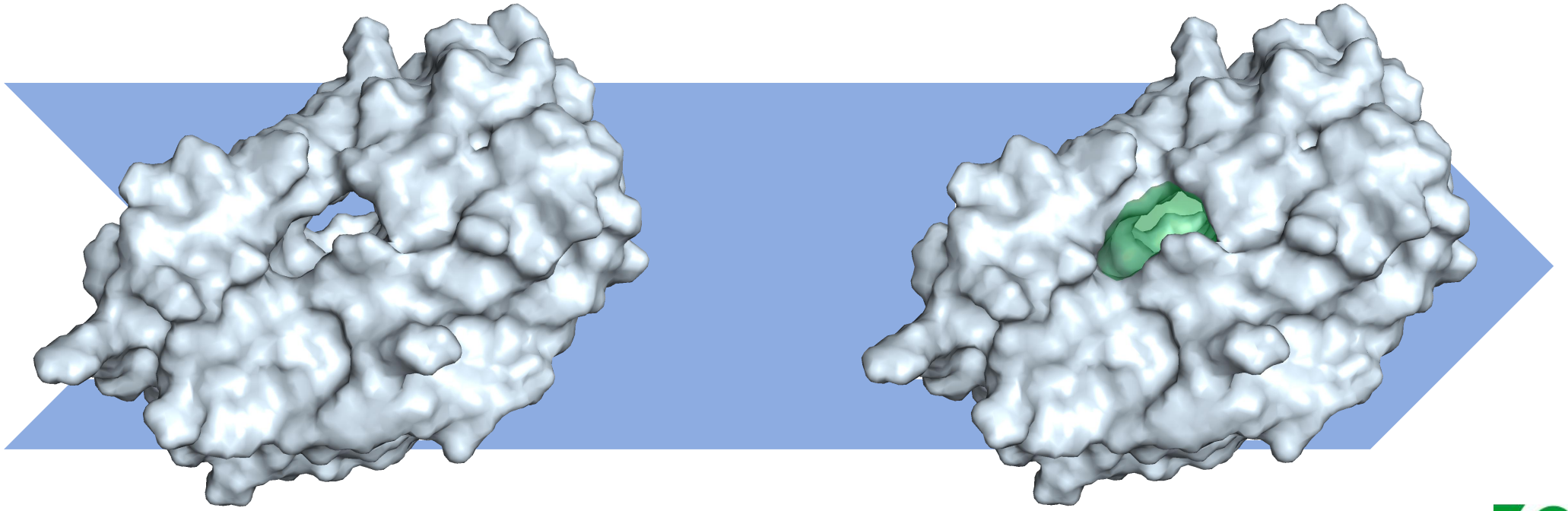
# Prototypical Explanations in an AI method for Protein Pocket Detection

Giovanni Bocchi [1], Alessandra Micheletti [1],
Carmen Gratteri [2] and Carmine Talarico [2]

[1] Department of Environmental Science and Policy, University of Milan

[2] Dompé Farmaceutici S.p.A.

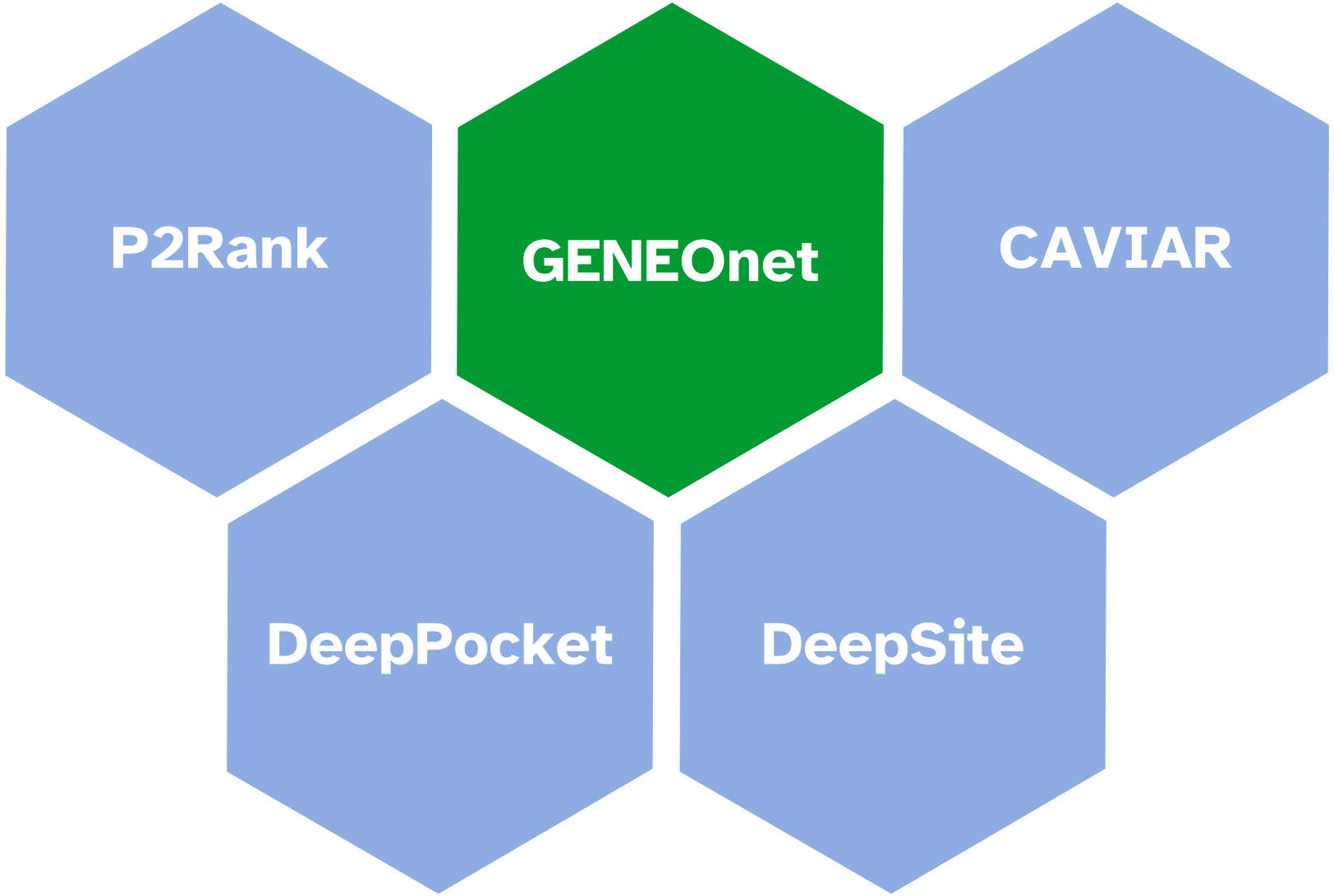
# Problem 1: Protein Pocket Detection



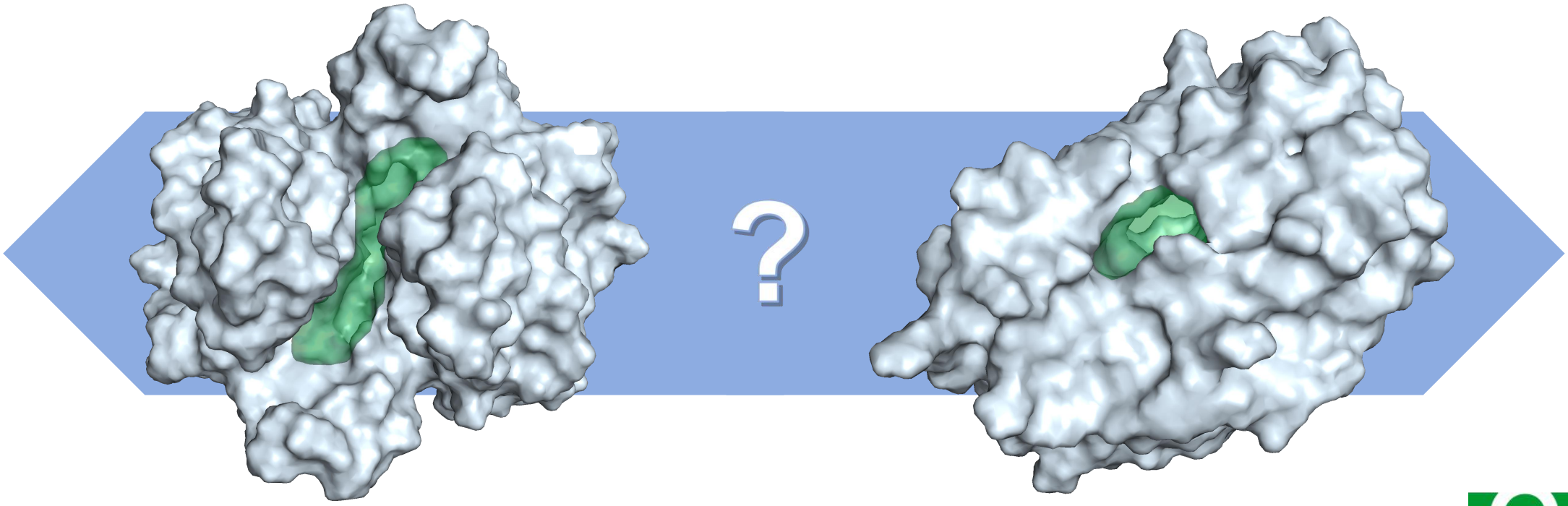
Protein 3BVB surface

Protein 3BVB pocket

# AI (not that SAFE) solutions



# Problem 2: Pocket Similarity




Protein 11GS pocket

Protein 3BVB pocket

# Group Equivariant Non-Expansive Operators

## GENEO (Group Equivariant Non-Expansive Operators)

Given two functional spaces $\Phi = \{\varphi: X \to \mathbb{R}\}$ and $\Psi = \{\psi: Y \to \mathbb{R}\}$, two groups $G$ and $H$ of transformations of the functions domains ($X$ and $Y$) and a fixed homomorphism $T: G \to H$, we define a Group Equivariant Non-Expansive Operator as a function $F$ from $\Phi$ to $\Psi$ with the following two properties:

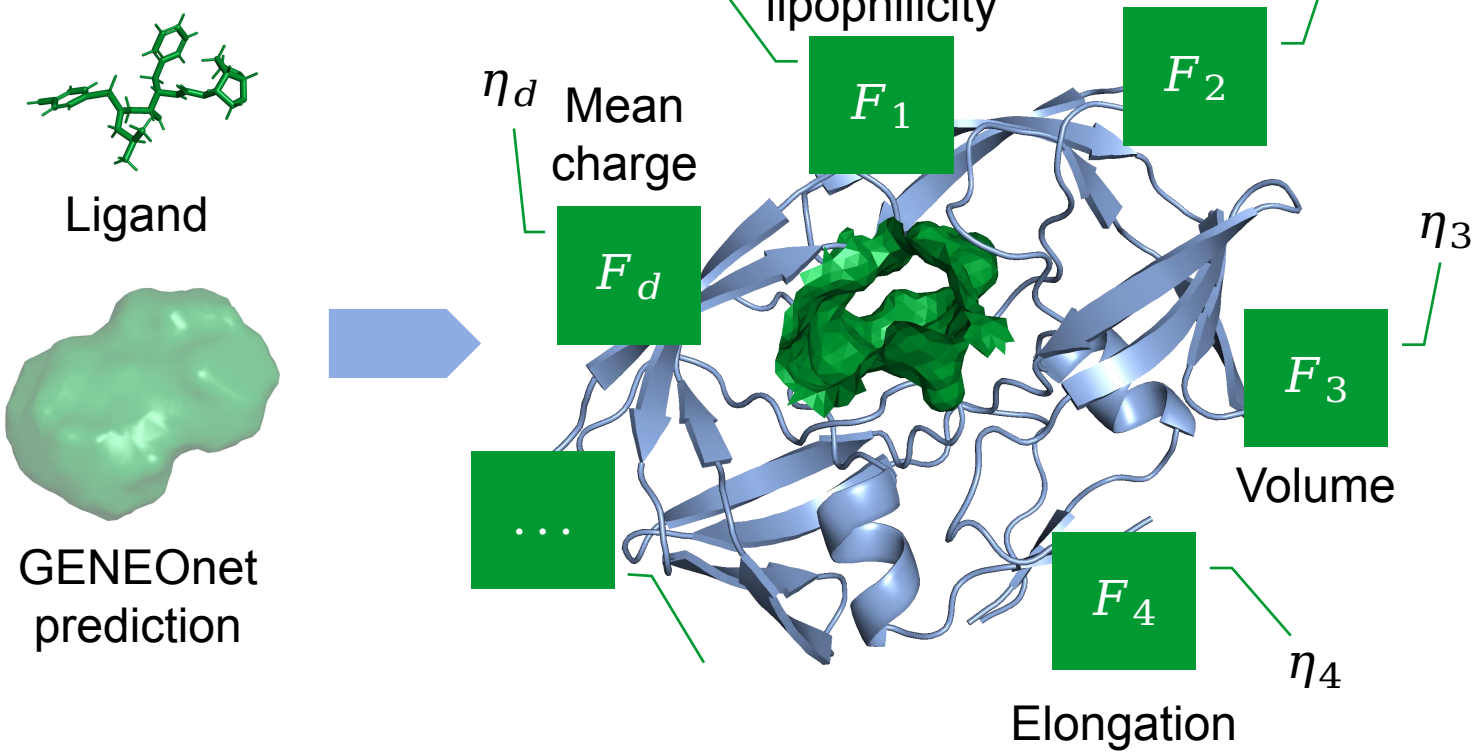
**Equivariance:** For every $\varphi \in \Phi$ and $g \in G$ it holds that

$$\boldsymbol{F(\varphi \circ g) = F(\varphi) \circ T(g)}$$

**Non-Expansivity:** For every $\varphi_1, \varphi_2 \in \Phi$ it holds that

$$\boldsymbol{\|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty}$$

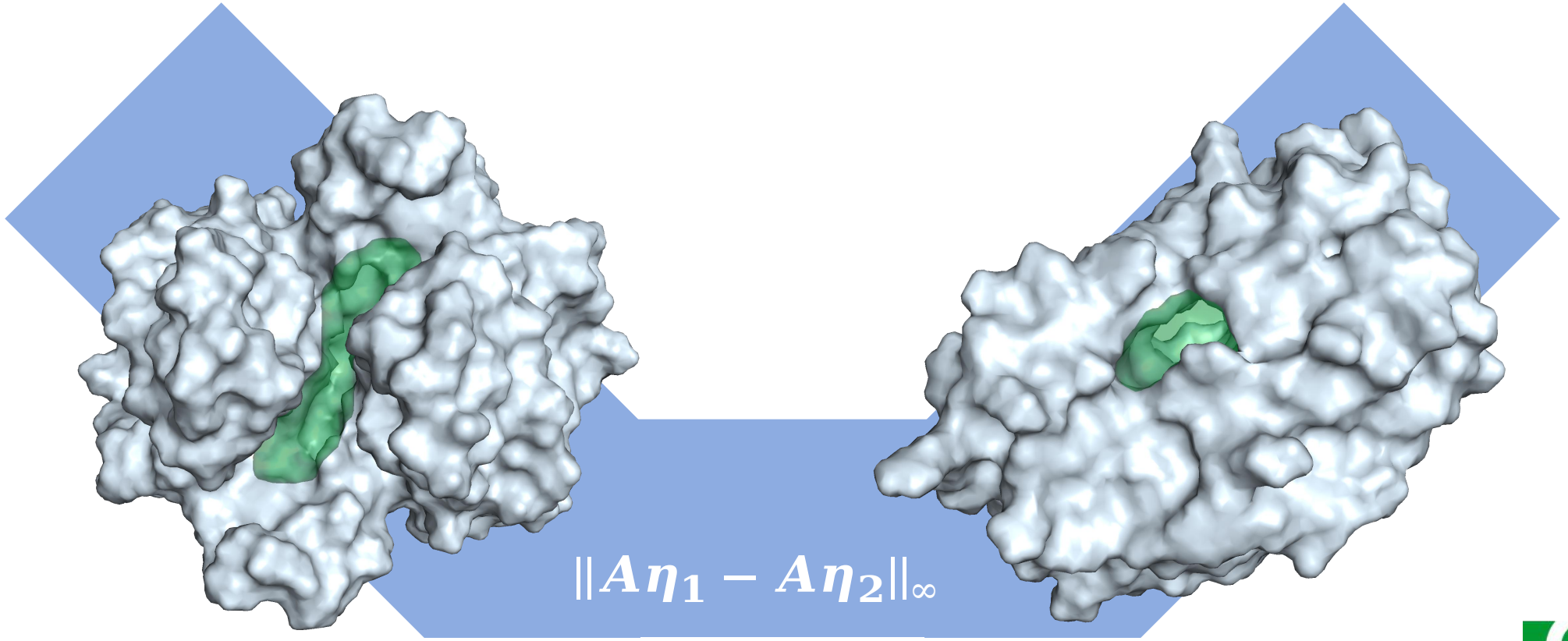
# GENEOmap




Each GENEO $F_i$ computes an isometry invariant coefficient $\eta_i$. Together, such coefficients, constitute the isometry invariant embedding of the pocket.

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_d \end{bmatrix} \in \mathbb{R}^d$$

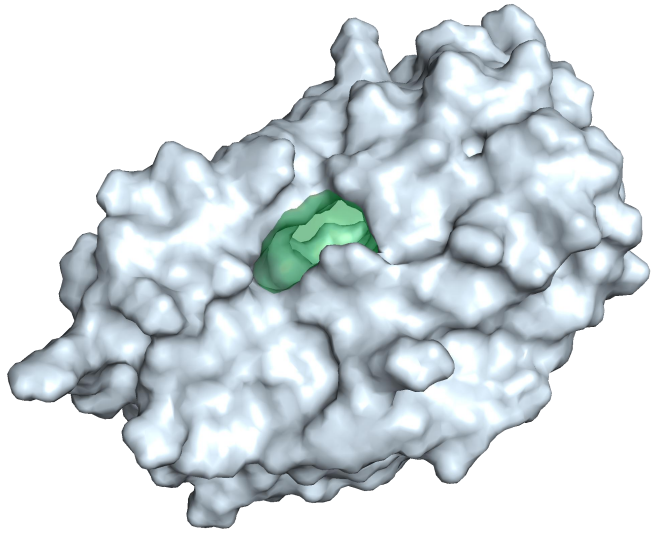
# GENEOmap: similarity measure




Protein 11GS pocket mapped to $\eta_1$

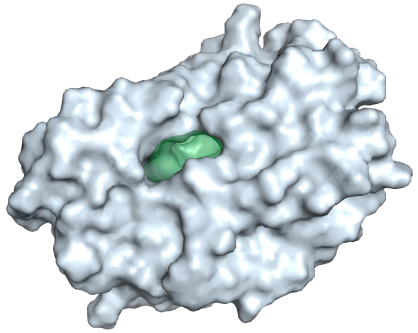
Protein 3BVB pocket mapped to $\eta_2$

# Similarity enables to define Prototypes



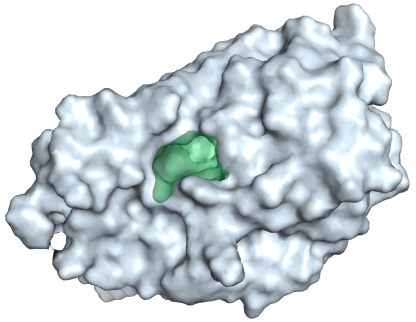
Protein: 3BVB Ligand: ???
Pocket mapped to $\tilde{\eta}$

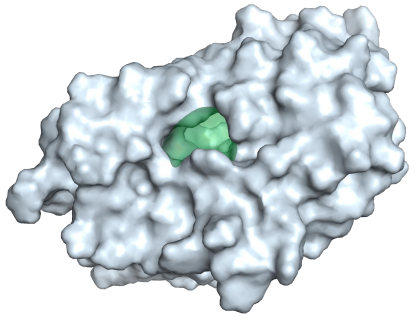
**Library of pockets mapped to $\eta_i$**



1st nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 4Q1X
Ligand: 017

2nd nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 5QKY
Ligand: 017





3rd nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 3TKW
Ligand: 017

# Results on ProSPECCTs benchmark

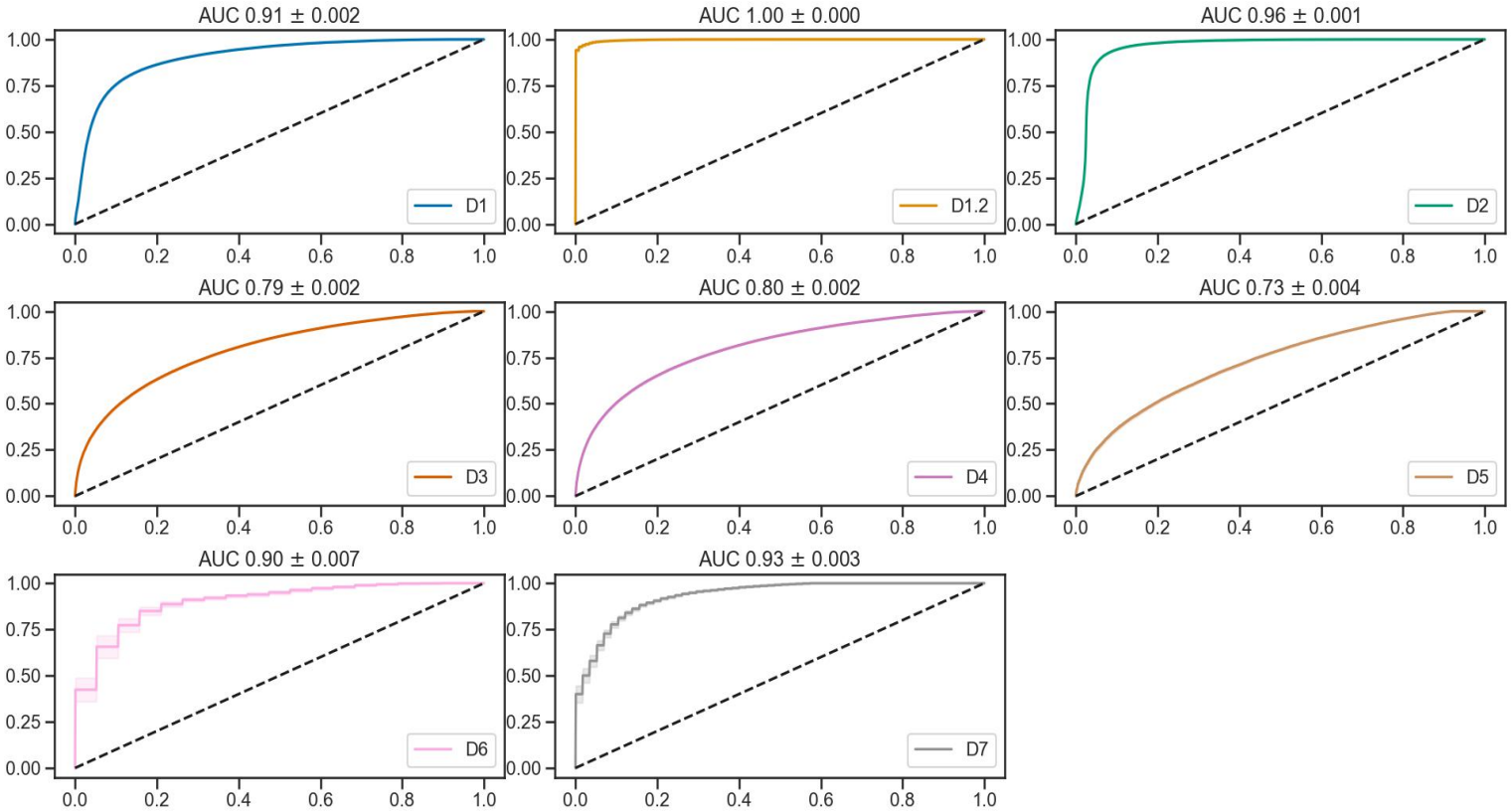
ProSPECTTs it's a well known benchmark for pocket similarity in a binary classification sense.

| Dataset | similar (y=0) | dissimilar (y=1) |
|---|---|---|
| D1 | 13430 | 92846 |
| D1.2 | 241 | 1784 |
| D2 | 7729 | 100512 |
| D3 | 13430 | 67150 |
| D4 | 13430 | 67150 |
| D5 | 920 | 5480 |
| D6 | 19 | 43 |
| D7 | 115 | 56284 |

Given two pocket embeddings $\eta_1, \eta_2$ we say that they are similar if $\|A\eta_1 - A\eta_2\|_\infty \leq \tau$ or dissimilar otherwise. The matrix $A$ can be learned using a subset of the data.

$$\widehat{\boldsymbol{y}} = \mathbf{1}_{[\boldsymbol{\tau},+\infty)}(\|\boldsymbol{A\eta_1} - \boldsymbol{A\eta_2}\|_\infty)$$

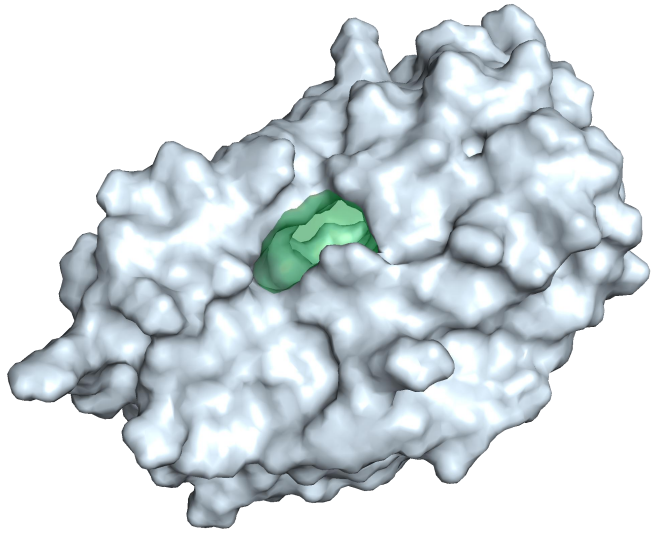
# Results on ProSPECCTs benchmark




ROC curves for GENEOmap on ProSPECCTs

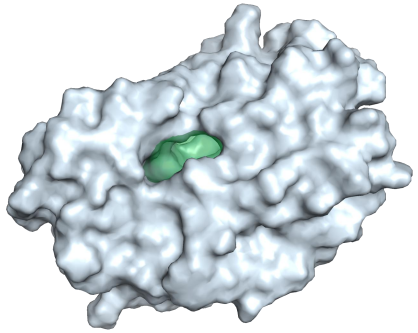
| Dataset | GENEOmap | Site2Vec |
|---|---|---|
| D1 | 0.91±0.002 | 1.00 |
| D1.2 | 1.00±0.000 | 0.94 |
| D2 | 0.96±0.001 | 1.00 |
| D3 | 0.79±0.002 | 0.99 |
| D4 | 0.80±0.002 | 0.99 |
| D5 | 0.73±0.004 | 0.86 |
| D6 | 0.90±0.007 | 0.53 |
| D7 | 0.93±0.003 | 0.66 |
| Mean | 0.88±0.001 | 0.87 |

# Prototypes for GENEOnet predictions



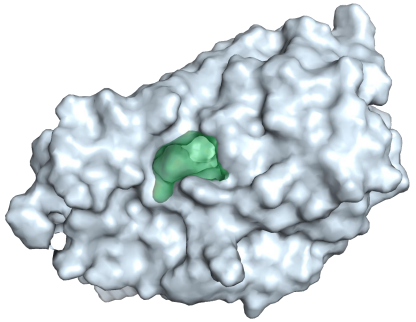
Protein: 3BVB Ligand: ???
**GENEOnet predicted pocket**
mapped to $\tilde{\eta}$

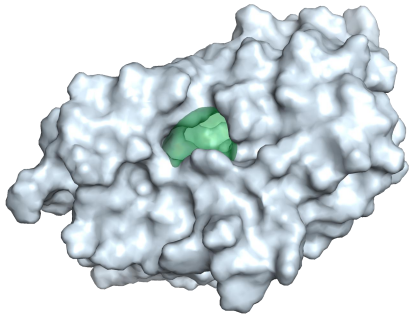
Library of
**GENEOnet predictions**
mapped to $\eta_i$



1st nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 4Q1X
Ligand: 017

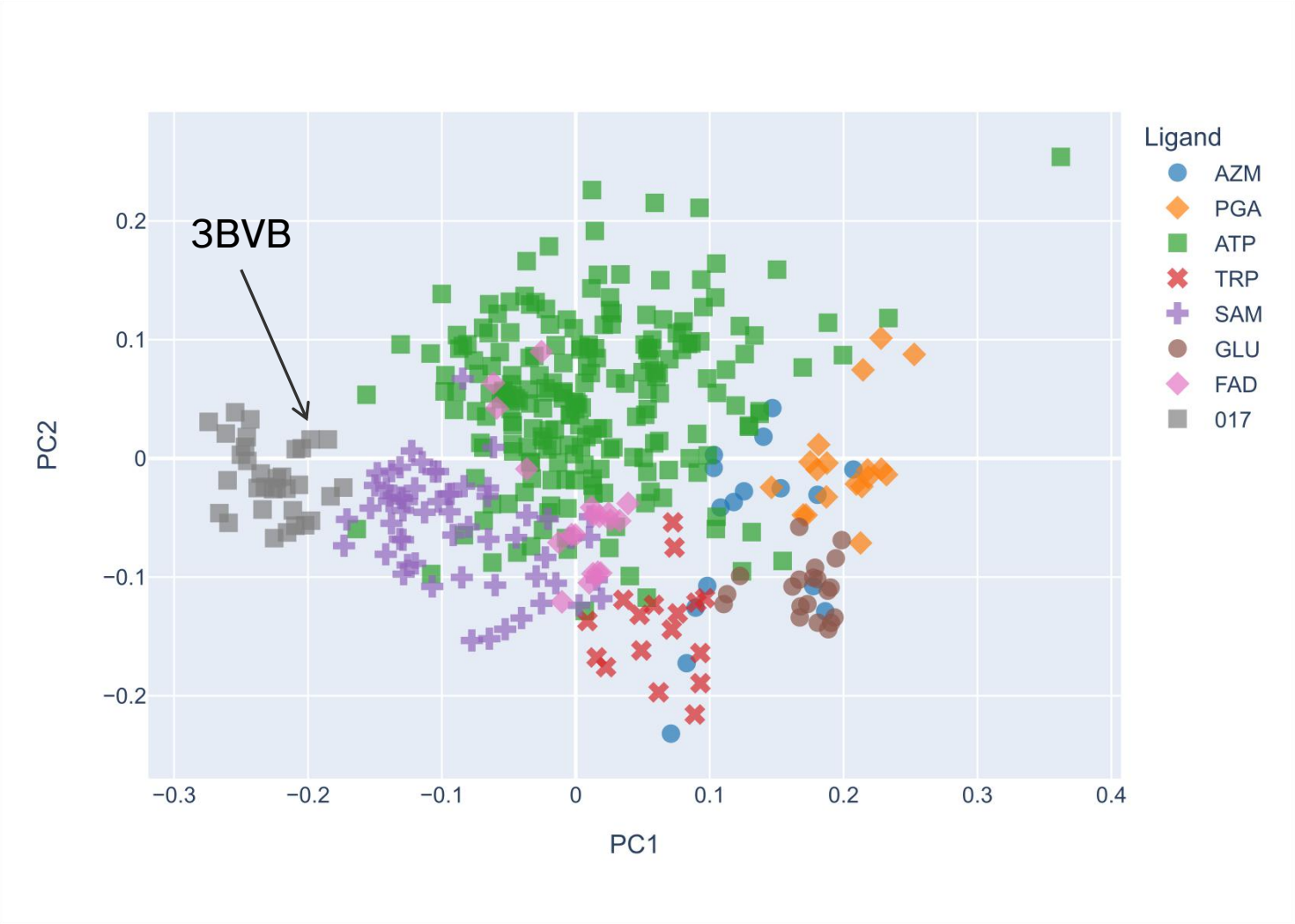
2nd nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 5QKY
Ligand: 017





3rd nn $\|\tilde{\eta} - \eta_i\|_\infty$
Protein: 3TKW
Ligand: 017

# Visualization of GENEOmap embeddings



# Thank you for the attention!

# Main references


- Ehrt, C., Brinkjost, T., Koch, O.: A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLOS Computational Biology* **14**(11), 1–50 (2018) https://doi.org/10.1371/journal.pcbi.1006483
- Bhadra, A., Yeturu, K.: Site2vec: a reference frame invariant algorithm for vector embedding of protein-ligand binding sites. *Machine Learning-Science and Technology* **2**(1) (2021) https://doi.org/10.1088/2632-2153/abad88
- 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)
- Bocchi, G., Frosini, P., Micheletti, A., et al.: GENEOnet: statistical analysis supporting explainability and trustworthiness. *Statistics* **0**(0), 1–26 (2025) https://doi.org/10.1080/02331888.2025.2478203