On the use of group equivariant non-expansive operators for topological data analysis and geometric deep learning

Patrizio Frosini

Department of Mathematics, ALMA-AI, AM^2 and ARCES, University of Bologna
patrizio.frosini@unibo.it

Big Data challenges for Mathematics: state of the art and future perspectives, February 25, 2022

Outline

Data analysis is not just about data

Topological and metric basics for the theory of GENEOs

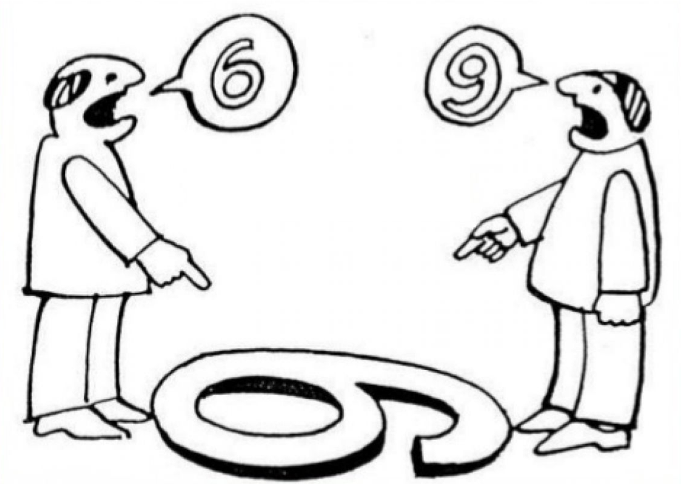
Compactness and convexity of the space of GENEOs

Methods to build GENEOs

How can we use GENEOs in applications?

Data analysis is not just about data

Data interpretation depends on the observer:



Observers are often more important than data

We are usually not directly interested in data, but in data observers. For example, a patient is usually interested not in the data representing a computerized axial tomography of her body, but in the diagnosis that her doctor can make from these data.

Data analysis strongly depends on the chosen observer. If data analysis were not dependant on the chosen observer, then physicians' diagnoses would always be identical, scientists would always see the same causes for each phenomenon, and all people would agree in judging who the heroes and villains in a movie or a political event are.

It is indeed well known that different agents can have different reactions in the presence of the same data, and this suggests that **data analysis should study the pairs (data, observer) instead of data alone**.

What are data?

Data are usually produced by measurements (or actions) made by observers. Before proceeding, we have to determine what measurements are in our mathematical model.

Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events.

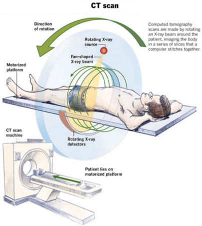
WIKIPEDIA

According to this definition, measurements (and hence data) can be often seen as functions φ associating a real number $\varphi(x)$ with each point x of a set X of characteristics. (This definition admits a natural extension to vector-valued functions but, for the sake of simplicity, we will treat here the case of scalar-valued functions).

Data are measurements made by observers

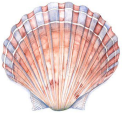
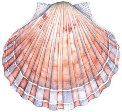
Some examples of data that can be seen as measurements (i.e., functions):

- An electrocardiogram (a function from $\mathbb{R}$ to $\mathbb{R}$);
- A gray-level image (a function from $\mathbb{R}^2$ to $\mathbb{R}$);
- A computerized tomography (CT) scan (a function from a helix to $\mathbb{R}$).

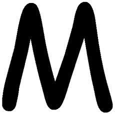
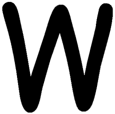


Observers are often associated with invariance groups

Observers often think that some data are equivalent to each other, according to an invariance group.

shape of  **= shape of** 

The group G is not established once and forever: when the observer changes, G changes too.

shape of  **≠ shape of** 

Data equivalence w.r.t. a group of permutations

Our data are represented by real-valued functions.

What do the expressions "data equivalence" and "data similarity" mean in our setting?

Two functions $\varphi_1, \varphi_2 : X \to \mathbb{R}$ *are* **equivalent** *with respect to a group* G *of permutations on* X *if a* $g \in G$ *exists, such that* $\varphi_1 = \varphi_2 \circ g$*.*

Two functions $\varphi_1, \varphi_2 : X \to \mathbb{R}$ *are* **similar** *with respect to a group* G *of permutations on* X *if a* $g \in G$ *exists, such that* $\|\varphi_1 - \varphi_2 \circ g\|_\infty$ *is small.*

These observations lead us to define the concept of *natural pseudo-distance* with respect to the group G.

The natural pseudo-distance d_G

Let X and G be a topological space and a subgroup of the group $\mathrm{Homeo}(X)$ of all homeomorphisms from X to X, respectively. Let us assume that φ_1, φ_2 are two continuous and bounded functions from X to $\mathbb{R}$, and consider the value $\boxed{\inf_{g \in G} \|\varphi_1 - \varphi_2 \circ g\|_\infty}$.

This value is called the *natural pseudo-distance* $d_G(\varphi_1, \varphi_2)$ between φ_1 and φ_2 with respect to the group G.

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

We could look at d_G as the ground truth in data comparison, when data equivalence is expressed by the group G.

The natural pseudo-distance d_G

If G is the trivial group Id, then d_G is the max-norm distance $\|\varphi_1 - \varphi_2\|_\infty$. Moreover, if G_1 and G_2 are subgroups of $\mathrm{Homeo}(X)$ and $G_1 \subseteq G_2$, then

$$d_{\mathrm{Homeo}(X)}(\varphi_1, \varphi_2) \leq d_{G_2}(\varphi_1, \varphi_2) \leq d_{G_1}(\varphi_1, \varphi_2) \leq \|\varphi_1 - \varphi_2\|_\infty$$

for every $\varphi_1, \varphi_2 \in C^0(X, \mathbb{R})$.

We usually restrict d_G to $\Phi \times \Phi$, where Φ is a bounded subset of $C^0(X, \mathbb{R})$.

Our general assumptions about data and observers

Our mathematical model is based on these assumptions:

- The space of observers is often more important than the space of data;
- The study of the space of observers requires the development of a new topological-geometric model.
- This new model could be of great use in data analysis, when the role of the observers is not negligible.

These assumptions suggest us to move from **Topological Data Analysis** to the new field of **Topological Observer Analysis**.

Observers can be seen as equivariant operators

Observers are structures able to change data into other data, and usually do that by respecting some data equivalences, i.e., by commuting with some transformations.

As a first approximation, observers can be represented as group equivariant operators (GEOs).

In this talk we will give some results on the theory of **Group Equivariant Non-Expansive Operators** (**GENEOs**).

Why "non-expansive"?
Because

1. observers are often assumed to simplify the metric structure of data in order to produce meaningful interpretations;
2. non-expansiveness guarantees good topological properties.

Data analysis is not just about data

Topological and metric basics for the theory of GENEOs

Compactness and convexity of the space of GENEOs

Methods to build GENEOs

How can we use GENEOs in applications?

Measurements as admissible functions

Let X be a nonempty set. Let Φ be a topological subspace of the set $\mathbb{R}_b^X$ of all bounded functions φ from X to $\mathbb{R}$, endowed with the topology induced by the metric

$$D_\Phi(\varphi_1, \varphi_2) := \|\varphi_1 - \varphi_2\|_\infty . \tag{0.1}$$

We can see X as the space where we can make our measurements, and Φ as the space of all possible measurements. We will say that Φ is the *set of admissible functions*. In other words, **Φ is the set of all functions from X to $\mathbb{R}$ that can be produced by our measuring instruments (or other observers)**. For example, a gray-level image can be represented as a function from the real plane to the interval $[0, 1]$ (in this case $X = \mathbb{R}^2$).

We recall that the initial topology τ_{in} on X with respect to Φ is the coarsest topology on X such that every function φ in Φ is continuous.

A pseudo-metric on X

Let us define on X the pseudo-metric

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

D_X induces a topology τ_{D_X} on X.

Theorem

The topology τ_{D_X} is finer than the initial topology τ_{in} on X with respect to Φ. If Φ is totally bounded, then τ_{D_X} coincides with τ_{in}.

The use of D_X implies that we can distinguish two points only if a measurement exists, taking those points to different values.

Theorem

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

Some magic happens: each bijection is an isometry

Let $\mathrm{Bij}(X)$ be the group of all bijections from X to X, and denote by $\mathrm{Bij}_{\Phi}(X)$ the subgroup of all $g \in \mathrm{Bij}(X)$ such that $\varphi \circ g \in \Phi$ and $\varphi \circ g^{-1} \in \Phi$ for every $\varphi \in \Phi$. Let $\mathrm{Homeo}(X)$ be the group of all homeomorphisms from X a X with respect to D_X, and denote by $\mathrm{Homeo}_{\Phi}(X)$ the subgroup of all $g \in \mathrm{Homeo}(X)$ such that $\varphi \circ g \in \Phi$ and $\varphi \circ g^{-1} \in \Phi$ for every $\varphi \in \Phi$. Let $\mathrm{Iso}(X)$ be the group of all isometries from X a X, and denote by $\mathrm{Iso}_{\Phi}(X)$ the subgroup of all $g \in \mathrm{Iso}(X)$ such that $\varphi \circ g \in \Phi$ and $\varphi \circ g^{-1} \in \Phi$ for every $\varphi \in \Phi$.

Proposition

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

A pseudo-metric on G

Let us now focus our attention on a subgroup G of $\mathrm{Homeo}_{\Phi}(X)$. We can define a pseudo-metric D_G on G by setting

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

Theorem

G is a topological group with respect to D_G and the action of G on Φ by right composition is continuous.

Theorem

If Φ is compact and G is complete then it is also compact with respect to D_G.

GEOs and GENEOs

Each pair (Φ, G) with $G \subseteq \mathrm{Homeo}_{\Phi}(X)$ is called a *perception pair*.

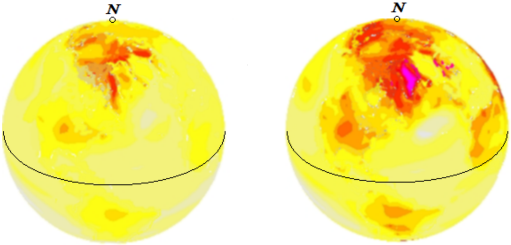
Let us assume that two perception pairs (Φ, G), (Ψ, H) are given, and fix a group homomorphism $T : G \to H$.

Each function $F : \Phi \to \Psi$ such that $\boxed{F(\varphi \circ g) = F(\varphi) \circ T(g)}$ 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., $\boxed{D_{\Psi}(F(\varphi_1), F(\varphi_2)) \leq D_{\Phi}(\varphi_1, \varphi_2)}$ for every $\varphi_1, \varphi_2 \in \Phi$), then F is called a *Group Equivariant Non-Expansive Operator (GENEO)* associated with the homomorphism T.

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 $(\Phi, G), (\Psi, H)$ 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 $[a, b]$
- $H =$ group of rotations of S^1

An example of GENEO

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

- $T(g)$ is the rotation $h \in H$ 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 $T : G \to H$.

In plain words, our GENEO transforms "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 $T : G \to H$ has been fixed.
Let us define a metric D_{GENEO} on $\mathrm{GENEO}((\Phi, G),(\Psi, H))$ 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 $\mathrm{GENEO}((\Phi, G),(\Psi, H))$ is compact with respect to D_{GENEO}.

Theorem

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

Elementary methods to build GENEOs

Proposition (Composition)

If $F_1 \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* $T_1 : G \to H$ *and* $F_2 \in \mathrm{GENEO}((\Psi, H), (\chi, K))$ *w.r.t.* $T_2 : H \to K$ *then* $F_2 \circ F_1 \in \mathrm{GENEO}((\Phi, G), (\chi, K))$ *w.r.t.* $T_2 \circ T_1 : G \to K$.

Proposition (Image by a 1-Lipschitz function)

If $F_1, \ldots, F_n \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* $T : G \to H$, *L is a 1-Lipschitz map from* $\mathbb{R}^n$ *to* $\mathbb{R}$, *and* $L^*(F_1, \ldots, F_n)(\Phi) \subseteq \Phi$ *(where* L^* *is the map induced by L), then* $L^*(F_1, \ldots, F_n) \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* T.

The next three statements follow from the last proposition.

Elementary methods to build GENEOs

Proposition (LATTICE OF GENEOS)

If $F_1, \ldots, F_n \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* $T : G \to H$ *and* $\max(F_1, \ldots, F_n)(\Phi), \min(F_1, \ldots, F_n) \subseteq \Phi$*, then* $\max(F_1, \ldots, F_n), \min(F_1, \ldots, F_n) \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* T.

Proposition (Translation)

If $F \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* $T : G \to H$*, and* $F_b(\Phi) \subseteq \Phi$ *for* $F_b(\varphi) := F(\varphi) - b$*, then* $F_b \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* T.

Proposition (Convex combination)

If $F_1, \ldots, F_n \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* $T : G \to H$*,* $(a_1, \ldots, a_n) \in \mathbb{R}^n$ *con* $\sum_{i=1}^{n} |a_i| \leq 1$ *and* $F_\Sigma(\Phi) \subseteq \Phi$ *for* $F_\Sigma(\varphi) := \sum_{i=1}^{n} a_i F_i(\varphi)$*, then* $F_\Sigma \in \mathrm{GENEO}((\Phi, G), (\Psi, H))$ *w.r.t.* T.

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 conjugacy action of G (i.e., $\mu(H) = \mu(gHg^{-1})$ for every $g \in G$).

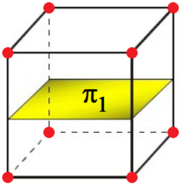
Proposition

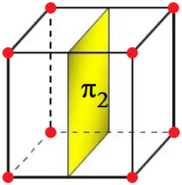
If μ is a permutant measure with respect to G, then the map $F_\mu : \mathbb{R}^X \to \mathbb{R}^X$ defined by setting $F_\mu(\varphi) := \sum_{h \in \mathrm{Bij}(X)} \varphi h^{-1} \, \mu(h)$ is a linear GEO.

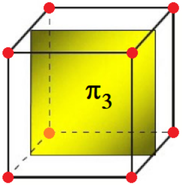
An example of permutant measure

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 $h \notin \{h_1, h_2, h_3\}$.







Building GENEOs by permutant measures

The following representation theorem holds.

Theorem

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 $(\mathbb{R}^X, G)$ to $(\mathbb{R}^X, G)$ (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 $\sum_{h \in \mathrm{Bij}(X)} |\mu(h)| \leq 1$.

Further details can be found in this preprint:

S. Botteghi, M. Brasini, P. Frosini and N. Quercioli, On the finite representation of group equivariant operators via permutant measures **https://arxiv.org/pdf/2008.06340.pdf**

Data analysis is not just about data

Topological and metric basics for the theory of GENEOs

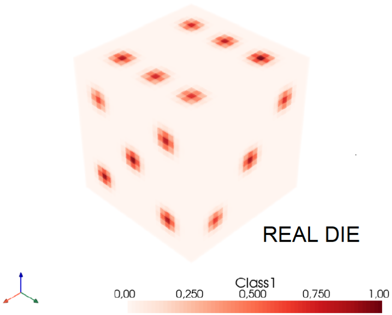
Compactness and convexity of the space of GENEOs

Methods to build GENEOs

How can we use GENEOs in applications?

What happens when we apply GENEOs to our data?

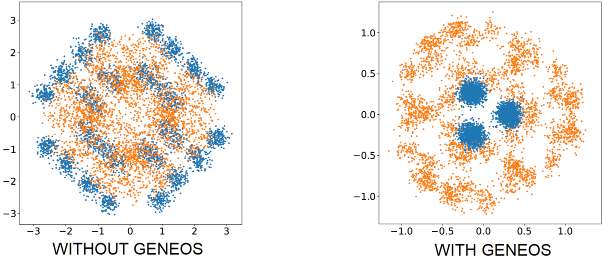
An example of use: comparison between real dice and fake dice.



(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:



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.

A real application: finding pockets in proteins

GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. https://arxiv.org/ftp/arxiv/papers/2202/2202.00451.pdf

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

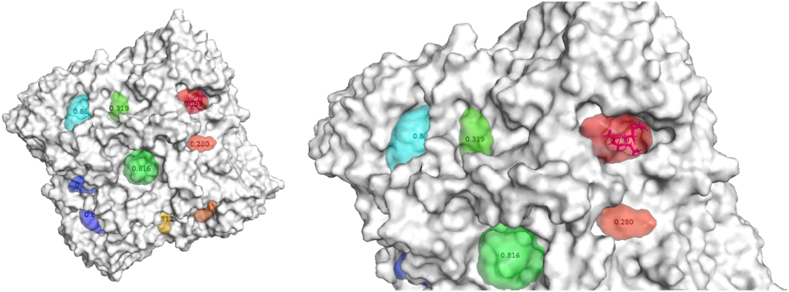
[2] Department of Mathematics, Università degli Studi di Bologna

[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

A real application: finding pockets in proteins



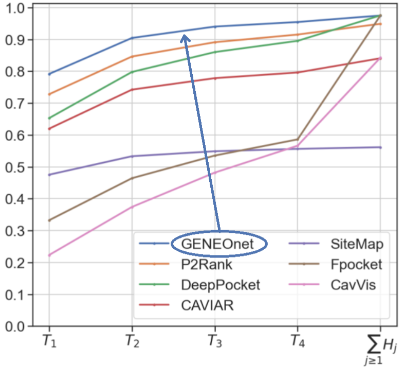
a) b)

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.

The search for the pockets was carried out by identifying an optimal GENEO in the convex hull of 8 GENEOs (each focused on a particular property of the pockets).

A real application: finding pockets in proteins

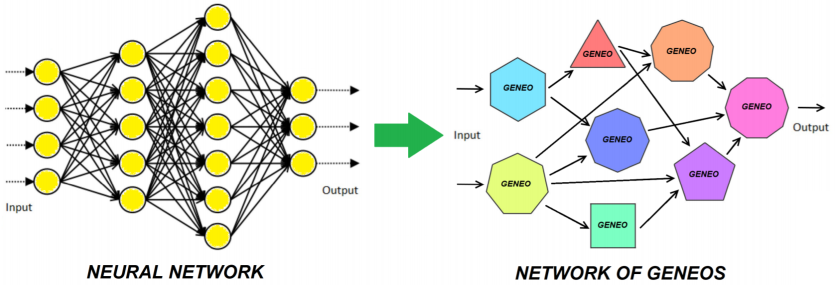
Here are the results of our experiments:



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

The main point in our approach

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.



Open questions

- How can we approximate a real observer (let us say, e.g., a physician) by GENEOs, in order to emulate her behaviour with respect to data?
- Can we devise constructive procedures, allowing us to build any possible GENEO with respect to a given equivariance group?
- What is the right way of comparing GENEOs in a topological-statistical setting?
- How should we select representative sets in a probability space of GENEOs?
- How can we compute the basic statistics for GENEOs?
- How can we predict the behaviour of networks of GENEOs and control their actions?
- How can we evaluate advantages and limits of an approach to data analysis based on the interaction of GENEOs and TDA?

SOME REFERENCES

- 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, n. 9, 423–433 (2 September 2019)
- F. Conti, P. Frosini, N. Quercioli, *On the construction of Group Equivariant Non-Expansive Operators via permutants and symmetric functions*, **Frontiers in Artificial Intelligence**, vol. 5 (2022), 1-11
- G. Bocchi, P. Frosini, A. Micheletti, A. Pedretti, C. Gratteri, F. Lunghini, A. R. Beccari, C. Talarico, *GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection*, 2022 https://arxiv.org/abs/2202.00451

THANKS FOR YOUR ATTENTION