3DOR 2022, 1-2 September, Florence, Italy

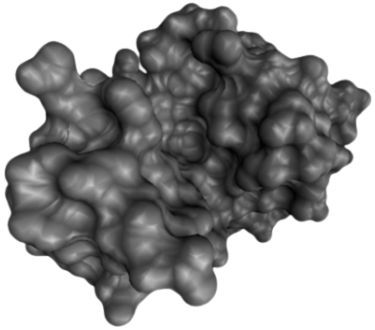
SHREC 2022:
Protein-Ligand Binding Site Recognition

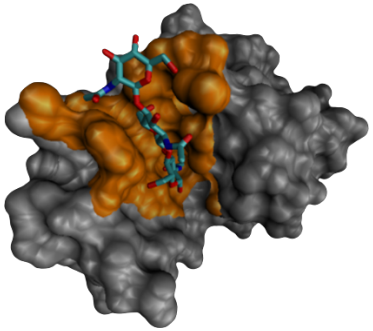
Luca Gagliardi, Andrea Raffo, Ulderico Fugacci, Silvia Biasotti, Walter Rocchia,
Hao Huang, Boulbaba Ben Amor, Yi Fang, Yuanyuan Zhang, Xiao Wang, Charles Christoffer,
Daisuke Kihara, Apostolos Axenopoulos, Stelios Mylonas, Petros Daras

Protein-Ligand Binding Site Recognition

Motivation:

*The **identification on protein surface of regions (called pockets) able to bind ligands** is one of the focal points of research activity in **computational biophysics** and **structural biology***





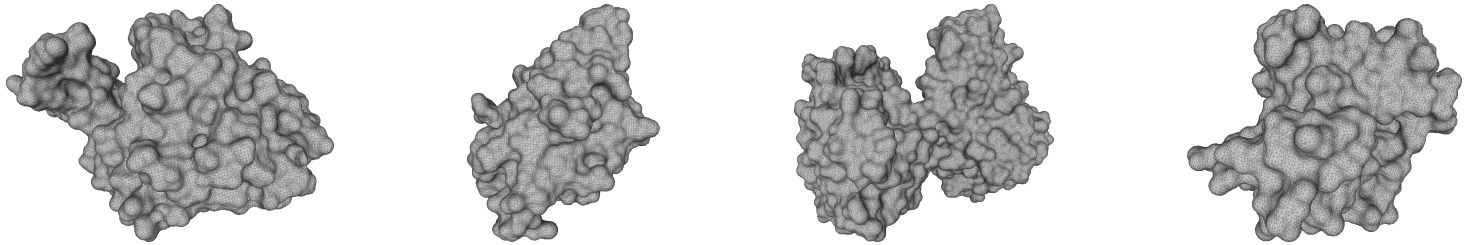
*which is **essential** and **preparatory** to*
***drug design**, **molecular docking**, development of **innovative therapeutic strategies**, …*

Goal:

*The general **objective of this SHREC track** is to*
*evaluate the effectiveness of computational methods in **recognizing***
most likely protein-ligand binding sites based on the geometrical structure of the protein

Protein-Ligand Binding Site Recognition

Dataset:



*We **extracted** protein-ligand complexes from **Binding MOAD** and processed by:*

- ***Considering** just ligands with **significative ligand molecular weight** and **resolution***
- ***Removing** **redundant structures***
- ***Creating** **PQR files** using the AMBER force field via the pdb2pqr software*
- ***Building** the triangulation (in OFF format) of the **SES molecular surfaces** via NanoShaper*
- ***Discarding** structures with **multiple connected components***
- ***Labeling** atoms and triangulation vertices in accordance with the **identified binding sites***
- ***Dropping** **highly overlapping binding regions***

Protein-Ligand Binding Site Recognition

Dataset:

The resulting dataset consists of:

- ***1091 protein structures***
- ***1721 binding sites***

Protein-Ligand Binding Site Recognition

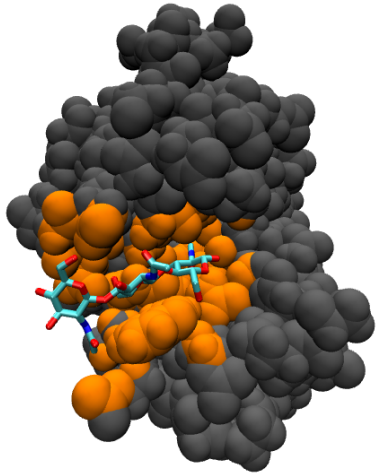
Dataset:

The resulting dataset consists of:

- ***1091* protein structures**
- ***1721* binding sites**

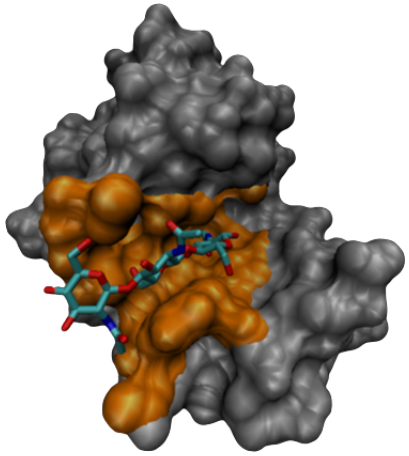
Provided in two different representations:

- ***Atom spheres* (PQR format)**
- ***SES surface* (OFF format)**



Atom spheres

SES surface



Protein-Ligand Binding Site Recognition

Dataset:

The resulting dataset consists of:

- ***1091 protein structures***
- ***1721 binding sites***

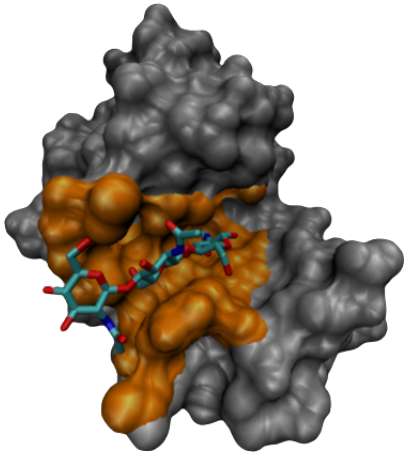
Provided in two different representations:

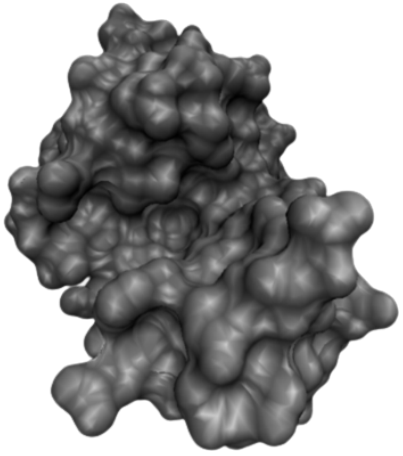
- ***Atom spheres (PQR format)***
- ***SES surface (OFF format)***

Subdivided into:

- ***Training set (85%)***
- ***(anonymized) Test set (15%)***

Training set





Test set

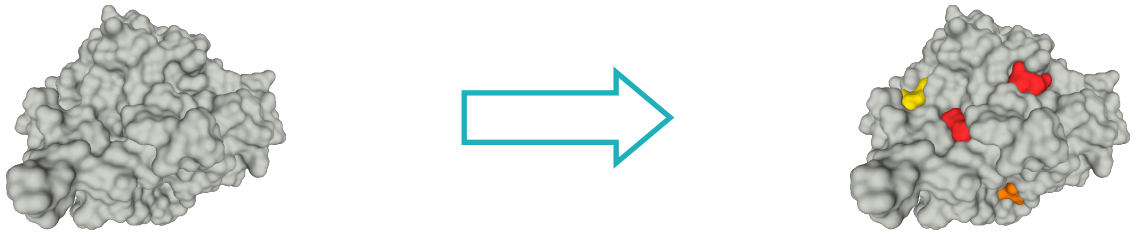
Protein-Ligand Binding Site Recognition

Task:

Given the dataset, we asked the participants to provide, for each protein in the test set,

a vector representing the 10 most likely binding sites they have identified in the model

either in terms of the vertices (if using the OFF files) or of the atoms (if using the PQR files)



*Further, we asked to provide **a ranking of the predicted sites**, from the most to least likely*

Remarks:

- *A single structure can contain **more than one binding site***
- *The training set does not imply any ranking*
  - *All provided pockets are **positive examples** and should be considered **equally important***

Protein-Ligand Binding Site Recognition

Methods:

Eight groups *from* ***four different countries*** *registered to the track*
Four *of them proceeded with the* ***submission*** *of their results:*

- ***Method M1 — Point Transformer***
  - *by H. Huang, B. Ben Amor, Y. Fang*
- ***Method M2 — GNN-Pocket***
  - *by Y. Zhang, X. Wang, C. Christoffer, D. Kihara*
- ***Method M3 — DeepSurf***
  - *by A. Axenopoulos, S. Mylonas, P. Daras*
- ***Method M4 — NS-Volume***
  - *by L. Gagliardi, W. Rocchia*

The ***organizers*** *of the track are L. Gagliardi, A. Raffo, U. Fugacci, S. Biasotti, W. Rocchia*

Protein-Ligand Binding Site Recognition

Methods:

Eight groups from four different countries registered to the track
Four of them proceeded with the submission of their results:

- **Method M1 — Point Transformer**
  - by H. Huang, B. Ben Amor, Y. Fang

OFF representation

- **Method M2 — GNN-Pocket**
  - by Y. Zhang, X. Wang, C. Christoffer, D. Kihara
- **Method M3 — DeepSurf**
  - by A. Axenopoulos, S. Mylonas, P. Daras
- **Method M4 — NS-Volume**
  - by L. Gagliardi, W. Rocchia

PQR representation

The organizers of the track are L. Gagliardi, A. Raffo, U. Fugacci, S. Biasotti, W. Rocchia

Protein-Ligand Binding Site Recognition

Methods:

Eight groups from four different countries registered to the track
Four of them proceeded with the submission of their results:

- **Method M1 — Point Transformer**
  - *by H. Huang, B. Ben Amor, Y. Fang*
- **Method M2 — GNN-Pocket**
  - *by Y. Zhang, X. Wang, C. Christoffer, D. Kihara*
- **Method M3 — DeepSurf**
  - *by A. Axenopoulos, S. Mylonas, P. Daras*

Learning approach

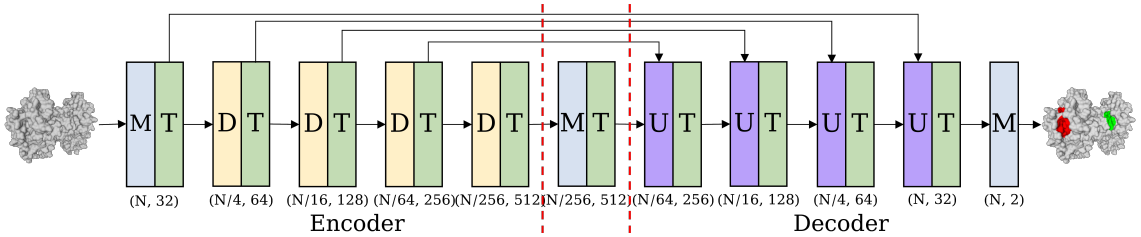
- **Method M4 — NS-Volume**
  - *by L. Gagliardi, W. Rocchia*

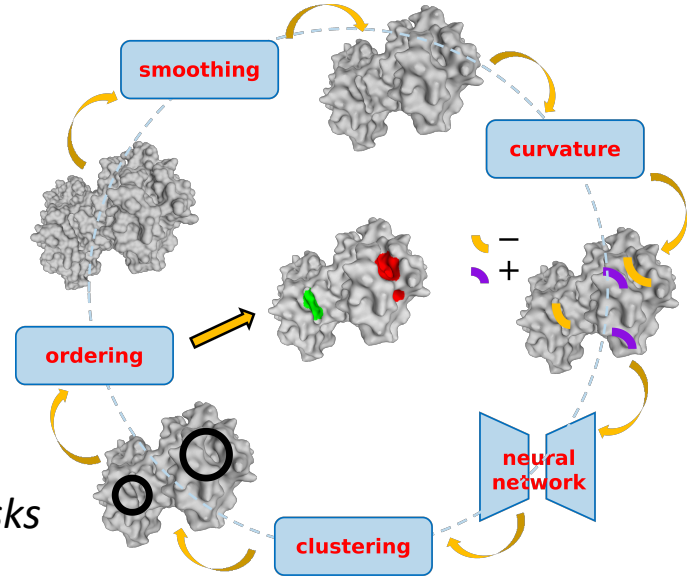
Direct approach

The organizers of the track are L. Gagliardi, A. Raffo, U. Fugacci, S. Biasotti, W. Rocchia

Protein-Ligand Binding Site Recognition

M1 — Point Transformer:

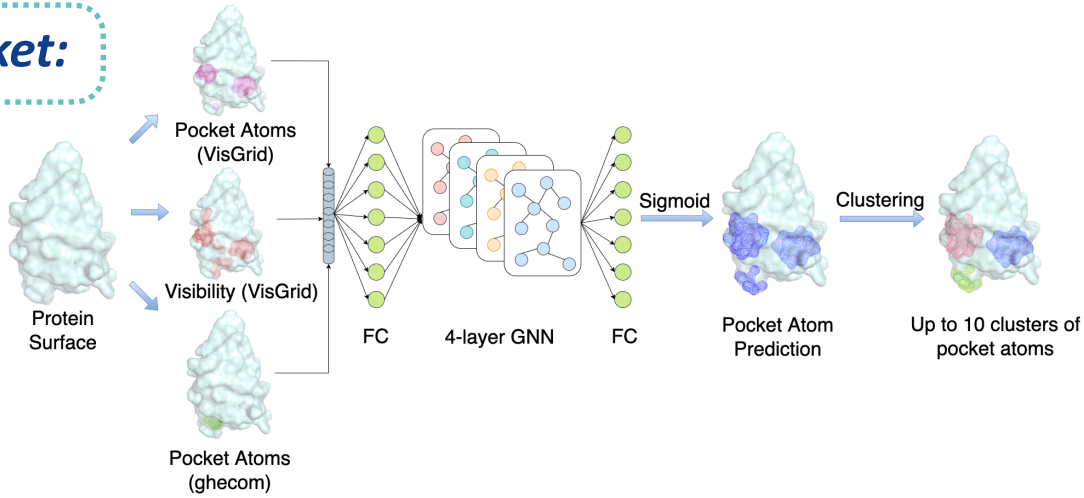




- *Originally proposed for machine translation, it has achieved notable performance on various computer vision tasks*
- *The **neural network** model:*
  - *Adopts a **U-Net architecture** consisting of an encoder and a decoder*
  - *Is adapted lo learn **per-vertex local shape geometric features***
  - *Is fed with a **5-dimensional vertex feature (coordinates and curvatures)** to predict a binary segmentation result as a ligandability score*
- *Binding regions are obtained by **clustering** the vertices with a high ligandability score through a **density-based algorithm***
  - *Regions are **filtered and ranked** on the basis of the **average squared ligandability score** of their vertices*

Protein-Ligand Binding Site Recognition

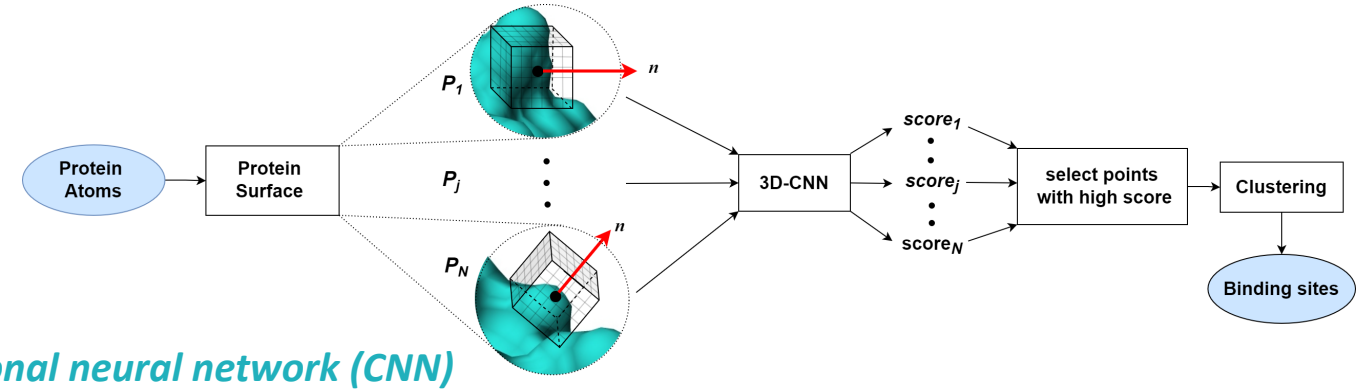
M2 — GNN-Pocket:



- *Based on a **graph neural network (GNN)** fed with a feature vector for each atom*
- *The vector is obtained by concatenating 3 features:*
  - *A binary output from VisGrid, which indicates if an atom has a **visibility** lower than a cutoff*
  - *The **number of closest grid points** that are **predicted as pockets by ghecom***
  - *The **number of grid points within 8 Å** that are **predicted as pockets by VisGrid***
- *A **bottom-up hierarchical clustering** method, which minimizes the distance between the closest pairs of clusters, is adopted to group pocket atoms into pocket regions*
  - *The top-10 pockets by the **sum of probability values of atoms** are selected*

Protein-Ligand Binding Site Recognition

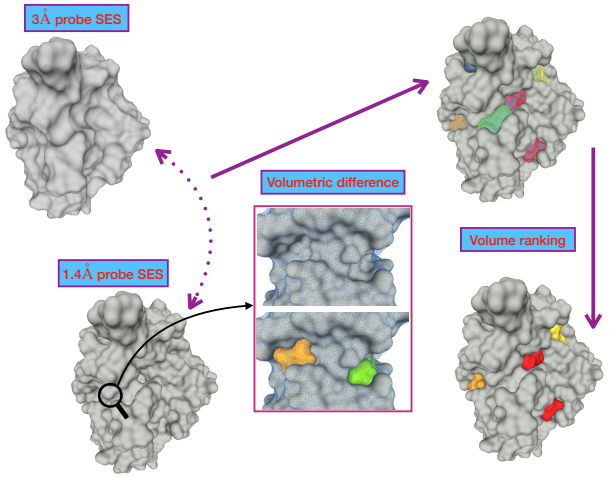
M3 — DeepSurf:



- *Based on a **convolutional neural network (CNN)***
- *A number of **local 3D voxelized grids** are placed on protein surface and used for extracting features with which feed the network*
- *For each protein atom, **18 chemical features** are calculated*
  - *Each grid voxel receives the features of the atoms inside it*
- *This requires information on the **atom types** that **lacks** **in the provided database***
  - *Such information was **inferred from the atom radii** (regarded as a highly confident indication of the atom type)*
- *Binding regions are obtained by **clustering** the vertices with a high ligandability score using the **mean-shift algorithm***
  - *Regions are **sorted** on the basis of the **average ligandability score** of their vertices*
- *DeepSurf was originally **trained on scPDB database***
  - *16034 entries corresponding to 4782 proteins with 17594 total binding samples*

Protein-Ligand Binding Site Recognition

M4 — NS-Volume:



- *Based on **NanoShaper**, an efficient software for the triangulation of molecular surfaces*
  - *NanoShaper **offers also a pocket detection function***
- ***Pockets** are defined as the **volumetric difference between** the space regions enclosed within the SESs of the protein obtained with **two different probe radii***
  - *Points are flagged if **simultaneously** inside the 3 Å SES and outside the 1.4 Å SES (water molecule effective radius)*
  - *A **filtering procedure** is adopted which preserves points which are*
    - *(i) within 1.4 Å from all flagged point or*
    - *(ii) within 1.4 Å from points fulfilling (i)*
  - *Pockets are defined as the **unconnected components** after the filtering by applying a **flood-fill procedure***
- *Obtained pockets are **sorted by volume***

Protein-Ligand Binding Site Recognition

Evaluation:

Inspired by state-of-the-art biophysical pocket detection methods, we adopted a figure of merit based on the ***combination of two scores***

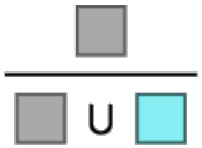
Protein-Ligand Binding Site Recognition

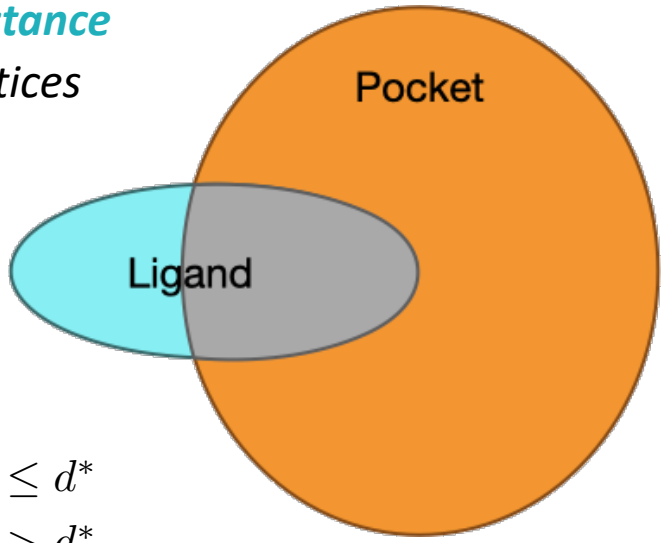
Evaluation:

Inspired by state-of-the-art biophysical pocket detection methods, we adopted a figure of merit based on the ***combination of two scores***

Ligand Coverage Score:

Fraction of ligand heavy atoms within a threshold distance from *the protein atoms (PQR file) or of the surface vertices (OFF file) that compose a* ***putative pocket***

LC := 



Formally,

$$\mathrm{LC} = \frac{1}{n_L}\sum_{j=1}^{n_L}\delta_{ij} \quad \text{for } \forall\, i \in \mathcal{P} \text{ with } \delta_{ij} = \begin{cases} 1 \text{ if } d(i,j) \leq d^* \\ 0 \text{ if } d(i,j) > d^* \end{cases}$$

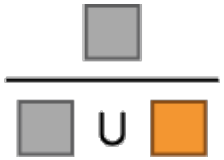
Protein-Ligand Binding Site Recognition

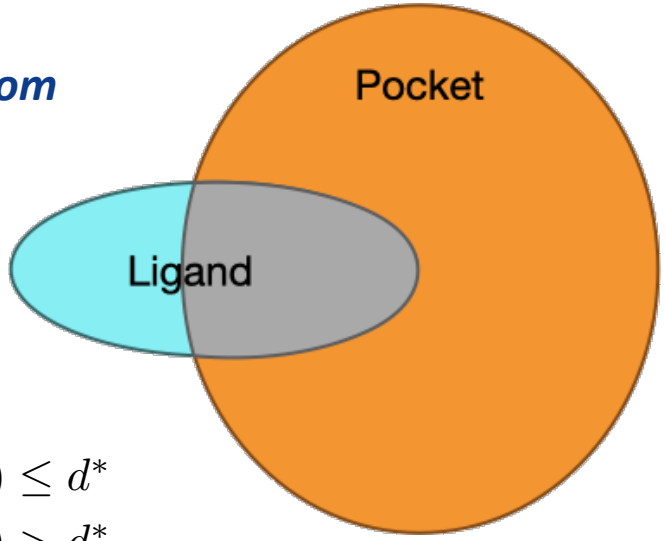
Evaluation:

Inspired by state-of-the-art biophysical pocket detection methods, we adopted a figure of merit based on the ***combination of two scores***

Pocket Coverage Score:

Fraction of the surface belonging to a pocket *which is* ***within a threshold distance from any ligand heavy atom***

PC := 



Formally,

$$\mathrm{PC} = \frac{1}{n_P} \sum_{i=1}^{n_P} \delta_{ij} \quad \text{for } \forall\, j \in \mathcal{L} \text{ with } \delta_{ij} = \begin{cases} 1 \text{ if } d(i,j) \leq d^* \\ 0 \text{ if } d(i,j) > d^* \end{cases}$$

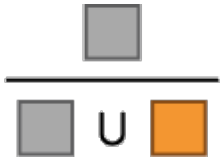
Protein-Ligand Binding Site Recognition

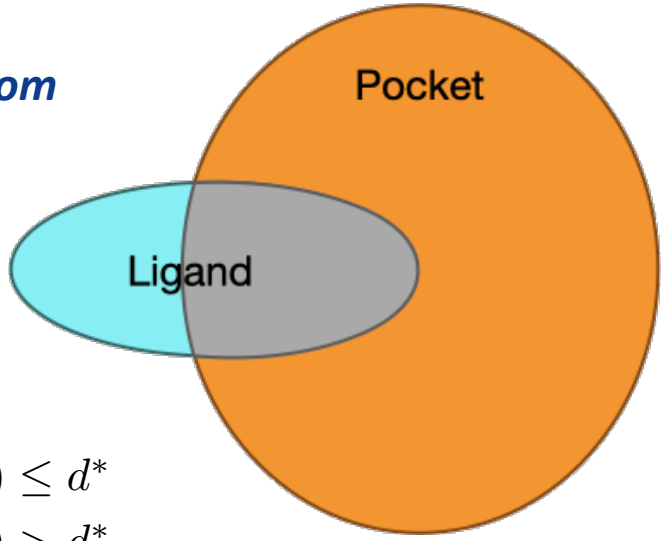
Evaluation:

Inspired by state-of-the-art biophysical pocket detection methods, we adopted a figure of merit based on the ***combination of two scores***

Pocket Coverage Score:

Fraction of the surface belonging to a pocket *which is* ***within a threshold distance from any ligand heavy atom***

PC := 



Formally,

$$\mathrm{PC} = \frac{1}{n_P} \sum_{i=1}^{n_P} \delta_{ij} \quad \text{for } \forall\, j \in \mathcal{L} \text{ with } \delta_{ij} = \begin{cases} 1 \text{ if } d(i,j) \leq d^* \\ 0 \text{ if } d(i,j) > d^* \end{cases}$$

A putative pocket is ***correctly matched*** *if it scores at least* ***50% in LC*** *and at least* ***20% in PC***

Protein-Ligand Binding Site Recognition

Comparison:

Method	Top1	Top3	Top10	LC	PC	nPockets
M1 - Point Transformer	69.1	75.9	75.9	96.4	60.4	2.1
M2 - GNN-Pocket	53.4	54.6	55.4	93.7	47.5	1.9
M3 - DeepSurf	87.6	89.2	89.2	95.0	67.9	1.6
M4 - NS-Volume	59.0	76.7	83.9	88.8	74.8	11.6
Fpocket	60.2	75.1	84.7	92.5	64.7	8.9

- *We report*
  - *Average ranking in terms of **Top1**, **Top3**, and **Top10** performance*
  - *Average **LC** and **PC** scores over successfully predicted pockets*
  - *Average number of generated **pockets per structure***
- *Results are expressed as the **percentage of success rate normalized** over the total number of structure-ligand pairs*
- *For sake of comparison, we report also the results obtained by **Fpocket** on the same dataset*
  - *A standard and **well established tool** for pocket detection*
  - ***Not eligible** as a competing method in the SHREC track since it considers also chemical features and not just geometrical ones*

Protein-Ligand Binding Site Recognition

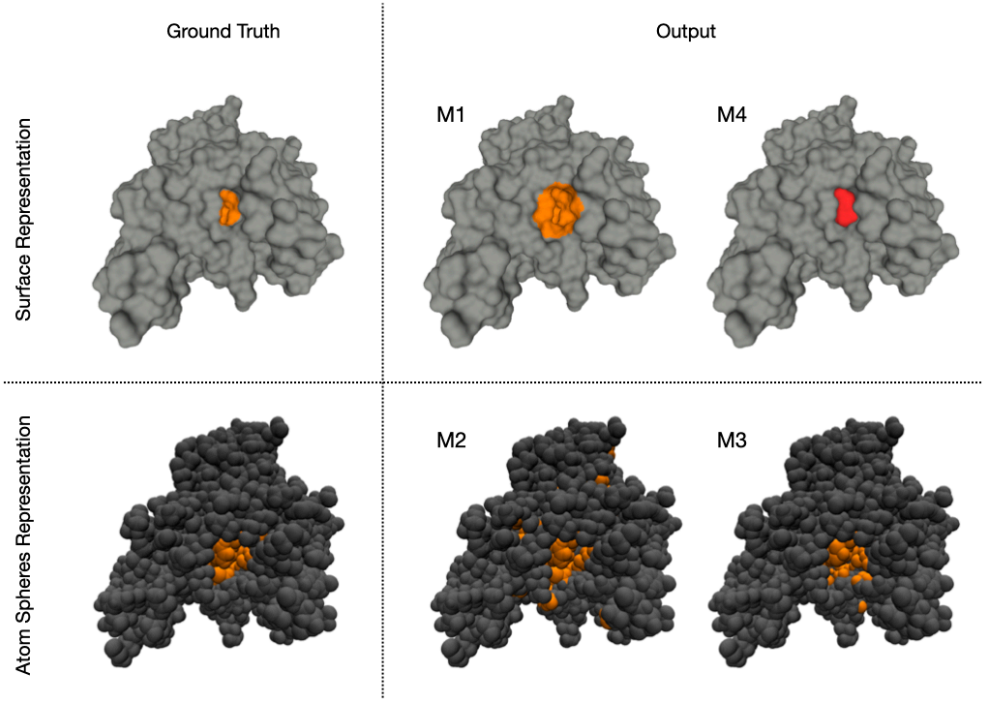
Comparison:

Method	Top1	Top3	Top10	LC	PC	nPockets
M1 - Point Transformer	69.1	75.9	75.9	96.4	60.4	2.1
M2 - GNN-Pocket	53.4	54.6	55.4	93.7	47.5	1.9
M3 - DeepSurf	87.6	89.2	89.2	95.0	67.9	1.6
M4 - NS-Volume	59.0	76.7	83.9	88.8	74.8	11.6
Fpocket	60.2	75.1	84.7	92.5	64.7	8.9

- ***M3** shows an **excellent performance***
  - *Obtained results **outstand also Fpocket***
  - *Despite the **small number of putative pockets generated**, these are extremely well predicted*
  - *Information leveraged goes **beyond pure geometry** and the training set is larger than the one provided*
- ***On Top10**, M4 and Fpocket obtain similar scores to M3*
- *Only M4 and Fpocket return **more than about 2 putative pockets per structure** on average*
- *All methods perform **very well in term of Ligand Coverage score***
- *A significantly **lower Pocket Coverage score** is measured*

Protein-Ligand Binding Site Recognition

An Example:



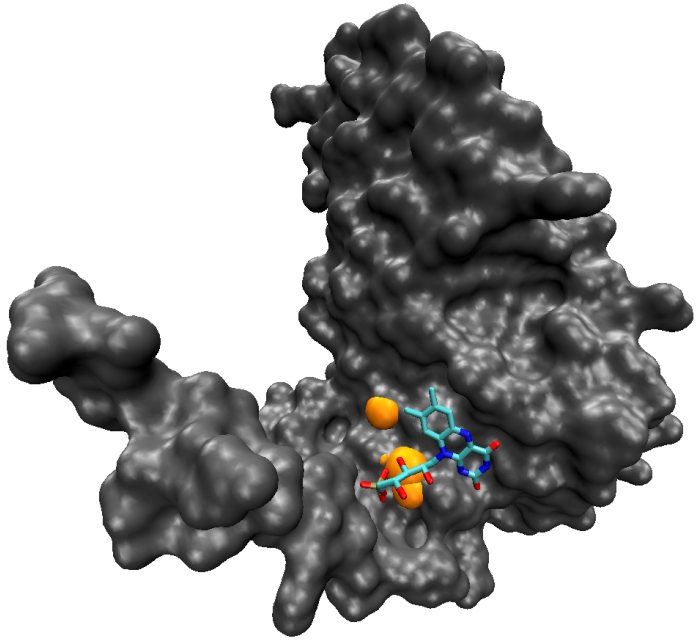
	LC	PC
M1	100	76.5
M2	100	36.2
M3	100	87.4
M4	100	88.6

- ***Low average PC score*** *indicates that a method is prone to generate* ***pockets which are larger than the binding ligand***
- *We observe that, in general,* ***M2*** *generates pockets which are* ***often larger*** *than the binding region and* ***scattered into disconnected segments***

Protein-Ligand Binding Site Recognition

Discussion:

*Proposed methods have **difficulties** in identifying particularly **shallow binding sites***



This is due to:

- *Methods completely relying on geometry for the generation of putative pockets are **optimized to recognize cleft and cavities** (which are often found to contain binding ligands)*
- *Shallow pockets attain the role of binding site mainly for their **chemical properties** rather than their shape*
- *Shallow sites are **rare in the training set***

Protein-Ligand Binding Site Recognition

Conclusions:

- *We proposed a SHREC track aimed at evaluating the effectiveness of computational methods based purely on **geometrical information** in the **detection of binding sites** on protein surfaces*
- *We created a **dataset of protein structures** expressed both in terms of **atom spheres and SES surface representations** and we enriched the training set with positive examples of **known ligandable pockets***
- *We analyzed and compared **four different proposed methods** on the basis of **two evaluation measures**:*
  - *Most of the proposed methods show **very good performance***
  - *All methods struggle in the recognition of **shallow binding sites***
  - *Proposed methods generally perform **low Pocket Coverage** score*
- *Future directions and possible improvements:*
  - *Problem in this SHREC track is an instance of a **one-class discrimination task***
  - *Low PC scores suggests the possibility of considering a **higher segmentation** of the returned sites into separate smaller pockets or sub-units*

Protein-Ligand Binding Site Recognition

Resources:

- *Dataset, benchmark, predictions of participants that originated the results:*
  https://github.com/concept-lab/shrec22_proteinLigandBenchmark
- ***M1 — Point Transformer:***
  https://github.com/aaron-h-code/Protein_SHREC2022/
- ***M2 — GNN-Pocket:***
  https://github.com/kiharalab/GNN_pocket
- ***M3 — DeepSurf:***
  https://github.com/stemylonas/DeepSurf_SHREC22
- ***M4 — NS-Volume:***
  https://github.com/concept-lab/NS_pocket

Thank you for the attention!

Ulderico Fugacci CNR - Imati

ulderico.fugacci@ge.imati.cnr.it