A new paradigm for Artificial Intelligence based on Group Equivariant Non-Expansive Operators - GENEOs

Giovanni Bocchi, Alessandra Micheletti

Abstract Artificial Intelligence (AI) is now pervasive in the everyday life. AI is quite often based on deep learning techniques. Deep learning continuosly shows to be very effective but on the other side its inherent opacity is also well known: deep learnig designers cannot always explain AI decisions that are not impossible to alter or counterfeit. thus the need for eXplainable Artificial Intelligence (XAI).

Key words: XAI, GENEOs, pocket detection.

1 Introduction

Equivariant operators are proving to be increasingly important in deep learning, in order to make neural networks more transparent and interpretable [2,7]. The use of such operators corresponds to the rising interest in the so called "explainable artificial intelligence" [6,14], which looks for methods and techniques whose functioning can be understood by humans. In accordance with this line of research, Group Equivariant Non-Expansive Operators (GENEOs) have been recently proposed as elementary components for building new kinds of networks [3-5].Their use is grounded in Topological Data Analysis (TDA) and guarantees good mathematical properties, such as compactness, convexity, and finite approximability, under suitable assumptions on the space of data and by choosing appropriate topologies.

More formally, a GENEO is a functional operator that transforms data into other data. By definition, it is assumed to commute with the action of given groups of

Giovanni Bocchi
Department of Environmental Science and Policy, University of Milan, via Saldini 50, 20133 Milano, Italy, e-mail: giovanni.bocchi1@unimi.it

Alessandra Micheletti
Department of Environmental Science and Policy, University of Milan, via Saldini 50, 20133 Milano, Italy, e-mail: alessandra.micheletti@unimi.it

transformations (equivariance) and to make the distance between data decrease (non-expansivity). The groups contain the transformations that preserve the "meaning" of our data, while the non-expansivity condition means that the operator must simplify the data metric structure. Both equivariance and non-expansivity are important: while equivariance reduces the computational complexity by exploiting symmetries of data, non-expansivity guarantees that the space of GENEOs can be finitely approximated.

In this paper we will introduce GENEOs and their main mathematical properties and we will show promising results obtained in an industrial application, namely protein pocket detection.

2 Basic definitions and properties of GENEO spaces

Let us now formalize the concept of GENEO, as was introduced in [3] . We assume that a space Φ of functions from a set X to $\mathbb{R}^k$ is given, together with a group G of transformations of X, such that if $\varphi \in \Phi$ and $g \in G$ then $\varphi \circ g \in \Phi$. We call the couple (Φ, G) *perception pair*. We also assume that Φ is endowed with the topology induced by the L_∞-norm $D_\Phi(\varphi_1, \varphi_2) = ||\varphi_1 - \varphi_2||_\infty,\ \varphi_1, \varphi_2 \in \Phi$. Let us assume that another perception pair (Ψ, H) is given, with Ψ endowed with the topology induced by the analogous L_∞-norm distance D_Ψ, and let's fix a homomorphism $T : G \to H$.

Definition 1. A map $F : \Phi \to \Psi$ is called a *group equivariant non-expansive operator (GENEO)* if the following conditions hold:

1. $F(\varphi \circ g) = F(\varphi) \circ T(g)$ for every $\varphi \in \Phi$, $g \in G$ (equivariance);
2. $\|F(\varphi) - F(\varphi')\|_\infty \leq \|\varphi - \varphi'\|_\infty$ for every $\varphi, \varphi' \in \Phi$ (non-expansivity).

If we denote by F_{all} the space of all GENEOs between (Φ, G) and (Ψ, H) and we introduce the metric

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

the following main properties of F_{all} can be proven (see [**?**] for the proofs).

Theorem 1. *If Φ and Ψ are compact, then F_{all} is compact with respect to the topology induced by D_{GENEO}.*

Theorem 2. *If Ψ is convex, then F_{all} is convex.*

Theorem 1 guarantee that if the spaces of data are compact, then also the space of GENEOs is compact, thus it can be well approximated by a finite number of representatives, reducing thus the complexity of the problem. Theorem 2 implies that if the space of data is also convex, then any convex combination of GENEOs is still a GENEO. Thus when both properties hold we have an easy instrument to obtain new GENEOs starting from a finite number.

3 GENEOnet

We used GENEOs to build *GENEOnet* [5] a geometrical explainable machine learning method to detect pockets on the surface of proteins which are likely to host ligans (i.e. drugs quite often). Protein pockets detection is a key problem in the context of drug development, since being able to identify only a small number of good sites, allows a scientist to restrict the action of virtual screening procedures, saving thus both computational resources and time. This research is ongoing, in collaboration with the Italian pharmaceutical company *Dompé Farmaucetici*.

This problem is particularly suitable to be treated with GENEOs: on one side there is some important empirical chemical-physical knowledge that can not be directly embedded in the usual machine learning techniques, but can be injected in a GENEO architecture, and, on the other side, the problem shows a natural equivariance property, since if we rotate or translate a protein, its pockets will be coherently transformed in the same way. This suggest that pocket detection is equivariant with respect to the group of spatial isometries.

For the application input data have been discretized by surrounding each molecule with a bounded region divided into a 3D grid of voxels. In this way the data are modelled as bounded functions from the Euclidean space $\mathbb{R}^3$ to $\mathbb{R}^d$. We chose $d = 8$, number of distinct geometrical, chemical and physical potential fields computed on each molecule and took into account for the analysis[1].

The input data are fed to a layer of GENEOs chosen from a set of parametric families of operators, each one parametrized by one shape parameter $\sigma_i, i = 1, \dots, 8$. These families were designed in order to include the a priori knowledge of the experts of medicinal chemistry. We opted for convolutional operators with L^1 normalized kernels: $F_k(\varphi) = \int_{\mathbb{R}^3} \varphi(x)k(x-y)\,dy$. The behavior of such operators is determined by their kernels, thus by making the i-th kernel dependent only on one shape parameter σ_i, we have direct control on the action of each operator. We mainly used Gaussian kernels[2]. Nonetheless all the kernels are rotationally invariant functions, this fact, together with the properties of convolution, guarantees that the corresponding operators satisfy the key requirement to be equivariant with respect to the group of isometries of $\mathbb{R}^3$.

In the second step the d operators are combined through a convex combination, with weights $\alpha_1, \dots, \alpha_d$, with $\alpha_i \in [0,1], \forall i$ and $\sum_{i=1}^{d} \alpha_i = 1$. The output of the convex combination operator is normalized to a function ψ from $\mathbb{R}^3$ to $[0,1]$. Here $\psi(x)$ can be read as the probability that a point $x \in \mathbb{R}^3$ belongs to a pocket. Finally, given a probability threshold $\theta \in [0,1]$, we get the different predicted pockets by taking the connected components of the superlevel set $\{\psi \geq \theta\} \subseteq \mathbb{R}^3$. The entire model pipeline is depicted in Figure 1. The model that was described so far has a total of 17

[1] See [5] for further details on the specific channels.

[2] $k(x) = C\exp\left(-\frac{||x||^2}{2\sigma_i}\right)$. Or kernels having shapes of spheres or of spherical crowns, assuming alternatively positive and negative values in different parts of the interior of the sphere or crown, and zero outside. See [5] for further details about the kernels.

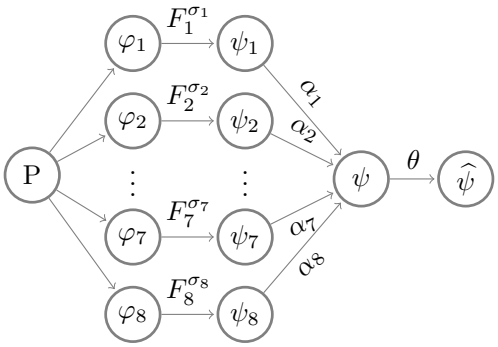

Fig. 1 Model workflow: input channels $\varphi_1,\dots,\varphi_8$ are fed to the GENEOs $F_1,\dots,F_8$ dependent on the shape parameters $\sigma_1,\dots,\sigma_8$. The intermediate outputs $\psi_1,\dots,\psi_8$ are combined through convex combination with weights $\alpha_1,\dots,\alpha_8$ to get the final result ψ. To get predictions a thresholding operation with a parameter θ is applied obtaining the binary function $\widehat{\psi}$

parameters. The fact that the model only employs convolutional operators, and their linear combinations, allowed us to set up an optimization pipeline quite similar to a 3D Convolutional Neural Network (CNN), but with two fundamental differences. First of all GENEOnet has a really tiny set of parameters[3]. Additionally the convolutional kernels of the GENEOs are not learned entry by entry as in classical CNNs (in this way equivariance would not be preserved), instead the kernels are computed at each step from the shape parameters that are updated during the optimization.

In order to identify the unknown parameters, we chose to optimize a cost function that evaluates the goodness of our predictions[4]. Eventually, after training, pockets are found as the connected components of the thresholded output of the model, resulting in a set of unranked pockets. Actually this representation is not much informative, since it is usually desirable to compute also the "druggability" of the identified cavities, that is a ranking score of the pockets on the basis of their fitness to host a ligand. Thus we devised a procedure to score pockets so that the output of the model consists in a list of pockets ranked by their corresponding scores[5].

That said, in order to identify the optimal model, we opted for a two-step optimization procedure: in the first step we generated $m = 200$ models $(\mathscr{M})_{k=1}^{m}$ optimized from $(T_k, IC_k)_{k=1}^{m}$, where T_k is a training set of size 200 subsampled from the whole dataset and IC_k are the randomly generated initial values of the parameters. In the second step each model was evaluated for it's scoring capabilities computing H_1 (see next paragraph) on a validation set in order to select the one with highest H_1. This final model was evaluated on an independent (both from the training and

[3] For comparison DeepPocket [1] , a recent approach that uses a 3D CNN to rescore fPocket [12] predictions, has 665 122 parameters.

[4] See [5] for further details about parameters optimization.

[5] See [5] for further details regarding pockets scoring.

the validation sets) test set to produce the results of next section.[6]

We compared the results of GENEOnet with other recent methods for protein pocket detection some of those based on ML techniques. We decided to base our comparison on the scores assigned by the different methods to the pockets. In this way performed a comparison based on the ability of the model to assign the highest scores to pockets that match the true ones. Given a dataset of proteins having only one ligand, and thus one "true pocket" each, we can compute the fraction of proteins whose true pocket is hit by the predicted one with highest score, by the one with second highest score and so on, obtaining coefficients[7] $H_j = \frac{1}{n}\sum_{k=1}^{n} \Delta_j(\tau_k, \mathscr{M}(P_k))$. Finally we computed the cumulative sum of these fractions $T_j = \sum_{i=1}^{j} H_i$. In this way different methods can be compared directly: if a model $\mathscr{M}$ is such that $T_j^{\mathscr{M}} \geq T_j^{\mathscr{M}'}$ for every $j \geq 1$ then $\mathscr{M}$ is better than $\mathscr{M}'$. The results reported in Table 1 show that GENEOnet has a better performance than all the other methods considered.

Table 1 T_j values for a subset of the methods in the comparison. See [5] for the full analysis.

Method	T_1	T_2	T_3	T_4	$\sum_{j\geq 1} H_j$
GENEOnet [5]	**0.792**	**0.905**	**0.941**	**0.955**	**0.975**
P2Rank [10]	0.728	0.846	0.891	0.915	0.949
DeepPocket [1]	0.653	0.798	0.860	0.895	**0.975**
CAVIAR [13]	0.620	0.743	0.779	0.797	0.841
SiteMap [9]	0.475	0.533	0.549	0.556	0.562
Fpocket [12]	0.332	0.464	0.535	0.586	**0.975**
CavVis [15]	0.223	0.374	0.482	0.566	0.843

Acknowledgements Funding from Dompè Farmaceutici S.p.A. to run this project is acknowledged by the authors.

References

1. Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Priyakumar, U. D.: DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. Journal of Chemical Information and Modeling (2021) doi: 10.1021/acs.jcim.1c00799
2. Anselmi, F., Evangelopoulos, G., Rosasco, L., and Poggio, T. : Symmetry-adapted representation learning. Pattern Recognition (2019) doi: 10.1016/j.patcog.2018.07.025

[6] Dataset size: 12995, Training set size: 200, Validation set size 3073 (48 proteins in the intersection with the training set), Test set size: 9070 (totally disjoint form the other sets). Protein data retrieved from PDBbind v2020 dataset [11].

[7] $\mathscr{M}(P)$ denotes the result of the model applied to a protein P. We say that the j-th scored predicted pocket hits the true pocket if it has the greatest overlap with the true one ($\iff \Delta_j(\tau, \mathscr{M}(P)) = 1$). If no predicted pocket has an intersection with the true one we say that the method failed on that protein

3. Bergomi, M. G., Frosini, P., Giorgi, D., and Quercioli, N.: Towards a topological–geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0087-3
4. Bocchi, G., Botteghi, S., Brasini, M., Frosini, P., and Quercioli, N.: On the finite representation of group equivariant operators via permutant measures. Annals of Mathematics and Artificial Intelligence (in press) (2023) doi: 10.1007/s10472-022-09830-1
5. Bocchi, G., Frosini, P., Micheletti, A., Pedretti, A. *et al.*: GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. (2022) preprint at arXiv:2202.00451.
6. Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.J. *et al.*: Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Scientific Reports (2021) doi: 10.1038/s41598-021-83922-6
7. Cohen, T. and Welling, M.: Group equivariant convolutional networks. In proceedings of the International Conference on Machine Learning (2016).
8. Conti, F., Frosini, P., and Quercioli, N.: On the Construction of Group Equivariant Non-Expansive Operators via Permutants and Symmetric Functions. Frontiers in Artificial Intelligence (2022) doi: 10.3389/frai.2022.786091
9. Halgren, T.: New method for fast and accurate binding-site identification and analysis. Chemical Biology & Drug Design (2007) doi: 10.1111/j.1747-0285.2007.00483.x
10. Krivak, R. and Hoksza, D.: P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics (2018) doi: 10.1186/s13321-018-0285-8
11. Liu, Z., Su, M., Han, L., Liu, J. *et al.*: Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Accounts of Chemical Research (2017) doi: 10.1021/acs.accounts.6b00491
12. Le Guilloux, V., Schmidtke, P., and Tuffery, P.: Fpocket: An open source platform for ligand pocket detection. BMC Bionformatics (2019) doi: 10.1186/1471-2105-10-168
13. Marchand, J.R., Pirard, B., Ertl, P., and Sirockin, F.: CAVIAR: a method for automatic cavity detection, description and decomposition into subcavities. Journal of Computer-Aided Molecular Design (2021) doi: 10.1007/s10822-021-00390-w
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0048-x
15. Simoes, T. M. C., and Gomes, A. J. P.: CavVis-A Field-of-View Geometric Algorithm for Protein Cavity Detection. Journal of Chemical Information and Modeling (2019) doi: 10.1021/acs.jcim.8b00572