Protein Pocket Detection Using Group Equivariant Non-Expansive Operators

Patrizio Frosini

Department of Computer Science, University of Pisa
patrizio.frosini@unipi.it

Knots & Proteins - Workshop in Bologna
May 19-20, 2025

Outline

Our epistemological assumptions

Some information about GENEOs

Finding pockets in proteins by applying GENEOs

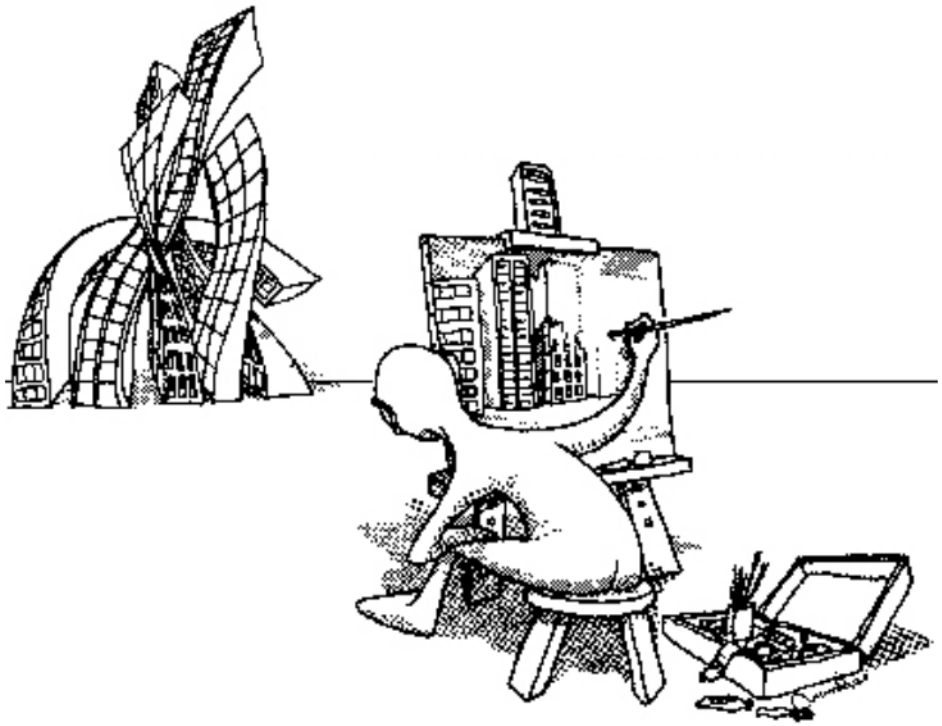
Our epistemological assumptions

Some information about GENEOs

Finding pockets in proteins by applying GENEOs

Assumption 1: Data are processed by observers

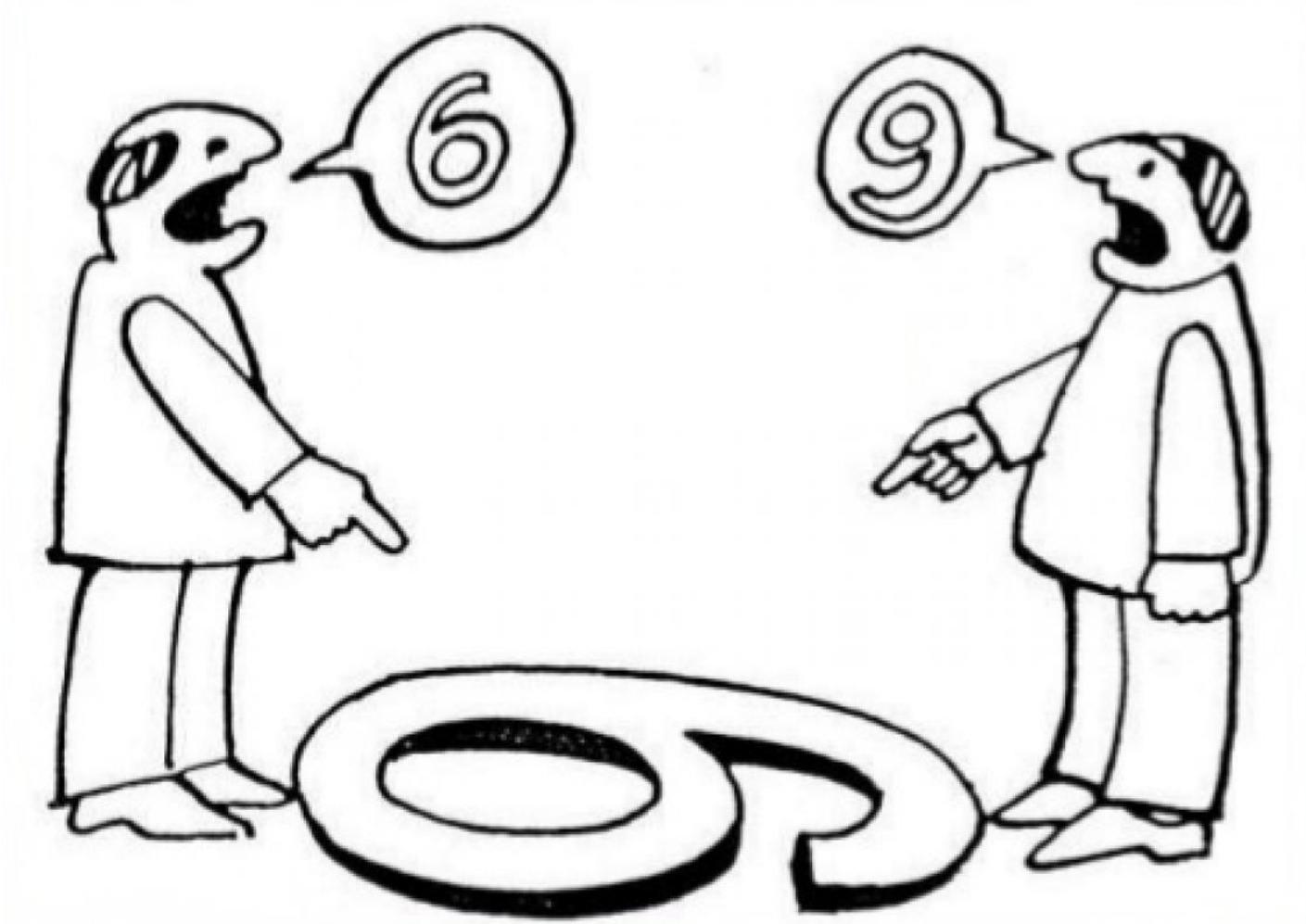
Data have no meaning without an observer to interpret them.



An observer is an agent that transforms data while preserving their symmetries.

Assumption 2: Observers are variables

Data interpretation strongly depends on the chosen observer.



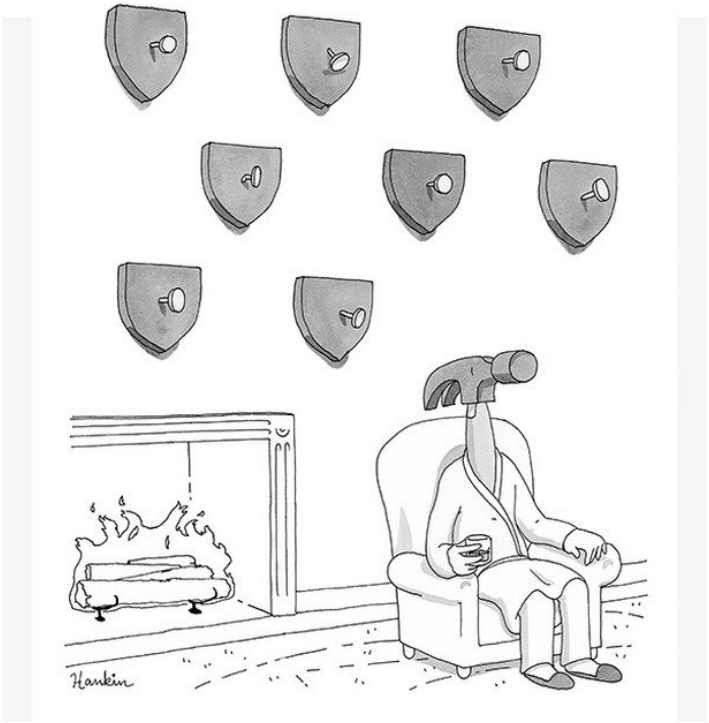
Assumption 3: Observers are important

We are rarely directly interested in the data, but rather in how observers react to their presence.

Consequently, we should focus more on the properties of the observers than on the properties of the data.

Assumption 4: There is no structure in the data

Generally speaking, data lack inherent structure. Instead, the structure of data reflects the observer's own structure.



The shape is not in the data but in the eyes of the observer.

How can we translate these ideas into mathematics?

?

?

Perception spaces and GENEOs

PERCEPTION SPACE

GROUP EQUIVARIANT NON-EXPANSIVE OPERATOR (GENEO)

Our epistemological assumptions

Some information about GENEOs

Finding pockets in proteins by applying GENEOs

Let's start by defining perception spaces

We recall that a pseudo-metric is just a metric d without the property $d(x_1, x_2) = 0 \implies x_1 = x_2$.

Definition

Let us consider

1. A nonempty set Φ endowed with a pseudo-metric D_Φ.
2. Let us denote by the symbol $*$ the left action of a group $(G, \circ)$ on Φ, and endow G with the pseudo-metric D_G defined by setting $D_G(g_1, g_2) := \sup_{\varphi \in \Phi} D_\Phi(g_1 * \varphi, g_2 * \varphi)$ for any $g_1, g_2 \in G$. We will also assume that the action of the group G on the metric space (Φ, D_Φ) is isometric, i.e., for every $\varphi_1, \varphi_2 \in \Phi$ and every $g \in G$, $D_\Phi(g * \varphi_1, g * \varphi_2) = D_\Phi(\varphi_1, \varphi_2)$.

We say that (Φ, G) is a perception space.

Perception spaces

The set Φ represents the data we may get from our measuring tools (functions, graphs, cloud of points,...). The group G represents the possible invariances of data the observer may be interested in.
For example, Φ can be a set of grey-level images represented as functions from $\mathbb{R}^2$ to $[0,1]$, while G can be the group of isometries of the real plane.
Another simple example can be given by the set of electrocardiograms represented as functions of the time variable, while G can be the group of time translations.
In any case, the following statement holds.

Proposition

$(G,\circ)$ is a topological group and the action of G on Φ is continuous.

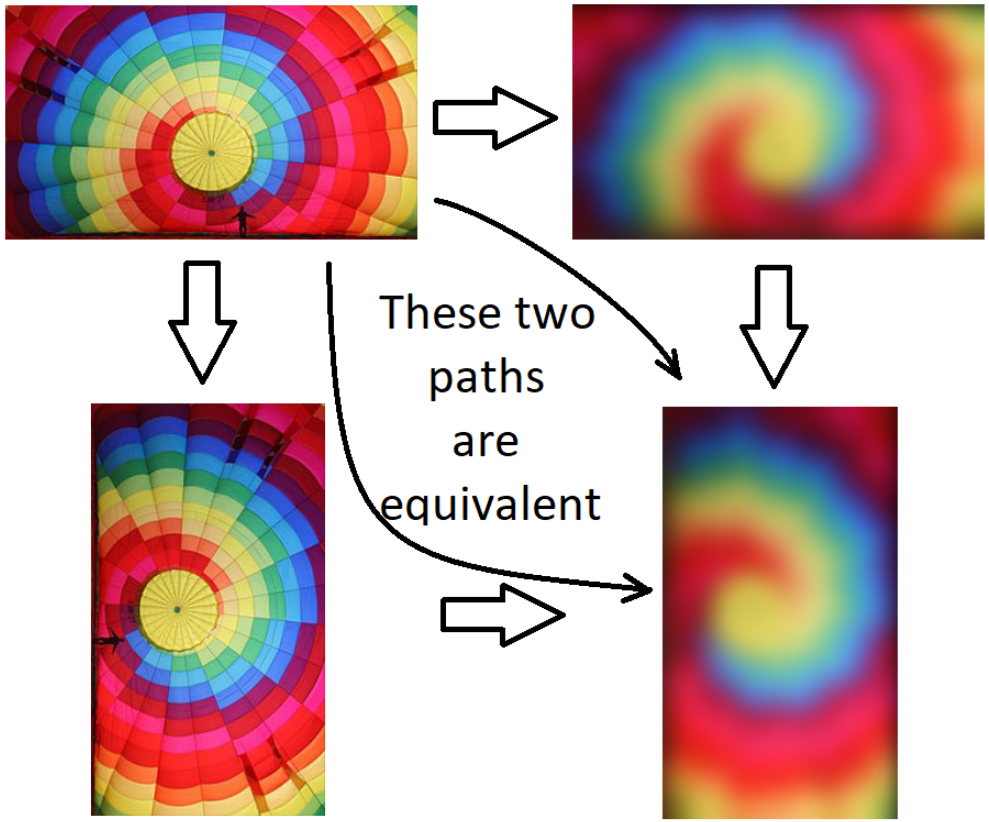
GEOs and GENEOs

Definition

- Let (Φ, G), (Ψ, K) be two perception spaces. If a map $F : \Phi \to \Psi$ and a group homomorphism $T : G \to K$ are given, such that $F(g * \varphi) = T(g) * F(\varphi)$ for every $\varphi \in \Phi$, $g \in G$, we say that (F, T) is an (extended) *group equivariant operator* (GEO).
- If (F, T) is non-expansive (i.e. $D_\Psi(F(\varphi_1), F(\varphi_2)) \leq D_\Phi(\varphi_1, \varphi_2)$ for every $\varphi_1, \varphi_2 \in \Phi$, and $D_K(T(g_1), T(g_2)) \leq D_G(g_1, g_2)$ for every $g_1, g_2 \in G$), we say that (F, T) is an (extended) *group equivariant non-expansive operator* (GENEO).

An example of GENEO

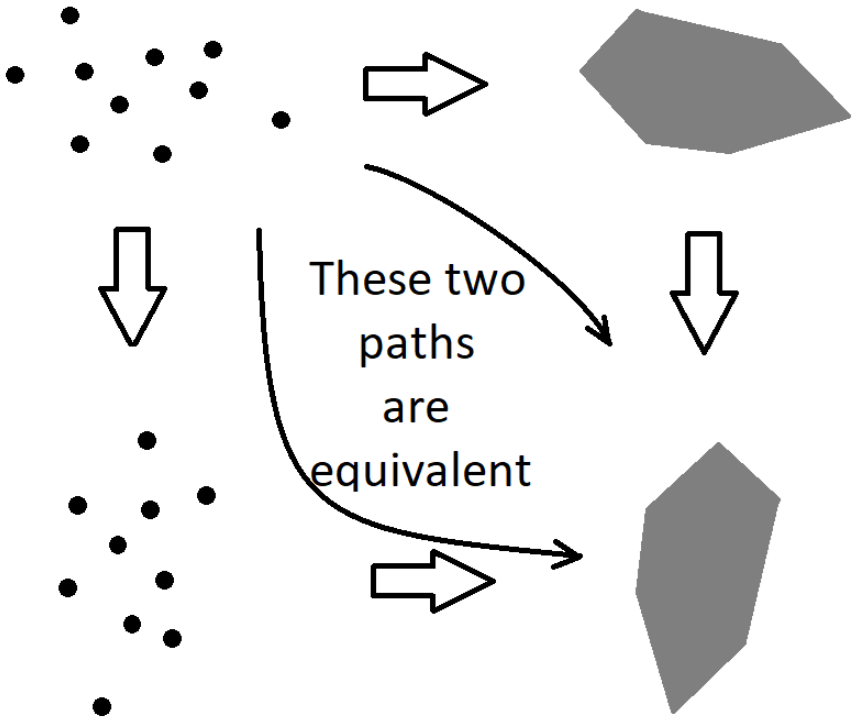
When we blur an image by applying a **convolution** with a rotationally symmetric kernel whose mass is less than 1 in L^1, we are applying a GENEO (here, we are considering the **group of isometries**).



Here, the maximum distance between functions is used.

Another example of GENEO

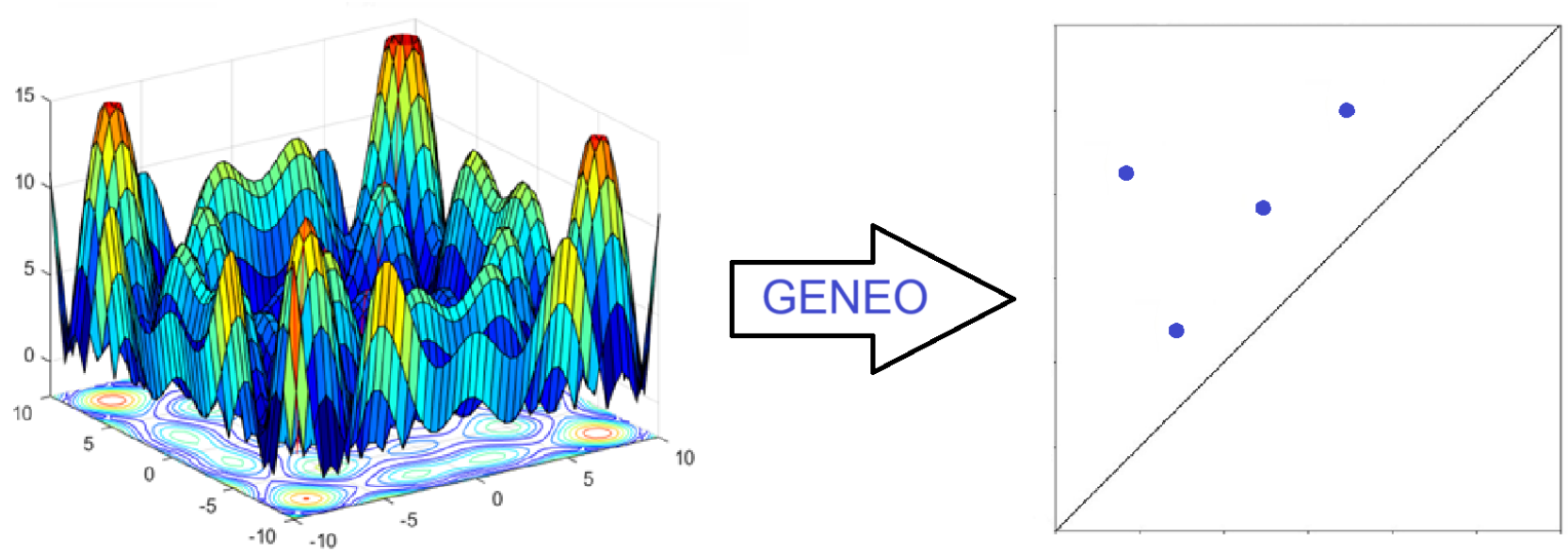
When we compute the **convex hull** of a cloud of points, we are applying a GENEO (here, we are considering the **group of isometries**).



Here, the Hausdorff distance between compact sets is used.

Another example of GENEO

For those familiar with topological data analysis and persistence diagrams, the operator that maps filtering functions to persistence diagrams constitutes another example of a GENEO.



Why are GENEOs interesting?

- GENEOs are based on a precise topological/geometric theory (guaranteeing compactness and convexity properties, representability by permutant measures, some relevant links with TDA, and much more).
- GENEOs allow us to represent the information we know about the chosen observer.
- GENEOs' non-expansiveness property is a strong constraint, allowing for relevant data simplification.
- GENEOs allow for a compositional approach to deep learning.
- Studying the shape of the observer space (representable by GENEOs) is often more important than studying the shape of the data space.

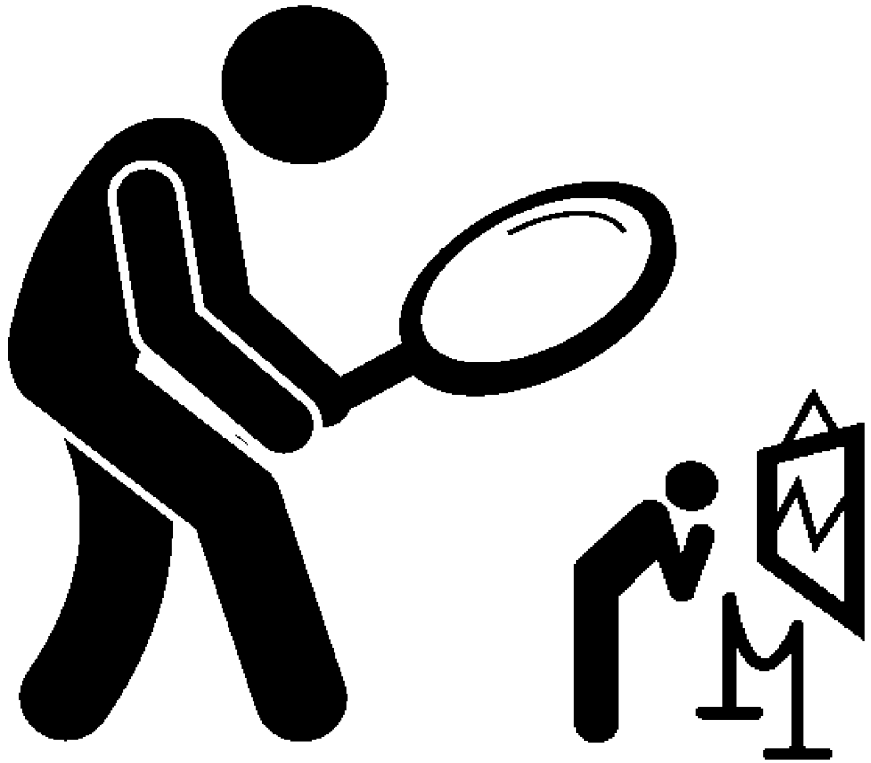
Two key observations (1)

Our main goal is to build a robust geometric and compositional theory for approximating an ideal observer through GENEOs and GEOs.

$$\approx \quad \begin{array}{c} (\Psi, K) \\ \uparrow \\ (F, T) \\ | \\ (\Phi, G) \end{array}$$

Two key observations (2)

GENEOs can be taken as inputs of higher-level GENEOs. Data obtained through measuring instruments can be seen as GENEOs of level 0. Therefore, hierarchies of GENEOs can be considered.



Construction of GENEOs

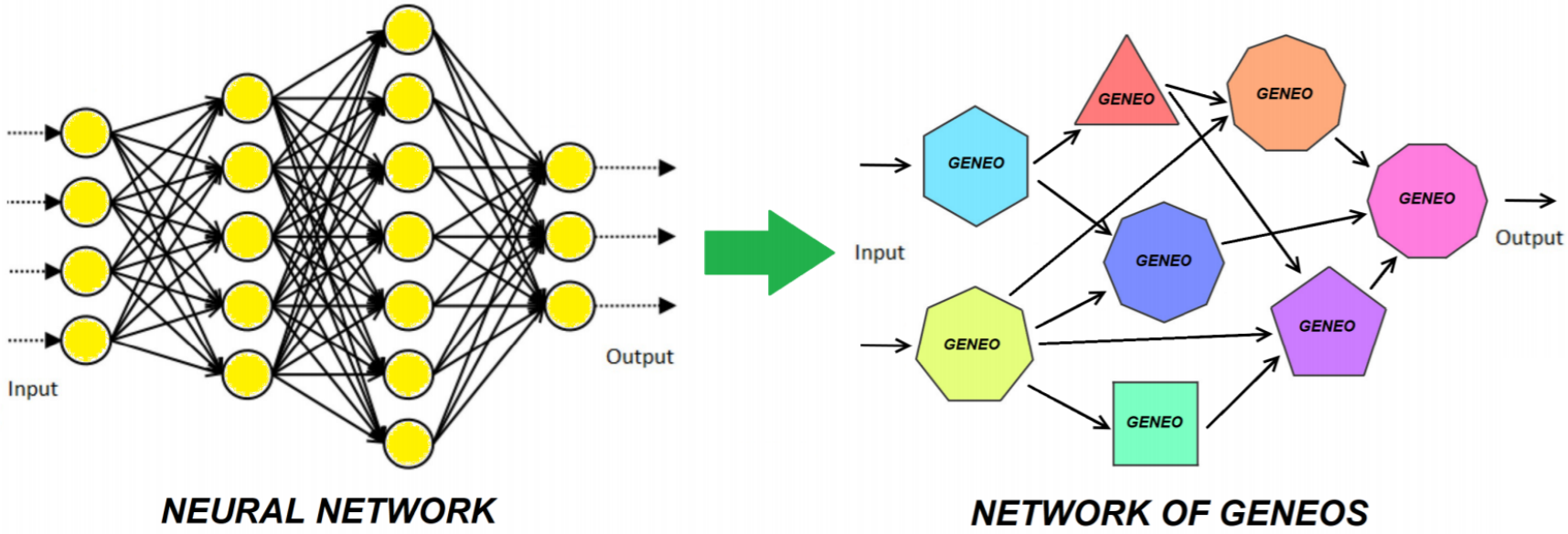
How can we build GENEOs?

When data are represented by real-valued functions, the space of GENEOs is closed under composition, computation of minimum and maximum, translation, direct product, and convex combination. (However there is much more than this...)

GENEOs are like LEGO bricks that can be combined together to form more complex GENEOs.

The main point in the approach based on GENEOs

In perspective, we are looking for a good compositional theory for building efficient and transparent networks of GENEOs.
Some preliminary experiments suggest that replacing neurons with GENEOs could make deep learning more transparent and interpretable and speed up the learning process.



A reference for the general setting

- Jacopo Joy Colombini, Filippo Bonchi, Francesco Giannini, Fosca Giannotti, Roberto Pellungrini and Patrizio Frosini,

  Mathematical Foundation of Interpretable Equivariant Surrogate Models,

  Proceedings of the World Conference on Explainable Artificial Intelligence (XAI-2025) (to appear), Novel Post-hoc & Ante-hoc XAI Approaches, 09-11 July, 2025 - Istanbul, Turkey.

An interesting case

An important case arises when the elements of the set Φ are functions and the group G consists of homeomorphisms of $\mathbb{R}^n$. In this setting, the action of G on Φ is defined via right composition.

Many types of data can be represented as functions:

Images, electrocardiograms, computerized tomography scans, and more.

Additionally:

- A point cloud C in $\mathbb{R}^n$ (where C is equivalent to the function $d_C : \mathbb{R}^n \to \mathbb{R}$ that expresses the distance from C).
- A graph Γ (where Γ is equivalent to its adjacency matrix, which can be interpreted as a function).

GENEOs and Machine Learning

If you are interested, you can find more details about the theory of GENEOs for functions in these papers:

- M. G. Bergomi, P. Frosini, D. Giorgi, N. Quercioli,

  Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning,

  Nature Machine Intelligence, vol. 1(9) (2019), 423–433.

  https://www.nature.com/articles/s42256-019-0087-3
- G. Bocchi, P. Frosini, M. Ferri,

  A novel approach to graph distinction through GENEOs and permutants,

  Scientific Reports, 15 (2025), 6259.

  https://www.nature.com/articles/s41598-025-90152-7

GENEOs and Machine Learning

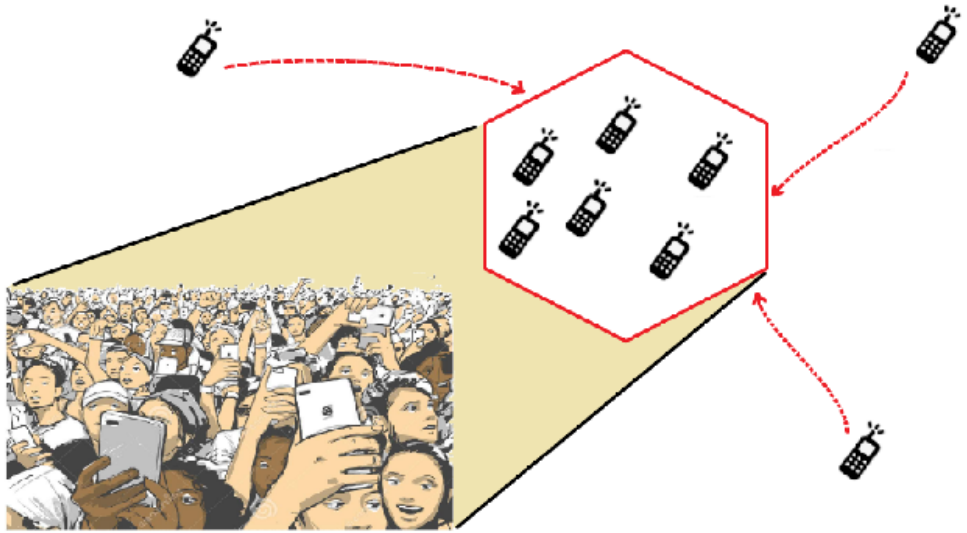
For more details about the use of GENEOs in Machine Learning, you can have a look at this paper:

https://ems.press/journals/mag/articles/10389352

Research projects (I)

CNIT / WiLab - Huawei Joint Innovation Center (JIC)

Project on GENEOs for 6G



WILAB HUAWEI

Research projects (II)

PANDORA

Horizon Europe (HORIZON)
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Project: 101135775 — PANDORA
Funding: approximately 9 million euros.

Task 3.3 - Leveraging domain knowledge for explainable learning:
This task aims to investigate the use of domain knowledge in the development of explainable AI models. Tools like GENEOs for applications in TDA and ML and new theoretical methods of GENEOs for explainable AI will be used.

The project has received funding form the European Union's Horizon Europe Framework Programme (Horizon) under grant agreement No 101135775

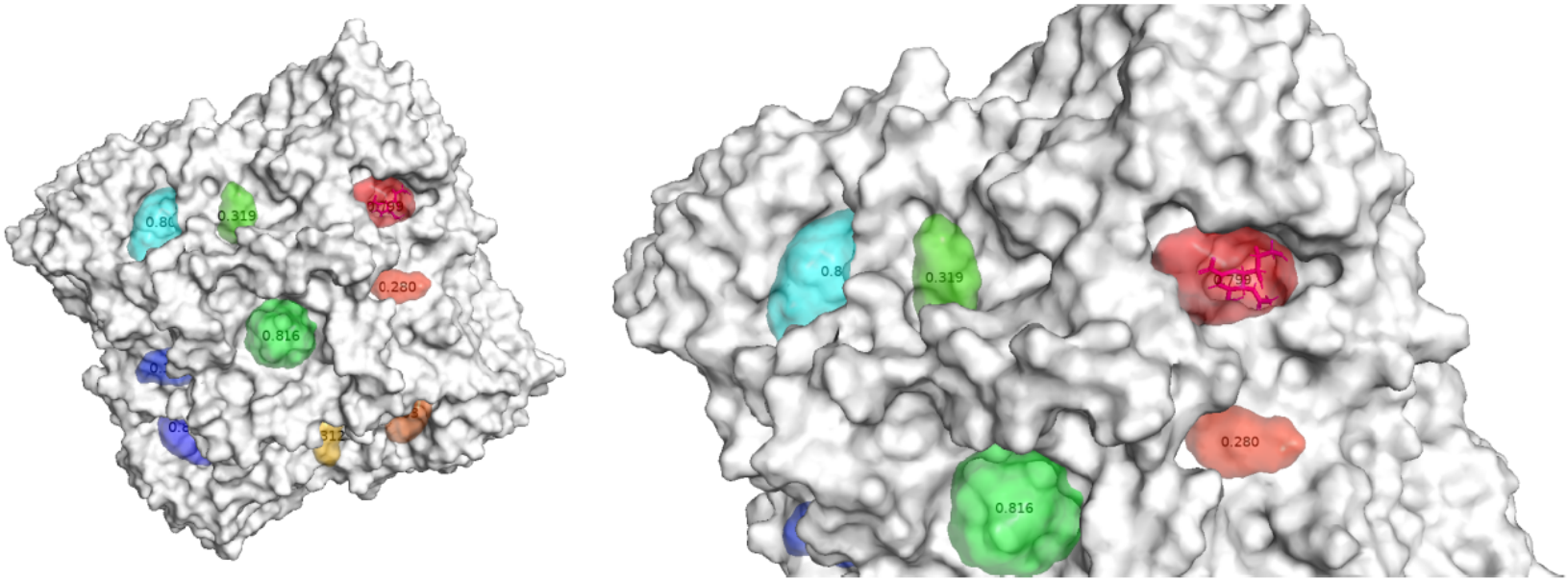
Our epistemological assumptions

Some information about GENEOs

Finding pockets in proteins by applying GENEOs

Finding pockets in proteins by applying GENEOs

GENEOs can be used for the detection of druggable protein pockets.



a) b)

Model predictions for protein 2QWE. In Figure a) the global view of the prediction is shown, where different pockets are depicted in different colors and are labelled with their scores. In Figure b) the zoomed of the pocket containing the ligand is shown.

Our data

Data sources: the PDBbind v.2020 database (Liu et al., 2017) and the RCSB PDB (Berman et al., 2003).

The protein structures were preprocessed using the Schrödinger Protein Preparation Wizard (Schrödinger, The Schrödinger Software. 2020). A total of 12295 protein-ligand complexes from PDBbind and 41519 from the RCSB PDB were retrieved.

The data from PDBbind were used to train a set of models, to select the best model in terms of scoring, and compare it with other methods.

Data preparation

We represent the protein-ligand complex stored in a PDB file as functions in a compact and convex space.

The space around the protein is discretized using a parallelepiped grid of cubic voxels. For each voxel, piecewise constant approximations of 8 input functions, or channels φ_i, are computed.

The functions φ_i describe a set of geometrical, physical, and chemical protein properties that are considered to be relevant for pocket detection by experts.

The co-crystallized ligand of a protein will be used in the evaluation step to define the true pocket (i.e. the ground truth) for the parameters identification.

The functions φ_i

Name	Type	Expression	Notes
Distance	Geometrical	$\varphi_1(x) = d(x, x_{a^*}) - r_{a^*}$	x_{a^*} and r_{a^*} are coordinates and radius of nearest atom to the point x.
Gravitational	Geometrical	$\varphi_2(x) = \sum_{a \in A} \frac{m(a)}{d(x,x_a)}$	$m(a)$ is the mass of atom a.
Electrostatic	Physical	$\varphi_3(x) = \sum_{a \in A} \frac{q(a)}{d(x,x_a)}$	$q(a)$ is the partial charge of atom a.
Lipophilic	Chemical	$\varphi_4(x) = \sum_{a \in A} \frac{l(a)}{d(x,x_a)}$	$l(a)$ is the lipophilic coefficient of the atom a if it is negative, 0 otherwise.
Hydrophilic	Chemical	$\varphi_5(x) = \sum_{a \in A} \frac{h(a)}{d(x,x_a)}$	$h(a)$ is the lipophilic coefficient of the atom a if it is positive, 0 otherwise.
Polar	Chemical	$\varphi_6(x) = \sum_{a \in A} \frac{p(a)}{d(x,x_a)}$	$p(a)$ is 1 if atom is polar, 0 otherwise.
HB Acceptor	Chemical	$\varphi_7(x) = \sum_{a \in A} -\varepsilon_a(R^6 - 2R^4)$	$R = R_{min}/d(x, x_a) + 0.96$ where ε_a and R_{min} are parameters of the specific type of atom.
HB Donor	Chemical	$\varphi_8(x) = \sum_{a \in A} -\varepsilon_a(R^6 - 2R^4) \cdot \cos^2 \phi_1 \cos^2 \phi_2$	$R = R_{min}/d(x, x_a) + 0.96$ where ε_a and R_{min} are parameters of the specific type of atom, ϕ_1 and ϕ_2 are angles defined by triples of points involved in the bond.

GENEOnet

The channels φ_i describing the protein are fed to a layer of 8 GENEOs, $F_1, \ldots, F_8$.

Each F_i is a convolutional operator defined by setting $F_i(\varphi_i) = \varphi_i * K_i$, where K_i is a normalized kernel in $L^1(\mathbb{R}^3)$, symmetric with respect to the origin. This fact ensures that all the operators under consideration are indeed non-expansive and equivariant with respect to rigid motions in $\mathbb{R}^3$. Every operator F_i is associated with a shape parameter $\sigma_i \in \mathbb{R}$ regulating the "amplitude" of the kernel K_i.

The set $\{F_1, \ldots, F_8\}$ reflects the experts' prior knowledge on the relevant properties to identify a pocket.

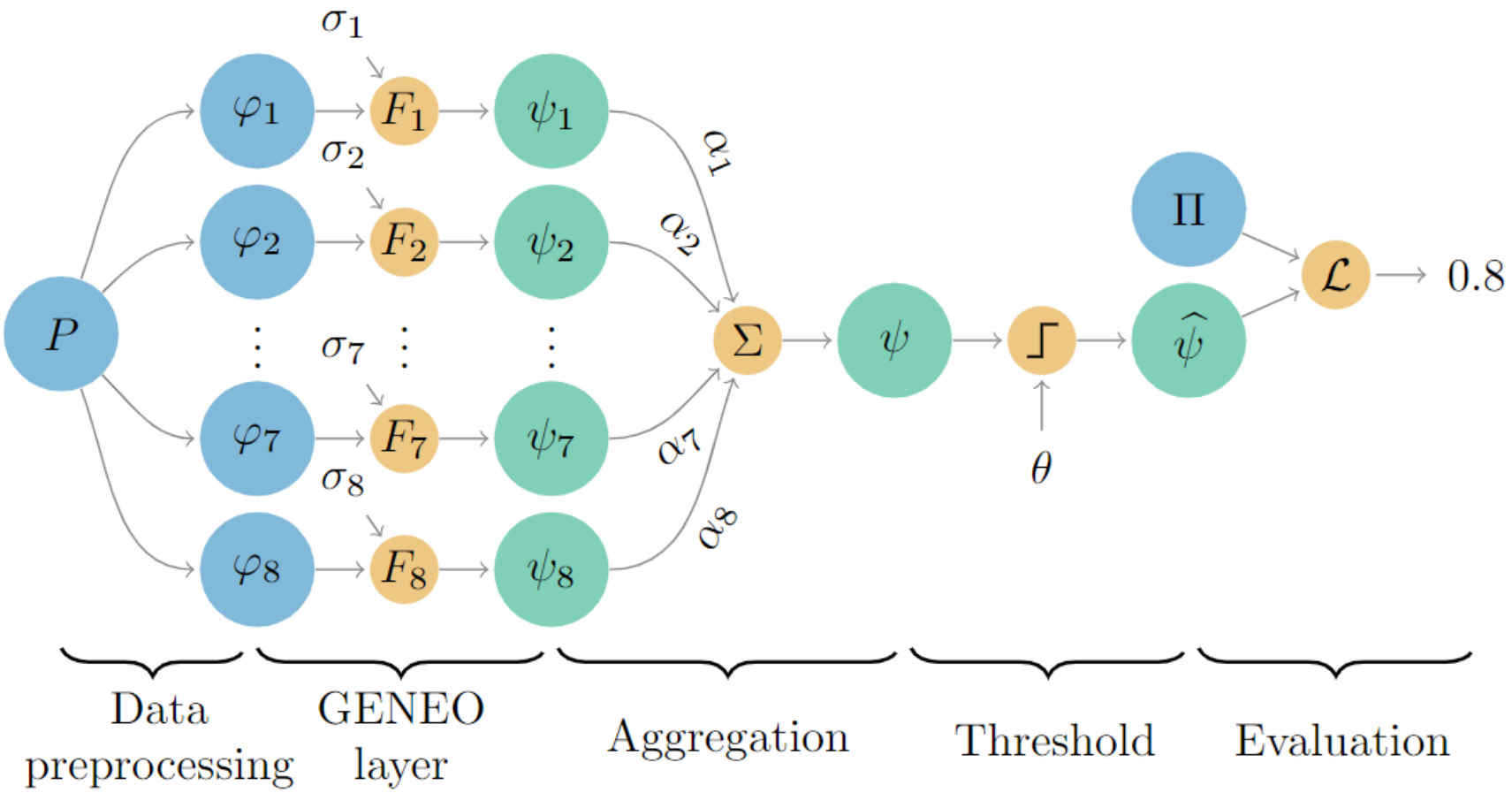
Convex combinations in GENEOnet

The GENEO outputs $\psi_i = F_i(\varphi_i)$ are combined through a convex combination, with weights $\alpha_1, \ldots, \alpha_8$ in order to obtain a composite operator $F\Big((\varphi_1, \ldots, \varphi_8)\Big) = \sum_{i=1}^{8} \alpha_i F_i(\varphi_i)$, which is a new GENEO. The output of the convex combination is then normalized to obtain a function ψ from $\mathbb{R}^3$ to $[0,1]$.

The function ψ can be interpreted as the probability that a voxel belongs to a pocket. The coefficients $\alpha_1, \ldots, \alpha_8$ can be regarded as weights, highlighting the importance of each channel in the pockets identification.

To obtain pockets, a thresholding operation with a parameter θ is applied to ψ, producing the binary function $\hat{\psi}$, which finally can be compared to the ground truth through a volumetric accuracy function that will be described later.

GENEOnet structure



Training

The data retrieved from the PDBbind were firstly used to train GENEOnet on the spatial recognition of the true pocket. In order to identify the unknown parameters, we choose to optimize an accuracy function evaluating the quality of the prediction.

For each crystallized complex, the ligand has been converted to the binary function τ that is equal to 1 in the voxels (possibly partially) overlapped to the ligand, and equal to 0 elsewhere.

The function τ represents our ground truth.

Training

We train GENEOnet by maximizing the following accuracy function with respect to our parameters:

$$\ell(\widehat{\psi}, \tau) = \frac{|\widehat{\psi} \wedge \tau| + \kappa|(\mathbf{1} - \widehat{\psi}) \wedge (\mathbf{1} - \tau)|}{|\tau| + \kappa|\mathbf{1} - \tau|} \in [0, 1].$$

Here $|\cdot|$ denotes the discretized volume, that is the number of voxels labelled with 1 inside the region, $\widehat{\psi} \wedge \tau$ is a function equal to 1 on the intersection between the predicted pockets $\widehat{\psi}$ and the true pocket τ, $\mathbf{1}$ is a constant function equal to 1. All these functions are defined on the voxelized bounding box built around the protein. They are binary and piecewise constant on each voxel. The hyperparameter κ ranges in $[0, 1]$ (we saw that the best values belong to $[0.01, 0.05]$).

Training

The optimization of $\ell(\widehat{\psi}, \tau)$ was performed using Adam optimizer.

A random set of 200 proteins from the PDBbind was used as a training set.

Training time for 50 epochs of the optimization algorithm is approximately 6 minutes with GPU acceleration (on a laptop equipped with an NVIDIA GeForce RTX 3060 GPU) and approximately 40 minutes with only CPU processing (on a laptop featuring an Intel® Core™ i7-10870H 8-core CPU).

Testing

After training, for each protein our trained network of GENEOs produces a set of predicted pockets, represented as connected components in the support of $\hat{\psi}$.

We order these predicted pockets according to a score obtained by computing the mean value of ψ on each pocket (the higher this value, the more reliable the predicted pocket according to GENEOnet).

This ordered list is the output of GENEOnet.

Comparing our results with the ground truth

Now we consider our ordered list and take the predicted pocket $\widehat{\Pi}_j$ that best overlaps the ground truth Π.

Method: We say that a predicted pocket $\widehat{\Pi}_j \subset \mathbb{R}^3$ *best overlaps* the true pocket $\Pi \subset \mathbb{R}^3$ if $\widehat{\Pi}_j$ maximizes the value $\frac{|\widehat{\Pi}_j \wedge \Pi|}{|\Pi|}$ in the set of predicted pockets. In this expression, $|\cdot|$ denotes the 3D discretized volume of a region, which corresponds to the number of voxels in that region.

We define H_j as the percentage of times that the best choice in the list is the j-th choice of GENEOnet.

NB: If no predicted pocket shares any intersection with the true one, we say that the method failed for that protein.

Evaluation parameters for testing

Moreover, by computing cumulative sums of the values H_j, we generate another sequence of coefficients $(T_j)_{j\geq 1}$ that represents the fraction of proteins whose true pocket has been successfully recognized *within* the j-th highest-scored predicted pocket, i.e.,

$$T_j = \sum_{i=1}^{j} H_i$$

Comparison of GENEOnet with other methods

GENEOnet has ben compared with the following state-of-the-art methods: Fpocket, P2Rank, DeepPocket, CAVIAR, SiteMap, CavVis.

These methods also evaluate the true pocket as the area outside the protein that contains the co-crystallized ligand.

Each method we consider orders the pockets it predicts, according to its scoring procedure. Therefore, we can define the values H_j and T_j for all those methods.

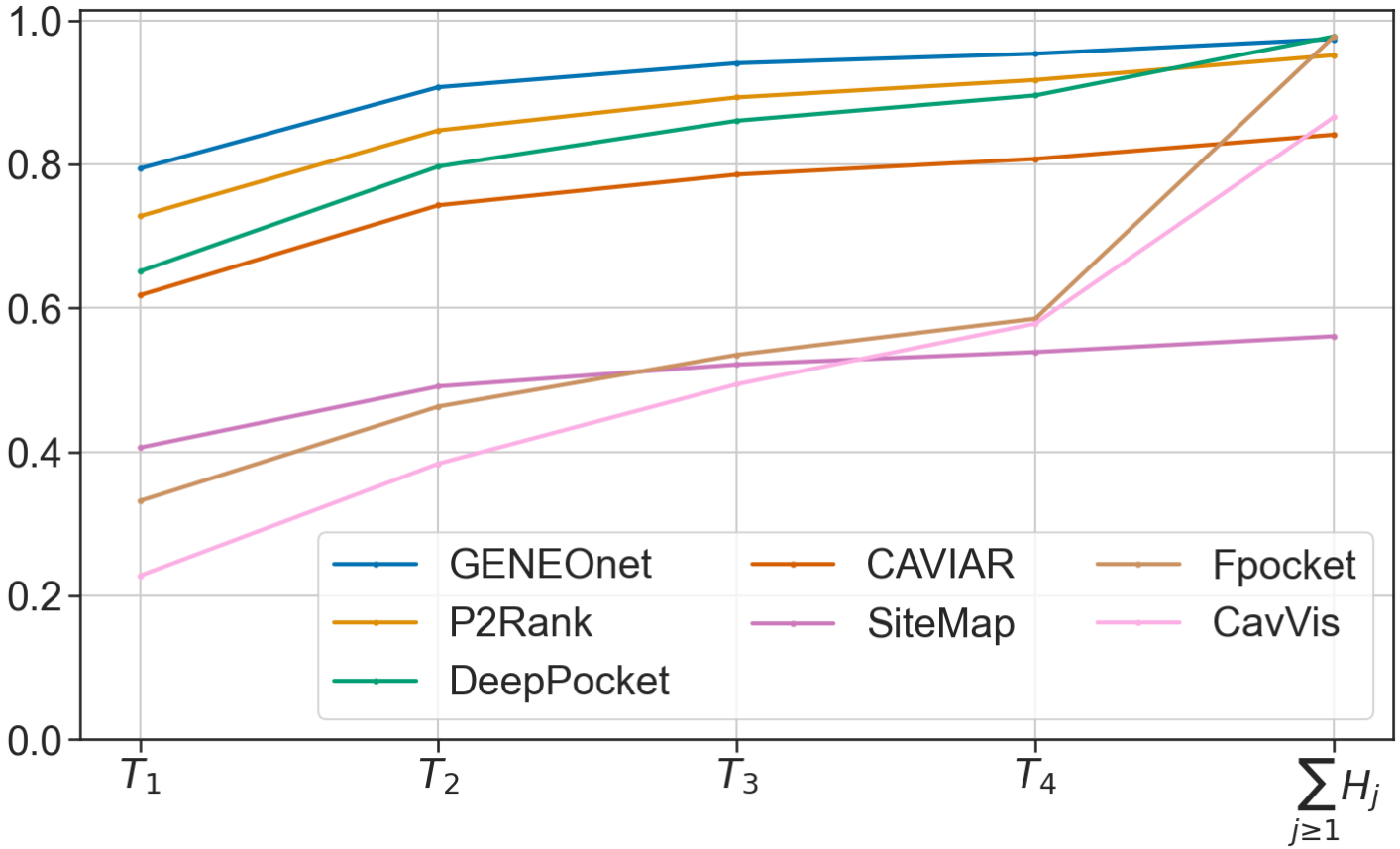
In the following table, we report estimates of H_j and T_j coefficients computed on a test set made of 9070 proteins from the PDBbind (this is about the 75% of the entire data extracted from the PDBbind).

Comparison of GENEOnet with other methods

Method	T_1	T_2	T_3	T_4	$\sum_{j\geq 1} T_j$
GENEOnet	**0.794**	**0.907**	**0.940**	**0.953**	0.974
P2Rank	0.728	0.847	0.893	0.917	0.952
DeepPocket	0.651	0.797	0.861	0.896	**0.978**
CAVIAR	0.618	0.743	0.786	0.808	0.841
SiteMap	0.406	0.491	0.521	0.538	0.561
Fpocket	0.332	0.463	0.535	0.585	**0.978**
CavVis	0.228	0.384	0.494	0.578	0.865

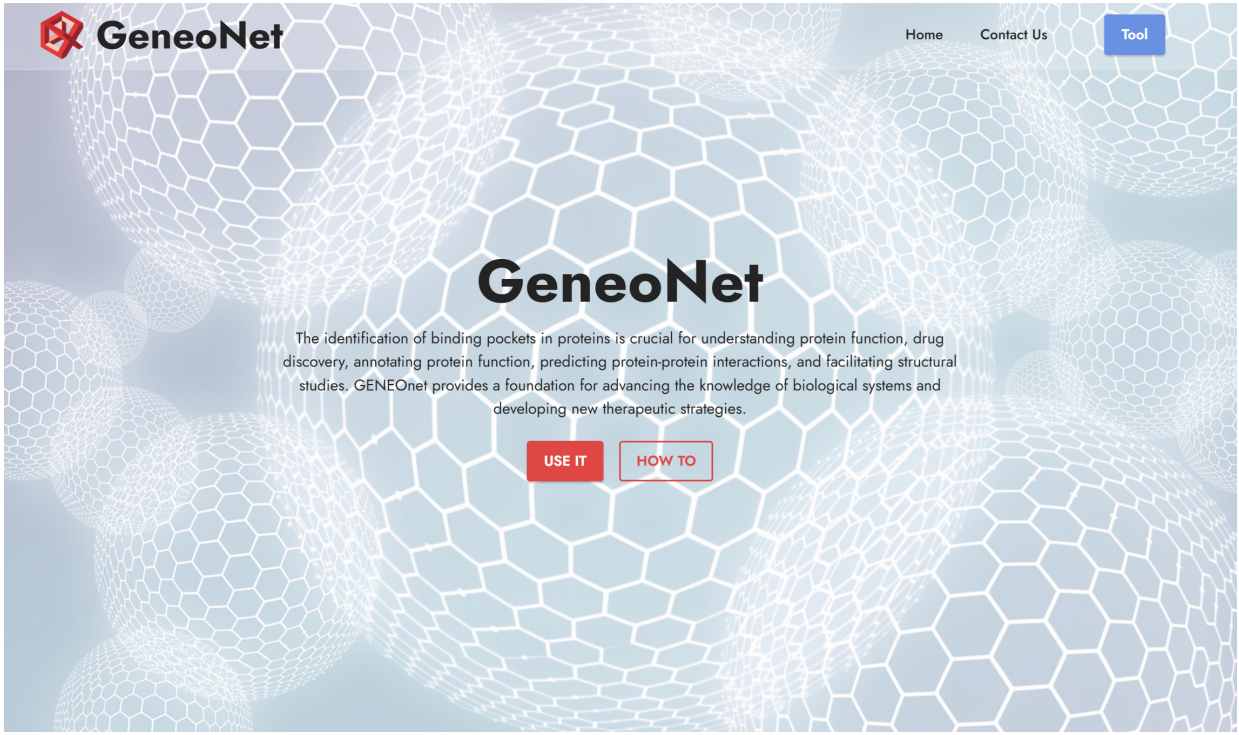
Table 1: **T_j values for the compared methods computed on BIND_TEST.** Bold numbers are used to highlight, for every coefficient, the method(s) that reached the best result.

Results



Please note that GENEOnet uses only 17 parameters, whereas a CNN like DeepPocket requires 665,122 parameters.

GeneoNet webservice

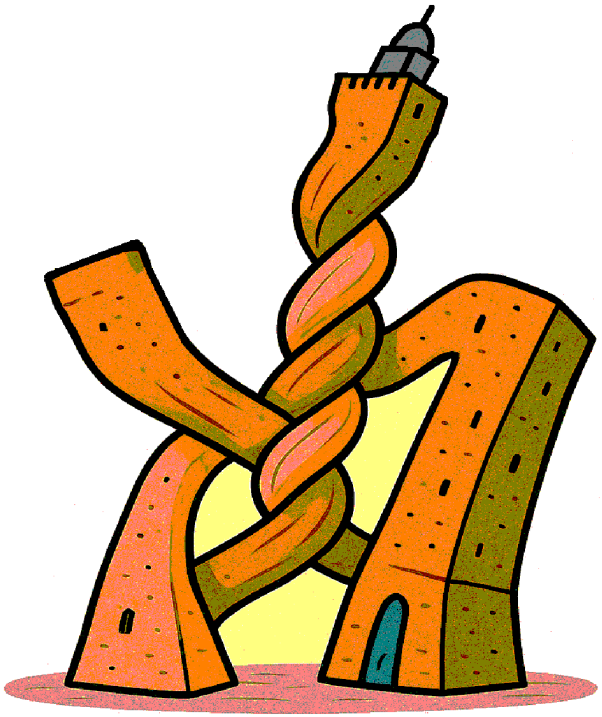


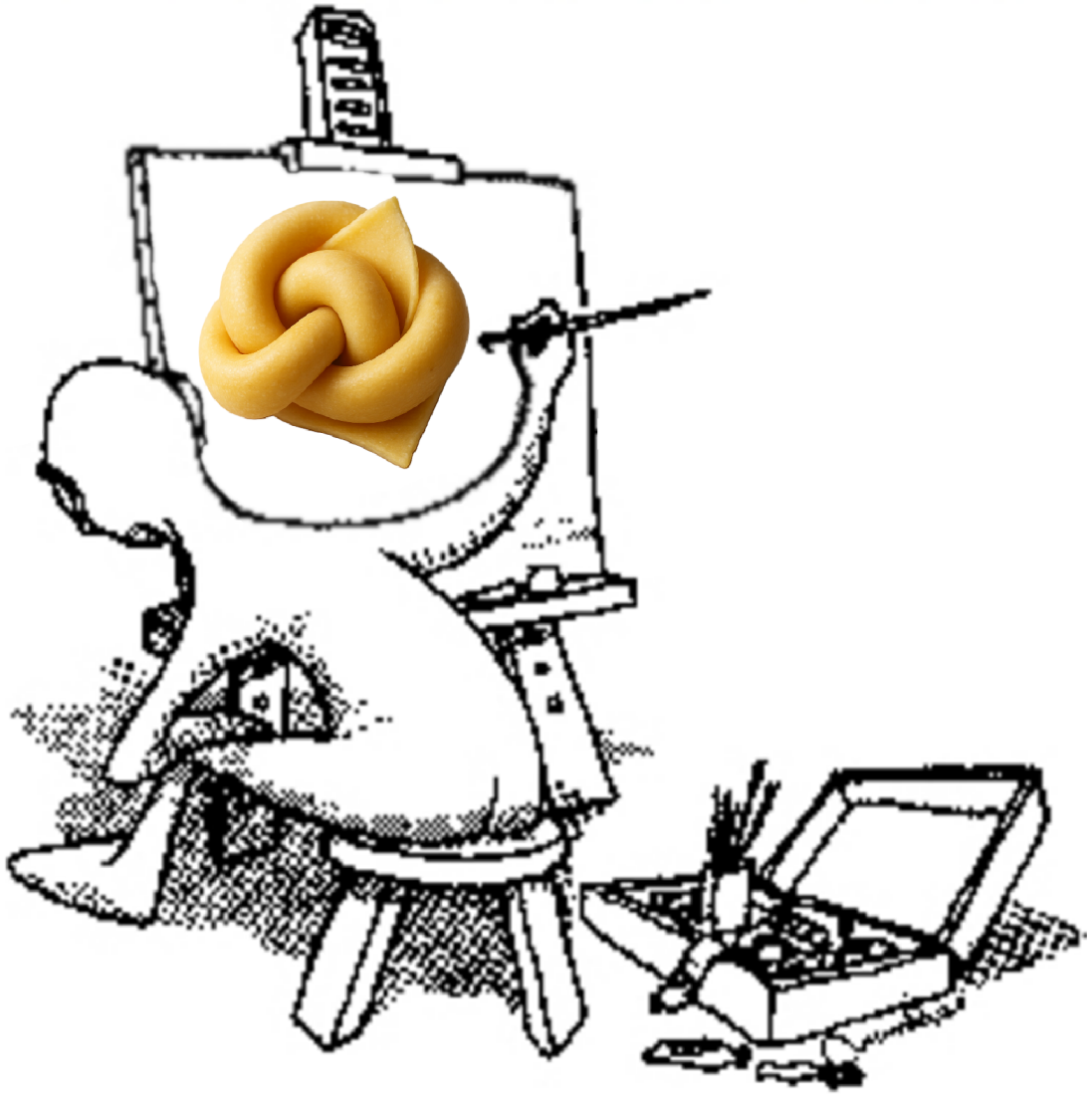
The GeneoNet webservice represents the outcome of our partnership with Italian Pharmaceutical Company Dompé Farmaceutici S.p.A.: https://geneonet.exscalate.eu/

GeneoNet webservice

More information about GeneoNet is available in this paper:

G. Bocchi, P. Frosini, A. Micheletti, A. Pedretti, G. Palermo, D. Gadioli, C. Gratteri, F. Lunghini, A. R. Beccari, A. Fava, C. Talarico, *A geometric XAI approach to protein pocket detection*, The 2nd World Conference on eXplainable Artificial Intelligence, Valletta, Malta, July 17-19, 2024, , vol. 3793, 217-224 (2024).
https://ceur-ws.org/Vol-3793/paper_28.pdf





THANKS FOR YOUR ATTENTION