GENEOnet: Statistical analysis supporting explainability and trustworthiness.

Giovanni Bocchi*[1], Patrizio Frosini[2], Alessandra Micheletti[1], Alessandro Pedretti[3], Carmen Gratteri[4], Filippo Lunghini[5], Andrea R. Beccari[5], and Carmine Talarico[5]

[1]Dept. of Environmental Science and Policy, University of Milan
[2]Dept. of Computer Science, University of Pisa
[3]Dept. of Pharmaceutical Sciences, University of Milan
[4]LIGHT S.c.a.r.l.
[5]Dompé Farmaceuici S.p.A.

Group Equivariant Non-Expansive Operators (GENEOs) have emerged as mathematical tools for constructing networks for Machine Learning and Artificial Intelligence. Recent findings suggest that such models can be inserted within the domain of eXplainable Artificial Intelligence (XAI) due to their inherent interpretability. In this study, we aim to verify this claim with respect to GENEOnet, a GENEO network developed for an application in computational biochemistry by employing various statistical analyses and experiments. Such experiments first allow us to perform a sensitivity analysis on GENEOnet's parameters to test their significance. Subsequently, we show that GENEOnet exhibits a significantly higher proportion of equivariance compared to other methods. Lastly, we demonstrate that GENEOnet is on average robust to perturbations arising from molecular dynamics. These results collectively serve as proof of the explainability, trustworthiness, and robustness of GENEOnet and confirm the beneficial use of GENEOs in the context of Trustworthy Artificial Intelligence.

Keywords— GENEOs; sensitivity analysis; equivariance; robustness; feature importance;

*Corresponding author. Email: giovanni.bocchi1@unimi.it

1. Introduction

The confluence of medicinal chemistry and computational chemistry, coupled with advancements in drug design and discovery [1,2], presents numerous opportunities to investigate and evaluate the transparency and trustworthiness of Artificial Intelligence (AI) applications. In recent years, there has been a significant surge in AI solutions addressing computational biochemistry challenges [3]. These have included the development of algorithms for generating three-dimensional protein structures, [4–6] predicting protein-protein interactions [7], and forecasting protein-ligand interactions [8–10]. However, the pace at which explainable models or validations of the trustworthiness of such AI systems have been developed has not kept up with these innovations. As a result, questions regarding the transparency of these systems have often gone unanswered, leading to skepticism within certain segments of the scientific community [11,12].

One pressing challenge in this field is protein pocket detection [13,14], which involves identifying locations within the three-dimensional structure of a protein where small molecules known as ligands, which usually have the role of drugs, are likely to bind. Identifying protein pockets is crucial in drug development, as pinpointing a limited number of potential binding sites on a molecule's surface enables scientists to streamline virtual screening processes. This approach conserves computational resources, reduces time, and accelerates the later stages of development, which are thus based on a reduced number of laboratory tests. Notably, numerous machine learning and AI solutions have been proposed to tackle this problem [15–22]. However, as far as we are aware, none of these approaches has prioritised the transparency or trustworthiness of their underlying algorithms, casting a significant shadow over the reasons behind methods' failures, and thereby hindering further reduction of the laboratory experimental component. In addition, transparency is crucial, since it ensures that independent researchers can reproduce and verify AI-driven discoveries, strengthening the scientific basis for new treatments. Furthermore, drug development is heavily regulated by agencies like the FDA and EMA. If AI plays a role in the process, regulators need clear explanations of how the AI reached its conclusions to ensure safety and efficacy before human trials.

Protein pocket detection is a problem characterized by a notable geometrical property: if a protein structure undergoes a rigid motion, its pockets should not change apart from their location and orientation. Although such rigid motions are unlikely to occur in real-world scenarios, due to the dynamic nature of proteins, this property would be highly desirable for any model predicting protein pockets, in order to trust its predictions. Moreover, the dynamical behavior of proteins can be simulated using Molecular Dynamics (MD), a powerful tool that allows researchers to model complex protein movements [23–25]. This capability enables us to evaluate the robustness and coherence of the predictions of a pocket detector in a biologically relevant context. Specifically, MD simulations can be used to simulate a sequence of small, non-rigid perturbations to the initial protein structure, effectively allowing to test the resilience of pocket detection algorithms in response to structural changes.

Recently, we proposed GENEOnet [26], a network-based approach for protein pocket detection that leverages Group Equivariant Non-Expansive Operators (GENEO). A key design feature of GENEOnet is its inherent geometrical coherence, known as equivariance, which ensures that the predicted binding sites behave coherently under rigid motions. Furthermore, because of the mathematical properties of GENEOs, we were able to develop GENEOnet with the aim of having an explainable and robust method for protein pocket detection. In particular, the relatively small number of learnable parameters in the network allows us to assign meaningful interpretations to each parameter, effectively treating them as feature importance indicators [27–30], which are one of the most prominent research areas inside the field of XAI. As elaborated further in this paper, GENEOnet aligns effectively within the framework of S.A.F.E. Artificial Intelligence. The

acronym S.A.F.E. represents Sustainable, Accurate, Fair, and Explainable Artificial Intelligence, which are the core requirements set forth by the recent European AI Act. For definitions, review papers, and proposals of metrics able to measure how much S.A.F.E. is an AI system refer to [31–36].

The paper is structured as follows: Section 2 presents the basic background of GENEOs, Section 3 briefly describes the GENEOnet model, Section 4 describes the statistical analyses that we performed to investigate the trustworthiness of GENEOnet. In particular in Section 4.1 we describe the data sources employed in the analyses; in Section 4.2 we detail the sensitivity analysis about GENEOnet parameters, showing that they can be easily interpreted from a statistical perspective, highlighting their potential to serve as global feature importance explanations; in Section 4.3 we test how frequently GENEOnet and some other AI models respect the equivariance property with respect to rigid motions of the inputs, marking a significant distinction of GENEOnet from the other considered methods; in Section 4.4, by exploiting molecular dynamics simulations, we also investigate the consistency of GENEOnet outputs in response to non-rigid perturbations, further fostering trust in its predictions. Lastly, in Section 5 we draw some final remarks.

2. Group Equivariant Non-Expansive Operators

Recent advancements have proposed Group Equivariant Non-Expansive Operators (GENEOs) as powerful mathematical instruments for constructing network models within machine learning and artificial intelligence applications. The current state of GENEO theory is particularly active, focusing on developing methods to generate GENEOs in various contexts [37–40] and exploring their generalizations [41]. For a comprehensive understanding of GENEOs and their properties, please refer to [42].

At a high level, GENEOs should be perceived as data processing agents, defined for working with functional data. The two key characteristics of such agents, equivariance and non-expansivity, are inherent in their name.

Equivariance assumes growing importance when handling data subject to geometric transformations [43–46]. Equivariance is defined in relation to a group of transformations of the data domain, known as an equivariance group. To be considered equivariant, an operator must commute with all transformations within the equivariance group. The most basic form of equivariance, named invariance, implies that the agent's outputs stay the same when inputs are obtained through a transformation belonging to the equivariance group. For example, if we have a point cloud derived from an object scan, the represented object remains unchanged regardless of the spatial orientation of the point cloud. This necessitates that an agent processing such point clouds should be invariant with respect to the group of rigid motions of the Euclidean space to produce reliable results. Ultimately, equivariance serves as a means to incorporate relevant data symmetries and previous knowledge within the processing pipeline for problems of interest, emulating, in a way, a Bayesian approach that emphasizes the geometric characteristics of data and evaluates their quality rather than focusing on distributions [47].

We must acknowledge that equivariance is not a prerogative of models built with GENEOs, other methods proposed in literature are constructed to possess this property. Examples are Group Equivariant Convolutional Networks [46,48] and Graph Neural Networks [49,50]. However, even if equivariance alone guarantees a big enhancement in trustworthiness, such models must still be considered black boxes due to their deep architecture and inherent complexity.

The other property of GENEOs that allows the building of simple and intrinsically explainable models such as GENEOnet, is non-expansivity, a trait typically lacking in methods such as Group Equivariant Convolutional Networks.

Non-expansivity, in particular, ensures that GENEOs are 1-Lipschitz operators with respect to appropriate distances on the functional input and output space. Non-expansivity carries significant implications when studying the topological properties of the GENEO space [see 42], but it can also be seen as a form of resistance to perturbations in the data. Stability is a rare property found in black box models, which have often been demonstrated to be unstable in ways that can be exploited to return inconsistent and unreliable results, such as in the case of adversarial attacks [51–54].

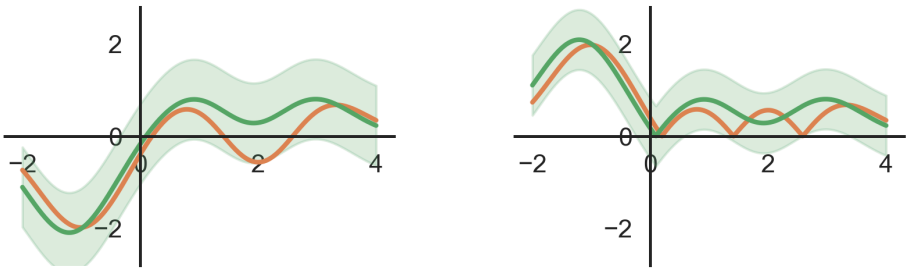
In combination, equivariance and non-expansivity can provide a high level of trustworthiness to models built using GENEOs [26,55,56]: the variability derived from geometric transformations is controlled alongside robustness to minor perturbations in the inputs.

For sake of clarity, we will now shortly formalize the concept of group-equivariant non-expansive operator. The reader can refer to [42] for a comprehensive understanding of GENEOs and their properties.

We assume that a space Φ of functions from a set X to $\mathbb{R}^k$ is given, together with a group G of permutations 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 distance defined by the L_∞-norm $||\varphi_1 - \varphi_2||_\infty$, where $||\varphi_1 - \varphi_2||_\infty = \sup_{x \in X} |\varphi_1(x) - \varphi_2(x)|$ for any $\varphi_1, \varphi_2 \in \Phi$. Let us assume that another perception pair (Ψ, H) is given, with Ψ endowed with the topology induced by the analogous L_∞-norm distance, and let's fix a homomorphism $T : G \to H$.

Definition 2.1. 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).



(a) φ_1, φ_2 (b) $F(\varphi_1), F(\varphi_2)$

Figure 1: Example of non-expansivity and its relationship with the use of the L_∞ distance. Two real-valued functions are shown (a) before and (b) after the application of the non-expansive operator $F(\varphi) = |\varphi|$. The green neighborhood highlights the L_∞ distance between the functions allowing to notice that such distance is smaller for $F(\varphi_1), F(\varphi_2)$ than for φ_1, φ_2.

Figure 1 provides an illustrative example of the second property of GENEOs and the rationale behind the choice of using the distance defined by the L_∞-norm $||\varphi_1 - \varphi_2||_\infty$.

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

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

the following main properties of spaces of GENEOs can be proven (see [42] for the proofs).

Theorem 2.2. *If Φ and Ψ are compact, then F_{all} is compact with respect to the topology induced by D_{GENEO}.*

Corollary 2.3. *If Φ and Ψ are compact with respect to the ∞-norms D_Φ and D_Ψ, respectively, then F_{all} can be ε-approximated by a finite set for any $\epsilon > 0$.*

Theorem 2.4. *If Ψ is convex, then F_{all} is convex.*

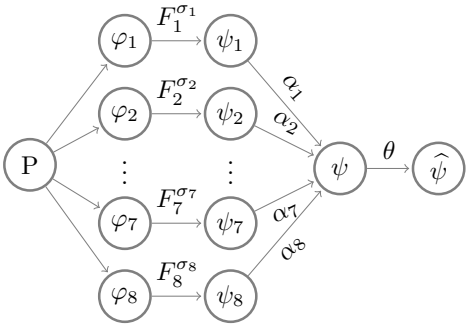
Theorem 2.2 and Corollary 2.3 demonstrate that when data spaces are compact, the space of GENEOs inherits this compactness, thus allowing for adequate approximation by a limited number of representatives, which simplifies the problem. According to Theorem 2.4, if the data space is convex, any convex combination of GENEOs will also be a GENEO. Consequently, when both compactness and convexity are present, one can easily construct any element of F_{all} using only a finite set of operators. Moreover, the convex nature of F_{all} assures a unique minimum for each strictly convex cost function within the GENEO space, facilitating the identification of an 'optimal GENEO.' These properties, when utilized wisely, enable the development of highly efficient neural networks characterized by reduced complexity—fewer nodes and layers—leading to models with enhanced interpretability due to their clear architecture. GENEOnet serves as an early example of constructing a neural network with the application of GENEOs.

3. GENEOnet

GENEOnet [26,57] (which can be freely tested as a webservice at https://geneonet.exscalate.eu) is a specialized GENEO network model designed for protein pocket detection. It features a shallow network architecture that is composed of a limited number of GENEO units.

A brief summary follows of the six steps executed by GENEOnet to detect protein pockets (for further details refer to [26]):

1. Data preprocessing: On a protein P, GENEOnet first computes a grid of voxels discretizing the bounding box surrounding the protein surface. Subsequently, it approximates eight potential functions φ_i, each of which models essential aspects of the protein structure from geometric, physical, and chemical viewpoints.
2. GENEO layer: A convolutional rigid motion equivariant operator $F_i^{\sigma_i}$, each based on a kernel depending on a shape parameter σ_i, is applied to each potential function. The resulting function is normalized between 0 and 1.
3. Convex combination: GENEO outputs are combined through a convex combination with weights α_j. The resulting combination output, denoted as ψ, normalized between 0 and 1, represents the likelihood that each voxel belongs to a pocket.
4. Thresholding: The final output $\hat{\psi}$ is obtained by taking the connected components of the spatial region in which ψ is above the parameter θ.
5. Evaluation: Having the ground truth (i.e. the ligand) available, the output can be compared and evaluated using a volumetric accuracy function. Such accuracy function has not the common structure of a precision or recall function, since we are not solving a problem



(a) GENEOnet architecture.

Unit	Channel	σ	α	θ
1	Distance	3.110	0.362	0.756
2	Gravitational	5.197	0.002	
3	Electrostatic	2.561	0.054	
4	Lipophilic	4.678	0.338	
5	Hydrophilic	3.545	0.001	
6	Polar	6.166	0.185	
7	HB Acceptor	4.186	0.056	
8	HB Donor	3.908	0.001	

(b) Optimal parameters' values.

Figure 2: Model architecture and optimal parameters obtained after training the model. Figure (a) visually describes the steps that GENEOnet performs to get from an input protein P to the final prediction $\hat{\psi}$. Additionally ψ is used for performing the scoring of the individual pockets. Table (b) provides the values of the optimal parameters of the model, obtained after a training on a set of 200 molecules and a validation procedure on a set of about 3000 molecules. Given the small number of involved parameters, it is possible to interpret their meaning. In particular, the α coefficients can be intended as feature importance for the input potentials.

of classification. It is evaluating a weighted proportion of the volumes of the regions which are correctly recognized as 'pockets', and those which are correctly recognized as 'non pockets'. See [26] for further details.

6. Scoring: Each predicted pocket is assigned a score, computed as a weighted average of ψ within its spatial region, enabling pocket ranking.

Figure (2a) provides a graphical representation of the GENEOnet architecture, starting from an input protein P to the binary output $\hat{\psi}$, as described in points (1)-(6).

Despite its relative simplicity and the quite small number of learnable parameters, GENEOnet has been shown [see 26] to exhibit slightly superior performance in identifying the true pockets with the top-ranked predicted pockets when compared to state-of-the-art methods. Table (2b) allows us to show one of the key advantages of GENEOnet, namely its minimal number of trainable parameters (the trainable parameters are only 17). This fact enables the assignment of meaningful interpretations for humans to each of its seventeen parameters. The shape parameters σ exert an influence over the operator kernels, while the convex combination coefficients α_i serve as feature importances for potentials. Lastly, the threshold coefficient θ determines the significance level for voxel activation.

The construction of the model with a reduced number of parameters has to be attributed to two primary factors: the equivariance property and the knowledge injection employed during the selection of potentials and corresponding operators. The employment of GENEOs that are equivariant with respect to rigid motions of the space significantly reduces the number of parameters because the model does not need to learn this symmetry from the data; instead, it is directly encoded in the choice of operators.

Furthermore, the convex constraint on the α coefficients may promote their sparsity during

learning, thereby further minimizing the number of significant parameters. The convex combination parameters merit additional attention because, as the intermediate outputs ψ_j are normalized between 0 and 1, the convex combination coefficients assume the role of feature importances for the various potentials and units within the GENEO layer. These feature importances constitute an *e*xplanation for this model yet independent of any specific instance of protein to which it is applied; thus, the α coefficients can be regarded as a *m*odel-specific global explanation for GENEOnet (for a complete taxonomy refer to [58]).

4. Statistical analyses

4.1. Data sources

Two publicly accessible datasets have been utilized for subsequent analyses: PDBbind [59] and ATLAS [60]. The PDBbind database consists of a curated collection of biomolecular complexes accompanied by experimentally determined binding affinities. The 2020 version contains approximately 19,000 protein-ligand complexes; nearly 3,000 have been chosen for the training and validation of GENEOnet, and about 9000 have been used for the comparison with other methods. This same subset will be utilized also in the analyses performed in this work. Conversely, ATLAS contains MD simulations for 1,390 proteins, along with analyses of the dynamics and data visualizations. From ATLAS, we obtained 37 simulations corresponding to proteins that are shared with the aforementioned set extracted from the PDBbind and will be employed in the following sections.

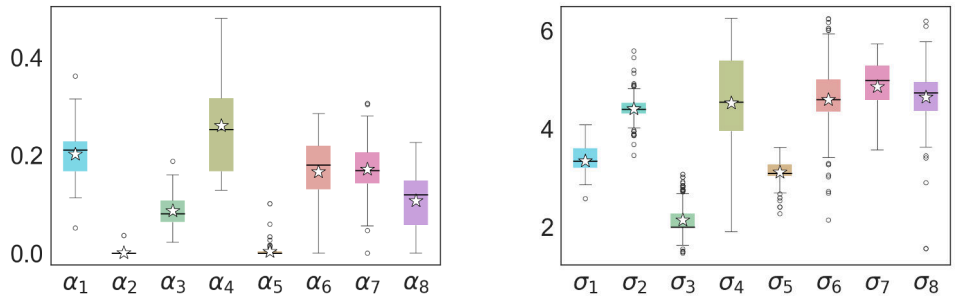
4.2. Sensitivity analysis

A first statistical analysis was conducted to examine the response of GENEOnet parameters to variations in the training dataset.

The training set for GENEOnet consisted of 200 protein-ligand complexes; accordingly, this analysis involved $n = 200$ independent repetitions of the model training, with each iteration featuring a training set of 200 complexes selected uniformly at random from the data collected from the PDBbind. The training optimization always started from the same initial parameter guess in each repetition, ensuring that the only source of variation between the different runs was the training set itself. The boxplots displayed in Figure (3a) and (3b) thus represent sensitivity analyses of the parameters with respect to variations in the training data.

The results reported in Figure (3a) show that α_1 and α_4, which are also the largest coefficients of the optimal set of convex combination parameters reported in Table (2b), have distributions which are quite far from zero, thus showing that the corresponding potentials, the Distance and the Lipophilic one, maintain a high importance in the pocket identification procedure, whatever the training set of the model. In contrast, the distributions of α_2 and α_5 are concentrated on values close to zero, so the corresponding potentials, the Gravitational and Hydrophilic ones, never play a relevant role in identifying pockets. This further strengthens the explanatory role of the parameters α compared to the mere point estimates shown in Table (2b).

Furthermore, a biochemical rationale underpins some of these findings, as evidenced by the frequent negative correlation between Hydrophilic and Lipophilic potentials. Typically, the exposed regions of a protein surface are hydrophilic, given their solvation primarily in aqueous environments. Conversely, less exposed and more sheltered areas may exhibit hydrophobic (and thus lipophilic) characteristics. Consequently, the model incorporates analogous information through both the Lipophilic and Hydrophilic potentials. Thus, it is appropriate that only one of these is deemed significant for the identification of pockets that are usually unlikely to be placed in the



(a) Distributions of convex combination coefficients.

(b) Distributions of convolutional kernel coefficients.

Figure 3: Empirical distributions of the estimated parameters of GENEOnet, obtained by randomly varying the training set. (a) shows distributions of the convex combination coefficients, that are interpreted as feature importance for the channels, (b) instead shows distributions of the shape coefficients of the convolutional kernels used in the GENEO layer of the architecture. The values of the σ_i coefficients is related with the amplitude of the region surrounding each voxel in which the corresponding channel contributes to the output of the model.

most exposed regions of the protein surface. Moreover, the presence of hydrophilic interactions can also be encoded in other potentials such as Polar, Hydrogen Bond (HB) Acceptor, and HB Donor; anyway, this is unlikely to happen for lipophilic interactions. This further justifies the high importance given to the Lipophilic potential and at the same time the Hydrophilic potential is given a lower importance because polar interactions are probably better captured by the other mentioned potentials.

All these considerations regarding the importance of potentials align with the empirical knowledge of medicinal chemistry. This is a direct and relevant outcome of incorporating prior knowledge, made possible by the unique architecture of GENEOnet. It demonstrates that GENEOnet can automatically perform the physicochemical analyses that a skilled expert would typically conduct manually on single proteins — an otherwise unfeasible task given the vast number of molecules analyzed in virtual screening experiments.

Moreover, the distributions of the coefficients of the convolutional kernels in Figure (3b) have a dispersion that is similar to the corresponding dispersion of the distributions of the convex combination coefficients, thus confirming the same uncertainty quantification of the point estimates of the parameters of each potential.

4.3. Equivariance analysis

Equivariance is pivotal in understanding the functioning of GENEOnet, as it makes its predictions consistent with respect to rigid motions s in $\mathbb{R}^3$ of the protein structure. It is fundamental in our example since the specific position in which the protein is observed should not change the pockets identification. However, upon discussing the matter with practitioners of medicinal chemistry, we found that equivariance is not a feature that an average pocket identification algorithm will guarantee. Consequently, we decided to evaluate the level of equivariance displayed by

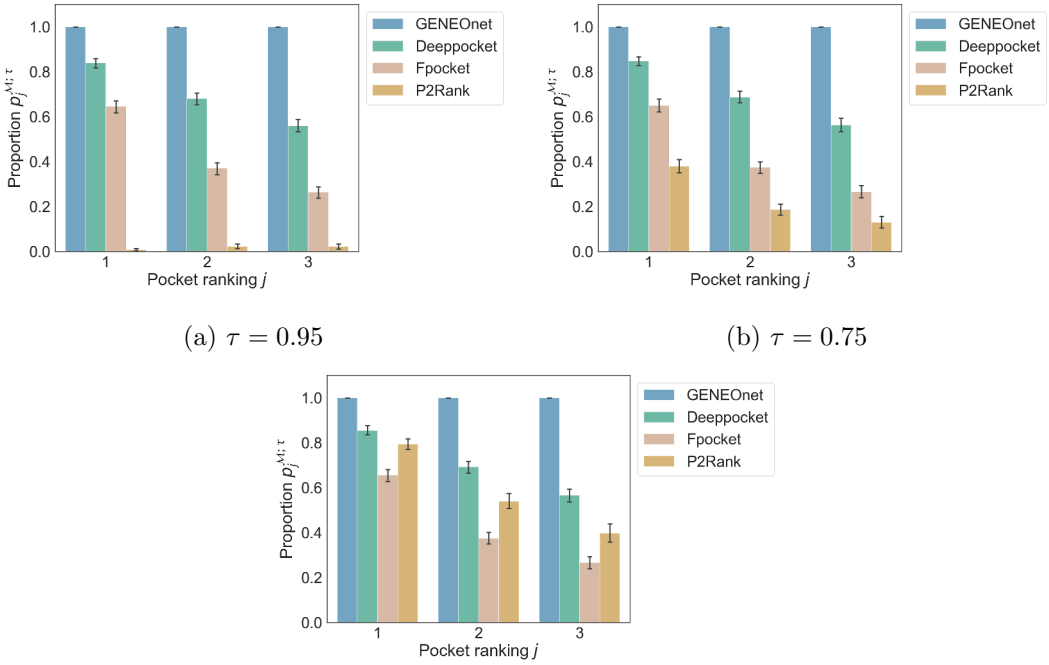
GENEOnet and compare it with several state-of-the-art algorithms considered in [26].

To this aim, the methodology applied is the following:

1. We systematically sampled $N = 2000$ proteins $(P_1, \dots, P_N)$ from the larger dataset extracted from PDBbind.
2. For each protein P_i we generated a rotated protein $\rho(P_i)$ obtained by applying to the structure a rotation ρ of $\pi/2$ around the x-axis.
3. For each method $\mathcal{M}$ under investigation, we computed predictions on both the original input $\mathcal{M}(P_i)$ and the rotated $\mathcal{M}(\rho(P_i))$.
4. The analysis was confined to the three top-ranked pockets. We will denote as
$$\mathcal{M}(P_i)_j \quad j \in \{1, 2, 3\}$$
the spatial region (reported on the GENEOnet grid so that there is a common reference for all methods) identifying the j-th predicted pocket by method $\mathcal{M}$.
5. By applying the inverse rotation ρ^{-1} to the predicted pockets returned on the rotated proteins, we can compute the proportion of overlap between the non-rotated and rotated predictions for protein P_i (here $|\cdot|$ stands for the volume).
$$O_j^{\mathcal{M}}(P_i) = \frac{|\mathcal{M}(P_i)_j \cap \rho^{-1}(\mathcal{M}(\rho(P_i))_j)|}{|\mathcal{M}(P_i)_j|} \quad j \in \{1, 2, 3\}$$
If method $\mathcal{M}$ didn't output a prediction for one index j on either the original or the rotated protein then the value of $O_j^{\mathcal{M}}(P_i)$ has been considered a missing value.
6. Finally for each $j \in \{1, 2, 3\}$ we estimated the proportion $p_j^{\mathcal{M};\tau}$ of proteins for which the overlap is non-missing and above a threshold τ as follows:
$$\hat{p}_j^{\mathcal{M};\tau} = \frac{1}{|I_j^{\mathcal{M}}|} \sum_{i \in I_j^{\mathcal{M}}} \mathbb{1}\Big(O_j^{\mathcal{M}}(P_i) \geq \tau\Big),$$
where we denoted by $I_j^{\mathcal{M}}$ the set of indices for which $O_j^{\mathcal{M}}(P_i)$ is non-missing. Empirical confidence intervals based on the sample have also been computed.

Setting $\tau = 1$ allows for the estimation of the proportion of detected pockets that show perfect equivariance, while considering lower values of τ allows for differences that may arise from the implementation or numerical approximation errors, resulting in the calculation of the proportion of pockets showing approximate equivariance. Concerning rotations with respect to the x-axis, Figure 4 reports estimates of $(p_1^{\mathcal{M};\tau}, p_2^{\mathcal{M};\tau}, p_3^{\mathcal{M};\tau})$ for values of $\tau \in \{0.95, 0.75, 0.5\}$ and for the four methods: GENEOnet, DeepPocket [15], Fpocket [61] and P2Rank [17] along with 99% confidence intervals.

The results allow us to conclude that GENEOnet exhibits a higher degree of equivariance (approximately equal to 1) compared to other methods under consideration. The 99% confidence intervals displayed in Figures (4a), (4b) and (4c) (and reported in Table A) demonstrate that the proportion of equivariance for GENEOnet is significantly greater than all other methods at a significance level of 0.01 for all the considered values of τ, thus ranging from perfect to approximate equivariance. This is a significant finding as P2Rank and DeepPocket were, respectively, the second and third-best performing models in the task of pocket identification [26]. However, their predictions are susceptible to some degree of instability with respect to the spatial pose of the input protein, which may also affect the pocket rankings when transitioning from the



(a) $\tau = 0.95$

(b) $\tau = 0.75$

(c) $\tau = 0.50$

Figure 4: Mean values and confidence intervals, obtained from different protein samples, of the proportions $\hat{p}_j^{\mathcal{M};\tau}$ computed for GENEOnet and other three methods for pocket detection, obtained for different values of the threshold τ. This Figure reports the results about rotations of $\pi/2$ around the x axis.

original to the rotated protein. This finding, which aligns with the claim made by medicinal chemistry experts, constitutes a significant source of unreliability for these models, as the frozen pose, which is retrieved from databases such as the PDBbind and may be used in the training data, is far from being canonical in any sense. For sake of completeness, we repeated the same analysis considering: rotations of $\pi/2$ first around the y-axis and then around the z-axis. We report the results in Figures G and H in Appendix A. The similarity of the results obtained in these cases with the results of Figure 4 on the x axis provides an even stronger guarantee about the resilience of GENEOnet to such types of transformations. We refer the reader to Section 5 for a discussion of considering other kinds of rotation as well.

4.4. Robustness analysis

In the previous Section, we delved into testing the equivariance on protein structures subjected to rigid perturbations. However, it is crucial to extend our considerations also to non-rigid transformations, given the inherent dynamism of proteins that should be viewed as moving entities rather than static objects. In the realm of medicinal chemistry, MD simulations offer a valuable tool for studying the dynamic behavior of a protein immersed in a solvent environment, typically water molecules.

Given that the time interval between successive frames in MD simulations is generally small,

we can presume that structural changes between consecutive frames are minimal, apart possibly for a negligible number of cases.

With this rationale and owing to the intrinsic biological significance, we opted to assess the resilience of GENEOnet to detect the same pockets on protein MD data.

For each of the 37 proteins sourced from the ATLAS database, we exploited two simulations spanning 100 ns each: the first composed of 1000 frames and the second consisting of 10000 (hence with a time step of 100 ps and 10 ps respectively).

Upon collecting the data, we proceeded to execute the following experiments:

1. For each protein P^i (where i ranges from 1 to 37) we considered the frames of the dynamics, denoted as P^i_t, where t spans from 1 to T (in case of the coarser simulations we processed the entire dynamics, thus $T = 1000$, while in the case of finer simulations we considered only the initial 250 frames, thus $T = 250$. This second choice is motivated by the related computational costs and is justified in detail in Section 5 and by the results reported in Table E of Appendix B.)

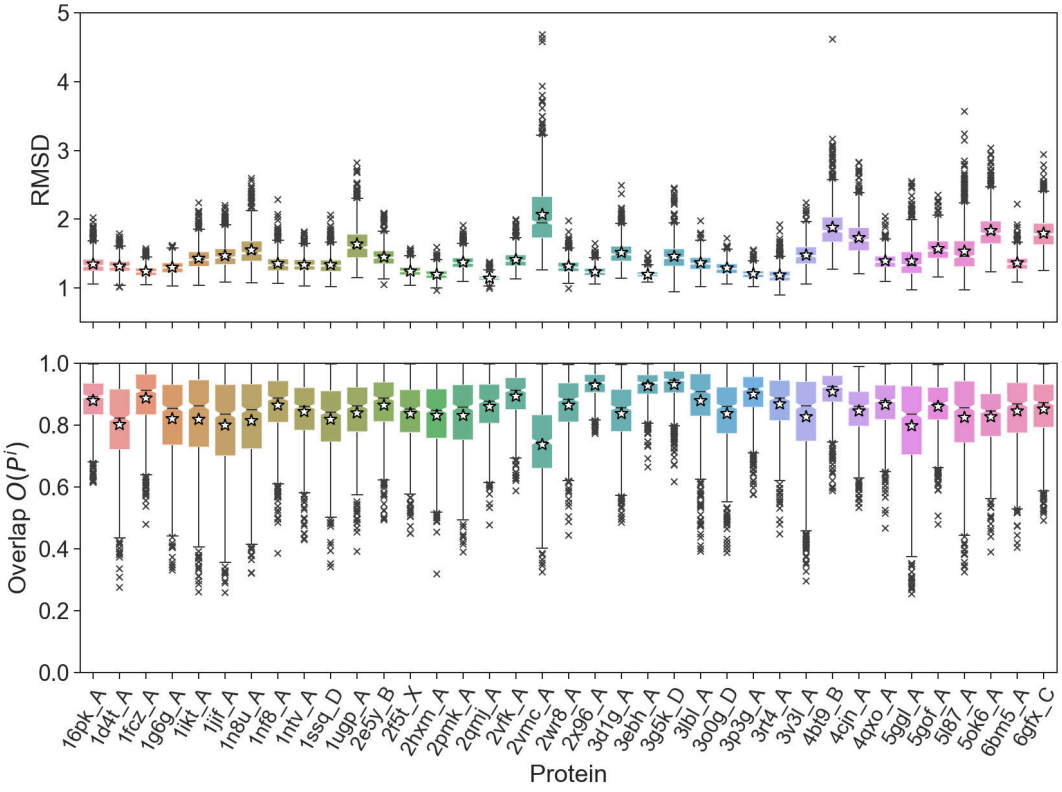
2. For each time step t, we computed the global prediction generated by GENEOnet (i.e. the union of all the pockets detected by GENEOnet on the surface of the protein), denoted as $\mathcal{G}(P^i_t)$, without segmenting it into distinct pockets, unlike the approach employed in the preceding Section.

3. For every time step t (where t ranges from 2 to T), we calculated the degree of overlap between the prediction generated for the current frame and that produced for the preceding frame:
$$O_t(P^i) = \frac{|\mathcal{G}(P^i_{t-1}) \cap \mathcal{G}(P^i_t)|}{|\mathcal{G}(P^i_{t-1})|}$$
We also computed the Root Mean Squared Deviation (RMSD) between the atomic positions of the two consecutive frames considering all atoms:
$$RMSD_t(P^i) = \sqrt{\frac{1}{|N_i|}\sum_{j=1}^{N_i}\left|\left|x^j_{t-1} - x^j_t\right|\right|^2}$$
where N_i refers to the total count of atoms within protein P^i, while x^j_t denotes the coordinate vector representing the position of atom j at frame t.

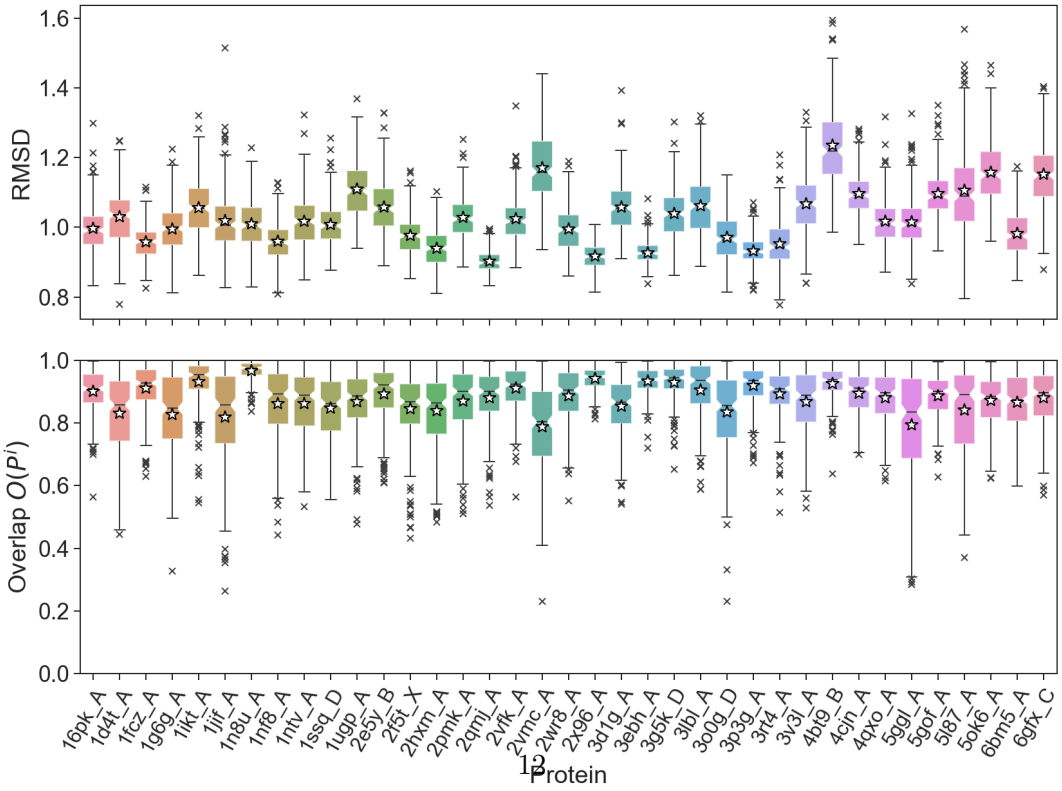
Assuming non-expansivity, we would expect to witness, to some degree, a negative correlation between the series representing RMSD and those depicting the overlap for the same protein. Actually, in some cases, a protein could move during the MD mainly in the parts where no pockets are present, thus it may happen to observe high RMSD values that do not correspond to small overlap values. The reverse should instead be observed more frequently, i.e. small RMSD values should correspond to high overlaps of the pockets.

Figures (5a) and (5b) offer a comparative analysis of the distributions of overlaps and RMSD values, respectively, for the selected proteins for both molecular dynamics.

The negative correlation between overlap and RMSD is moderately clear in the simulations with a time delta of 100 ps (in Figure (5a)) since RMSD values are rarely below 1Å. However, in some cases, higher RMSD distribution values correspond to small overlap distribution values, such as in the case of protein 2vmc_A; but this is not always the case, as can be observed for proteins 1ugp_A or 4bt9_B. Reversely, when the distribution of RMSD is concentrated around small values, the corresponding distribution of the overlap is mainly located in correspondence of values close to 1, like, for example, for proteins 2x96_A or 3ebh_A.



(a) Distributions of overlap and RMSD for the dynamics with time delta 100 ps, considering all the 1000 frames.



(b) Distributions of overlap and RMSD for the dynamics with time delta 10 ps, considering the first 250 frames.

Figure 5: Robustness analysis

The results are then in accordance with the definition of non-expansivity, which prescribes that the distance between the model outputs should be smaller than the distance between the inputs; thus pockets that are similar and close in two subsequent frames should be both detected and have a big overlap.

Moreover, in the case of finer MD simulations with a time delta of 10 ps we can make similar observations with the difference that the mean overlaps are generally higher and the mean RMSD tends to be lower as shown in Figure (5b) (see also Appendix B and Table D for a quantitative evaluation of overlaps).

5. Concluding remarks

In this study, we conducted a comprehensive evaluation of GENEOnet, a network architecture comprising GENEOs designed for protein pocket detection. To assess its explainability and trustworthiness, we employed several statistical analyses.

Initially, a sensitivity analysis was conducted to assess the effect of variations in the training dataset on the model coefficients. This examination provided an enhanced comprehension of the parameters involved, facilitating improved interpretability and allowing for an evaluation of their statistical significance.

Secondly, we compared the equivariance properties of GENEOnet with those of other state-of-the-art models for protein pocket detection. The results of this experiment demonstrated that GENEOnet's employment of GENEOs as building blocks yields a substantially higher proportion of equivariance than other methods, indicating that its outputs are significantly more reliable when subjected to rigid transformations of the input data.

We acknowledge that we considered only rotations ρ of angle $\pi/2$ around the coordinate axes. Even if this was enough to obtain the results that were shown, it was done primarily in order to easily return the rotated voxel grid to the original grid for comparison. More general rotations (those with axes different from one of the coordinate ones, for example, or of angles not multiple of $\pi/2$), instead, would have been harder to take into account because of the more challenging alignment of the grids that would be required in that case (a motivating example of such a rotation and the difficulties that may arise is provided in Figure F in Appendix A). Given the consistency of outcomes across three coordinate axes, we assert that this sufficiently demonstrates GENEOnet enhanced resilience to rotations and, more generally, to rigid motions due to its property of equivariance.

Lastly, we tested GENEOnet robustness against non-rigid perturbations of the input data with the help of MD simulations. Our findings indicate that GENEOnet exhibits strong resilience to such perturbations, particularly when the perturbations are smaller, as in the case of 10 ps MD simulations.

Regarding the robustness analysis, the reader could criticize that, for the finer dynamics with a time delta of 10 ps, we considered only the initial segment of each MD simulation and that this could hide wider rearrangements of the protein structure that might happen later in the dynamics. We recognize this potential concern, and for this reason, we have computed the RMSD values based on the full dynamics of 10000 frames. We subsequently compared the RMSD distributions derived from just the initial 250 frames to those from the complete set of dynamics. The results can be found in the Appendix B Table E. The observed differences—both in terms of differences in empirical means and the Wasserstein distances between empirical distributions—are minimal enough to support our decision to focus solely on the first 250 frames.

Considering all these aspects, the findings collectively illustrate that GENEOnet is a model that not only offers transparency but also maintains substantial reliability when subjected to geometric transformations and perturbations. Furthermore, in a broader sense, they highlight

the value of GENEOs as a resource for creating innovative and trustworthy AI models with a focus on interpretability. Moreover, it is important to emphasize that GENEOnet has become a commonly utilized tool by private companies, aiding their efforts in drug design with encouraging outcomes. Although GENEOnet has not yet led to the discovery of new drugs, due to its relatively recent inception of around two years and the inherently lengthy process of drug development, we are optimistic that GENEOnet will soon excel in addressing numerous unresolved challenges in drug discovery.

Acknowledgement(s)

Scientific support is acknowledged by the Italian GNAMPA - INDAM group. Computational resources were partially provided by the INDACO core facility for HPC at Università degli Studi di Milano.

Data availability

The two data sources used in the analyses are open and freely available at http://pdbbind.org.cn/ (for PDBbind) and https://www.dsimb.inserm.fr/ATLAS/ (for ATLAS).

Disclosure statement

Filippo Lunghini, Andrea Rosario Beccari, and Carmine Talarico are employees of Dompé Farmaceutici S.p.A.

Funding

The authors acknowledge partial funding from Dompé Farmaceutici S.p.A. for developing this research.

Acronyms

AI Artificial Intelligence

GENEOs Group Equivariant Non-Expansive Operators

HB Hydrogen Bond

MD Molecular Dynamics

RMSD Root Mean Squared Deviation

XAI eXplainable Artificial Intelligence

References

[1] Torres PHM, Sodero ACR, Jofily P, et al. Key Topics in Molecular Docking for Drug Design. International Journal of Molecular Sciences. 2019;20(18).

[2] Lionta E, Spyrou G, Vassilatis DK, et al. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. Current Topics in Medicinal Chemistry. 2014;14(16):1923–1938.

[3] Nag S, Baidya ATK, Mandal A, et al. Deep learning tools for advancing drug discovery and development. 3 Biotech. 2022;12(5).

[4] Marcu SB, Tabirca S, Tangney M. An Overview of Alphafold's Breakthrough. Frontiers in Artificial Intelligence. 2022;5.

[5] Yim J, Staerk H, Corso G, et al. Diffusion models in protein structure and docking. Wiley Interdisciplinary Reviews-Computational Molecular Science. 2024;14(2).

[6] Loeffler HH, He J, Tibo A, et al. Reinvent 4: Modern AI-driven generative molecule design. Journal of Cheminformatics. 2024;16(1).

[7] Tsuchiya Y, Yamamori Y, Tomii K. Protein-protein interaction prediction methods: from docking-based to AI-based approaches. Biophysical Reviews. 2022;14(6, SI):1341–1348.

[8] Crampon K, Giorkallos A, Deldossi M, et al. Machine-learning methods for ligand-protein molecular docking. Drug Discovery Today. 2022;27(1):151–164.

[9] Clyde A, Liu X, Brettin T, et al. AI-accelerated protein-ligand docking for SARS-CoV-2 is 100-fold faster with no significant change in detection. Scientific Reports. 2023;13(1).

[10] Sauer S, Matter H, Hessler G, et al. Optimizing interactions to protein binding sites by integrating docking-scoring strategies into generative AI methods. Frontiers in Chemistry. 2022 OCT 19;10.

[11] Ivanenkov YA, Evteev SA, Malyshev AS, et al. AlphaFold for a medicinal chemist: tool or toy? Russian Chemical Reviews. 2024;93(3).

[12] Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chemical Science. 2024; 15(9):3130–3139.

[13] Laurie ATR, Jackson RM. Methods for the prediction of protein-ligand binding sites for Structure-Based Drug Design and virtual ligand screening. Current Protein & Peptide Science. 2006;7(5):395–406.

[14] Di Palma F, Abate C, Decherchi S, et al. Ligandability and druggability assessment via machine learning. Wiley Interdisciplinary Reviews-Computational Molecular Science. 2023; 13(5).

[15] Aggarwal R, Gupta A, Chelur V, et al. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. Journal of Chemical Information and Modelling. 2022;62(21):5069–5079.

[16] Jimenez J, Doerr S, Martinez-Rosell G, et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33(19):3036–3042.

[17] Krivak R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics. 2018;10.

[18] Nazem F, Ghasemi F, Fassihi A, et al. 3D U-Net: A voxel-based method in binding site prediction of protein structure. Journal of Bioinformatics and Computational Biology. 2021; 19(2).

[19] Zhang Y, Qiao S, Ji S, et al. DeepSite: bidirectional LSTM and CNN models for predicting DNA-protein binding. International Jornal of Machine Learning and Cybernetics. 2020; 11(4):841–851.

[20] Zhang H, Saravanan KM, Lin J, et al. DeepBindPoc: a deep learning method to rank ligand binding pockets using molecular vector representation. PEERJ. 2020;8.

[21] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Improving detection of protein-ligand binding sites with 3d segmentation. Scientific Reports. 2020 MAR 19;10(1).

[22] Jiang M, Wei Z, Zhang S, et al. FRSite: Protein drug binding site prediction based on faster R-CNN. Journal of Molecular Graphics & Modelling. 2019;93.

[23] Cerutti DS, Case DA. Molecular dynamics simulations of macromolecular crystals. Wiley Interdisciplinary Reviews-Computational Molecular Science. 2019 JUL;9(4).

[24] Bernardi RC, Melo MCR, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochimica et Biophysica Acta-General Subjects. 2015; 1850(5, SI):872–877.

[25] Meuwly M. Reactive molecular dynamics: From small molecules to proteins. Wiley Interdisciplinary Reviews-Computational Molecular Science. 2019;9(1).

[26] 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; 2022. Preprint at arXiv.

[27] Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations; 2016. p. 97–101.

[28] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017. p. 4765–4774.

[29] Giudici P, Raffinetti E. Shapley-Lorenz eXplainable Artificial Intelligence. Expert Systems with Applications. 2021;167:114104.

[30] Cinquini M, Giannotti F, Guidotti R, et al. Handling Missing Values in Local Post-hoc Explainability. In: eXplainable Artificial Intelligence, XAI 2023, PT II; (Communications in Computer and Information Science; Vol. 1902); 2023. p. 256–278. 1st World Conference on Explainable Artificial Intelligence (XAI).

[31] Giudici P. Safe machine learning. Statistics. 2024;58(3):473–477.

[32] Giudici P, Centurelli M, Turchetta S. Artificial intelligence risk measurement. Expert Systems with Applications. 2024;235:121220. Available from: `https://www.sciencedirect.com/science/article/pii/S0957417423017220`.

[33] Babaei G, Giudici P, Raffinetti E. A Rank Graduation Box for SAFE AI. Expert Systems with Applications. 2025;259:125239.

[34] Raffinetti E. A rank graduation accuracy measure to mitigate artificial intelligence risks. Quality and Quantity. 2023;57(2):131–150.

[35] Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. Decision analytics journal. 2023; 7:100230.

[36] Giudici P, Raffinetti E. Safe artificial intelligence in finance. Finance Research Letters. 2023; 56:104088.

[37] Conti F, Frosini P, Quercioli N. On the Construction of Group Equivariant Non-Expansive Operators *via* Permutants and Symmetric Functions. Frontiers in Artificial Intelligence. 2022;5.

[38] Camporesi F, Frosini P, Quercioli N. On a New Method to Build Group Equivariant Operators by Means of Permutants. In: Machine Learning and Knowledge Extraction, CD-MAKE 2018; (Lecture Notes in Computer Science; Vol. 11015); 2018. p. 265–272.

[39] Bocchi G, Botteghi S, Brasini M, et al. On the finite representation of linear group equivariant operators via permutant measures. Annals of Mathematics and Artificial Intelligence. 2023;91(4):465–487.

[40] Ahmad F, Ferri M, Frosini P. Generalized permutants and graph GENEOs. Machine Learning and Knowledge Extraction. 2023;5(4):1905–1920.

[41] Ferrari L, Frosini P, Quercioli N, et al. A topological model for partial equivariance in deep learning and data analysis. Frontiers in artificial intelligence. 2023;6.

[42] Bergomi MG, Frosini P, Giorgi D, et al. Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence. 2019;1(9):423–433.

[43] Cohen TS, Welling M. Group Equivariant Convolutional Networks. In: 33rd International Conference on Machine Laerning; Vol. 48; 2016.

[44] Worrall DE, Garbin SJ, Turmukhambetov D, et al. Harmonic Networks: Deep Translation and Rotation Equivariance. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017); 2017. p. 7168–7177.

[45] Worrall DE, Garbin SJ, Turmukhambetov D, et al. Harmonic Networks: Deep Translation and Rotation Equivariance. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017); 2017. p. 7168–7177.

[46] Cohen TS, Welling M. Group equivariant convolutional networks. In: International Conference on Machine Learning, Vol 48; (Proceedings of Machine Learning Research; Vol. 48); 2016. 33rd International Conference on Machine Learning.

[47] Ferrari L, Manzi G, Micheletti A, et al. Pandemic data quality modelling: a bayesian approach in the italian case. Quality and Quantity. 2024;.

[48] Cohen TS, Welling M. Steerable CNNs. In: Interantional Conference on Learning Representations; 2017.

[49] Morris C, Ritzert M, Fey M, et al. Weisfeiler and leman go neural: Higher-order graph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence. 2019 7; 33(01):4602–4609.

[50] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.; Vol. 2; 2005. p. 729–734 vol. 2.

[51] Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287–1289.

[52] Tsai MJ, Lin PY, Lee ME. Adversarial Attacks on Medical Image Classification. Cancers. 2023;15(17).

[53] Wang D, Dong L, Wang R, et al. Targeted Speech Adversarial Example Generation With Generative Adversarial Network. IEEE Access. 2020;8:124503–124513.

[54] Alcorn MA, Li Q, Gong Z, et al. Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019); 2019. p. 4840–4849.

[55] Lavado D, Soares C, Micheletti A, et al. Low-Resource White-Box Semantic Segmentation of Supporting Towers on 3D Point Clouds via Signature Shape Identification; 2023. Preprint at arXiv.

[56] Bocchi G, Ferri M, Frosini P. A novel approach to graph distinction through GENEOs and permutants; 2024. Preprint at arXiv.

[57] Bocchi G, Frosini P, Micheletti A, et al. A geometric XAI approach to protein pocket detection. In: Longo L, Liu W, Montavon G, editors. Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), Valletta, Malta, July 17-19, 2024; (CEUR Workshop Proceedings; Vol. 3793). CEUR-WS.org; 2024. p. 217–224. Available from: `https://ceur-ws.org/Vol-3793/paper_28.pdf`.

[58] Guidotti R, Monreale A, Pedreschi D, et al. Principles of explainable artificial intelligence; 2021. p. 9–31.

[59] Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015;31(3):405–412.

[60] Vander Meersche Y, Cretin G, Gheeraert A, et al. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. Nucleic Acids Research. 2023 11; 52(D1):D384–D392.

[61] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10.

Appendices

A. Appendix on equivariance analysis

In this appendix, we report Tables A, B, C containing equivariance proportion estimates, related to rotations around the coordinate axes, for values of $\tau \in \{0.99, 0.95, 0.75, 0.5\}$ which are shown in Figures 4, G, H. An examination of such Figures reveals that the estimates of equivariance proportions across the considered axes exhibit only minor variations. This observation further corroborates the findings of the equivariance analysis detailed in Section 4.3. Additionally, it has been highlighted in the main text that rotations involving angles not multiples of $\pi/2$, or around axes other than the coordinate ones, complicate the alignment of grids pre- and post-rotation. Such scenarios possibly necessitate function interpolation, leading to an unwarranted approximation level, especially since the aim is to demonstrate GENEOnet's enhanced robustness against these types of transformations. In particular, Figure F exemplifies some of the difficulties of grids alignment when considering the prediction for the original structure and for a $\pi/3$ rotation around the x-axis of protein ID 1W83. The voxel centers of the inverse rotation of the rotated output (darker) and the original structure (lighter) do not align since the grids are calculated independently for each structure. This involves discretizing the bounding box of each molecule where the box's axes are aligned with the coordinate axes. When the molecule is rotated by an angle not equal to a multiple of $\pi/2$, the bounding box of the rotated molecule changes compared to that of the unrotated molecule, potentially leading to discrepancies during overlap calculations. In addition, some voxels that are considered internal to the protein for one structure might be considered external for the rotation and vice versa. Nonetheless, while acknowledging these difficulties, we would like to emphasize that they arise solely from the way the grid calculation was implemented by GENEOnet and do not depend on the algorithm itself. When exact grids alignment is possible, as in the case of the already considered rotations of $\pi/2$ around one of the coordinate axes, such problems disappear.

B. Appendix on perturbation analysis

In this Appendix, we report Table D containing sample mean overlaps for the various proteins analyzed in Section 4.4. The table shows the average overlap estimates for both the 100 ps simulations, denoted as $\overline{O}_{100}$, for which all the 1000 frames of the molecular dynamic were taken into account, and the 10 ps simulations, denoted as $\overline{O}_{10}$, for which only the initial 250 frames were used for the estimates. The Table also reports the differences between the two estimates and the p-values of independent two samples T-tests for testing the hypotheses:

$$H_0 \colon \mu_{100} = \mu_{10}$$
$$H_1 \colon \mu_{100} < \mu_{10}$$

where μ_{100} and μ_{10} denote the population means for the overlap distributions of coarser and finer simulations respectively. We acknowledge that the two samples in these tests are not totally independent, but since for each protein the two simulations are different and are not a subsampling of each other, we can assume they are sufficiently independent to perform a T-test, even if they start from the same initial frame. Thus we can trust the p-values since the hypothesis of independence is not too heavily violated. We remark that we choose this type of test for simplicity, just to give a rough idea of the differences in the two time sampling strategies.

The last column of the Table reports `R` style significance codes relative to the p-values for readers' ease of interpretation. Indeed, it is easy to note that there is moderate to strong evidence (** or ***) of an increase in the mean overlaps for 24 proteins out of 37.

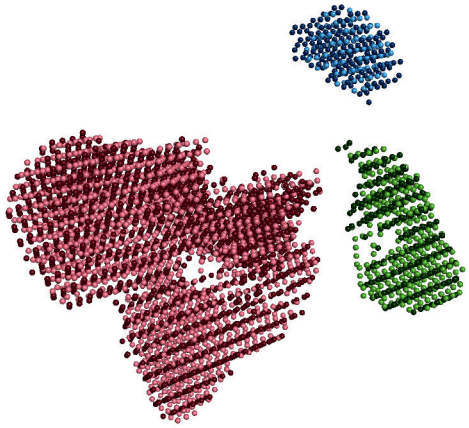
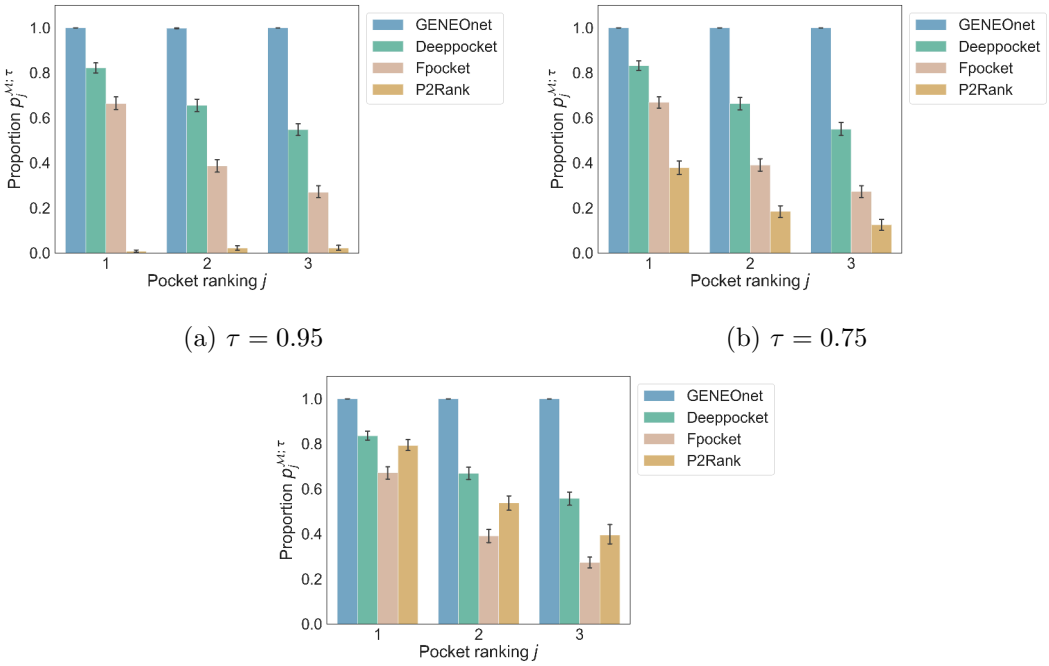


Figure F: Plots of the first three predicted pockets (respectively red, green and blue) by GENEOnet when applied to protein ID 1W83. Lighter dots represent voxel centers for the prediction on the original structure, while darker dots refer to the prediction on a $\pi/3$ rotation around the x-axis of the same structure. The prediction for the rotated structure is then rotated back to better highlight the difficulties of aligning the grids.

Moreover, Table E reports a comparison of RMSD distributions for the finer simulations having a time delta of 10 ps when considering the initial 250 frames or the complete simulation of 10000 frames. The results provide an important piece of experimental evidence that we believe justifies our choice of considering only the first 250 frames of the finer simulations in Section 4.4.

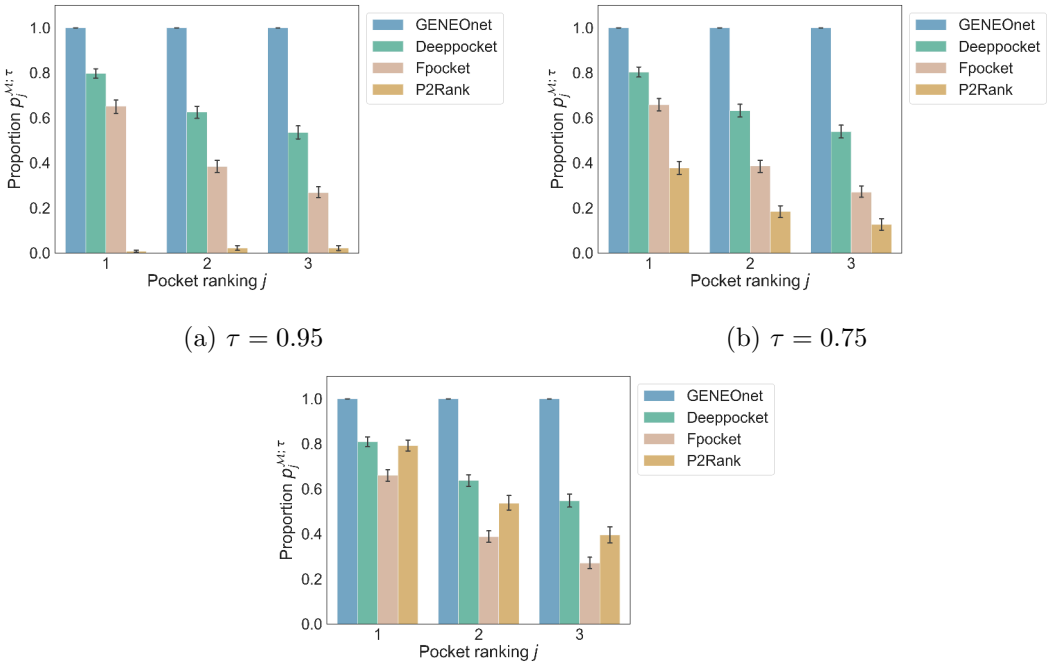


(a) $\tau = 0.95$

(b) $\tau = 0.75$

(c) $\tau = 0.50$

Figure G: Mean values and confidence intervals, obtained from different protein samples, of the proportions $\hat{p}_j^{\mathcal{M};\tau}$ computed for GENEOnet and other three methods for pocket detection, obtained for different values of the threshold τ. This Figure reports the results about rotations of $\pi/2$ around the y axis.



(a) $\tau = 0.95$

(b) $\tau = 0.75$

(c) $\tau = 0.50$

Figure H: Mean values and confidence intervals, obtained from different protein samples, of the proportions $\hat{p}_j^{\mathcal{M};\tau}$ computed for GENEOnet and other three methods for pocket detection, obtained for different values of the threshold τ. This Figure reports the results about rotations of $\pi/2$ around the z axis.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000, 1.000000]
	2	1918	0.996350	0.001377	[0.992803, 0.999898]
	3	1852	0.999460	0.000540	[0.998069, 1.000851]
Deeppocket	1	1997	0.834251	0.008323	[0.812812, 0.855691]
	2	1984	0.675907	0.010510	[0.648834, 0.702980]
	3	1969	0.556120	0.011200	[0.527271, 0.584968]
Fpocket	1	1998	0.642142	0.010727	[0.614511, 0.669773]
	2	1987	0.367891	0.010821	[0.340018, 0.395764]
	3	1971	0.263825	0.009929	[0.238249, 0.289401]
P2Rank	1	1866	0.007503	0.001998	[0.002356, 0.012650]
	2	1441	0.022207	0.003883	[0.012204, 0.032209]
	3	1078	0.022263	0.004496	[0.010683, 0.033844]

(a) $\tau = 0.99$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000, 1.000000]
	2	1918	1.000000	0.000000	[1.000000, 1.000000]
	3	1852	1.000000	0.000000	[1.000000, 1.000000]
Deeppocket	1	1997	0.838257	0.008242	[0.817028, 0.859487]
	2	1984	0.679940	0.010476	[0.652955, 0.706924]
	3	1969	0.559167	0.011192	[0.530339, 0.587995]
Fpocket	1	1998	0.646146	0.010700	[0.618584, 0.673708]
	2	1987	0.370408	0.010836	[0.342495, 0.398320]
	3	1971	0.263825	0.009929	[0.238249, 0.289401]
P2Rank	1	1866	0.008039	0.002068	[0.002712, 0.013365]
	2	1441	0.022207	0.003883	[0.012204, 0.032209]
	3	1078	0.022263	0.004496	[0.010683, 0.033844]

(b) $\tau = 0.95$

Table A: Estimates of equivariance proportions regarding rotations around x axis

.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000, 1.000000]
	2	1918	1.000000	0.000000	[1.000000, 1.000000]
	3	1852	1.000000	0.000000	[1.000000, 1.000000]
Deeppocket	1	1997	0.846770	0.008063	[0.826002, 0.867538]
	2	1984	0.686492	0.010418	[0.659657, 0.713327]
	3	1969	0.562722	0.011182	[0.533920, 0.591525]
Fpocket	1	1998	0.649149	0.010679	[0.621641, 0.676657]
	2	1987	0.373931	0.010857	[0.345964, 0.401897]
	3	1971	0.265348	0.009948	[0.239724, 0.290971]
P2Rank	1	1866	0.379421	0.011236	[0.350479, 0.408364]
	2	1441	0.186676	0.010268	[0.160227, 0.213125]
	3	1078	0.128942	0.010212	[0.102638, 0.155247]

(c) $\tau = 0.75$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000, 1.000000]
	2	1918	1.000000	0.000000	[1.000000, 1.000000]
	3	1852	1.000000	0.000000	[1.000000, 1.000000]
Deeppocket	1	1997	0.854782	0.007886	[0.834469, 0.875095]
	2	1984	0.692540	0.010362	[0.665849, 0.719232]
	3	1969	0.566277	0.011171	[0.537502, 0.595053]
Fpocket	1	1998	0.655155	0.010636	[0.627758, 0.682553]
	2	1987	0.374937	0.010863	[0.346956, 0.402918]
	3	1971	0.266362	0.009960	[0.240708, 0.292017]
P2Rank	1	1866	0.793676	0.009370	[0.769540, 0.817813]
	2	1441	0.538515	0.013137	[0.504676, 0.572354]
	3	1078	0.397032	0.014909	[0.358628, 0.435435]

(d) $\tau = 0.50$

Table A: Estimates of equivariance proportions regarding rotations around x axis

.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	0.997393	0.001165	[0.994393,1.000393]
	3	1852	0.998380	0.000935	[0.995972,1.000788]
Deeppocket	1	1998	0.815315	0.008683	[0.792948,0.837682]
	2	1988	0.655433	0.010661	[0.627971,0.682894]
	3	1968	0.545732	0.011226	[0.516814,0.574649]
Fpocket	1	1998	0.660160	0.010599	[0.632858,0.687462]
	2	1988	0.384306	0.010912	[0.356197,0.412414]
	3	1968	0.268801	0.009996	[0.243053,0.294549]
P2Rank	1	1866	0.007503	0.001998	[0.002356,0.012650]
	2	1441	0.022207	0.003883	[0.012204,0.032209]
	3	1078	0.022263	0.004496	[0.010683,0.033844]

(a) $\tau = 0.99$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	0.998957	0.000737	[0.997058,1.000856]
	3	1852	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1998	0.822322	0.008554	[0.800290,0.844355]
	2	1988	0.656439	0.010654	[0.628997,0.683881]
	3	1968	0.548272	0.011221	[0.519369,0.577176]
Fpocket	1	1998	0.664164	0.010568	[0.636942,0.691387]
	2	1988	0.386821	0.010926	[0.358678,0.414964]
	3	1968	0.269817	0.010008	[0.244038,0.295596]
P2Rank	1	1866	0.008039	0.002068	[0.002712,0.013365]
	2	1441	0.022207	0.003883	[0.012204,0.032209]
	3	1078	0.022263	0.004496	[0.010683,0.033844]

(b) $\tau = 0.95$

Table B: Estimates of equivariance proportions regarding rotations around y axis.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1852	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1998	0.832332	0.008360	[0.810800,0.853865]
	2	1988	0.663481	0.010600	[0.636176,0.690786]
	3	1968	0.550305	0.011217	[0.521413,0.579197]
Fpocket	1	1998	0.669670	0.010525	[0.642559,0.696780]
	2	1988	0.390845	0.010946	[0.362649,0.419041]
	3	1968	0.273374	0.010049	[0.247489,0.299259]
P2Rank	1	1866	0.378885	0.011233	[0.349951,0.407820]
	2	1441	0.185288	0.010239	[0.158915,0.211661]
	3	1078	0.126160	0.010117	[0.100099,0.152220]

(c) $\tau = 0.75$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1852	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1998	0.835836	0.008289	[0.814484,0.857187]
	2	1988	0.669014	0.010557	[0.641822,0.696206]
	3	1968	0.557927	0.011198	[0.529083,0.586771]
Fpocket	1	1998	0.672172	0.010504	[0.645114,0.699230]
	2	1988	0.391851	0.010951	[0.363642,0.420060]
	3	1968	0.273374	0.010049	[0.247489,0.299259]
P2Rank	1	1866	0.793676	0.009370	[0.769540,0.817813]
	2	1441	0.537821	0.013138	[0.503979,0.571663]
	3	1078	0.395176	0.014897	[0.356804,0.433549]

(d) $\tau = 0.50$

Table B: Estimates of equivariance proportions regarding rotations around y axis.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1853	0.999460	0.000540	[0.998070,1.000850]
Deeppocket	1	1996	0.792084	0.009086	[0.768681,0.815487]
	2	1983	0.622289	0.010890	[0.594239,0.650340]
	3	1970	0.532995	0.011243	[0.504034,0.561956]
Fpocket	1	1998	0.647648	0.010690	[0.620113,0.675183]
	2	1989	0.382604	0.010901	[0.354526,0.410682]
	3	1970	0.268020	0.009982	[0.242309,0.293732]
P2Rank	1	1866	0.007503	0.001998	[0.002356,0.012650]
	2	1441	0.022207	0.003883	[0.012204,0.032209]
	3	1078	0.022263	0.004496	[0.010683,0.033844]

(a) $\tau = 0.99$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1853	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1996	0.797595	0.008996	[0.774424,0.820766]
	2	1983	0.625819	0.010870	[0.597821,0.653818]
	3	1970	0.535025	0.011240	[0.506072,0.563979]
Fpocket	1	1998	0.652152	0.010658	[0.624699,0.679606]
	2	1989	0.384113	0.010909	[0.356014,0.412211]
	3	1970	0.269036	0.009994	[0.243293,0.294778]
P2Rank	1	1866	0.008039	0.002068	[0.002712,0.013365]
	2	1441	0.022207	0.003883	[0.012204,0.032209]
	3	1078	0.022263	0.004496	[0.010683,0.033844]

(b) $\tau = 0.95$

Table C: Estimates of equivariance proportions regarding rotations around z axis.

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1853	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1996	0.803607	0.008894	[0.780697,0.826517]
	2	1983	0.631871	0.010833	[0.603966,0.659776]
	3	1970	0.539086	0.011234	[0.510151,0.568022]
Fpocket	1	1998	0.658659	0.010610	[0.631328,0.685989]
	2	1989	0.386626	0.010922	[0.358493,0.414760]
	3	1970	0.270558	0.010012	[0.244770,0.296347]
P2Rank	1	1866	0.377814	0.011227	[0.348895,0.406732]
	2	1441	0.184594	0.010224	[0.158259,0.210929]
	3	1078	0.127087	0.010149	[0.100945,0.153230]

(c) $\tau = 0.75$

Method $\mathcal{M}$	Pocket j	Non-missing	$\hat{p}_j^{\mathcal{M};\tau}$	$\mathrm{SE}(\hat{p}_j^{\mathcal{M};\tau})$	$CI(p_j^{\mathcal{M};\tau})$
GENEOnet	1	2000	1.000000	0.000000	[1.000000,1.000000]
	2	1918	1.000000	0.000000	[1.000000,1.000000]
	3	1853	1.000000	0.000000	[1.000000,1.000000]
Deeppocket	1	1996	0.809619	0.008790	[0.786978,0.832260]
	2	1983	0.637418	0.010799	[0.609603,0.665233]
	3	1970	0.547208	0.011218	[0.518313,0.576103]
Fpocket	1	1998	0.660661	0.010595	[0.633369,0.687953]
	2	1989	0.388638	0.010932	[0.360478,0.416797]
	3	1970	0.271066	0.010017	[0.245263,0.296869]
P2Rank	1	1866	0.792605	0.009388	[0.768422,0.816787]
	2	1441	0.537127	0.013140	[0.503281,0.570973]
	3	1078	0.396104	0.014903	[0.357716,0.434492]

(d) $\tau = 0.50$

Table C: Estimates of equivariance proportions regarding rotations around z axis.

Protein	$\overline{O}_{100}$	$\overline{O}_{10}$	$\overline{O}_{100} - \overline{O}_{10}$	p-value	
16pk_A	0.879065	0.901982	-0.022916	0.000003	***
1d4t_A	0.802809	0.830869	-0.028060	0.001863	**
1fcz_A	0.887851	0.910334	-0.022483	0.000046	***
1g6g_A	0.822016	0.827410	-0.005394	0.285229	
1ikt_A	0.819245	0.931655	-0.112410	0.000000	***
1jif_A	0.800431	0.818471	-0.018040	0.053890	.
1n8u_A	0.814256	0.967725	-0.153469	0.000000	***
1nf8_A	0.865969	0.861657	0.004312	0.699292	
1ntv_A	0.844556	0.861024	-0.016469	0.013426	*
1ssq_D	0.819931	0.846528	-0.026597	0.000288	***
1ugp_A	0.840670	0.868306	-0.027637	0.000033	***
2e5y_B	0.865180	0.892541	-0.027362	0.000015	***
2f5t_X	0.838101	0.845194	-0.007093	0.173906	
2hxm_A	0.832058	0.840013	-0.007955	0.160914	
2pmk_A	0.831945	0.870040	-0.038095	0.000002	***
2qmj_A	0.861590	0.879843	-0.018254	0.001882	**
2vfk_A	0.893659	0.910764	-0.017104	0.000369	***
2vmc_A	0.738895	0.788653	-0.049757	0.000000	***
2wr8_A	0.864296	0.887062	-0.022766	0.000169	***
2x96_A	0.930616	0.941810	-0.011194	0.000011	***
3d1g_A	0.838189	0.853228	-0.015039	0.011526	*
3ebh_A	0.928322	0.933405	-0.005083	0.050812	.
3g5k_D	0.931966	0.929183	0.002783	0.752090	
3lbl_A	0.880286	0.905585	-0.025300	0.000101	***
3o0g_D	0.837158	0.834251	0.002907	0.627081	
3p3g_A	0.900774	0.919091	-0.018317	0.000064	***
3rt4_A	0.868931	0.893306	-0.024375	0.000017	***
3v3l_A	0.828454	0.868580	-0.040126	0.000000	***
4bt9_B	0.909766	0.926493	-0.016726	0.000015	***
4cjn_A	0.846983	0.894977	-0.047994	0.000000	***
4qxo_A	0.867791	0.881104	-0.013313	0.012176	*
5ggl_A	0.798100	0.793500	0.004600	0.644964	
5gof_A	0.861765	0.886590	-0.024824	0.000001	***
5l87_A	0.824715	0.841245	-0.016530	0.049069	*
5ok6_A	0.827920	0.871197	-0.043277	0.000000	***
6bm5_A	0.845692	0.865143	-0.019452	0.002517	**
6gfx_C	0.852719	0.880447	-0.027728	0.000011	***

Table D: Mean overlap differences between 100 ps (complete) and 10 ps (first 250 frames) simulations. p-values are relative to independent two samples T-tests for testing $H_0\colon \mu_{100} = \mu_{10}$ against $H_1\colon \mu_{100} < \mu_{10}$.

Protein	Mean 250	Mean 10000	Diff mean	Wass dist
16pk_A	0.992395	0.977577	0.014818	0.021810
1d4t_A	1.026452	0.989962	0.036491	0.043165
1fcz_A	0.954021	0.956506	-0.002486	0.006645
1g6g_A	0.992114	0.984260	0.007854	0.014901
1ikt_A	1.052467	1.017474	0.034994	0.041526
1jif_A	1.015065	1.022812	-0.007747	0.012325
1n8u_A	1.006749	1.040980	-0.034231	0.034231
1nf8_A	0.956974	0.988713	-0.031739	0.031739
1ntv_A	1.014137	0.997826	0.016311	0.022683
1ssq_D	1.003967	0.977806	0.026161	0.032606
1ugp_A	1.105593	1.073041	0.032552	0.040903
2e5y_B	1.054653	1.049093	0.005560	0.012588
2f5t_X	0.972836	0.952307	0.020529	0.026929
2hxm_A	0.936547	0.927659	0.008888	0.015240
2pmk_A	1.023954	1.016749	0.007205	0.014003
2qmj_A	0.899421	0.905052	-0.005631	0.006076
2vfk_A	1.019915	1.018705	0.001210	0.010206
2vmc_A	1.166225	1.162091	0.004134	0.022087
2wr8_A	0.992018	0.985540	0.006477	0.013619
2x96_A	0.913458	0.932021	-0.018563	0.018563
3d1g_A	1.053743	1.048730	0.005013	0.012630
3ebh_A	0.923923	0.926086	-0.002163	0.005716
3g5k_D	1.034933	1.005734	0.029199	0.035757
3lbl_A	1.057018	1.008158	0.048860	0.055188
3o0g_D	0.967058	0.978287	-0.011228	0.011341
3p3g_A	0.929515	0.936809	-0.007294	0.008842
3rt4_A	0.948901	0.902198	0.046703	0.052591
3v3l_A	1.062354	1.027277	0.035077	0.041480
4bt9_B	1.230349	1.154759	0.075590	0.082899
4cjn_A	1.092455	1.070507	0.021947	0.029452
4qxo_A	1.012820	1.012443	0.000376	0.007283
5ggl_A	1.011885	0.963687	0.048197	0.054318
5gof_A	1.092878	1.053894	0.038984	0.046002
5l87_A	1.101899	1.052890	0.049010	0.055158
5ok6_A	1.153471	1.133152	0.020319	0.028025
6bm5_A	0.977482	0.985086	-0.007604	0.009148
6gfx_C	1.147431	1.128881	0.018550	0.026664

Table E: Comparison of RMSD distributions in the case of finer simulations with a time delta of 10 ps. The first column shows the average RMSD considering the first 250 frames, while the second one reports the average RMSD value using the entire 10000 frames. The last two columns contain the difference between the sample means and the Wasserstein distance computed using the empirical distributions, in order to compare also the distributions instead of the means only. All such differences and distances are of the order of 10^{-2} or even smaller.