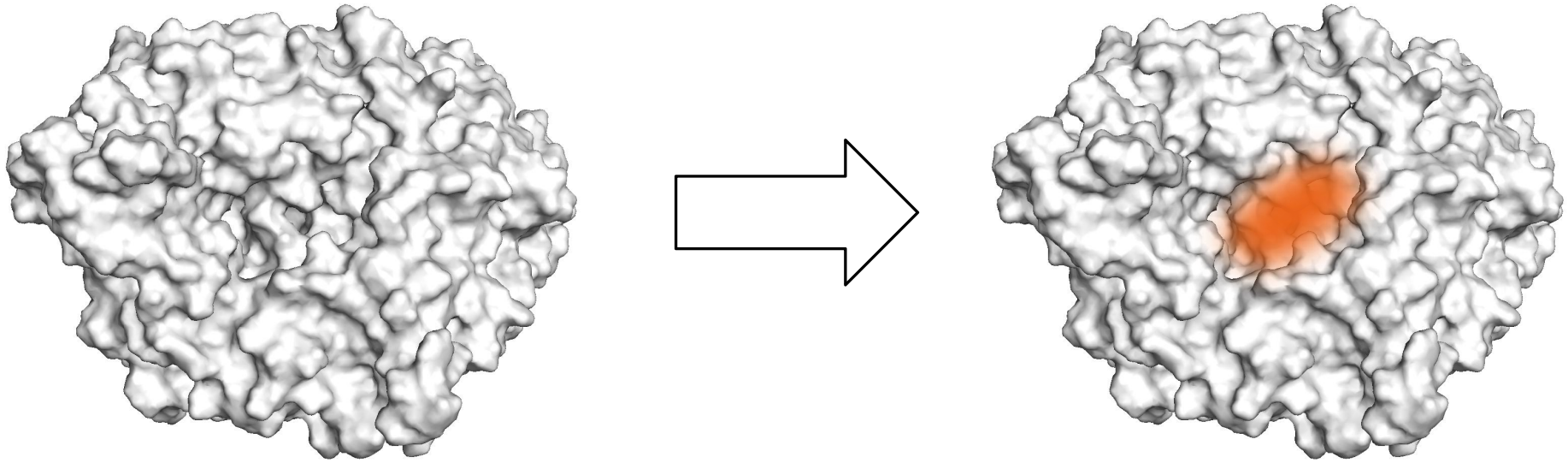
Networking

By combining different families of GENEOs, with one the allowed operations, it's possible to obtain networks of GENEOs.



Problem:
Protein Pocket Detection

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



Data & GENEOs

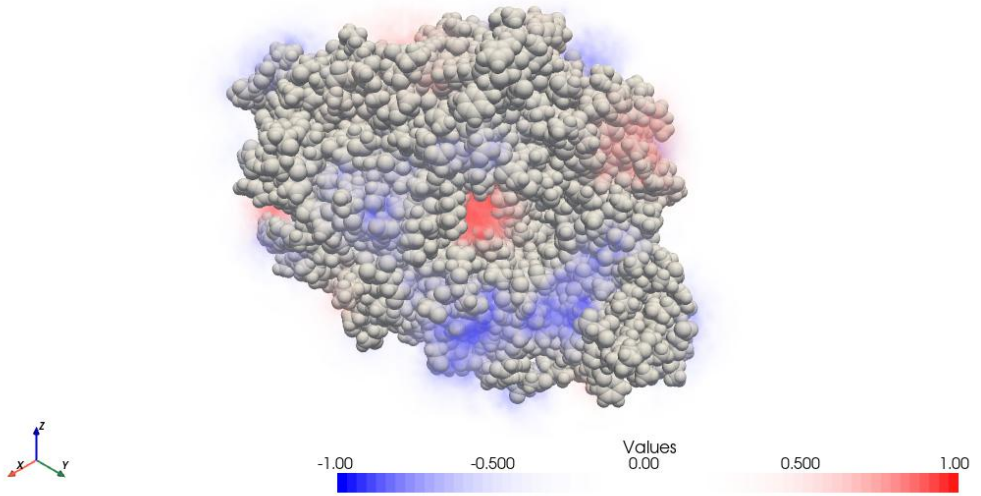
Complexes from the **PDBbind** dataset, were used to compute 8 functions

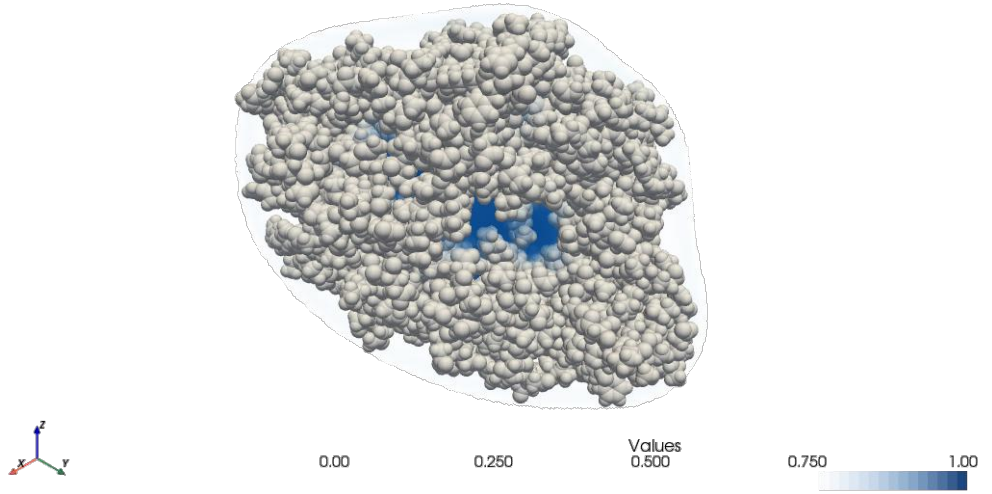
$$\varphi_i : B \subseteq \mathbb{R}^3 \to \mathbb{R}$$

called **potentials**. They describe the geometrical, physical and chemical properties of a protein.

The GENEOs F_i that are applied to φ_i are all **convolutional operators** with rotationally invariant kernels.This guarantees equivariance w.r.t rigid motions of the space.

Each kernel is designed to react to a specific property of the corresponding potential and depends on a **shape parameter** σ_i.





Aggregation & Thresholding

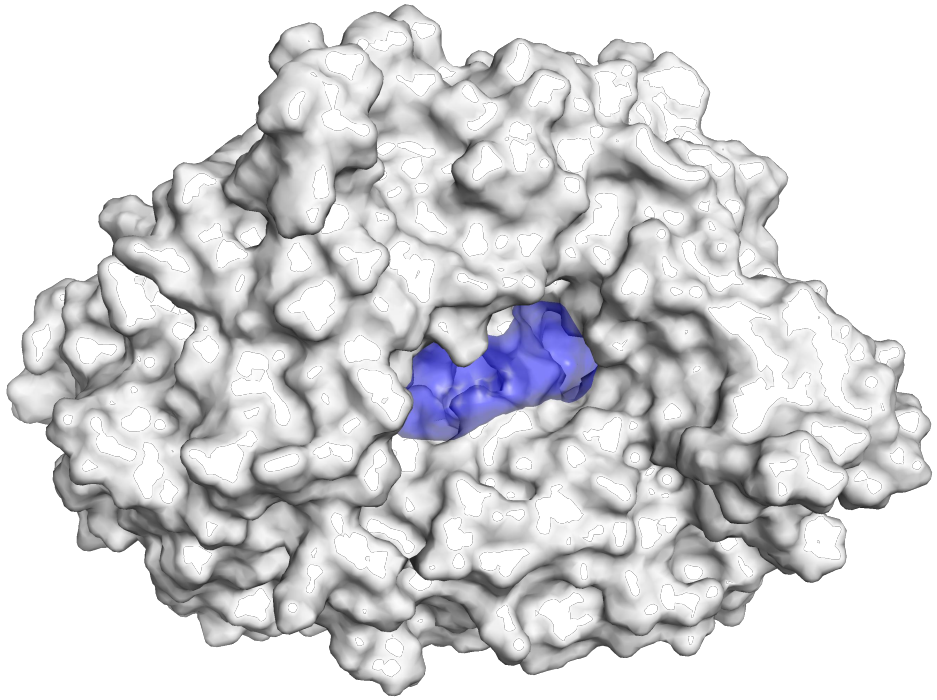
The 8 parametric families of GENEOs are then networked via a **convex combination** with weights α_i.

This “aggregated” GENEO blends the information of the potentials returning a single output function

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

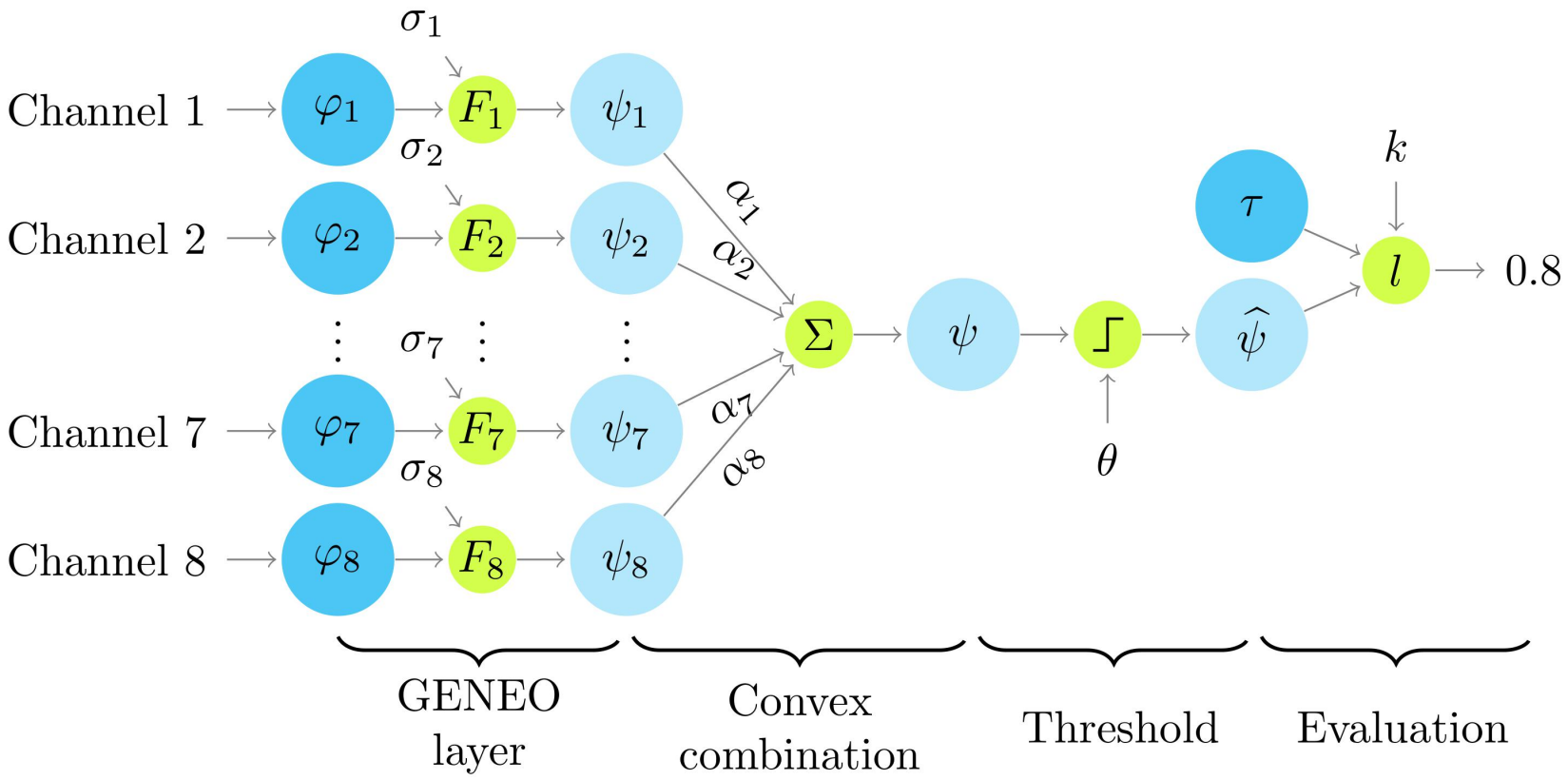
ψ assigns to each point of the space the probability of belonging to some pocket.

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



The Model

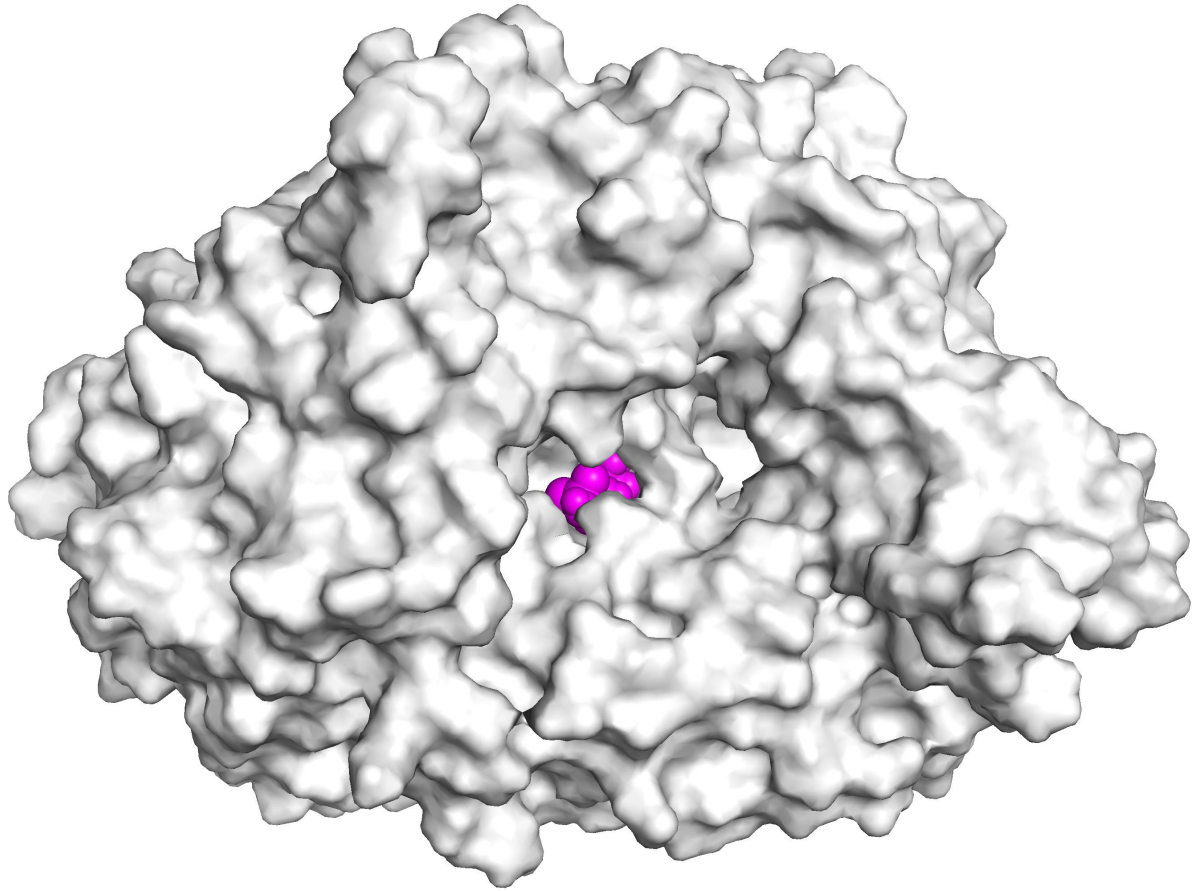
The GENEO-based model that was introduced is called **GENEOnet** and has the following architecture.



How to Train the Model?

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

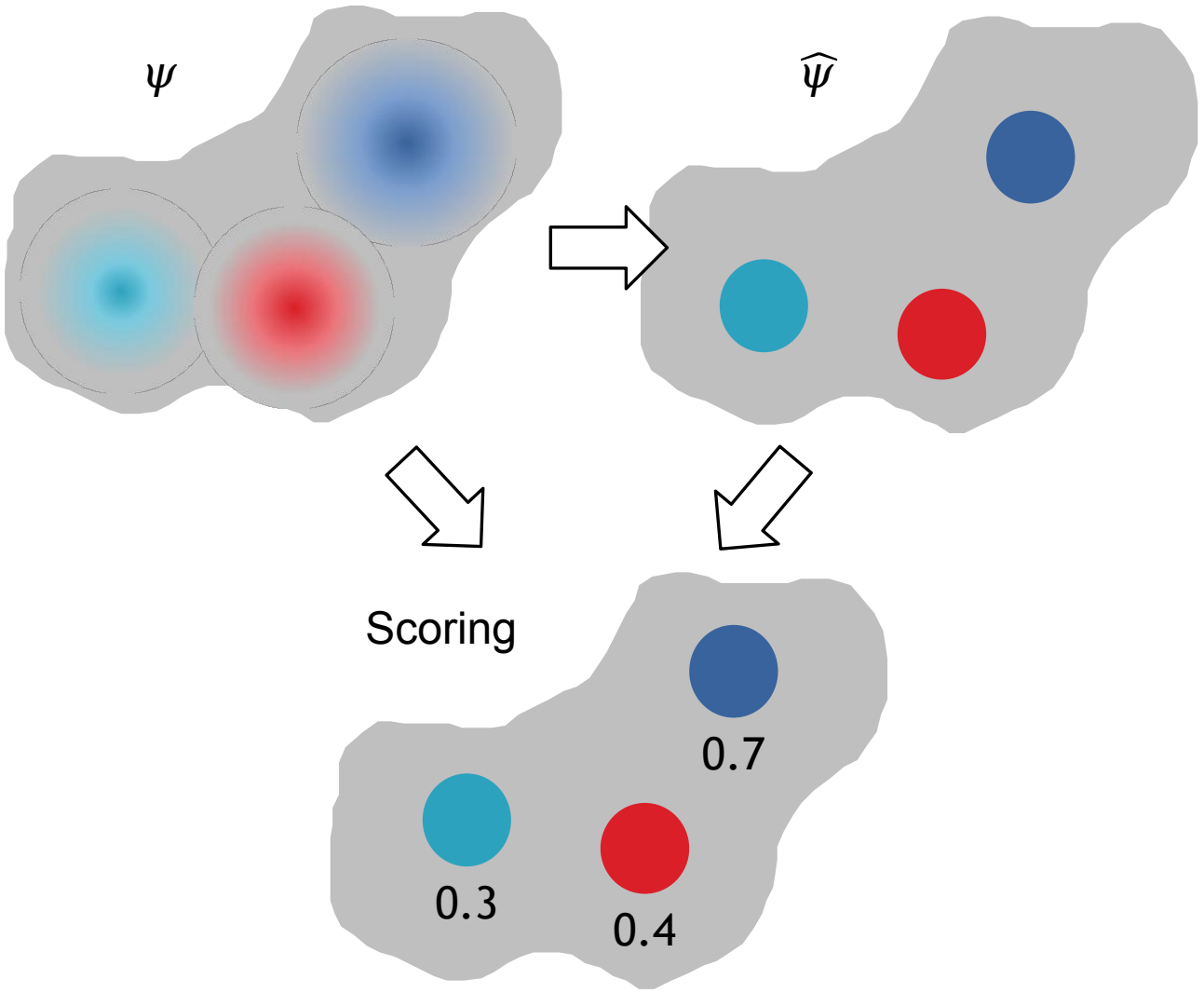
The loss function was designed in order to detect the matching between the ground truth and the prediction.



Scoring

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

Thus we derived a scoring function that assigns to 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.

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

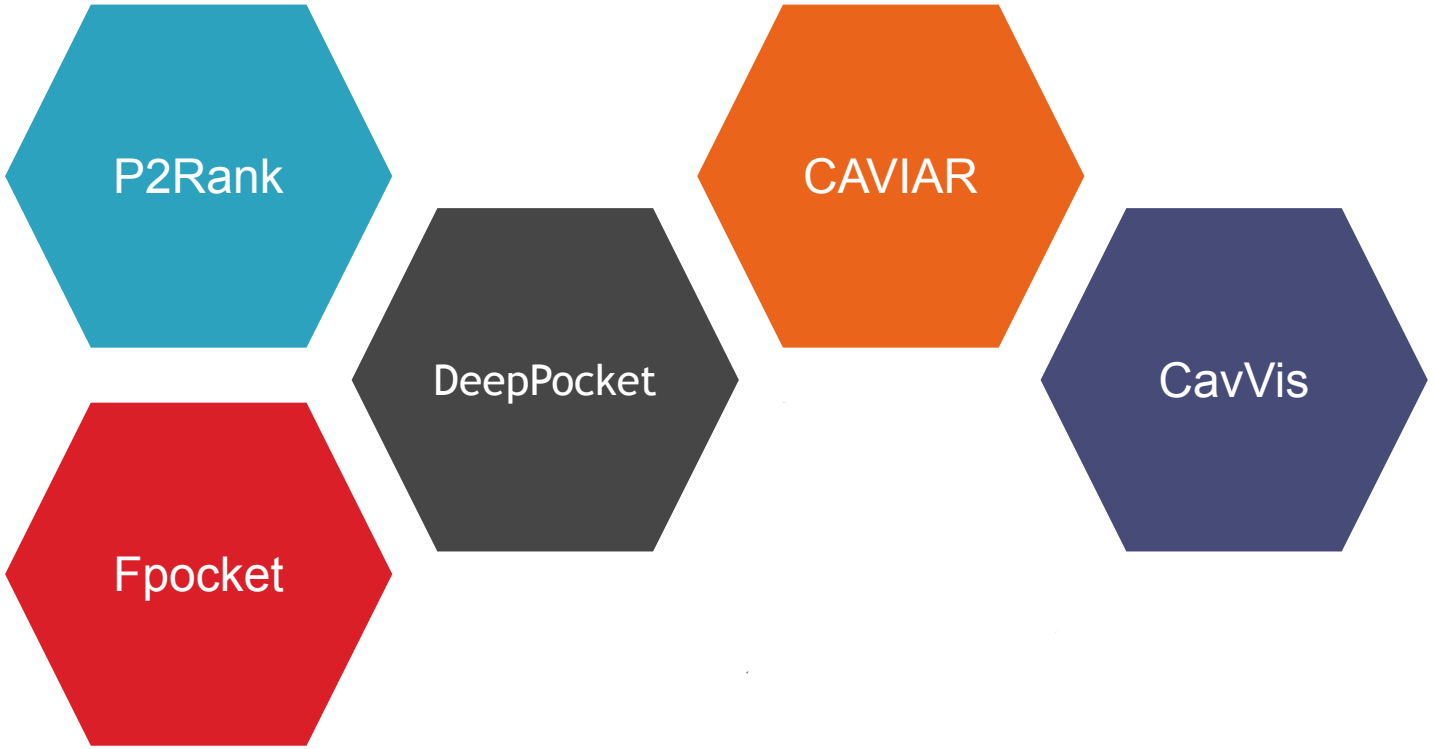
$$T_i = \frac{\#\ \text{matchings within the i} - \text{th top ranked}}{\#\ \text{proteins}} = \sum_{j=1}^{i} H_j$$

Model Selection

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

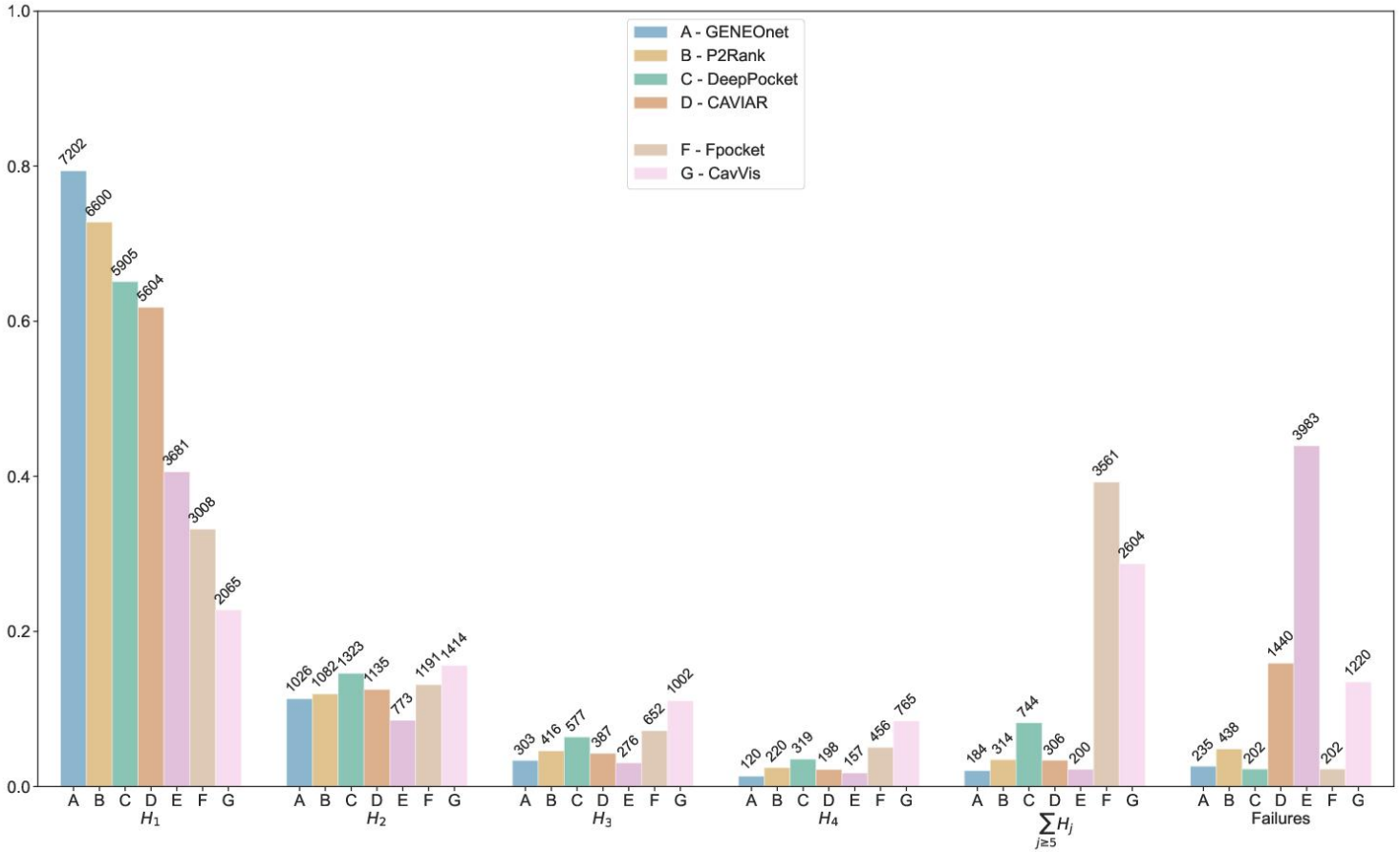
Comparison

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



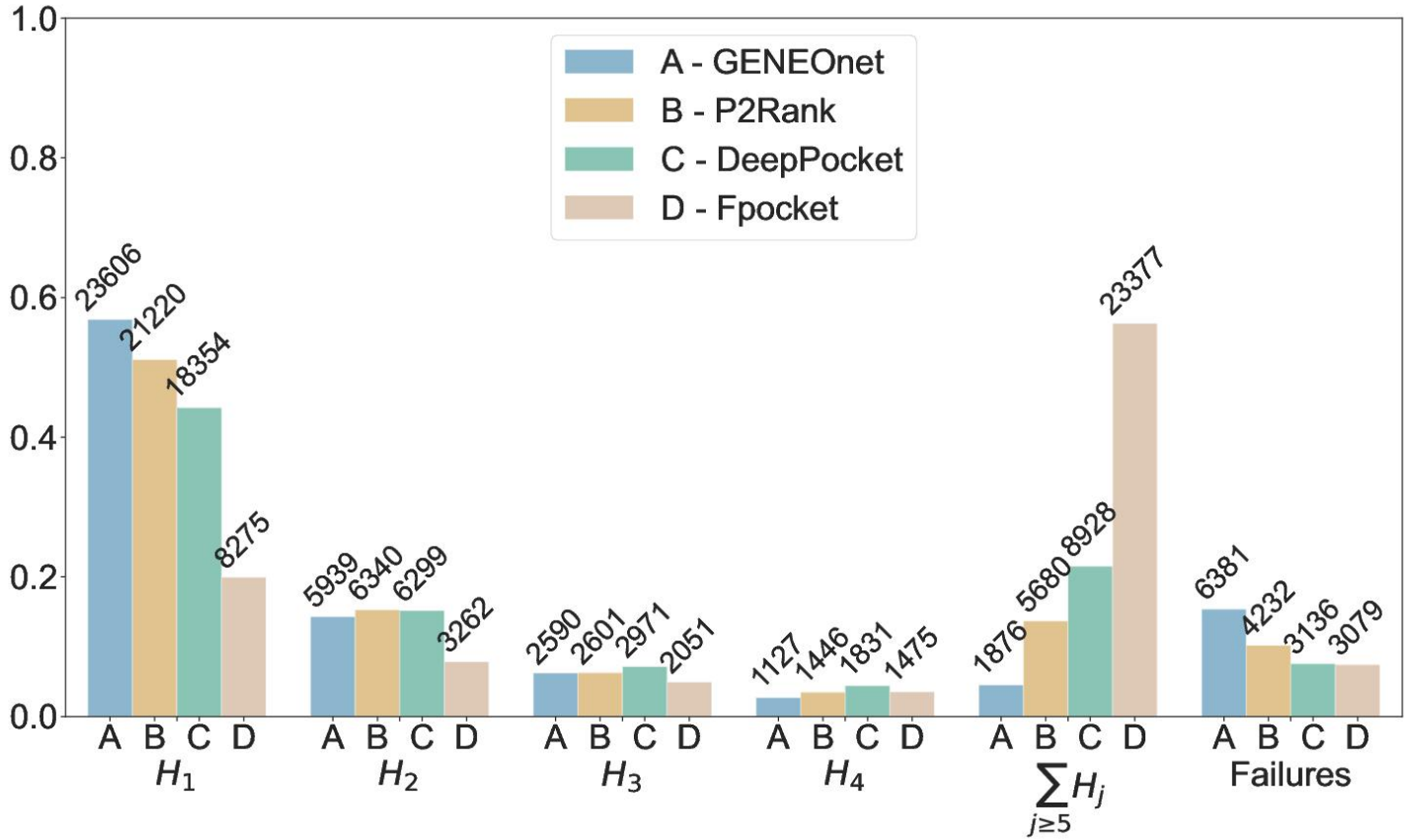
Results

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



Results

Wider evaluation on the whole **PDB dataset** made up of 41519 protein/ligand complexes.



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