Group Equivariant Non-Expansive Operators:
a pathway towards Explainable Machine Learning

Giovanni Bocchi

Department of Environmental Science and Policy
University of Milan, Italy
PhD Student in Mathematics

09/03/2023

Collaborators

Patrizio Frosini [1]
[1] Department of Mathematics
University of Bologna, Italy.

Alessandra Micheletti [2]
[2] Department of Environmental Science and Policy
University of Milan, 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.

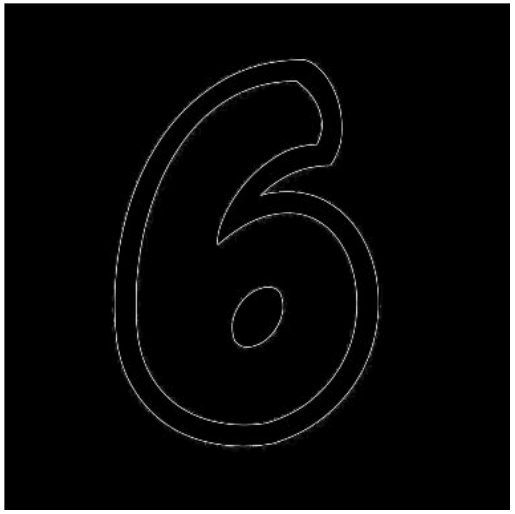
Epistemological approach

The motivation for introducing GENEOs lies in the following assumptions relative to a precise epistemological approach to data analysis:

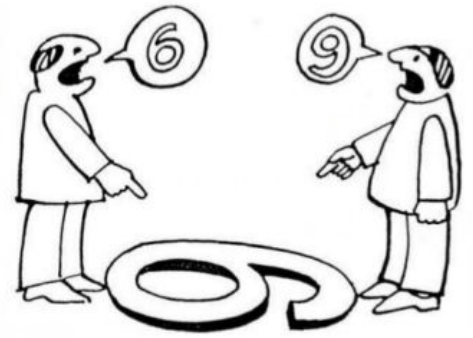
- Data can be quite often represented as functions defined on topological spaces.
- Data are only knowable when processed by an agent.
- Agents are defined by the way in which they act on functions and by some property of invariance (i.e. they commute with some transformations).
- Data similarity can be defined only with respect to an agent.

Epistemological approach (cont.)

(a) Gray level image can be seen as a function from $\mathbb{R}^2$ to $\mathbb{R}$.



(b) Sobel edge detection. Convolution operator equivariant w.r.t. isometries of $\mathbb{R}^2$.



(c) Observer dependent similarity.

Figure: Epistemological framework.

Ingredients - Data

To formally introduce GENEOs, let's consider two functional spaces whose functions are defined on some domains X and Y.

- $\Phi = \{\varphi \colon X \to \mathbb{R}\}$
- $\Psi = \{\psi \colon Y \to \mathbb{R}\}$

The space Φ represents the data domain in this framework, usually we assume at least to deal with bounded functions $\Phi \subseteq B(X, \mathbb{R})$.

Ingredients - Invariance

After the data we need to define the geometrical transformations we will require our GENEOs to commute with.

Let's consider a subgroup G (resp. H) of the group of all homeomorphisms of X (resp. Y) that are Φ-preserving (resp. Ψ-preserving):

$$\mathrm{Homeo}_{\Phi}(X) = \{g \in \mathrm{Homeo}(X) \colon \varphi \circ g \in \Phi \, \forall \, \varphi \in \Phi\}$$

$$G \trianglelefteq \mathrm{Homeo}_{\Phi}(X)$$

Finally fix a group homomorphism $T \colon G \to H$.

GENEOs definition

Definition (GENEO)

A Group Equivariant Non-Expansive Operator F is a map between Φ and Ψ that, for a fixed homomorphism of groups T, has these two properties:

- **Equivariance:** $F(\varphi \circ g) = F(\varphi) \circ T(g)$ for all $\varphi \in \Phi$ and for all $g \in G$.
- **Non-Expansivity:** $||F(\varphi_1) - F(\varphi_2)||_\infty \leq ||\varphi_1 - \varphi_2||_\infty$ for all $\varphi_1, \varphi_2 \in \Phi$.

GENEOs definition (cont.)

1. Equivariance encodes the fact that a GENEO must commute with a specific group of transformations of the data domain. In some sense we can say that GENEOs are able to filter out those transformations.

2. Non-expansivity implies that GENEOs tend to simplify the metric structure of data, so in some sense they provide a simpler representation of data. Moreover it is important to derive some topological properties of the space of GENEOs.

An example

If we consider

$$X = Y = \mathbb{R}^2$$

$$\Phi = \Psi \subseteq B(X, \mathbb{R})$$

$$G = H = \{\tau_a \colon X \to X \mid \tau_a(x) = x - a\} \quad T = Id_G$$

Thus since it's well known that $(\varphi \circ \tau) * k = (\varphi * k) \circ \tau$ then a convolutional operator with kernel k

$$F(\varphi) = \varphi * k$$

is a Translation Equivariant Operator. However in many problems equivariance just to translations is not enough.

Topological properties of GENEOs

Before introducing the most relevant topological properties of the space of GENEOs, let's do a brief excursus of the properties of the spaces introduced so far. For more details see[1]

- Φ is endowed with the topology induced by the sup norm distance.
- X is endowed with the topology induced by the data-dependent pseudo metric $d_X(x_1, x_2) = \sup_{\varphi \in \Phi} |\varphi(x_1) - \varphi(x_2)|$.
- If Φ is assumed to be compact and X is complete than X is also compact.

[1] 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.

Topological properties of the space of GENEOs (cont.)

More importantly it's possible to endow the space $\mathscr{F}$ of all GENEOs between (Φ, G) and (Ψ, H) w.r.t. T with a topology that is induced by the following metric:

$$D_{\text{GENEO}}(F_1, F_2) = \sup_{\varphi \in \Phi} ||F_1(\varphi) - F_2(\varphi)||_\infty$$

This definition brings to the two most relevant results regarding the space $\mathscr{F}$.

Compactness of the space of GENEOs

Theorem (Compactness)

If both Φ and Ψ are compact in the topology induced by the sup norm distance then also $\mathscr{F}$ is compact in the topology induced by the metric D_{GENEO}.

Compactness in the case of a metric space implies total boundedness: we now that for every ε it exists a finite collection of GENEOs $\{F_1, \ldots, F_n\}$ s.t. for every $F \in \mathscr{F}$ it holds that $D_{\text{GENEO}}(F, F_i) \leq \varepsilon$ for some $i = 1, \ldots, n$. Thus $\{F_1, \ldots, F_n\}$ can be interpreted as representatives of $\mathscr{F}$.

Convexity of the space of GENEOs

Given n GENEOs $F_1, \ldots, F_n$ their convex combination with non negative coefficients $\alpha_1, \ldots, \alpha_n$ such that $\sum_{i=1}^{n} \alpha_i = 1$ is defined as:

$$F(\varphi) = \sum_{i=1}^{n} \alpha_i F_i(\varphi)$$

If Ψ is convex than F is still a GENEO with the same properties of equivariance of $F_1, \ldots, F_n$, as stated by the following theorem.

Theorem (Convexity)

If Ψ is convex than the space $\mathscr{F}$ is also convex.

How to combine GENEOs ?

Convexity of the space of GENEOs guarantees that the convex hull of a finite number of GENEOs is fully contained in $\mathscr{F}$. This fact provides the first rule to derive new GENEOs from the ones available.

However convexity is definitely not the only way to obtain new GENEOs, indeed there are other results, mainly described in[2], that show further techniques for combining existing GENEOs.

[2] P. Frosini and N. Quercioli, "Some remarks on the algebraic properties of group invariant operators in persistent homology," in *1st International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, vol. LNCS-10410, Springer International Publishing, Aug. 2017, pp. 14–24. DOI: 10.1007/978-3-319-66808-6_2. [Online]. Available: https://hal.inria.fr/hal-01677132.

Elementary compositional rules

The following are all feasible ways of combining GENEOs.

- **Composition**: If F_1 is a GENEO from (Φ, G) to (Ψ, H) w.r.t. $T_1 \colon G \to H$ and F_2 is a GENEO from (Ψ, H) to (χ, K) w.r.t. $T_2 \colon H \to K$ then $F_2 \circ F_1$ is a GENEO from (Φ, G) to (χ, K) w.r.t. $T_2 \circ T_1 \colon G \to K$.
- **Minimum and Maximum**: If $F_1, \ldots, F_n$ are GENEOs from (Φ, G) to (Ψ, H) w.r.t. T then both $\min(F_1, \ldots, F_n)(\varphi)$ and $\max(F_1, \ldots, F_n)(\varphi)$ are GENEOs provided that their output belong to Ψ.
- **Translation**: If F is a GENEO from (Φ, G) to (Ψ, H) w.r.t. T then F_b defined as $F_b(\varphi) = F(\varphi) - b$ with $b \in \mathbb{R}$ is a GENEO provided that $F(\varphi) - b \in \Psi$ for all $\varphi \in \Phi$.

Other ways to obtain new GENEOs

Other, less elementary, ways of generating GENEOs are under study: for example it has been shown that linear GENEOs admit a representation by mean of permutant measures[3].

Theorem (Representation of linear GENEOs)

Assume that $G \subseteq Aut(X)$ transitively acts on the finite set X and F is a map from $\mathbb{R}^X$ to $\mathbb{R}^X$. The map F is a linear group equivariant non-expansive operator for $(\mathbb{R}^X, G)$ if and only if a permutant measure μ exists such that $F(\varphi) = \sum_{h \in Aut(X)} \varphi \circ h^{-1} \mu(h)$ for every $\varphi \in \mathbb{R}^X$ and $\sum_{h \in Aut(X)} |\mu(h)| \leq 1$.

[3] G. Bocchi, S. Botteghi, M. Brasini, *et al.*, "On the finite representation of linear group equivariant operators via permutant measures," *Annals of Mathematics and Artificial Intelligence*, Feb. 2023, ISSN: 1573-7470. DOI: 10.1007/s10472-022-09830-1. [Online]. Available: https://doi.org/10.1007/s10472-022-09830-1.

Building networks of GENEOs

This evolving but rich enough compositional theory of GENEOs allows to develop models in which several GENEO units are combined using the aforementioned techniques. This line of research aims to build efficient and transparent networks as an alternative to efficient but obscure Neural Networks.

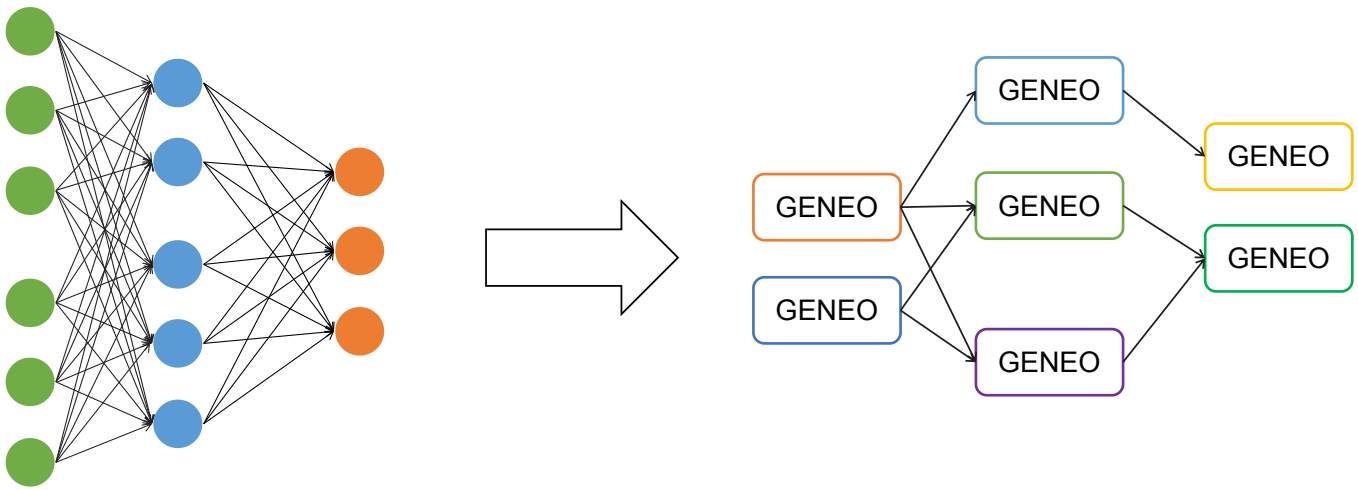


Figure: From neural networks to GENEO networks

How GENEOs and TDA are connected ?

At this stage of this exposition it's not clear if there is any relationship between GENEOs and Topological Data Analysis. Surely GENEOs allow to perform an analysis of data that takes into account their topological properties by choosing the equivariance group and how the operator will act, but can we say something more ?

The answer is yes, as partially explained in the following.

Definition of the pseudo metric $D^{\Phi,\mathcal{F}}_{\text{match}}$

A first fact is that, given a set $\mathcal{F}$ of GENEOs, the bottleneck distance d_{match} between the k^{th} persistence diagrams of their outputs defines the following pseudo-metric on the data

$$D^{\Phi,\mathcal{F}}_{\text{match}}(\varphi_1,\varphi_2) = \sup_{F\in\mathcal{F}} d_{\text{match}}(\text{Dgm}_k(F(\varphi_1)),\text{Dgm}_k(F(\varphi_2)))$$

This pseudo-metric is relevant mainly for two reasons. First we can compare it to the pure bottleneck distance between the k^{th} persistence diagrams of the data

$$D^{\text{TDA}}_{\text{match}}(\varphi_1,\varphi_2) = d_{\text{match}}(\text{Dgm}_k(\varphi_1),\text{Dgm}_k(\varphi_2))$$

Comparison between $D^{\Phi,\mathcal{F}}_{\text{match}}$ and $D^{\text{TDA}}_{\text{match}}$

It is a well known fact that $D^{\text{TDA}}_{\text{match}}$ is a strongly Homeo(X)-invariant metric. Indeed for all $h \in \text{Homeo}(X)$

$$D^{\text{TDA}}_{\text{match}}(\varphi_1, \varphi_2) = D^{\text{TDA}}_{\text{match}}(\varphi_1, \varphi_2 \circ h) = D^{\text{TDA}}_{\text{match}}(\varphi_1 \circ h, \varphi_2)$$

This invariance is defined w.r.t. a very large group of transformations. However, in many situations, it could be useful to restrict the invariance to a smaller subgroup $G \subseteq \text{Homeo}(X)$. It's not difficult to check that if the set $\mathcal{F}$ contains G-equivariant operators then $D^{\Phi,\mathcal{F}}_{\text{match}}$ is a strongly G-invariant pseudo-metric, thus for every $g \in G$

$$D^{\Phi,\mathcal{F}}_{\text{match}}(\varphi_1, \varphi_2) = D^{\Phi,\mathcal{F}}_{\text{match}}(\varphi_1, \varphi_2 \circ g) = D^{\Phi,\mathcal{F}}_{\text{match}}(\varphi_1 \circ g, \varphi_2)$$

Efficient computation of D_{GENEO}

The metric

$$D_{\mathrm{GENEO}}(F_1, F_2) = \sup_{\varphi \in \Phi} ||F_1(\varphi) - F_2(\varphi))||_\infty$$

has already been introduced to define the topology of the space of GENEOs. Nevertheless, from a computational point of view, this is not an easy computable metric between GENEOs since it involves a supremum of a possibly very large set of sup-norm values.

Efficient computation of D_{GENEO} (cont.)

To deal with the supremum the compactness of Φ can be exploited in order to reduce the computation to a finite set of functions (introducing an approximation), while the sup norm can be replaced by the bottleneck distance in the following way.

$$\Delta_{\mathrm{GENEO},k}(F_1, F_2) = \sup_{\varphi \in \Phi} d_{\mathrm{match}}(\mathrm{Dgm}_k(F_1(\varphi)), \mathrm{Dgm}_k(F_2(\varphi)))$$

Furthermore stability of the bottleneck distance ensures that

$$\Delta_{\mathrm{GENEO},k}(F_1, F_2) \leq D_{\mathrm{GENEO}}(F_1, F_2)$$

Thus Topological Data Analysis helps to provide a good proxy for computing D_{GENEO}.

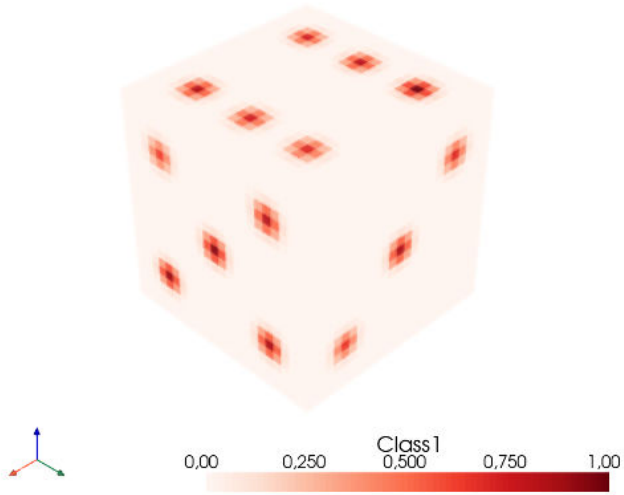
Applications of GENEOs

After introducing GENEOs and some connections with TDA, some applications to case studies will now be examined. We start with a naive toy example and proceed with a real case study

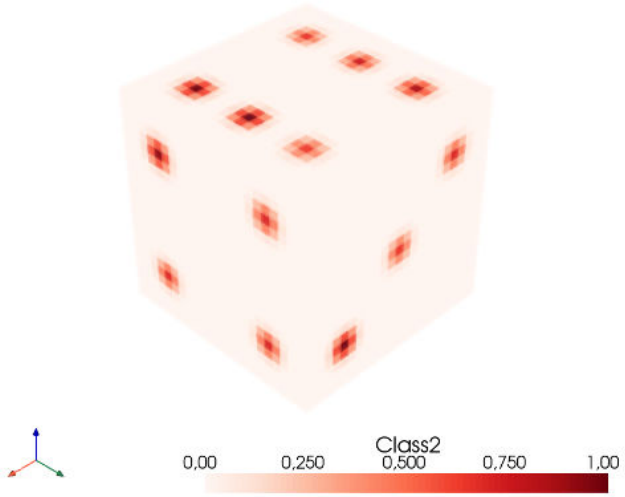
1. Dice classification
2. Protein pocket detection

Dice

The toy example involves classification of synthetic data of 3D scans of commercial and fake dice both seen as functions from a finite cubical grid X to $\mathbb{R}$.



(a) Commercial die.



(b) Fake die.

In the first case opposite faces add up to seven while not in the second case as seen in the Figure.

Dice (cont.)

The main objective here was to demonstrate the effectiveness of permutant measures for obtaining GENEOs. Indeed a convex combination of GENEOs in the form

$$F(\varphi) = \frac{1}{|H|} \sum_{h_i \in H} \varphi \circ h_i^{-1}$$

has been used where H is now a permutant set used to define a permutant measure. One of such permutants is composed of the orthogonal reflections by the planes shown in Figure.

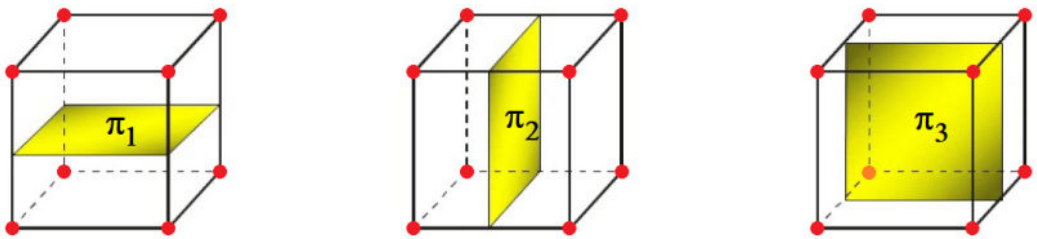
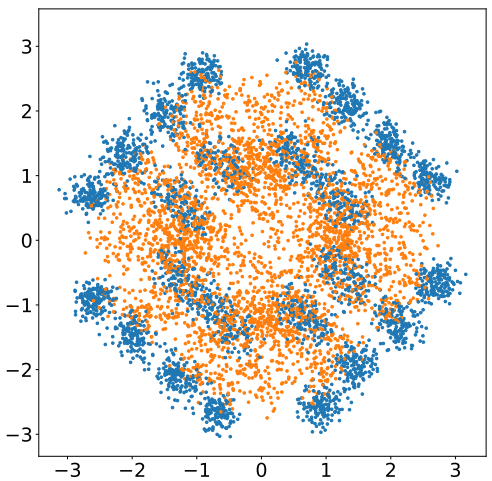


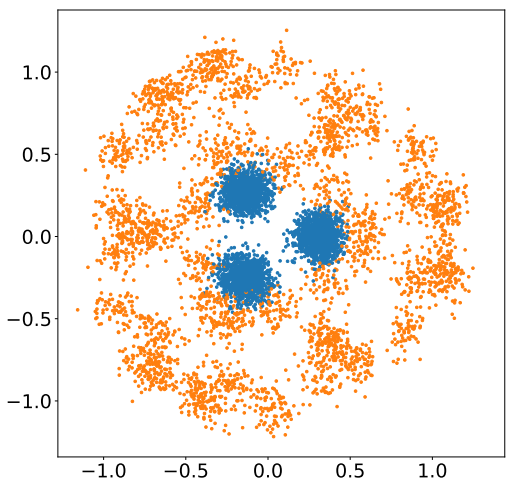
Figure: Example of permutant

Dice (cont.)

The selected GENEO was applied to the data followed by a PCA for the two classes of processed and unprocessed data:



(a) Unprocessed.



(b) GENEO processed.

The processed data show a much clearer separation between the two classes. Moreover a Support Vector Machine classifier obtained better results using as input GENEO processed data.

GENEOs networks

Fix n parametric families of GENEOs

$$\mathcal{F}_i = \{F^i_{\theta^i}\} \quad \theta^i \in \Theta_i \subseteq \mathbb{R}^{p_i}$$

Then derive a new family of operators by convex combination

$$\mathcal{F}_w = \left\{ \sum_{i=1}^{n} \alpha_i F^i_{\theta^i} \right\} \quad \alpha \in D^n$$

A GENEO $F \in \mathcal{F}_w$ is parametrized by the real vector $w = (\theta^1, \ldots, \theta^n, \alpha) \in \mathbb{R}^{\sum p_i + n}$.

GENEOs networks (cont.)

Consider a problem with functional data $(\varphi_i, \xi_i)_{i=1}^n$ and a loss function $\ell(\xi, \psi)$ then we can search for the GENEO $F \in \mathcal{F}_w$ (i.e. for the parameters w) such that

$$F = F_{w^*} \quad w^* = \arg\min_w \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, F_w(\varphi_i))$$

Moreover if all the GENEOs F^i depend in a differentiable way from the parameters then the optimization problem can be solved with Gradient Descent techniques.

A real example

The main question regarding GENEO networks deals with their accuracy.

We expect that equivariance and the possibility to inject prior knowledge in the definition of GENEOs will help to reduce model complexity and the need for training examples. But usually a gain in explainability is associated to a loss in accuracy.

Protein pocket detection

The problem comes from a very hot topic in medicinal chemistry: the aim is to identify, given the 3D structure of a protein, areas of the surface that are likely to host a ligand (i.e. a drug).

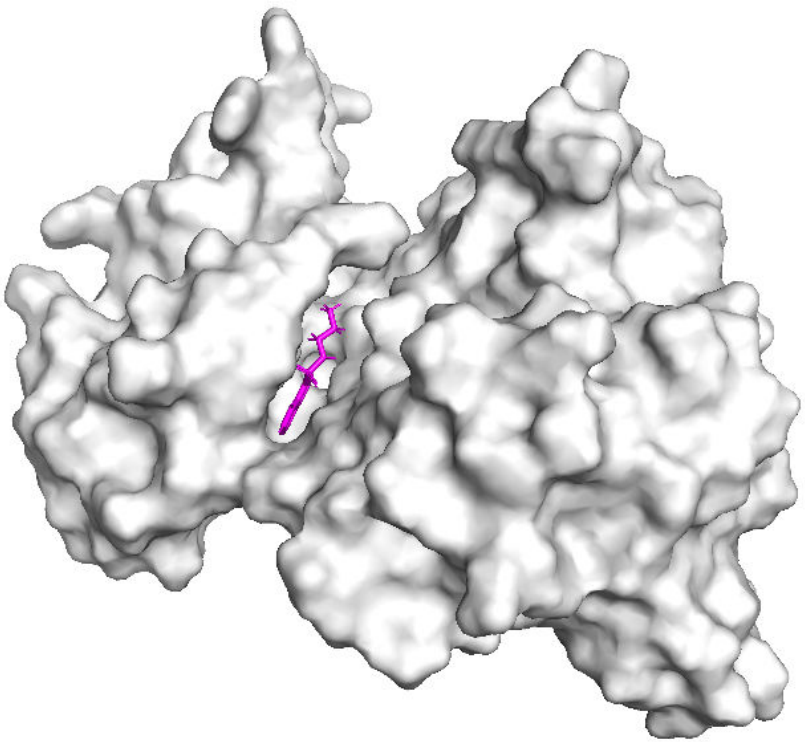
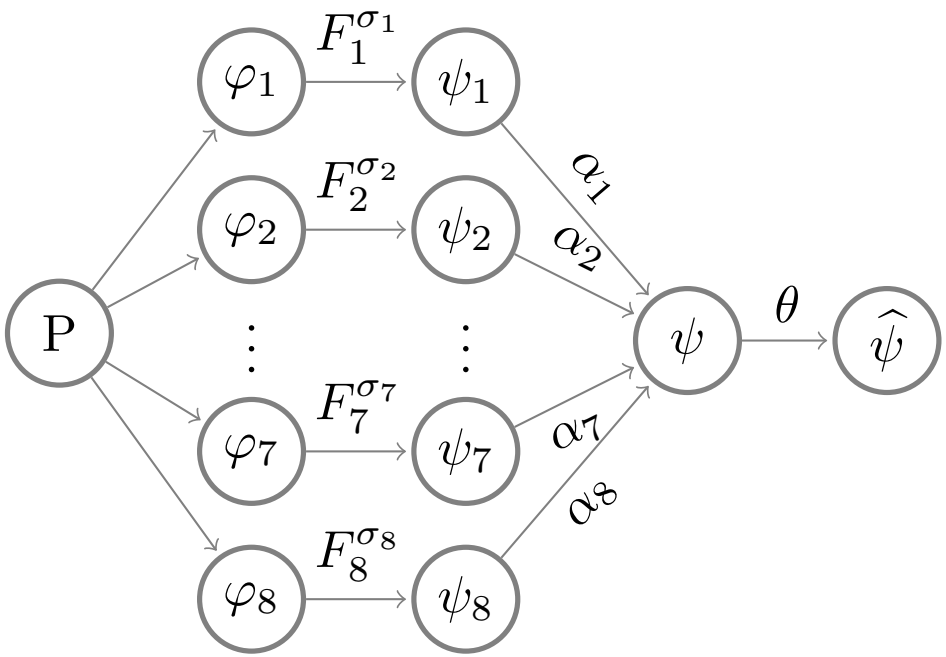


Figure: Protein ID 2CKE and corresponding ligand

Protein pocket detection (cont.)

This case study brought to the very first prototype of a network of GENEOs: **GENEOnet**[4]. The φ_i data were chosen as a reasoned selection of geometrical, physical and chemical potentials of the protein seen as functions $\varphi_i \colon B \subseteq \mathbb{R}^3 \to \mathbb{R}$.



[4] G. Bocchi, P. Frosini, A. Micheletti, *et al.*, "GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection," 2022. arXiv: 2202.00451.

Protein pocket detection (cont.)

GENEO units are composed of parametric families of convolutional operators with radial kernels. This fact, combined with the native equivariance of convolutions w.r.t. translations, guarantees that they are all GENEOs with respect to isometries of $\mathbb{R}^3$.

$$F_\theta(\varphi) = \varphi * k_\theta \quad k_\theta(x) = k_\theta(||x||)$$

The model has just 17 parameters: 8 shape parameters of convolutional kernels, 8 parameters of the convex combination and one threshold parameter.

Here the optimization has been carried out using a form of Gradient Descent.

Protein pocket detection (cont.)

The output of the model is a list of predicted pockets ranked by a score coefficient. The Figure shows the predictions for an example protein.

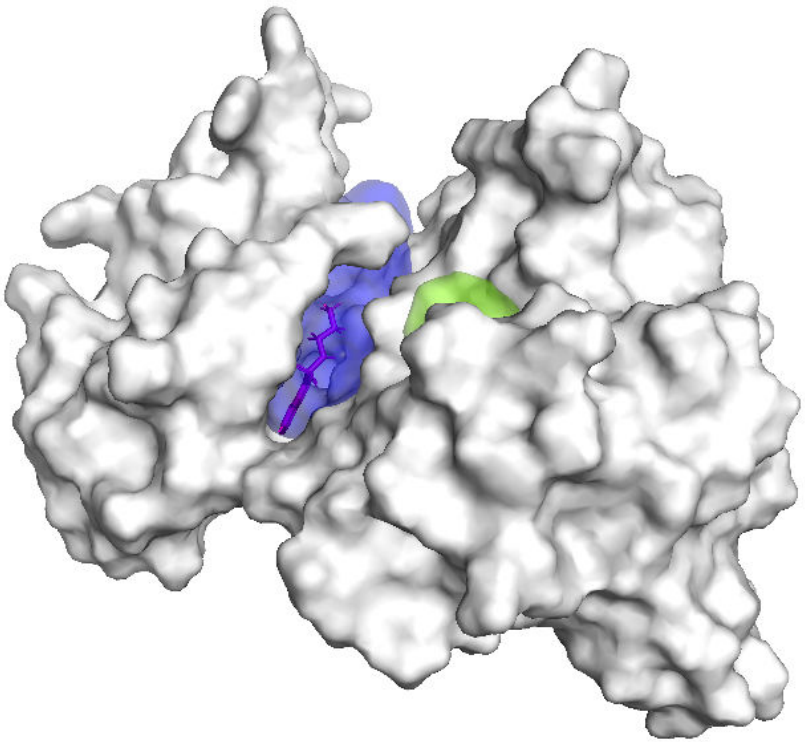
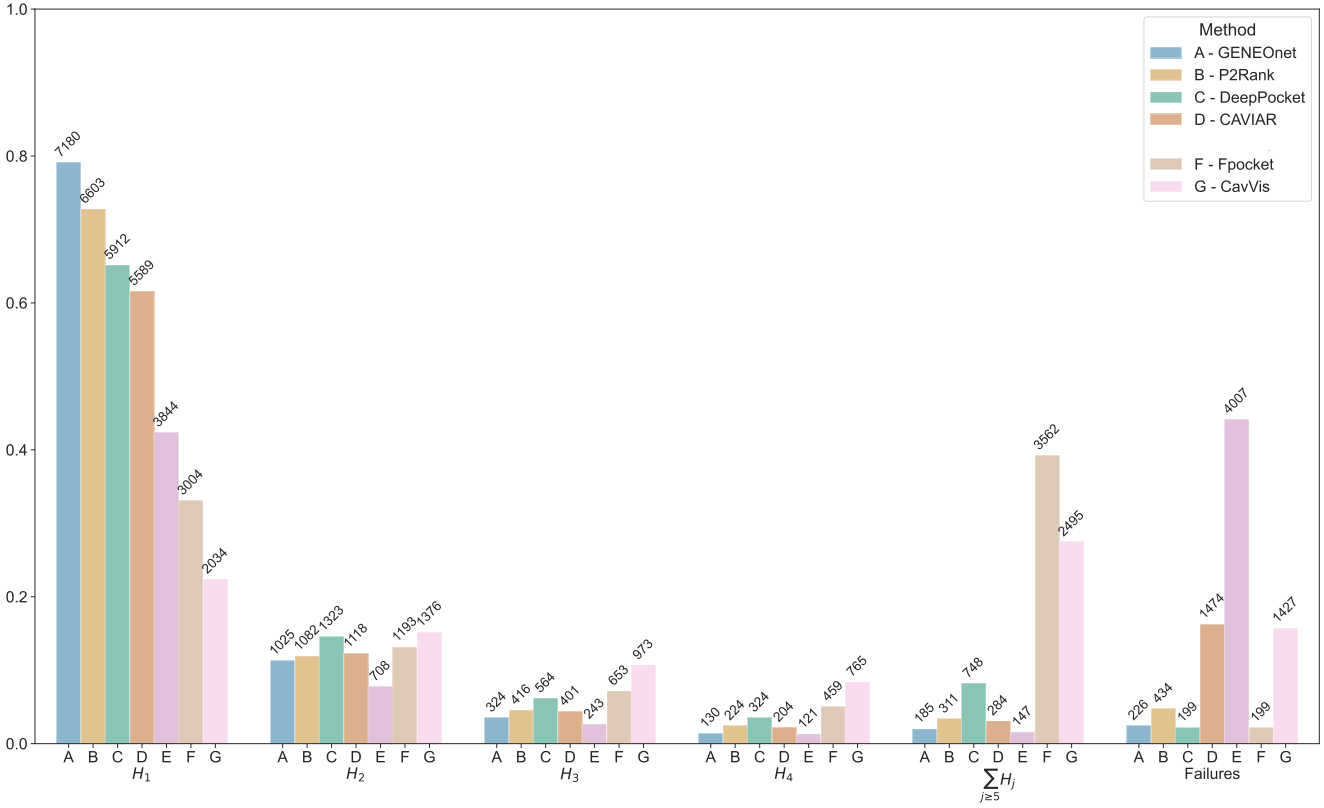


Figure: Prediction for protein 2KCE with ligand matching

Protein pocket detection (cont.)

Finally a comparison with other state-of-the-art methods. The bar chart shows the fraction of proteins in the test set for which each method identified the true pocket as the one with n^{th} highest score.



Take home message

GENEOs are under study both from a theoretical and applicative point of view since they allow to build networks such that:

1. It's possible to incorporate prior knowledge.
2. There's a reduction of model complexity and hunger of training examples.
3. There's the possibility to interpret the model parameters.
4. There's no significant loss in accuracy.

Thank you for your attention!

References I

Theory of GENEOs

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`.

P. Frosini and N. Quercioli, "Some remarks on the algebraic properties of group invariant operators in persistent homology," in *1st International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, vol. LNCS-10410, Springer International Publishing, Aug. 2017, pp. 14–24. DOI: `10.1007/978-3-319-66808-6_2`. [Online]. Available: `https://hal.inria.fr/hal-01677132`.

References II

Links with TDA

 A. Cerri, M. Ethier, and P. Frosini, "On the geometrical properties of the coherent matching distance in 2d persistent homology," *Journal of Applied and Computational Topology*, vol. 3, pp. 381–422, 4 2019.

 P. Frosini and G. Jabłoński, "Combining persistent homology and invariance groups for shape comparison," *Discrete Computational Geometry*, vol. 55, pp. 373–409, 2 2016.

References III

Case studies

G. Bocchi, S. Botteghi, M. Brasini, *et al.*, "On the finite representation of linear group equivariant operators via permutant measures," *Annals of Mathematics and Artificial Intelligence*, Feb. 2023, ISSN: 1573-7470. DOI: 10.1007/s10472-022-09830-1. [Online]. Available: https://doi.org/10.1007/s10472-022-09830-1.

G. Bocchi, P. Frosini, A. Micheletti, *et al.*, "GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection," 2022. arXiv: 2202.00451.