On the use of group equivariant non-expansive operators for protein pocket detection

Patrizio Frosini

Department of Mathematics, University of Bologna
patrizio.frosini@unibo.it

IMS-NTU joint workshop on Biomolecular Topology: Modelling and Data
June 27, 2024

Outline

Some basics on the theory of GENEOs

Some links between GENEOs and TDA

Finding pockets in proteins by applying GENEOs

Some basics on the theory of GENEOs

Some links between GENEOs and TDA

Finding pockets in proteins by applying GENEOs

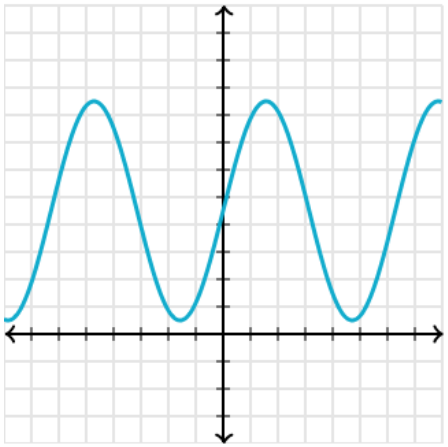
Assumption 1: Data are often represented by functions

Many kinds of data can be represented as functions:

Images, electrocardiograms, computerized tomography scans...

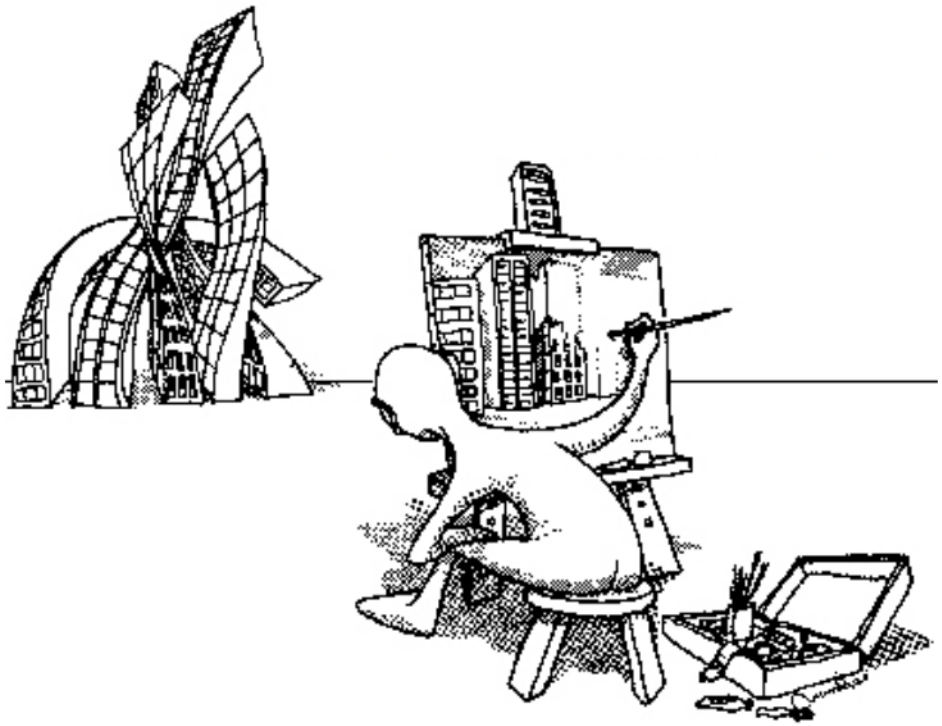
But also:

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



Assumption 2: Data are processed by observers

Data have no meaning if no observer elaborates them.



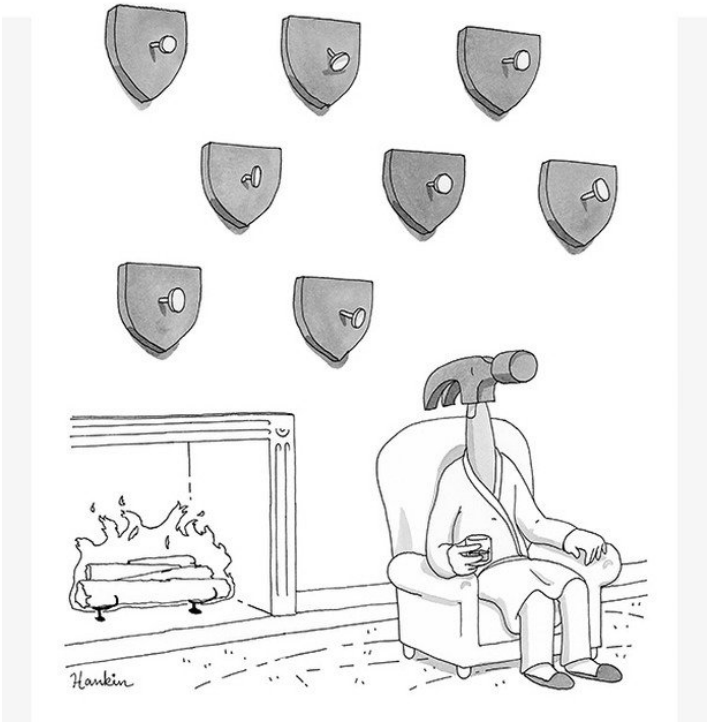
An observer is an agent that transforms data while respecting their symmetries.

Asm. 3: Observers are often more relevant than data

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

Asm. 4: No data structure

Generally speaking, there is no structure in data. The structure of data is a projection of the structure of the observer.



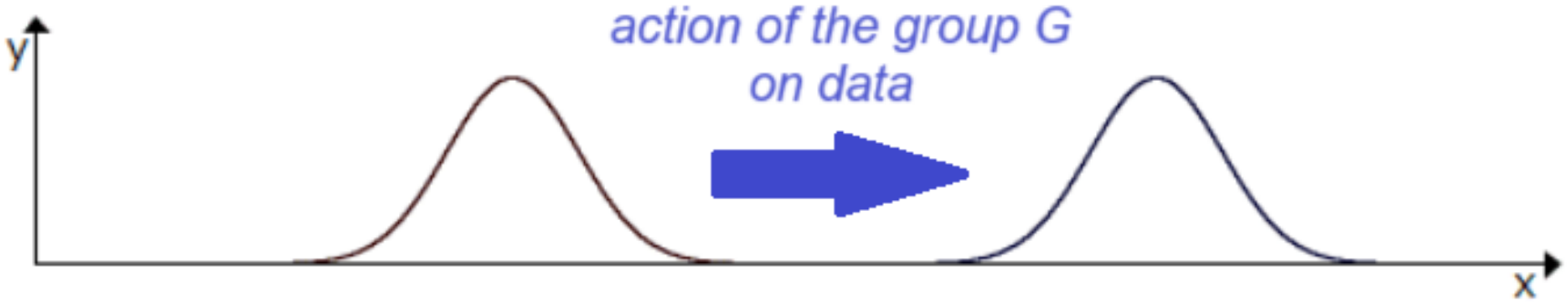
Let's start by defining perception pairs

Let us consider

1. A collection Φ of functions from a set X to $\mathbb{R}$;
2. A group G of bijections $g : X \to X$ such that $\varphi \in \Phi \implies \varphi \circ g \in \Phi$ for every $\varphi \in \Phi$.

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

The choice of a perception pair states which data can be considered as legitimate measurements (the functions in Φ) and which group represents the admissible symmetries between data (the group G).



An important remark

Another important assumption in our model is that any topological or metric structure we employ must be grounded in the functions we recognize as valid measurements. No reference to mathematical structures is allowed unless they are represented through these admissible measurements. Therefore, in our model, adding structure requires the expansion of the set of admissible functions.

What metric can we consider on Φ, X and G?

We endow Φ with the sup-norm metric:

$$D_\Phi(\varphi_1, \varphi_2) = \sup_{x \in X} |\varphi_1(x) - \varphi_2(x)|.$$

NB: What other metric could we put on Φ, given that X is not endowed with any measure or structure?

Then, we endow X with the pseudo-metric

$$D_X(x_1, x_2) = \sup_{\varphi \in \Phi} |\varphi(x_1) - \varphi(x_2)|.$$

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

Finally, we put on G the pseudo-metric

$$D_G(g_1, g_2) := \sup_{\varphi \in \Phi} D_\Phi(\varphi \circ g_1, \varphi \circ g_2).$$

Some useful propositions

Proposition

Every function in Φ is non-expansive and hence continuous.

Proposition

If Φ is compact and X is complete, then X is compact.

Proposition

G is a topological group for the topology induced by D_G, and the action of G on Φ by composition on the right is continuous.

Proposition

If Φ is compact and G is complete, then G is compact.

An interesting remark

If Φ is totally bounded, we can complete Φ, X, and G.

In other words, because of the previous results, if Φ is totally bounded, we can assume that Φ, X, and G are compact.

- F. Ahmad, *Compactification of perception pairs and spaces of group equivariant non-expansive operators*, `https://arxiv.org/pdf/2210.04043`

Some magic happens: each bijection is an isometry

- $\mathrm{Bij}_\Phi(X) = \{\textbf{bijections } g : X \to X \text{ s.t. } \Phi \circ g, \Phi \circ g^{-1} \subseteq \Phi\}$;
- $\mathrm{Homeo}_\Phi(X) = \{\textbf{homeomorphisms } g : X \to X \text{ s.t. } \Phi \circ g, \Phi \circ g^{-1} \subseteq \Phi\}$;
- $\mathrm{Iso}_\Phi(X) = \{\textbf{isometries } g : X \to X \text{ s.t. } \Phi \circ g, \Phi \circ g^{-1} \subseteq \Phi\}$.

Proposition

$\mathrm{Bij}_\Phi(X) = \mathrm{Homeo}_\Phi(X) = \mathrm{Iso}_\Phi(X)$.

GEOs and GENEOs

Let us assume that two perception pairs (Φ, G), (Ψ, K) are given, and fix a group homomorphism $T : G \to K$.

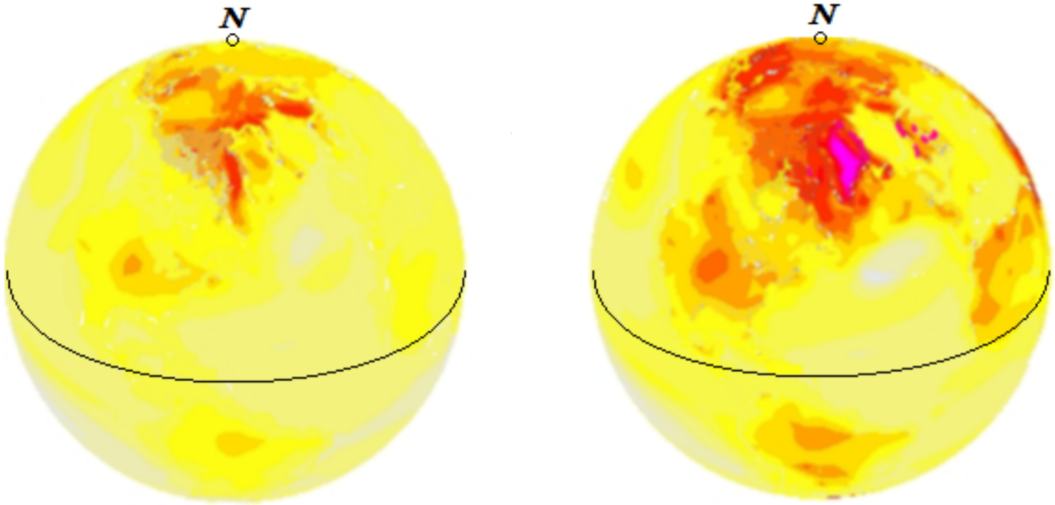
Each function $F : \Phi \to \Psi$ such that $F(\varphi \circ g) = F(\varphi) \circ T(g)$ for every $\varphi \in \Phi, g \in G$ is called a *Group Equivariant Operator (GEO)* associated with the homomorphism T.

If F is also non-expansive (i.e., $D_\Psi(F(\varphi_1), F(\varphi_2)) \leq D_\Phi(\varphi_1, \varphi_2)$ for every $\varphi_1, \varphi_2 \in \Phi$), then F is called a *Group Equivariant Non-Expansive Operator (GENEO)* associated with the homomorphism T.

GENEOs represent observers in our setting.

An example of GENEO

Let us assume to be interested in the comparison of the distributions of temperatures on a sphere, taken at two different times:



Let us also assume that only two opposite points N, S can be localized on the sphere.

An example of GENEO

Let us introduce two perception pairs $(\Phi, G), (\Psi, K)$ by setting

- $X = S^2$
- $\Phi =$ set of 1-Lipschitz functions from S^2 to a fixed interval $[a, b]$
- $G =$ group of rotations of S^2 around the axis $N - S$

and

- $Y =$ the equator S^1 of S^2
- $\Psi =$ set of 1-Lipschitz functions from S^1 to $[a, b]$
- $K =$ group of rotations of S^1

An example of GENEO

This is a simple example of GENEO from (Φ, G) to (Ψ, K):

- $T(g)$ is the rotation $h \in K$ of the equator S^1 that is induced by the rotation g of S^2, for every $g \in G$.
- $F(\varphi)$ is the function ψ that takes each point y belonging to the equator S^1 to the average of the temperatures along the meridian containing y, for every $\varphi \in \Phi$;

We can easily check that F verifies the properties defining the concept of group equivariant non-expansive operator with respect to the isomorphism $T : G \to K$.

In plain words, our GENEO simplifies the data by transforming "temperature distributions on the earth" into "temperature distributions on the equator".

Two key results (and two good news for applications)

Let us assume that a homomorphism $T : G \to K$ has been fixed.
Let us define a metric D_{GENEO} on $\mathrm{GENEO}((\Phi, G), (\Psi, K))$ (the space of all GENEOs from (Φ, G) to (Ψ, K) w.r.t. $T : G \to K$) by setting

$$D_{\mathrm{GENEO}}(F_1, F_2) := \sup_{\varphi \in \Phi} D_{\Psi}(F_1(\varphi), F_2(\varphi)).$$

Theorem

If Φ and Ψ are **compact**, *then $\mathrm{GENEO}((\Phi, G), (\Psi, K))$ is* **compact** *with respect to D_{GENEO}.*

Theorem

If Ψ is **convex**, *then $\mathrm{GENEO}((\Phi, G), (\Psi, K))$ is* **convex**.

Two key observations (1)

- While the space of data is often non-convex (and hence averaging data does not make sense), the assumption of convexity of Ψ implies the convexity of the space of observers and allows us to consider the "average of observers".

Two key observations (2)

- Our main goal is to develop a good geometric and compositional theory to approximate an ideal observer. In our model, "to approximate an observer" means to look for a GENEO F that minimizes a suitable "cost function" $c(F)$. The cost function quantifies the error that is committed by taking the GENEO F instead of the ideal observer. Since the space of GENEOs is compact and convex (under the assumption that the data spaces are compact and convex), if the cost function $c(F)$ is strictly convex we have that there is one and only one GENEO that best approximates the ideal observer.

Construction of GENEO with elementary methods

How can we build GENEOs?

Without going into technical details, we only observe here that, under reasonable assumptions,

- the composition of GENEOs is still a GENEO;
- the maximum and minimum of GENEOs is still a GENEO;
- the translation of a GENEO is still a GENEO;
- the convex combination of GENEOs is still a GENEO;

However there is much more than this...

Permutant measures

Let us consider the set $\Phi = \mathbb{R}^X \cong \mathbb{R}^n$ of all functions from a finite set $X = \{x_1, \ldots, x_n\}$ to $\mathbb{R}$, and a subgroup G of the group $\mathrm{Bij}(X)$ of all permutations of X.

Definition

A finite (signed) measure μ on $\mathrm{Bij}(X)$ is called a *permutant measure* with respect to G if every subset H of $\mathrm{Bij}(X)$ is measurable and μ is invariant under the conjugation action of G (i.e., $\mu(H) = \mu(gHg^{-1})$ for every $g \in G$).

Representation theorem for linear GENEOs

Theorem (Representation Theorem for linear GENEOs)

Let us assume that $G \subseteq \mathrm{Bij}(X)$ *transitively acts on the finite set* X *and that* F *is a map from* $\mathbb{R}^X$ *to* $\mathbb{R}^X$*. The map* F *is a linear GENEO from* $\mathbb{R}^X$ *to* $\mathbb{R}^X$ *with respect to the identical homomorphism* $\mathrm{id}_G : g \mapsto g$ **if and only if** *a permutant measure* μ *with respect to* G *exists, such that* $F(\varphi) = \sum_{h \in \mathrm{Bij}(X)} \varphi h^{-1} \, \mu(h)$ *for every* $\varphi \in \mathbb{R}^X$*, and* $\sum_{h \in \mathrm{Bij}(X)} |\mu(h)| \leq 1$.

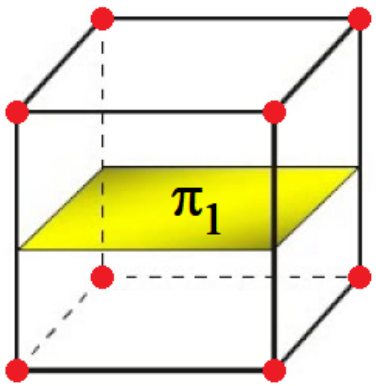
Two remarks:

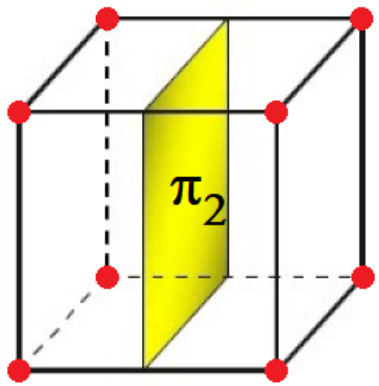
1. The set $\mathrm{PM}(G)$ of permutant measures with respect to G is a lattice and a real vector space.
2. The method for building GENEOs based on permutant measures can be generalized by replacing the weighted mean with a symmetric function, so building **non-linear** GENEOs.

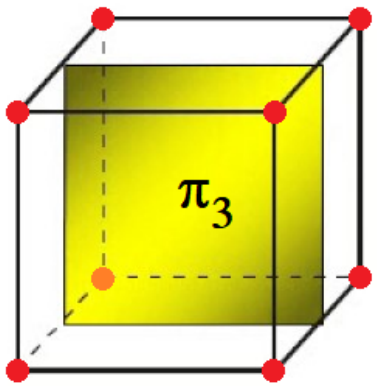
An example of permutant measure

Let us consider the set X of the vertices of a cube in $\mathbb{R}^3$, and the group G of the orientation-preserving isometries of $\mathbb{R}^3$ that take X to X. Let π_1, π_2, π_3 be the three planes that contain the center of mass of X and are parallel to a face of the cube. Let $h_i : X \to X$ be the orthogonal symmetry with respect to π_i, for $i \in \{1, 2, 3\}$.

We can now define a permutant measure μ on the group $\mathrm{Bij}(X)$ by setting $\mu(h_1) = \mu(h_2) = \mu(h_3) = c$, where c is a positive real number, and $\mu(h) = 0$ for any $h \in \mathrm{Bij}(X)$ with $h \notin \{h_1, h_2, h_3\}$.







Why build GENEOs via permutant measures?

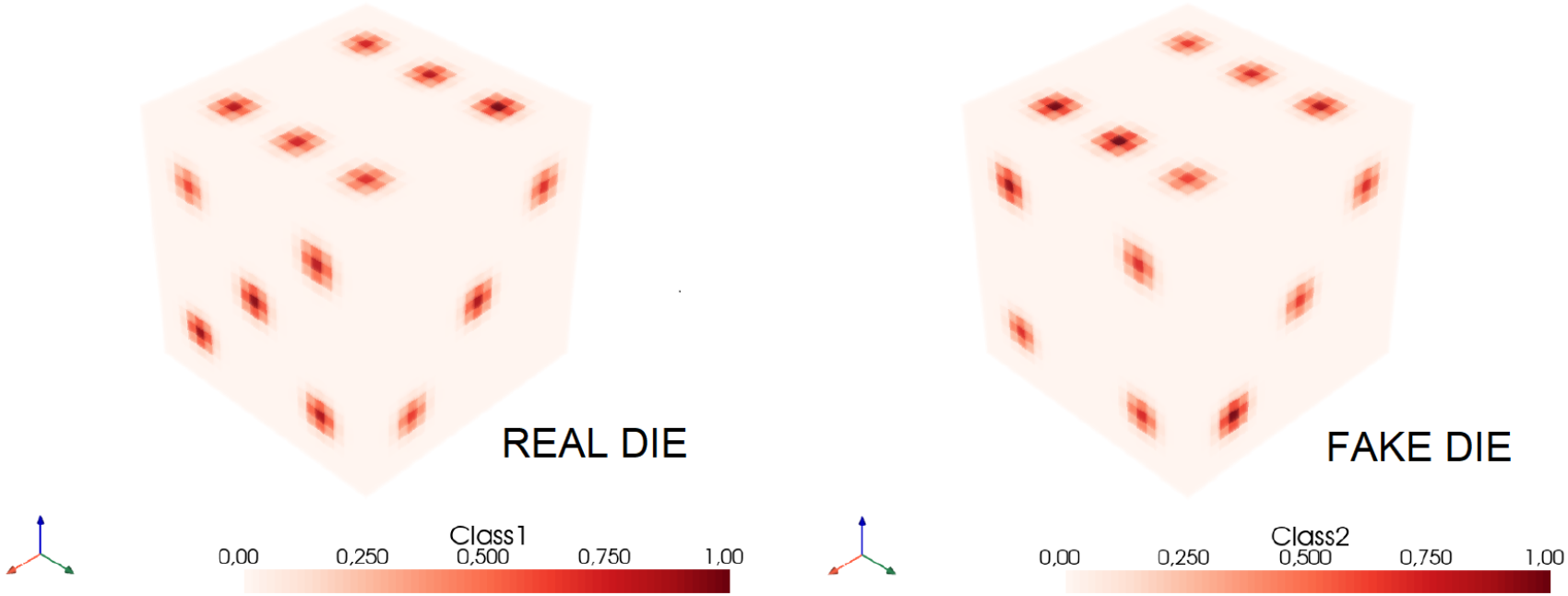
We observe that the smaller the support of a permutant measure μ, the more the summation $F_\mu(\varphi) := \sum_{h \in \mathrm{Bij}(X)} \varphi h^{-1} \, \mu(h)$ which defines the associated GENEO is simple to calculate.

In the example just seen, the group $\mathrm{Bij}(X)$ has cardinality 40,320, the equivariance group G contains 24 elements, while the permutant measure support contains only 3 permutations of X.

The usefulness of the construction method based on permutant measures lies in the fact that we can often use rather small permutants.

What happens when we apply GENEOs to our data?

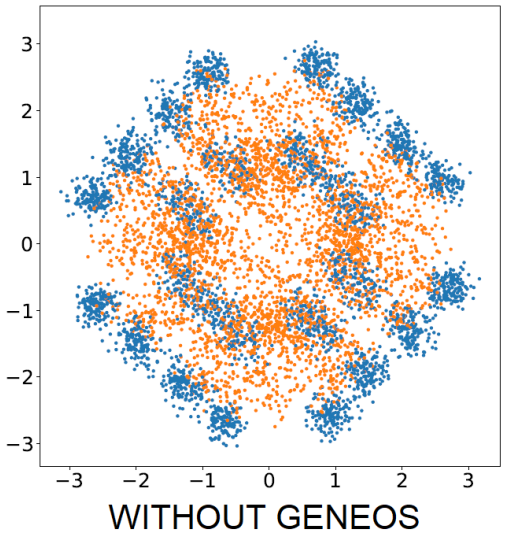
An example of use: comparison between real dice and fake dice.



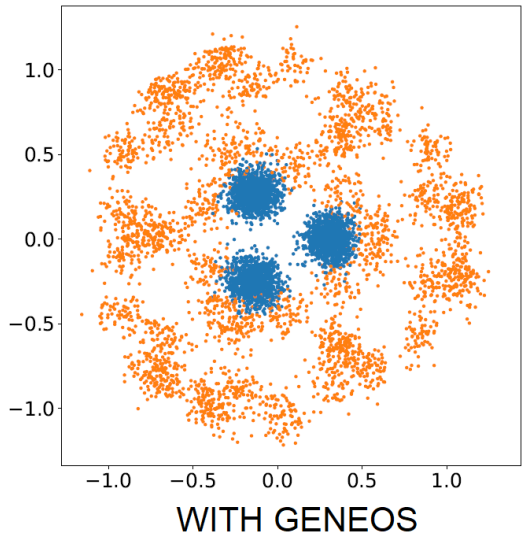
(Experiment and computations by Giovanni Bocchi)

What happens to data when we apply GENEOs?

We produced 10000 dice (a training set of size 7000 and a test set of size 3000), then we applied PCA to the test set and to the test set transformed by a suitable GENEO, optimized on the training set:



WITHOUT GENEOS

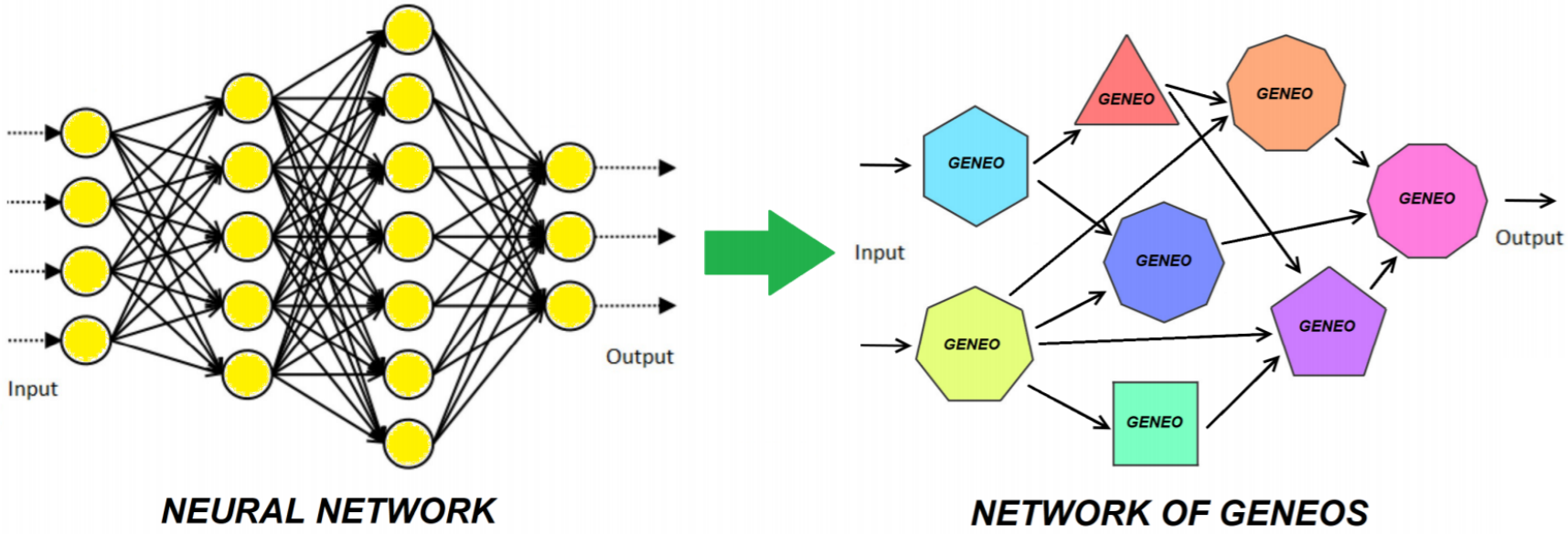


WITH GENEOS

For each die the first two principal components are plotted. Blue points are associated with **real dice**, while orange ones with **fake dice**. The GENEO we use was built by a convex combination of 3 GENEOs defined by permutant measures.

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.



GENEOs and Machine Learning

If interested, you can find more details about the theory of GENEOs in this paper:

- 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://rdcu.be/bP6HV

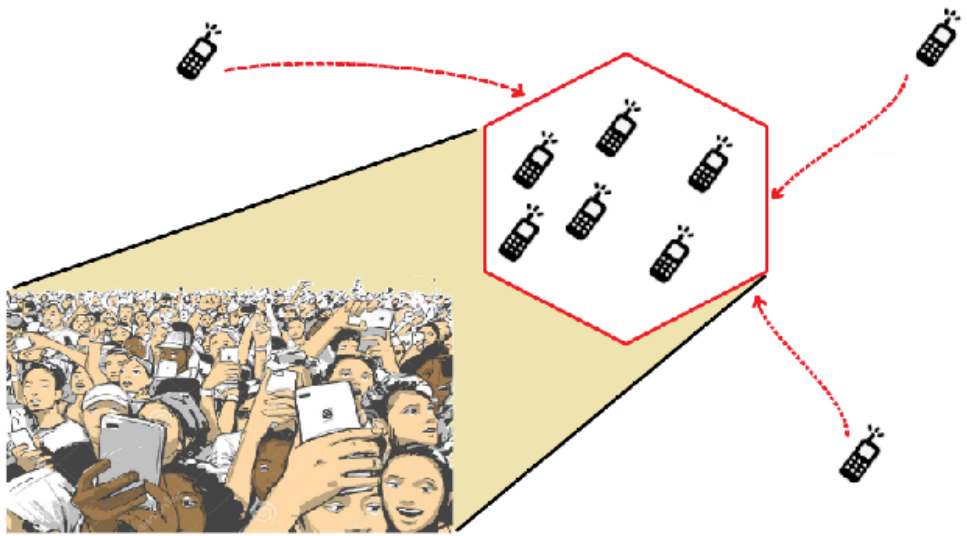
GENEOs and Machine Learning

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

Current research projects (I)

CNIT / WiLab - Huawei Joint Innovation Center (JIC)

Project on GENEOs for 6G



WILAB HUAWEI

Current 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

Some basics on the theory of GENEOs

Some links between GENEOs and TDA

Finding pockets in proteins by applying GENEOs

GENEOs restrict the invariance of TDA

Let $\mathscr{F}$ be a set of GENEOs between the perception pairs (Φ, G) and (Ψ, K), for a homomorphism $T : G \to K$. Call X the domain of the functions in Φ, and Y the domain of the functions in Ψ.
The use of GENEOs allows us to restrict the invariance of TDA by considering the following pseudo-metric:

$$\mathscr{D}^{\mathscr{F},\Phi}_{\text{match}}(\varphi_1, \varphi_2) = \sup_{F \in \mathscr{F}} d_B\Big(\mathrm{Dgm}_k(F(\varphi_1)), \mathrm{Dgm}_k(F(\varphi_2))\Big).$$

where $\mathrm{Dgm}_k(F(\varphi_1))$ and $\mathrm{Dgm}_k(F(\varphi_2))$ are the persistence diagrams in degree k of the functions $F(\varphi_1), F(\varphi_2)$, respectively.

The pseudo-metric $\mathscr{D}^{\mathscr{F},\Phi}_{\text{match}}$ is strongly invariant with respect to G, i.e., $\mathscr{D}^{\mathscr{F},\Phi}_{\text{match}}(\varphi_1, \varphi_2 \circ g) = \mathscr{D}^{\mathscr{F},\Phi}_{\text{match}}(\varphi_1 \circ g, \varphi_2) = \mathscr{D}^{\mathscr{F},\Phi}_{\text{match}}(\varphi_1, \varphi_2)$ for every $\varphi_1, \varphi_2 \in \Phi$ and every $g \in G$, but not invariant with respect to every $g \in \mathrm{Homeo}_\Phi(X)$, in general.

The computation of persistence diagrams is a GENEO

Let us endow the extended half-plane $\{x \leq y\} \subseteq (\mathbb{R} \cup \{\infty\})^2$ with the usual pseudo-metric δ, defined by setting $\delta((x,y),(x',y'))$ as

$$\min\left\{\max\left\{|x-x'|,|y-y'|\right\},\max\left\{\frac{|x-y|}{2},\frac{|x'-y'|}{2}\right\}\right\}$$

(by agreeing that $\infty - y = \infty$, $y - \infty = -\infty$ for $y \neq \infty$, $\infty - \infty = 0$, $\infty/2 = \infty$, $|\pm\infty| = \infty$, $\min\{\infty, c\} = c$, $\max\{\infty, c\} = \infty$).

Then, we take a compact metric space X and assume that:

- $\Phi = C^0(X, \mathbb{R})$ and $\Psi = C^0(\{x \leq y\}, \mathbb{R})$ (with "C^0" defined w.r.t. the metric and pseudo-metric that we have considered).
- G is the group of all homeomorphisms from X to X.
- K is the trivial group containing only the identity of $\{x \leq y\}$.
- $T : G \to K$ is the trivial homomorphism.

The computation of persistence diagrams is a GENEO

Let us consider the operator $F : \Phi \to \Psi$ defined by setting $F(\varphi) = \psi$, where $\psi(p)$ is the distance of $p \in \mathbb{R}^2$ from the support of the multiset $\mathrm{Dgm}_k(\varphi)$. In plain words, if we discard multiplicities, F associates each function φ with its persistence diagram, represented as a function ψ. The key point is that

F is a GENEO from (Φ, G) to (Ψ, K) with respect to T.

Equivariance follows from the invariance of persistence diagrams under the action of the homeomorphisms from X to X.
Non-expansiveness is a direct consequence of the stability of persistence diagrams with respect to the Hausdorff distance.

GENEOs interact with biparameter PH

The definition of the matching distance between two bifiltrations $\varphi, \psi : X \to \mathbb{R}^2$ of the topological space X can be seen as the supremum of the classical bottleneck distance between the persistence diagrams associated with the filtrations $F_{a,b}(\varphi), F_{a,b}(\psi) : X \to \mathbb{R}$, where the operator $F_{a,b}$ is defined by setting, for $a \in]0,1[$ and $b \in \mathbb{R}$,

$$F_{a,b}(\varphi) = \varphi^*_{(a,b)} = \max\left\{ \frac{\min\{a, 1-a\}}{a} \cdot (\varphi_1 - b), \frac{\min\{a, 1-a\}}{1-a} \cdot (\varphi_2 + b) \right\}.$$

The operator $F_{a,b}$ is a GENEO for any value of a and b

(provided that we consider the natural extension of the concept of GENEO to operators acting on vector-valued functions).

GENEOs can be compared by means of TDA

Persistent homology can be used to define a computable and stable pseudo-metric $\Delta_{\mathrm{GENEO},k}$ between GENEOs by setting

$$\Delta_{\mathrm{GENEO},k}(F_1,F_2) := \sup_{\varphi\in\Phi} d_B\left(\mathrm{Dgm}_k(F_1(\varphi)),\mathrm{Dgm}_k(F_2(\varphi))\right)$$

for every $F_1, F_2 \in \mathrm{GENEO}((\Phi,G),(\Psi,K))$.

Remark

Persistent homology also gives a shortcut to compare elements of each equivariance group G, by the pseudo-distance

$$\Delta_{G,k}(g_1,g_2) := \sup_{\varphi\in\Phi} d_B\left(\mathrm{Dgm}_k(\varphi\circ g_1),\mathrm{Dgm}_k(\varphi\circ g_2)\right).$$

Some basics on the theory of GENEOs

Some links between GENEOs and TDA

Finding pockets in proteins by applying GENEOs

Finding pockets in proteins by applying GENEOs

GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection.

Giovanni Bocchi [1], Patrizio Frosini [2], Alessandra Micheletti [1], Alessandro Pedretti [3] Carmen Gratteri [4], Filippo Lunghini [5] , Andrea Rosario Beccari [5] and Carmine Talarico [5]

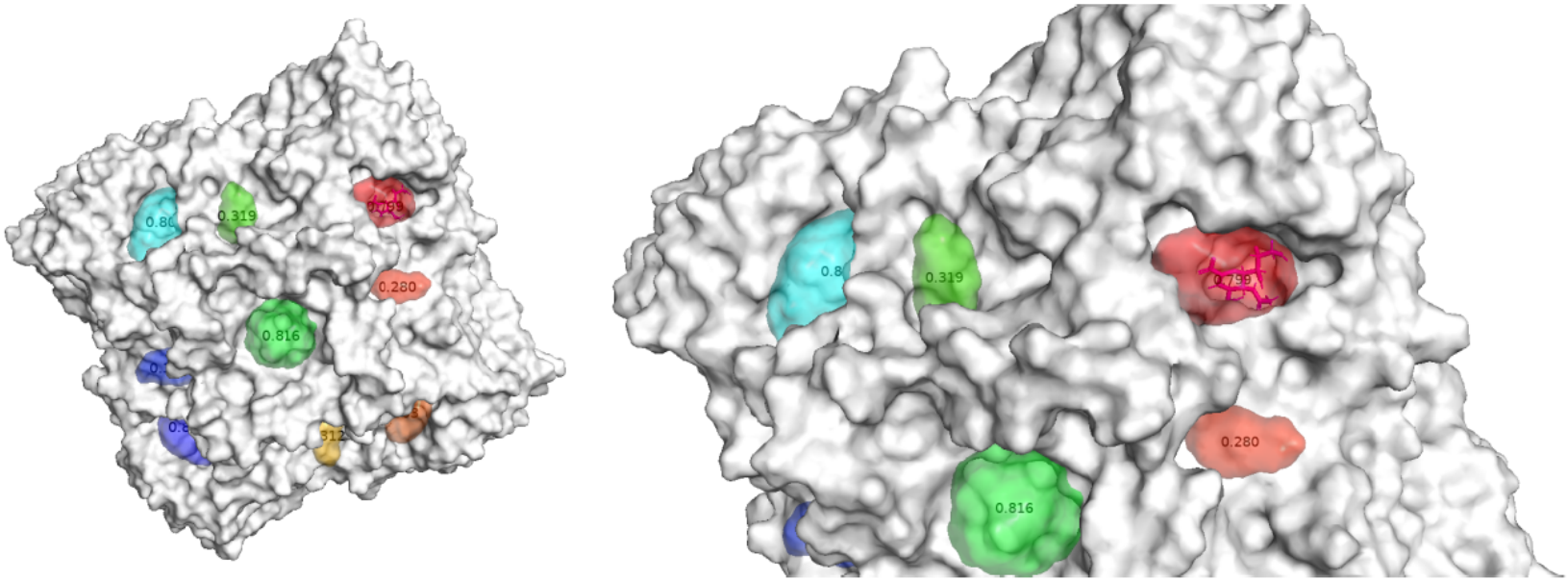
[1] Department of Environmental Science and Policy, Università degli Studi di Milano
[2] Department of Mathematics, Università degli Studi di Bologna
[3] Department of Pharmaceutical Sciences, Università degli Studi di Milano
[4] Dipartimento di Scienze della Salute, Università degli Studi "Magna Græcia di Catanzaro"
[5] Dompé Farmaceutici SpA

https://arxiv.org/ftp/arxiv/papers/2202/2202.00451.pdf

Updated results of this research will be presented at xAI-2024 (The 2nd World Conference on eXplainable Artificial Intelligence). Giovanni Bocchi has produced the data shown in these slides.

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.

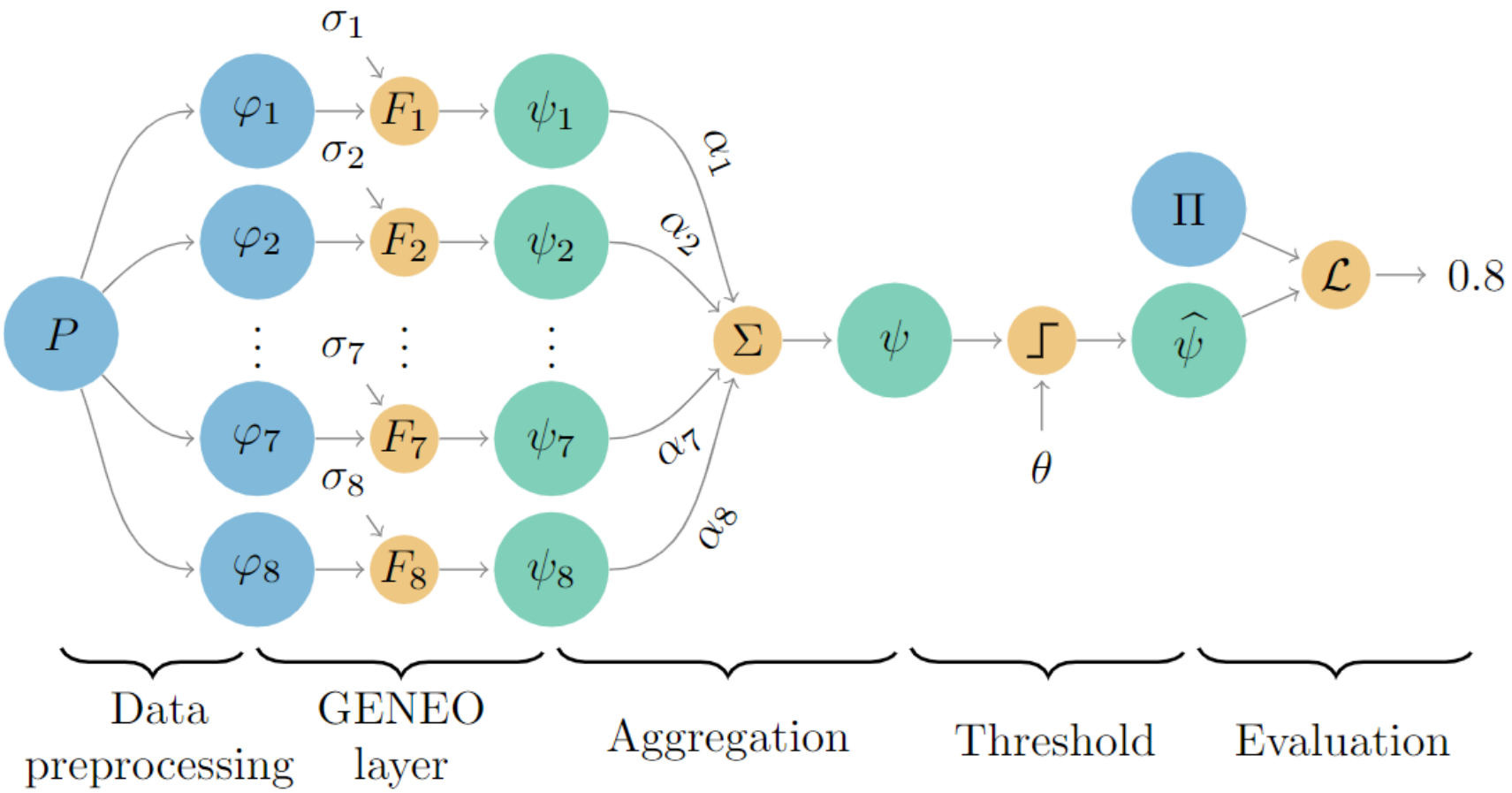
Convex combinations in GENEOnet

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.

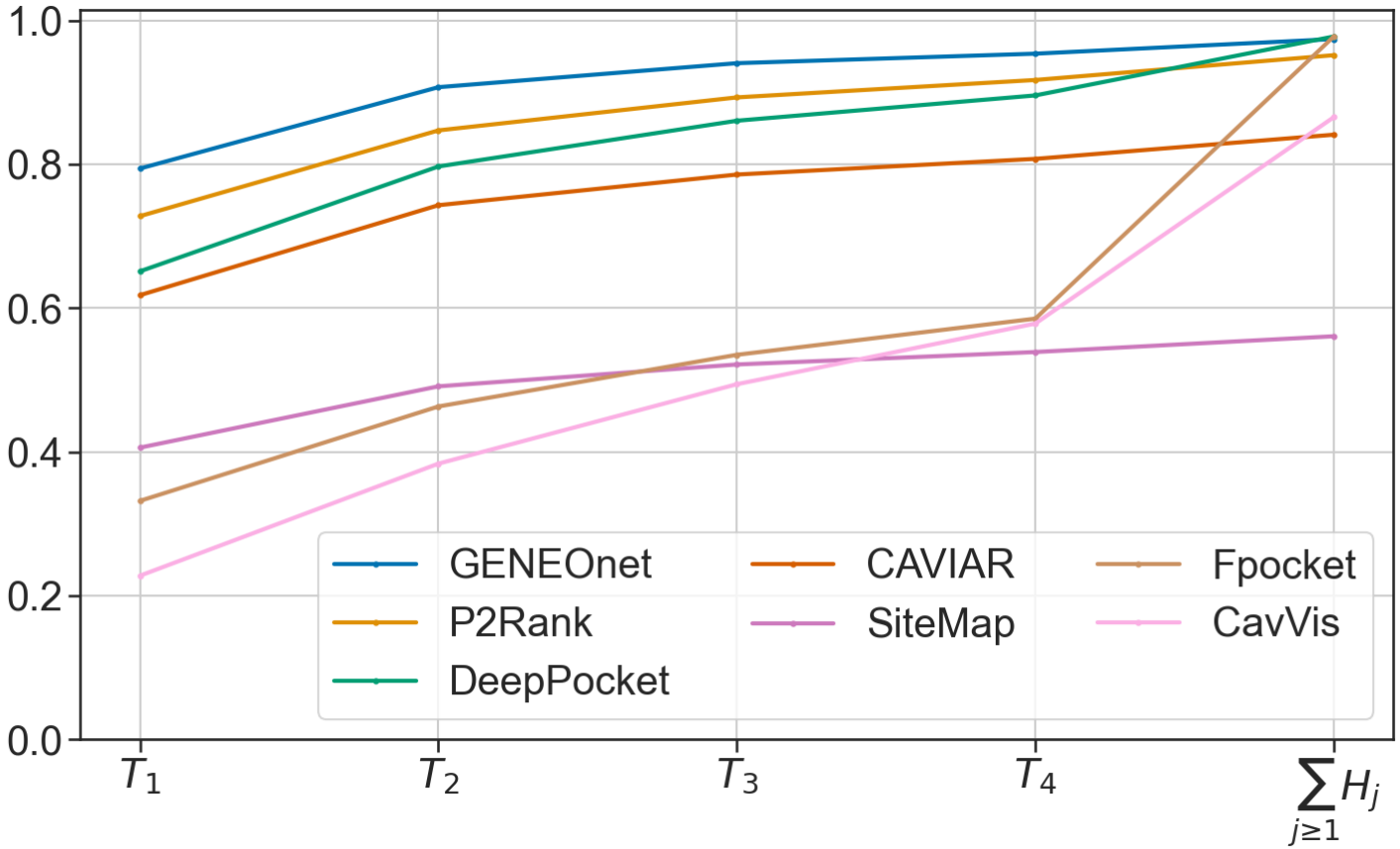
In the following table, we report estimates of H_j and T_j coefficients computed on a test set made of 9070 proteins from the PDBbind (this is about the 75% of the entire data extracted from the PDBbind).

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



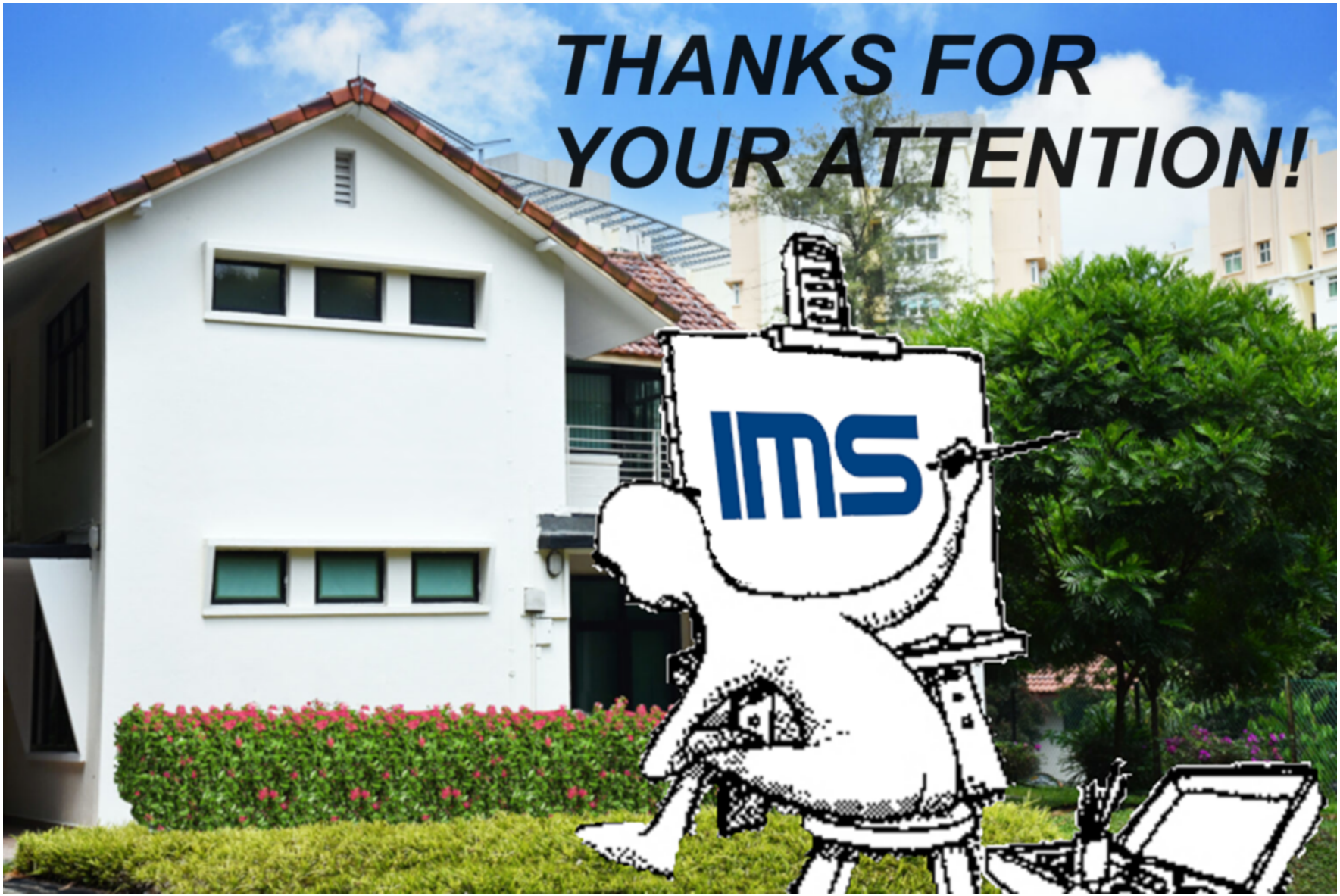
Please note that GENEOnet uses 17 parameters, while a CNN such as DeepPocket requires 665122 parameters.

GENEOnet webservice

The GENEOnet webservice represents the outcome of our partnership with Italian Pharmaceutical Company Dompé Farmaceutici S.p.A., and can be accessed via the URL: `https://geneonet.exscalate.eu`.

Webservice Developer: Anna Fava (EXSCALATE, Dompé Farmaceutici SpA)

The webservice has been designed for open access within the scientific community to test and evaluate our model. It has a user-friendly interface and delivers results in mere seconds. Upon submitting the code, the protein structure and all associated annotations from the Protein Data Bank are retrieved. Following submission, protein pockets are identified via GENEOnet, and the results are subsequently presented in a results table.


THANKS FOR
YOUR ATTENTION!
IMS