

Prototypical Explanations in an AI method for Protein Pocket Detection

Giovanni Bocchi¹, Alessandra Micheletti¹, Carmen Gratteri², and Carmine Talarico³

¹ University of Milan, Via Giovanni Celoria 10, Milano, 20124, Italy
`giovanni.bocchi1@unimi.it, alessandra.micheletti@unimi.it`

² LIGHT S.c.a.r.l., Via Branze 45, Brescia, 25123, Italy
`carmen.gratteri.ext@exscalate.eu`

³ Dompé Farmaceutici S.p.A., Via Tommaso de Amicis 95, Napoli, 80123, Italy
`carmine.talarico@dompe.com`

Abstract. Prototype-based explanations are a well-established technique in eXplainable Artificial Intelligence (XAI), commonly used for classification problems like image classification. This study presents a novel approach to find prototypical explanations for a protein pocket detection algorithm previously developed by the authors. The method aims to identify pockets that were predicted in a way similar to a specific instance of interest, thereby providing insights into relevant information for medicinal chemists. To validate our approach, we tested it as a binary classification problem, distinguishing between similar and dissimilar pocket pairs using the ProSPECCTs benchmark. The results showed that our method outperformed other state-of-the-art methods, taking into account the uncertainty of the predictions due to variations in the training data. The proposed approach uses k -nearest neighbors in a d -dimensional latent space of pocket descriptors to identify prototypes linked to the instance being explained.

Keywords: Prototypes, XAI, pocket similarity, GENEIO.

1 Introduction

Prototypical explanations [9] have become increasingly crucial in eXplainable Artificial Intelligence (XAI) and S.A.F.E. AI [10, 7] for elucidating the workings of black box algorithms. These prototypes are examples from the dataset that closely resemble the instance for which a prediction needs explanation. By presenting similar, understandable instances, prototypes help reveal the rationale behind a model’s decision-making. Strategies for identifying prototypes include locating cluster centroids, finding nearest neighbors based on specific distances, or identifying proximate data when mapped in a latent space [11].

Recently, Group Equivariant Non-Expansive Operators (GENEOs) have been introduced as mathematical tools to model “observers of data” and to build machine learning models [1]. Moreover, GENEOs guarantee the explainability and

trustworthiness of the resulting models. In particular, one such model called GENEOnet [4] was developed as an application to the medicinal chemistry problem of protein pocket detection (for comprehensive reviews of protein pocket detection, also from a machine learning perspective, refer to [6]). The use of GENEOnet allowed us to make GENEOnet an intrinsically explainable model and, moreover, its robustness and trustworthiness have been already investigated in [5, 3]. In this paper, we present a new method to assess the similarity between pocket pairs. Once this method, which can be adapted to GENEOnet, can successfully distinguish similar pairs from dissimilar ones, it will be used to improve the explainability of GENEOnet by identifying prototypes for its predictions. Such prototypes are predicted pockets that are closest to the instance in question, according to their positions in a specifically designed latent descriptor space.

2 GENEOnet method

In this Section, we will briefly introduce GENEOnet and the GENEOnet method for protein pocket detection, and then we will describe the procedure we used to map binding sites to points in a latent space. Furthermore, we will show how to use such embedded points to compare pockets. We will then evaluate our technique against other state-of-the-art methods used to find pocket similarities. Finally, we will detail how to use the embedded points to obtain prototypical explanations for GENEOnet.

2.1 GENEOnet and GENEOnet

Group Equivariant Non-Expansive Operators are proving as powerful mathematical tools used in AI to build network models that maintain coherence when data undergoes geometric transformations. This property, called equivariance, ensures that the output remains consistent if the input data is transformed by a specific group of transformations. Another key characteristic of GENEOnet is non-expansivity, which guarantees they are stable and robust to minor changes in the data. Together, equivariance and non-expansivity thus provide high trustworthiness to models built using GENEOnet. Such models are usually quite powerful in spite of having simple architectures. To formally define GENEOnet, we start with a space Φ of functions from a set X to $\mathbb{R}$ and a group G of transformations of X . We also assume that Φ is equipped with the topology induced by the distance defined by the L_∞ -norm. Next, we have another pair (Ψ, H) and a group homomorphism $T: G \rightarrow H$, which connects the two groups in a way that allows us to use GENEOnet. (For a more detailed understanding of GENEOnet and their properties, please refer to [1]).

Definition 1. A map $F: \Phi \rightarrow \Psi$ 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).

GENEOnet [4] (available as a web service at <https://geneonet.exscalate.eu>) serves as a paradigmatic example of constructing a network employing GENEOS. It is a specialized network model, designed to detect protein pockets, featuring a shallow architecture composed of a small number of GENEIO units. Despite its relative simplicity and a small number of learnable parameters, GENEOnet has been shown to perform slightly better than state-of-the-art methods in identifying true pockets with top-ranked predictions. Simultaneously with the performances, we conducted an initial assessment of GENEOnet’s explainability and robustness [5].

2.2 GENEOnet

We will now briefly describe the procedure used to embed a protein pocket to a point in the Euclidean space $\mathbb{R}^d$. The procedure consists of the following steps:

1. Read the input protein file.
2. Isolate the binding site using the crystallized ligand or the GENEOnet prediction.
3. Represent the molecular surface as a mesh.
4. For each vertex, compute a set of local chemical features.
5. Calculate shape descriptors that remain invariant under rigid motions. Assemble all descriptors into a single vector $\eta \in \mathbb{R}^d$.
6. Standardize the vectors.
7. Apply a learnable linear map $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$ with $\|A\|_\infty = 1$ to the vectors.

The infinity norm of a $d \times d$ matrix $A = (a_{ij})_{i,j=1,\dots,d}$ is defined as: $\|A\|_\infty = \max_{i=1}^d \sum_{j=1}^d |a_{ij}|$. Notably, we emphasize that we are employing the infinity norm because the condition $\|A\|_\infty = 1$ means that for two sites represented by vectors η and $\tilde{\eta}$, the inequality

$$\|A\eta - A\tilde{\eta}\|_\infty \leq \|A\|_\infty \|\eta - \tilde{\eta}\|_\infty \leq \|\eta - \tilde{\eta}\|_\infty \quad (1)$$

is satisfied. Thus since the vector η is, by construction, invariant with respect to rigid motions of the space and thanks to Inequality 1, the operator that takes a binding site to its vector $A\eta$ is invariant with respect to rigid motions and non-expansive, and thus it is a GENEIO.

2.3 Evaluation on ProSPECCTs benchmark

The output of our method can be used to evaluate pocket similarity: let’s assume to have two proteins $P, \tilde{P}$ and call $A\eta, A\tilde{\eta}$ the corresponding outputs. We can consider the sup norm between $A\eta, A\tilde{\eta}$ as a score for pocket similarity: $s(P, \tilde{P}) = \|A(\eta - \tilde{\eta})\|_\infty$. We aim to design a method where proteins with similar pockets have a value of $s(P, \tilde{P})$ close to zero, and those with dissimilar pockets have a large value of $s(P, \tilde{P})$. In this way, by choosing a threshold value $\tau > 0$, we can derive a binary classifier by setting

$$\hat{g} = \mathbb{1}_{[\tau, +\infty)}(s(P, \tilde{P})), \quad (2)$$

where $\mathbb{1}_A(x) = 1$ if $x \in A$, and is 0 otherwise, so that $\hat{y} = 0$ corresponds to similar pockets, and $\hat{y} = 1$ corresponds to dissimilar pockets.

Thus, by using a benchmark collection of datasets, like ProSPECTTs [8], that contains previously labelled similar and dissimilar pockets, it is possible to compute conventional metrics for binary classification. A convenient evaluation metric is the Area Under Curve (AUC), which has been already used in previous studies involving ProSPECTTs. We recall that a random classifier has an AUC equal to 0.5 while larger values denote better performances than a random choice, having 1.0 as the optimal case.

We took as a reference the extensive comparison done in [2]. To include our method in their comparative study, we applied the approach outlined in Section 2.2 to learn an appropriate linear transformation A . We undertook the following steps on the 8 datasets contained in ProSPECTTs:

1. For each dataset, we split pocket pairs into a training and a validation part, where the latter consisted of 20% of the total.
2. From each training set, 500 similar and 500 dissimilar pairs were sampled (with possible replacement, in case the training split has fewer than 500 examples of the same class).
3. This set of $N = 8000$ pairs was the final training set for our method. All those pairs that were not sampled to build the global training set have been merged into the validation parts of each dataset.
4. Using only the global training set we optimized the following loss function with Adam optimizer:

$$\mathcal{L}(A) = \frac{1}{N} \sum_{i=1}^N \left(s((P, \tilde{P})_i) - y_i \right)^2 + \lambda \left(\|A\|_\infty - 1 \right)^2 \quad (3)$$

the first term is the data loss and the second one is a penalty term due to the constraint $\|A\|_\infty = 1$.

Moreover, steps from 1. to 4., were repeated 50 times with learning rate 1e-4, batch size 32, and for 600 epochs. The value of λ was fixed to 1/2 for each simulation. For each model and each dataset, we computed the ROC curve for both the corresponding validation set and the entire dataset (including the training part). To ensure an equitable comparison with the outcomes reported in a prior study [2], we opted to evaluate all methods utilizing the complete datasets.

Table 1 shows the comparison results. For brevity, we only report the comparison with Site2Vec, which was the best-performing method among the 21 compared in [2]. The results show that our approach has a slightly higher average AUC value across all datasets compared to Site2Vec. Furthermore, the mean AUC of Site2Vec is below the lower confidence interval for the mean AUC of our method, indicating that our approach may be more reliable. It is worth noting that uncertainty quantification is lacking in all methods compared in [2], making it difficult to assess their robustness. Interestingly, we achieved an improved outcome on D7, which pertains to effective pharmaceutical uses of site similarities within the context of drug design. As previously reported in [12], Site2Vec’s low

performance on this dataset raises significant concerns about its applicability in real-world drug discovery scenarios.

Table 1: AUC values for the 8 datasets in the ProSPECCTs benchmark.

Model	Dataset								Mean
	D1	D1.2	D2	D3	D4	D5	D6	D7	
GENEMap	0.91 ± 0.002	1.00 ± 0.000	0.96 ± 0.001	0.79 ± 0.002	0.80 ± 0.002	0.73 ± 0.004	0.90 ± 0.007	0.93 ± 0.003	0.88 ± 0.001
Site2Vec	1.00	0.94	1.00	0.99	0.99	0.86	0.53	0.66	0.87

2.4 Finding prototypes with GENEMap

Once our model for pocket similarity has been trained using the ProSPECCTs datasets, we can leverage it to obtain prototypes for predictions made by GENEMap. Before introducing prototypes, we want to stress that, unlike other methods for pocket similarity, GENEMap does not necessarily require the crystallized ligand as input, but it can run using GENEMap predictions. From a pharmaceutical perspective, this is significant because it enables us to accelerate drug design processes when dealing with proteins that do not yet have an annotated ligand. Indeed, by mapping GENEMap’s predicted pocket onto a database of known complexes of proteins and ligands (which are often drugs), we can identify, by similarity, potential new uses for existing and approved medications - a process called drug repurposing.

On the other hand, our method has significant value in the context of XAI because it enables us to provide further explanations for why GENEMap makes its predictions. When we have a pocket generated by GENEMap, we can use our method to identify the k nearest neighbors between proteins whose pockets are mapped in the same latent space as the prediction. These neighboring proteins, which may also have ligand annotations, can then be used to provide additional insights into why GENEMap generated that specific outcome. To illustrate this, consider the example shown in Fig. 1, where we applied our method to protein 3BVB, a mutation of HIV-1 protease. We found that all five nearest neighbors identified by GENEMap are different mutations of the same protein structure and bind to the same ligand (017); this finding would be particularly informative in analyzing the prediction for 3BVB if no additional information were available. Indeed, this example demonstrates how our approach can be used to provide further explanations for GENEMap’s predictions, making it more transparent and useful for medicinal chemistry experts.

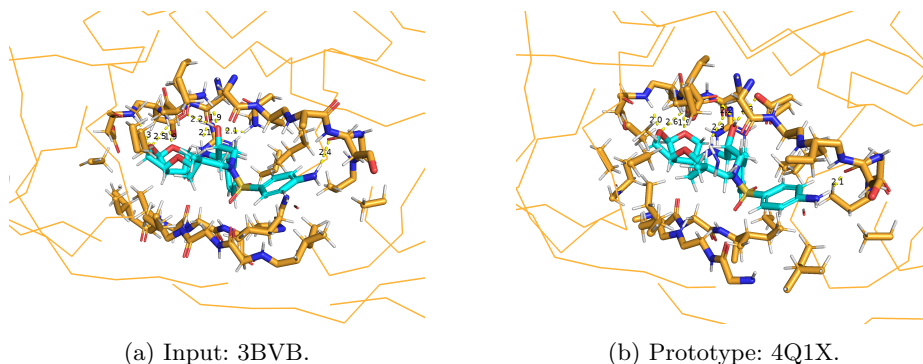


Fig. 1: First nearest neighbor of protein 3BVB's pocket, used as prototype

References

1. Bergomi, M.G., Frosini, P., Giorgi, D. et al. Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. *Nat. Mach. Intell.*, 1(9):423–433, 2019.
2. Bhadra, A. and Yeturu, K. Site2vec: a reference frame invariant algorithm for vector embedding of protein-ligand binding sites. *Mach. Learn. Technol.*, 2(1), 3 2021.
3. Bocchi, G., Frosini, P., Micheletti, A. et al. GENEOnet: statistical analysis supporting explainability and trustworthiness. *Statistics*, 0(0):1–26, 2025.
4. Bocchi, G., Frosini, P., Micheletti, A. et al. GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. preprint at arXiv, 2022.
5. Bocchi, G., Frosini, P., Micheletti, A. et al. A geometric XAI approach to protein pocket detection. In *Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium*, volume 3793, pages 217–224, 2024.
6. Di Palma, F., Abate, C., Decherchi, S. et al. Ligandability and druggability assessment via machine learning. *Wiley Interdiscip. Rev. Mol. Sci.*, 13(5), 2023.
7. Duke, T., Giudici, P. Responsible AI in Practice. Apress, Berkeley, CA, 2025.
8. Ehrt, C., Brinkjost, T. and Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (prospeccts). *PLOS Comput. Biol.*, 14(11):1–50, 11 2018.
9. Frixione, M. and Lieto, A. Prototypes vs exemplars in concept representation. In *KEOD*, 2012.
10. Giudici, P. Safe machine learning. *Statistics*, 58(3):473–477, 2024.
11. Guidotti, R., Monreale, A., Matwin, S. et al. Black box explanation by learning image exemplars in the latent feature space. In *ECML PKDD*, pages 189–205, 2020.
12. Reim, T., Ehrt, C., Graef, J. et al. SiteMine: Large-scale binding site similarity searching in protein structure databases. *Arch. Pharm.*, 357(5):2300661, 2024.