Group Equivariant Non-Expansive Operators

From TDA to Neural Networks.

Giovanni Bocchi

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

11/25/2022

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.

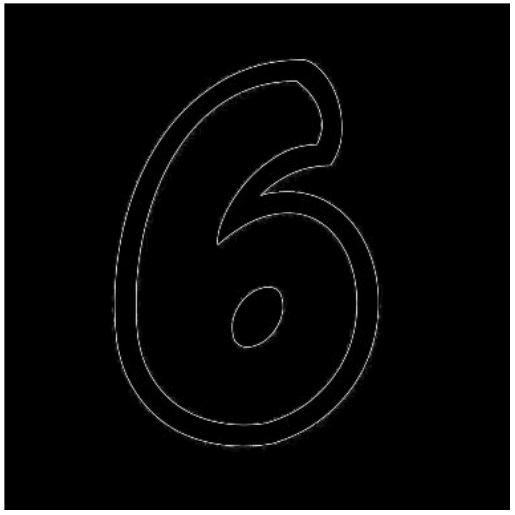
Epistemological approach

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

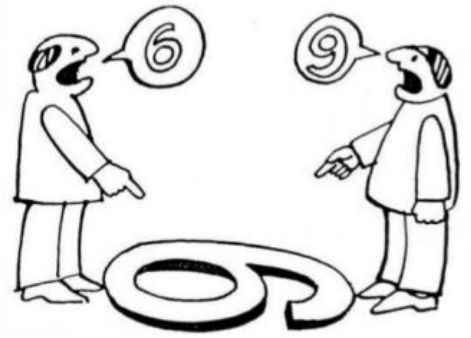
- Data can be quite often represented as functions defined on topological spaces.
- Data are only knowable when processed by an agent.
- Agents are defined by the way in which they act on functions and by some property of invariance (i.e. they commute with some transformations).
- Data similarity can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

Preliminaries

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

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

Then 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), i.e. those $g \in \text{Homeo}(X)$ such that $\varphi \circ g \in \Phi$ and $\varphi \circ g^{-1} \in \Phi$ for all $\varphi \in \Phi$. Finally fix a homomorphism T between G and 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.

Topological properties of the space of GENEOs

Let's consider the space $\mathscr{F}$ of all GENEOs between (Φ, G) and (Ψ, H) w.r.t. T. Then it is possible to define the following metric:

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

and to consider the topology on $\mathscr{F}$ induced by D_{GENEO}. 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 is quite relevant from an applicative point of view: it justifies the fact that to approximate the space of GENEOs arbitrarily well, it is sufficient to consider only a finite subset of **Representative** operators.

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 [P. Frosini et al][1], that show further techniques for combining existing GENEOs.

[1] 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 $\varphi \in \Phi$.

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.

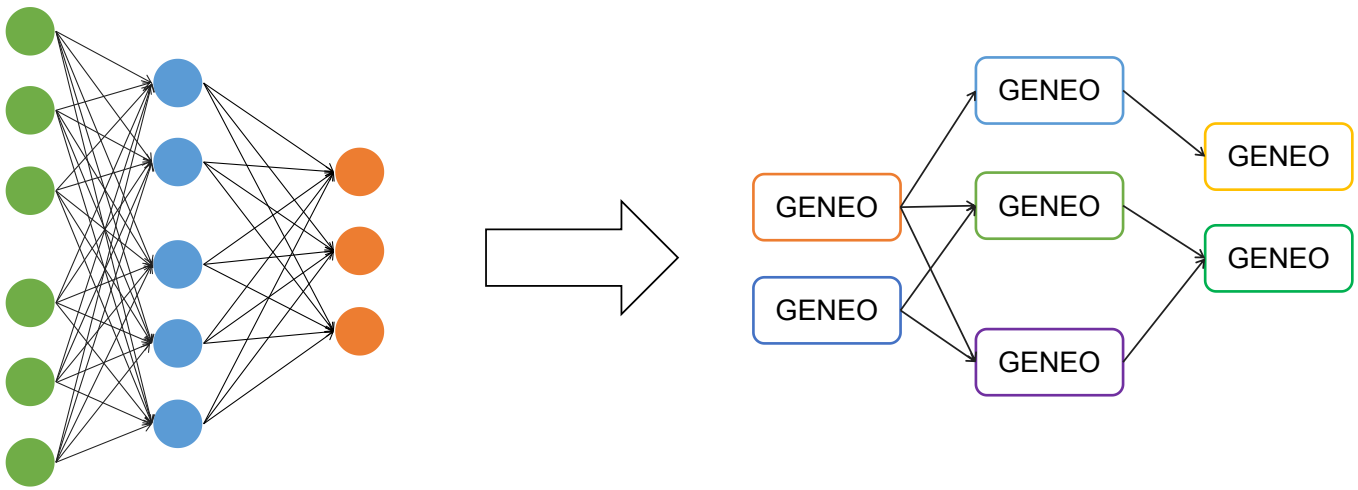


Figure: From neural networks to GENEO networks

Connections with Topological Data Analysis

For those familiar with Topological Data Analysis here there's a short and informal list of some connections between GENEOs and TDA:

- GENEOs allow to restrict the invariance of TDA to proper subgroups of the group of all homeomorphisms.
- Under suitable hypothesis the operator that computes persistence diagrams can be seen as a GENEO.
- The 2D matching distance between persistent Betti numbers functions can be described in terms of GENEOs.

Applications of GENEOs

After introducing GENEOs and some insights into connections with TDA, let's introduce the first application of a network of GENEOs to a real world problem.

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).

Protein pocket detection

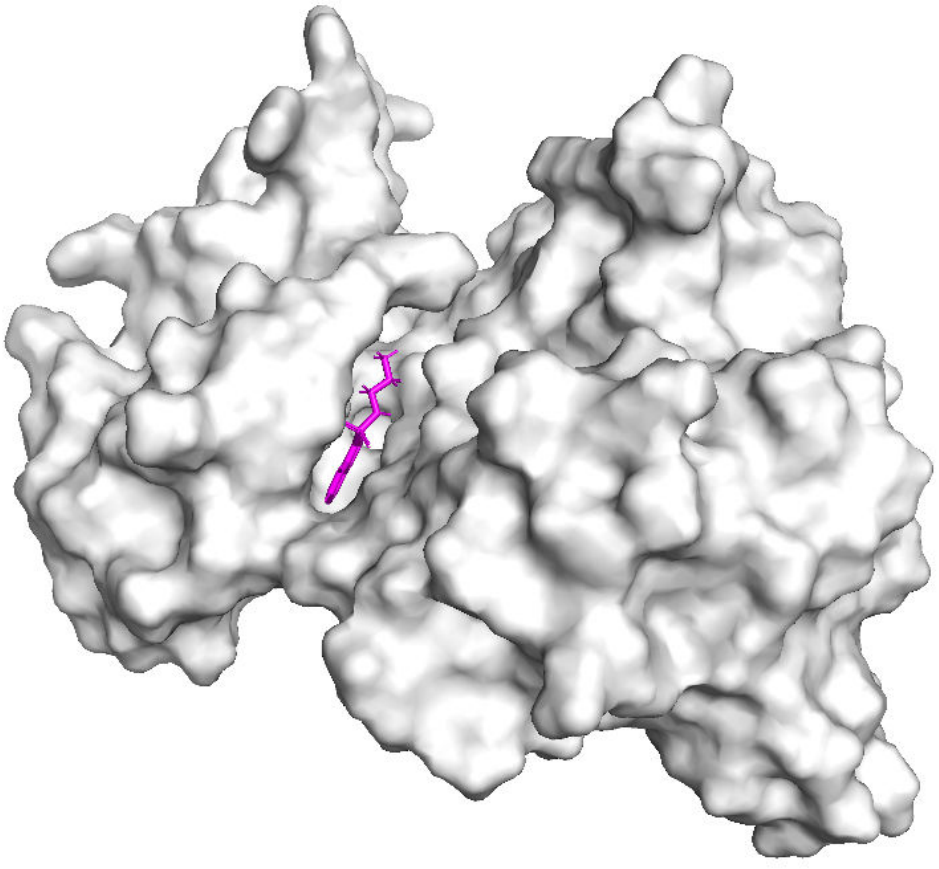


Figure: Protein ID 2CKE and corresponding ligand

Protein pocket detection (cont.)

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

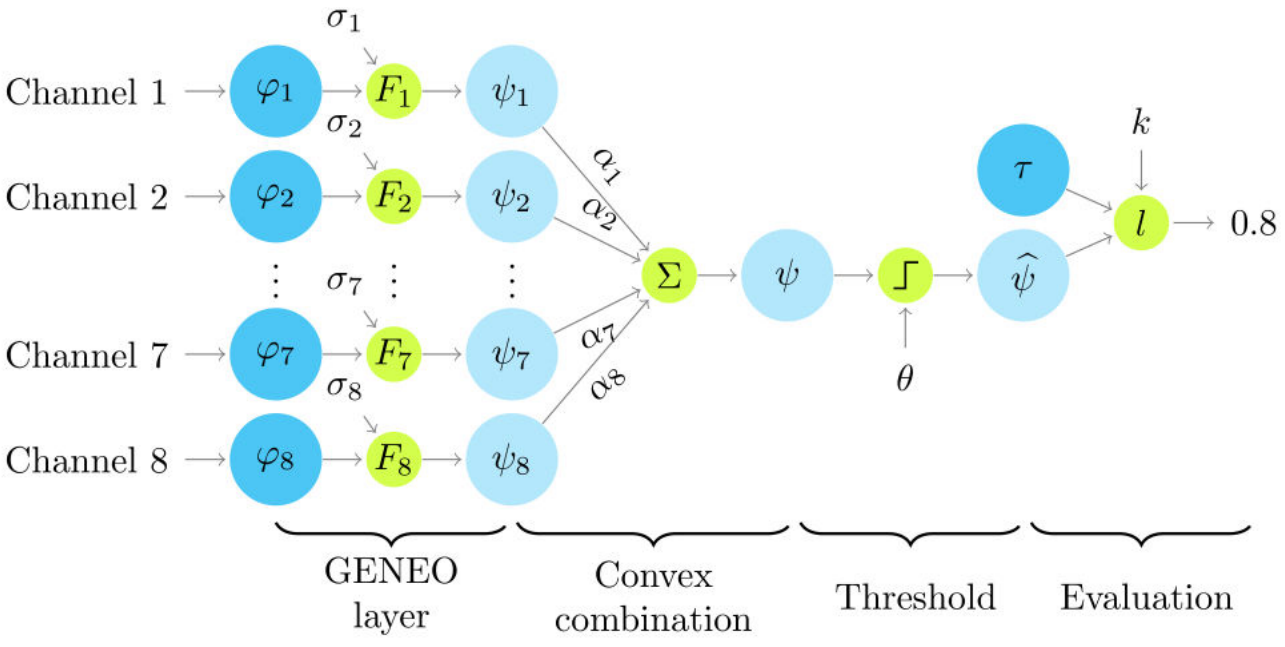


Figure: GENEOnet model

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 $\mathbb{R}^3$.

The model, called GENEOnet[2], has just 17 parameters: 8 shape parameters of convolutional kernels, 8 parameters of the convex combination and one threshold parameter.

Optimization has been carried out in a neural network style using a form of Gradient Descent.

[2] 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.)

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.

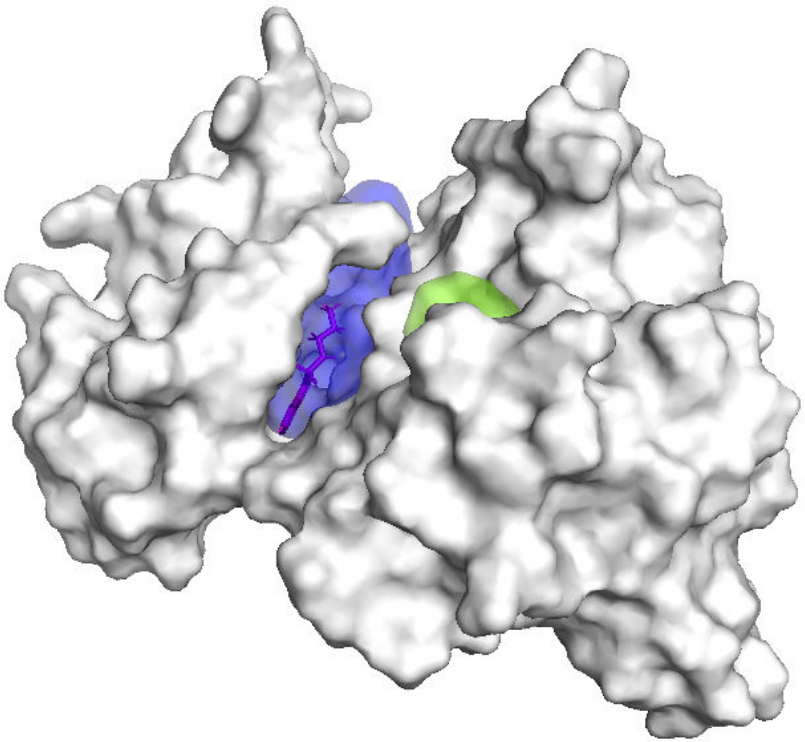


Figure: Prediction for protein 2KCE with ligand matching

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.

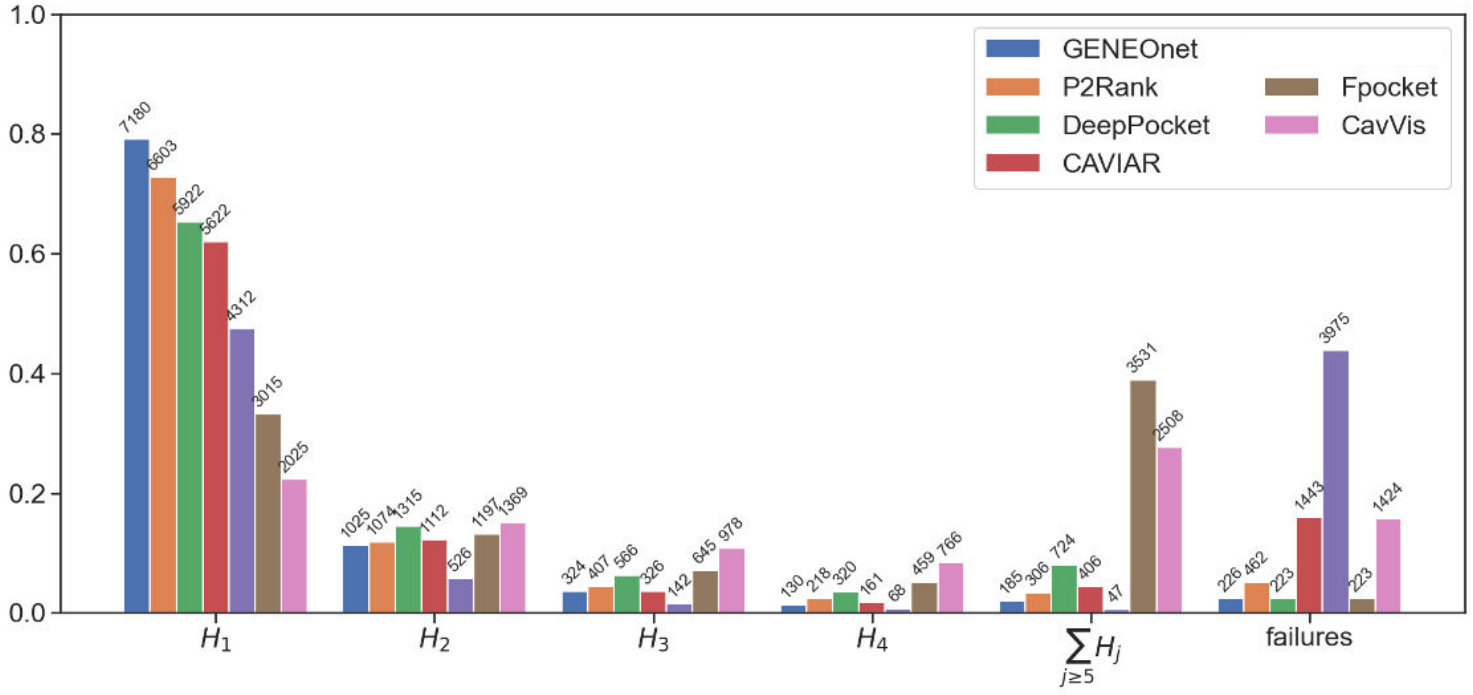
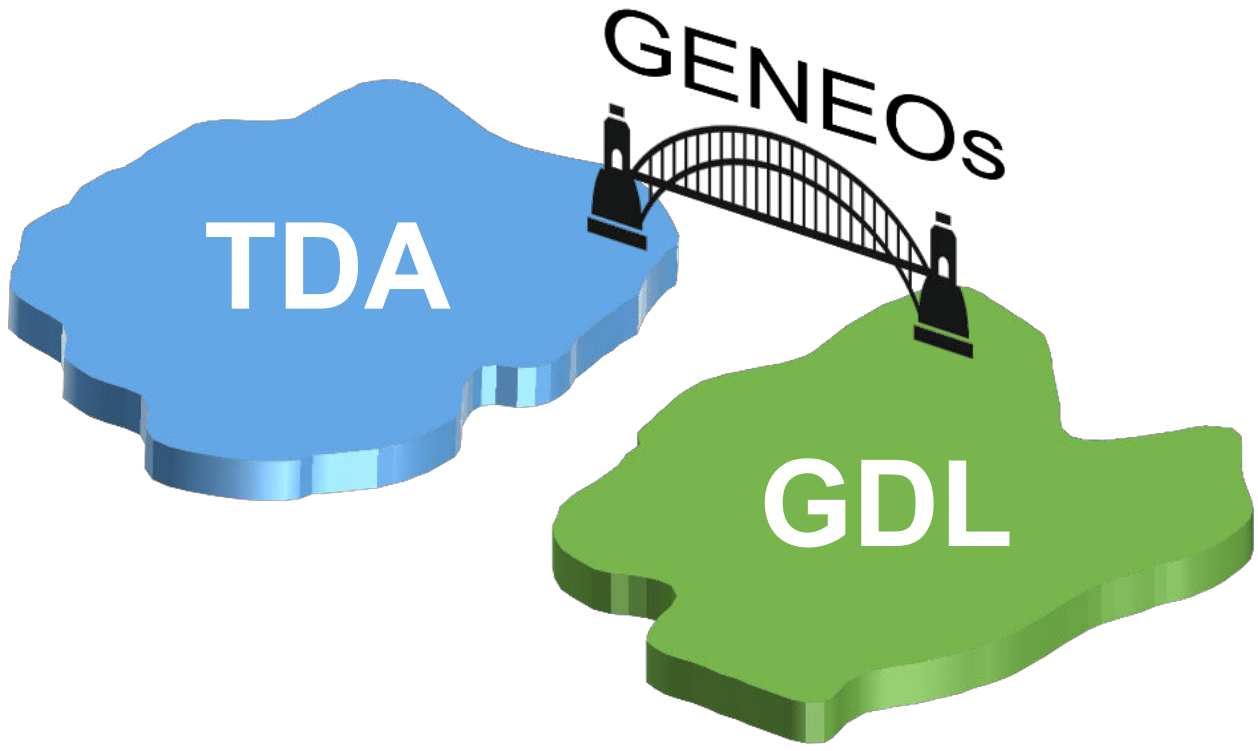


Figure: Comparison of different methods

Take home message



Thank you for your attention!

References I

Theory of GENEOs

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.

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.

References II

Case studies


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.


G. Bocchi, S. Botteghi, M. Brasini, *et al.*, "On the finite representation of group equivariant operators via permutant measures," 2022. arXiv: 2008.06340v2.