Explainability of neural networks through the use of GENEOs

Patrizio Frosini

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

Workshop on Applied Geometry and Topology for Data Sciences
February 24-28, 2025, Shanghai

Outline

Collaborators

Joint work with:

- Filippo Bonchi (University of Pisa)
- Jacopo Joy Colombini (Scuola Normale Superiore, Pisa)
- Francesco Giannini (Scuola Normale Superiore, Pisa)
- Roberto Pellungrini (Scuola Normale Superiore, Pisa)

What is a GENEO?

Some basics on the theory of GENEOs

GENEOs and XAI

What is a GENEO?

- A Group Equivariant Non-Expansive Operator (GENEO) is a mathematical tool used to approximate observers that act on data.
- The theory of GENEOs is based on the idea that the geometric characteristics of observers significantly influence the interpretation of data.
- In this talk, we will explore the core properties of GENEOs, examine their role in machine learning, and discuss their promising applications in explainable artificial intelligence.

Let's start by defining perception spaces

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 an (extended) 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.*

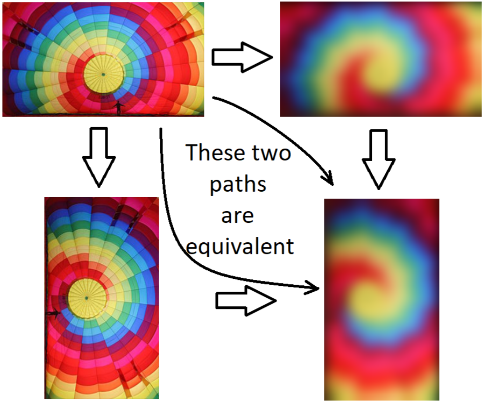
GEOs and GENEOs

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).

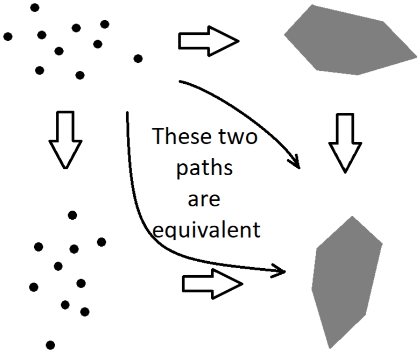
An example of GENEO

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 (T is the identity taking the **group of isometries** to itself).



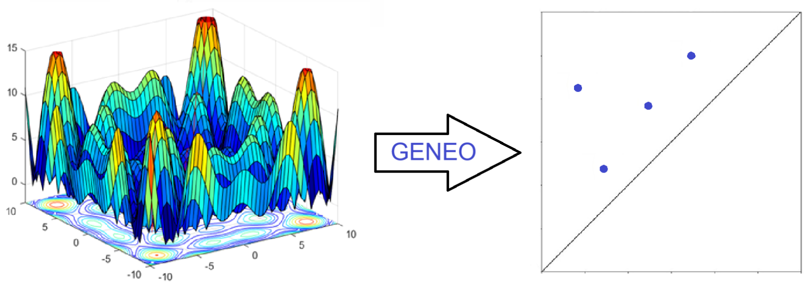
Another example of GENEO

When we compute the **convex hull** of a cloud of points, we are applying a GENEO (here, T is the identity taking the **group of isometries** to itself).



Another example of GENEO

The operator taking filtering functions to persistence diagrams is another example of GENEO.

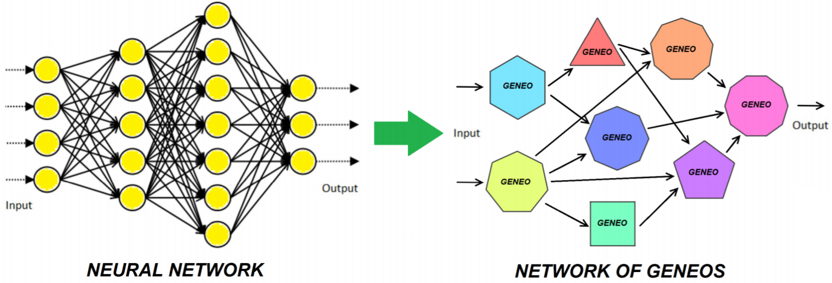


Why are GENEOs interesting?

- 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.

The main point in the approach based on GENEOs

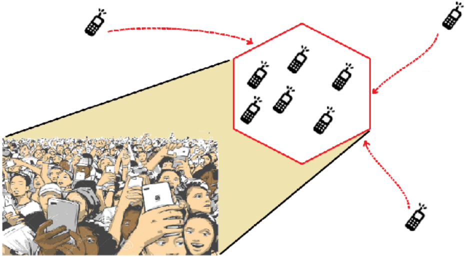
In perspective, we are looking for a good compositional theory for building efficient and transparent networks of GENEOs.
Some preliminary experiments suggest that replacing neurons with GENEOs could make some applications in deep learning more transparent and interpretable and speed up the learning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



Current research projects (I)

CNIT / WiLab - Huawei Joint Innovation Center (JIC)

Project on GENEOs for 6G



WILAB HUAWEI

Current research projects (II)

PANDORA

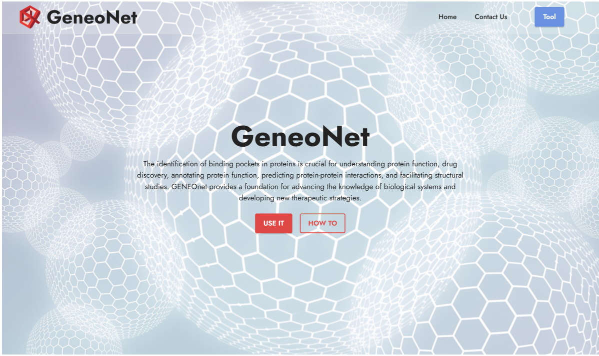
Horizon Europe (HORIZON)
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Project: 101135775-PANDORA
Funding: approximately 9 million euros.

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

https://pandora-heu.eu/consortium/

Current research projects (III)



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

Some references about GENEOs (I)

- P. Frosini, G. Jabłoński, *Combining persistent homology and invariance groups for shape comparison*, **Discrete & Computational Geometry**, vol. 55 (2016), n. 2, pp. 373-409.
- 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 (2019), pp. 423-433.
- Giovanni Bocchi, Stefano Botteghi, Martina Brasini, Patrizio Frosini and Nicola Quercioli, *On the finite representation of linear group equivariant operators via permutant measures*, **Annals of Mathematics and Artificial Intelligence**, vol. 91 (2023), n. 4, pp. 465-487.

Some references about GENEOs (I)

- Alessandra Micheletti, *A new paradigm for artificial intelligence based on group equivariant non-expansive operators*, **European Mathematical Society Magazine**, 128 (2023), pp. 4–12.
- 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*, **xAI-2024 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings**, Valletta, Malta, July 17-19, 2024, (Edited by Luca Longo, Weiru Liu, Grégoire Montavon), pp. 217-224.
- G. Bocchi, M. Ferri, P. Frosini, *A novel approach to graph distinction through GENEOs and permutants*, **Scientific Reports**, 15, 6259 (2025).

Basic idea

How can we mathematically and generally formalize the concept of an explanation provided by an agent, viewed as an operator?

Informal idea: We could say that the action of an agent A is explained by another agent B from the perspective of an agent C if:

1. C perceives A and B as similar to each other;
2. C perceives B as less complex than A.

E.g., let's consider two neural networks represented as two GEOs.

Note that a GEO can take another GEO as an input.

Basic idea

How can we transform our informal idea into a precise mathematical model?

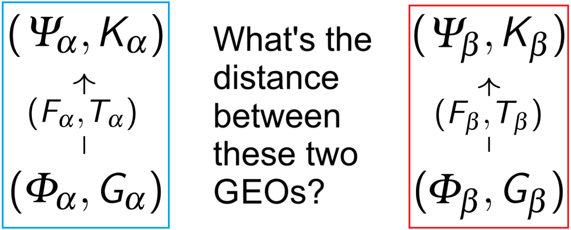
Let us begin by formalizing property 1.

Informal idea: We could say that the action of an agent A is explained by another agent B from the perspective of an agent C if:

1. C perceives A and B as similar to each other;
2. C perceives B as less complex than A.

An extended pseudo-metric for *ALL* GEOs

We have to introduce a pseudo-metric between GEOs that remains well-defined even when the GEOs operate on different domains and produce outputs in distinct codomains. This is a non-trivial challenge.



In other words, what does it mean for two GEOs to behave approximately the same way?

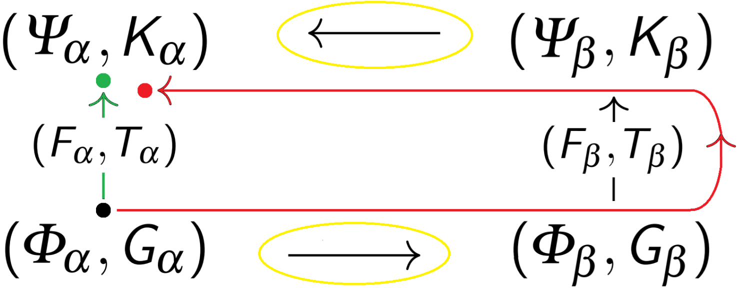
Our main goal: observer approximation

The previous pseudo-metric is necessary to build a geometric theory for approximating an ideal observer through GENEOs and GEOs.

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

An extended pseudo-metric for *ALL* GEOs

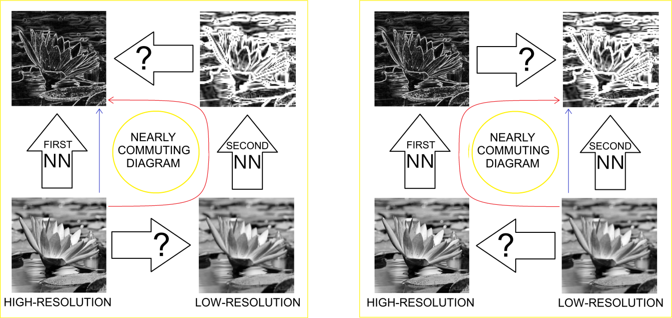
Informally speaking, two GEOs are considered similar if there exist two horizontal GENEOs that make this diagram "nearly commutative", with the same holding true in the opposite direction:



We can measure the non-commutativity of each diagram by a **cost function**.

An example

Suppose we have two neural networks for edge detection in images, represented as GEOs.

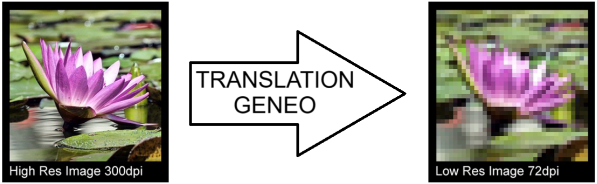


The two neural networks are considered close if there exist two pairs of horizontal GENEOs that make these diagrams "nearly commutative".

An extended pseudo-metric for *ALL* GEOs

To formalize our new pseudo-metric d_E between GEOs, let us consider the category $\mathbf{S}_{all}$ whose objects are all perception spaces, and whose morphisms $(F, T) : (\Phi, G) \to (\Phi', G')$ are GENEOs.
The morphisms in $\mathbf{S}_{all}$ are called *translation GENEOs*. These morphisms describe the possible "logical correspondences" between data represented by different perception spaces.

For example, a translation GENEO might transform high-resolution images into low-resolution images.



An extended pseudo-metric for *ALL* GEOs

Let us choose a set $\mathscr{G}$ of GEOs. Therefore,

$$\mathscr{G} = \{(F_\alpha, T_\alpha) : (\Phi_\alpha, G_\alpha) \to (\Psi_\alpha, K_\alpha)\}_{\alpha \in A}.$$

To proceed with the definition of our pseudo-metric on $\mathscr{G}$, we need to specify which logical correspondences between data we consider admissible. To this end, let us consider a small subcategory $\mathbf{S}$ of the category $\mathbf{S}_{all}$.

$\mathscr{G}$ will be the set of GEOs where we will define our pseudo-metric, while the morphisms in $\mathbf{S}$ will be the translation GENEOs considered admissible.

Definition of the explainability distance

Let

$$(F_\alpha, T_\alpha) : (\Phi_\alpha, G_\alpha) \to (\Psi_\alpha, K_\alpha)$$
$$(F_\beta, T_\beta) : (\Phi_\beta, G_\beta) \to (\Psi_\beta, K_\beta)$$

be two GEOs in the given set of GEOs $\mathscr{G}$.
Let us consider a pair

$$\pi = \Big((L_{\alpha,\beta}, P_{\alpha,\beta}), (M_{\beta,\alpha}, Q_{\beta,\alpha})\Big)$$

of morphisms in **S**, with

- $(L_{\alpha,\beta}, P_{\alpha,\beta})$ a morphism from (Φ_α, G_α) to (Φ_β, G_β),
- $(M_{\beta,\alpha}, Q_{\beta,\alpha})$ a morphism from (Ψ_β, K_β) to (Ψ_α, K_α),

Note that the two GENEOs have opposite directions. We say that π is a crossed pair of translation GENEOs from (F_α, T_α) to (F_β, T_β).

Definition of the explainability distance

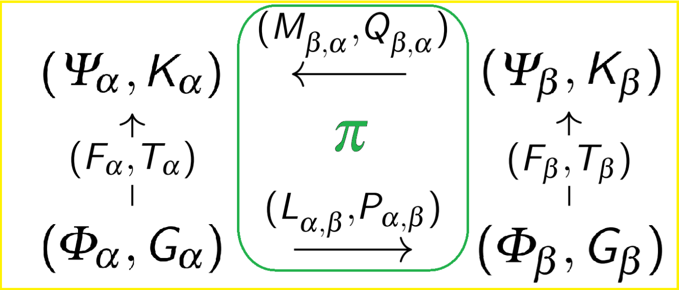


Figure: A crossed pair of translation GENEOs.

Definition of the explainability distance

To proceed, we need to equip each metric space Φ_α with a Borel probability measure μ_α. In simple terms, the measure μ_α represents the probability of the data points in Φ_α appearing in our experiments.

We will assume that all GENEOs in **S** are not just distance-decreasing (i.e., non-expansive) but also measure-decreasing, i.e., if $(L_{\alpha,\beta}, P_{\alpha,\beta}) : (\Phi_\alpha, G_\alpha) \to (\Phi_\beta, G_\beta)$ belongs to **S** and the set $A \subseteq \Phi_\alpha$ is measurable for μ_α, then $L_{\alpha,\beta}(A)$ is measurable for μ_β, and $\mu_\beta(L_{\alpha,\beta}(A)) \leq \mu_\alpha(A)$ (We recall that GENEOs are not surjective, in general).

Definition of the explainability distance

We also assume that the function that takes each $\varphi \in \Phi_\alpha$ to $f_{\alpha,\beta}(\varphi) := D_\Psi\Big((M_{\beta,\alpha} \circ F_\beta \circ L_{\alpha,\beta})(\varphi), F_\alpha(\varphi)\Big)$ is integrable with respect to the probability measure μ_α defined on the dataset Φ_α. The functional cost of π is defined by setting

$$\mathrm{cost}(\pi) := \int_{\Phi_\alpha} D_\Psi\Big((M_{\beta,\alpha} \circ F_\beta \circ L_{\alpha,\beta})(\varphi), F_\alpha(\varphi)\Big)\, d\mu_\alpha.$$

The value $\mathrm{cost}(\pi)$ quantifies how far the two paths in the next figure are from being equivalent, on average, when φ is randomly selected in Φ_α according to the probability measure μ_α.

Definition of the explainability distance

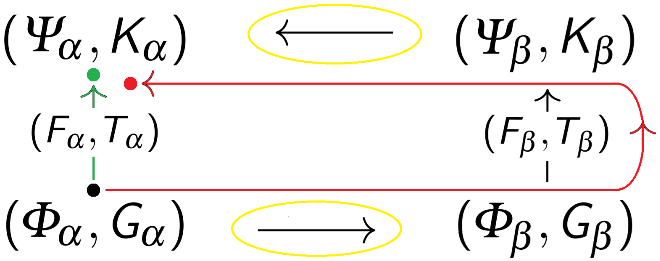
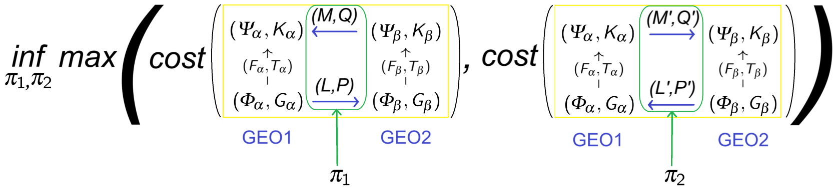


Figure: The explainability distance we are going to define measures how far the green path and the red path are from being equivalent, on average.

Definition of the explainability distance

We can formalize the new pseudo-metric d_E on $\mathscr{G}$ by defining $d_E(GEO1, GEO2)$ as the infimum of the maximum between the cost of π_1 and the cost of π_2, over all crossed pairs π_1 of admissible translation GENEOs from GEO1 to GEO2 and all crossed pairs π_2 of admissible translation GENEOs from GEO2 to GEO1.

Formally, $d_E(GEO1, GEO2)$ is equal to



Definition of the explainability distance

Proposition

d_E is an extended pseudo-distance.

The non-expansiveness of GENEOs is a key component in the definition of d_E.

In simple terms, the value $d_E((F_\alpha, T_\alpha),(F_\beta, T_\beta))$ measures the *cost* of changing (F_α, T_α) into (F_β, T_β).

When $d_E((F_\alpha, T_\alpha),(F_\beta, T_\beta))$ is small, it indicates that the GEOs (F_α, T_α) and (F_β, T_β) act approximately in the same way on the data they process, on average.

Back to the basic idea of explanation

Let us recall our informal idea.

Informal idea: We could say that the action of an agent A is explained by another agent B from the perspective of an agent C if:

1. C perceives A and B as similar to each other; ✓
2. C perceives B as less complex than A.

The formalization of 1 is completed using the pseudo-metric d_E.
How about the formalization of 2?

Complexity of GEOs

Let us assume a set $\Gamma = \{(F_i, T_i) : (\Phi_i, G_i) \to (\Psi_i, K_i)\}$ of GEOs is given. We will say that Γ is our internal library. For each GEO $(F_i, T_i) \in \Gamma$ we arbitrarily choose a value c_i representing the complexity $\mathrm{comp}((F_i, T_i))$ of (F_i, T_i).

Let us now consider the closure of Γ, i.e., the minimal set $\bar{\Gamma}$ such that

- $\bar{\Gamma} \supseteq \Gamma$;
- $\bar{\Gamma}$ is closed under composition (i.e., if $(F, T), (F', T') \in \bar{\Gamma}$ are composable, then $(F', T') \circ (F, T) \in \bar{\Gamma}$);
- $\bar{\Gamma}$ is closed under direct product (i.e., if the GEOs $(F, T), (F', T') \in \bar{\Gamma}$, then $(F, T) \otimes (F', T') \in \bar{\Gamma}$).

Each composition and direct product is associated with a complexity.

The complexity of each GEO $(F, T) \in \bar{\Gamma}$ is obtained by minimizing the sum of the complexities of the GEOs (F_i, T_i) that we use and the complexities of the compositions and direct products that we apply to build (F, T).

Back to the basic idea of explanation

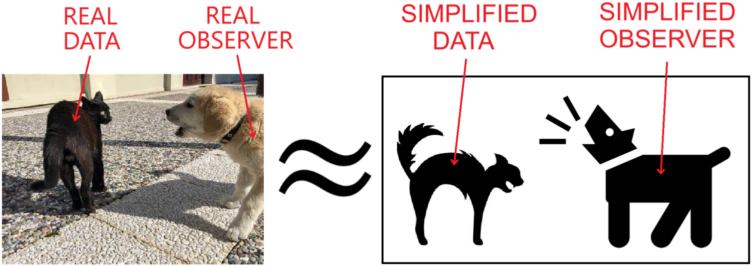
Informal idea: We could say that the action of an agent A is explained by another agent B from the perspective of an agent C if:

1. C perceives A and B as similar to each other; ✓
2. C perceives B as less complex than A. ✓

Our theoretical construction is now complete.

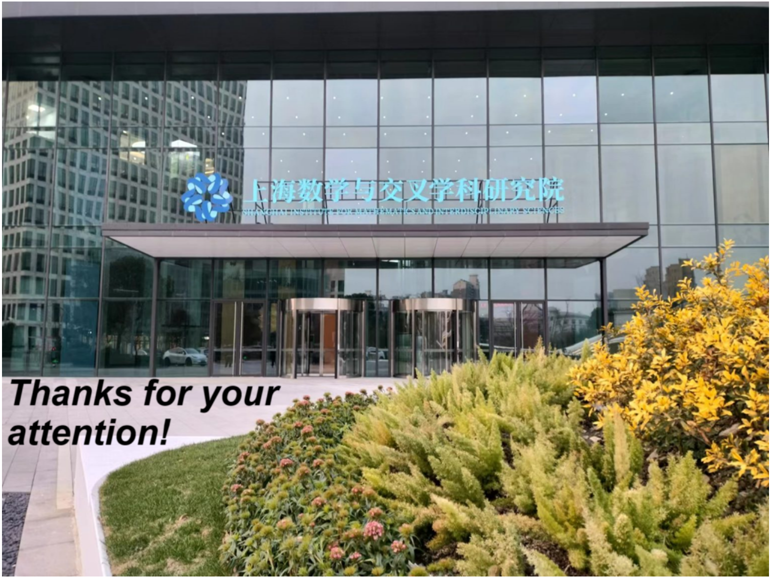
A mathematical concept of explanation

Now we can formalize our mathematical concept of **explanation**. Specifically, we can define it as follows: The action of an agent represented by a GEO (F_α, T_α) is explained **at a level ε** by the action of another agent of **complexity less than k** represented by a GEO $(F_\beta, T_\beta) \in \bar{\Gamma}$ when $d_E((F_\alpha, T_\alpha), (F_\beta, T_\beta)) \leq \varepsilon$.



Summary

To sum up, GENEOs are novel mathematical tools designed to approximate equivariant neural networks using a compositional approach. GENEOs are generally interpretable, making them potentially beneficial for explainable artificial intelligence (XAI).


上海数学与交叉学科研究院
Thanks for your attention!