Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study

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

[1]Dept. of Environmental Science and Policy, University of Milan

[2]Dept. of Mathematics, University of Bologna

[3]Dept. of Pharmaceutical Sciences, University of Milan

[4]Dompé Farmaceutici S.p.A.

Framework

Definition (GENEO)

A **Group Equivariant Non-Expansive Operator** F is a map between two functional spaces $\Phi = \{\varphi\colon X \to \mathbb{R}\}$ and $\Psi = \{\psi\colon Y \to \mathbb{R}\}$ that, given two subgroups $G \trianglelefteq \text{Homeo(X)}$, $H \trianglelefteq \text{Homeo(X)}$ and a fixed homomorphism of groups $T : G \to H$, has two properties:

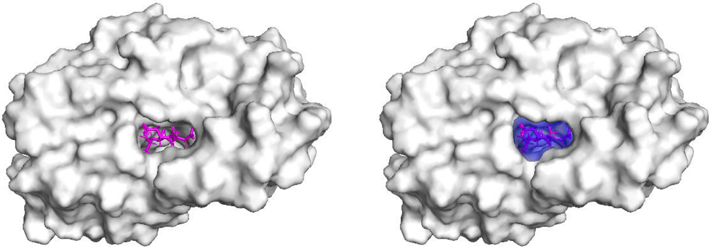
- **Equivariance:** $F(\varphi \circ g) = F(\varphi) \circ T(g)$ for all $\varphi \in \Phi$ and for all $g \in G$.
- **Non-Expansivity:** $||F(\varphi_1) - F(\varphi_2)||_\infty \leq ||\varphi_1 - \varphi_2||_\infty$ for all $\varphi_1, \varphi_2 \in \Phi$.

Why GENEOs ?

They have good **mathematical properties** and it's possible to create **networks** of GENEOs through some admissible operations.

Research Problem

The problem we investigated with the help of GENEOs comes from Medicinal Chemistry: the Problem of **Protein Pocket Detection**. The goal 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).

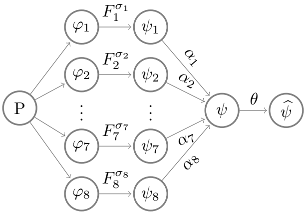


(a) Protein ID 1A30 (b) Protein ID 1A30, Pocket

Figure: Example of Protein Pocket Detection

Main Results

To solve the problem we developed a network of GENEOs called **GENEOnet**. The model is **simple** (very few trainable parameters) and thus **interpretable**, it allowed to incorporate **prior knowledge** about the problem and it's **equivariant** by design with respect to rigid motions of the space. It achieved comparable and slightly better results than other SOTA methods.



(a) Model architecture

Method	T_1	T_2	T_3	$\sum_{j\geq 1} H_j$
GENEOnet	**0.792**	**0.905**	**0.941**	0.975
P2Rank	0.728	0.847	0.892	0.952
DeepPocket	0.652	0.798	0.860	**0.978**
CAVIAR	0.616	0.739	0.783	0.837
Fpocket	0.331	0.462	0.534	**0.978**
CavVis	0.224	0.376	0.483	0.842

(b) Results

Figure: Model and Results

Conclusions

GENEOs constitute powerful mathematical tools for **Explainable Machine Learning** as they allow to build networks with desirable properties:

1. Very few trainable parameters.
2. Intrepretability of the parameter's values.
3. Possibility to incorporate prior knowledge.
4. Equivariance by design.
5. Fewer data are necessary for training.
6. Lower computational complexity compared to similar deep networks.

Main References

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.

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.