Persistence Homological Statistical Summaries for Ligand-Based Virtual Screening

Aras Asaad (PhD)

Joint work with Dr. Richard Cooper and Prof. Paul Finn

Methods Development Team,

Oxford Drug Design, Oxford, UK.

3D based Molecule Representation

Super-positional methods:

- ✓ ROCS
- ✓ Brutus
- ✓ EON
- ✓ Phase-Shape
- ✓ Shape-it
- ✓ Align-it
- ✓ ShaEP
- ✓ SHAFTS
- ✓ WEGA
- ✓ LIGSIFT
- ✓ LS-Align
- ✓ ESPsim ... etc

Non Super-positional methods:

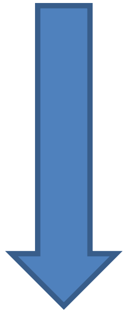
- ✓ Electroshape (Oxford Uni. / Oxford Drug Design)
- ✓ USR-CAT (Cambridge Uni.)
- ✓ Whales (Weighted Holistic Atom Localization and Entity Shape) – Zurich/Milan.
- ✓ MOLSG & E-MOLSG – Sheffield Uni.
- ✓ RGMolSA & KQMolSA (based on RIEMANNIAN GEOMETRY) - Newcastle Uni.
- ✓ Morse-Theory based (In progress- Oxford Uni.)
- ✓TDA (Topological Data Analysis)

Outline

- Topological Data Analysis: Brief introduction.
- Featurising the space of Persistence Barcodes: Statistical Summaries.
- Ligand based virtual screening
- Results using Internal and DUD-E datasets
- Comparison with state-of-the-art (SOTA)
- Future Research Directions

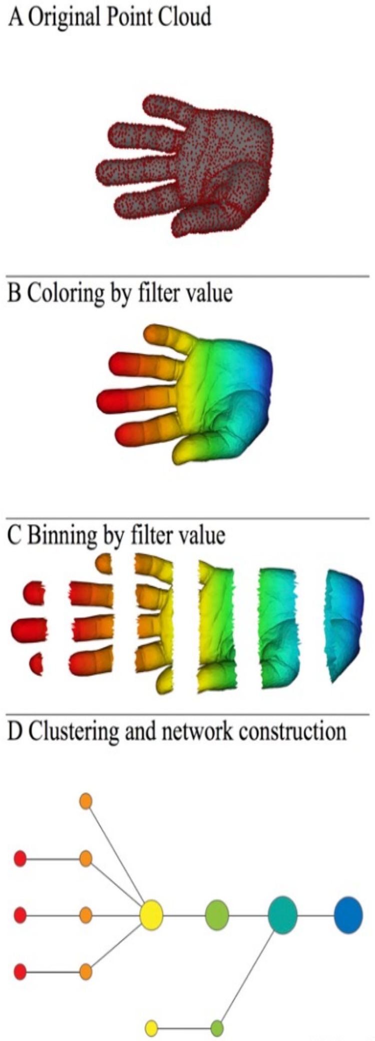
Topological Data Analysis

Persistent Homology



Algebraic Topology

Mapper Algorithm


A Original Point Cloud
B Coloring by filter value
C Binning by filter value
D Clustering and network construction

A) Data Set

Example: Point cloud data representing a hand.

B) Function f : Data Set → **R**

Example: x-coordinate

$f : (x, y, z) \rightarrow x$

C) Put data into overlapping bins.

Example: $f^{-1}(a_i, b_i)$

D) Cluster each bin & create network.

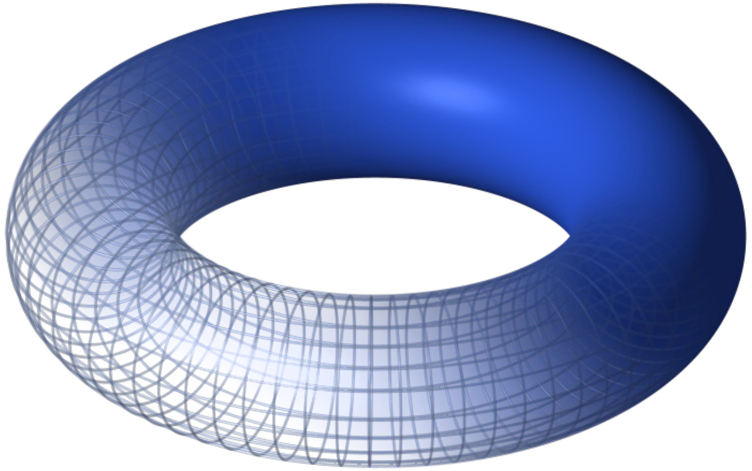
Vertex = a cluster of a bin.

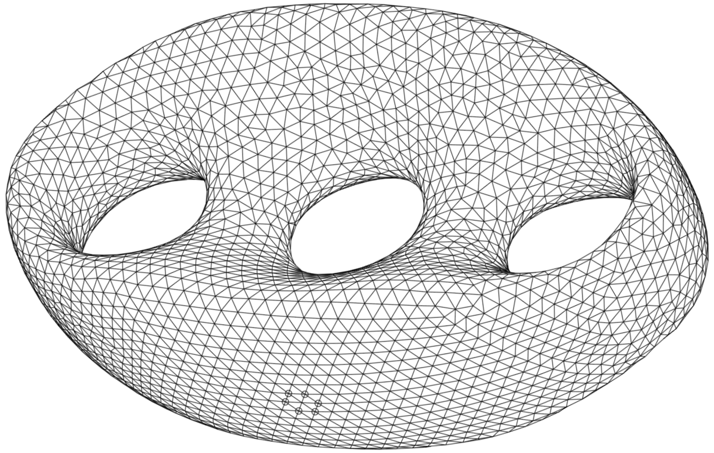
Edge = nonempty intersection between clusters

http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html

Algebraic Topology → Persistent Homology

Algebraic Topology is a collection of tools from Abstract Algebra (Groups, Rings, Fields, Ideals … etc) used to study algebraic invariants of topological spaces (up to homotopy equivalent).





Algebraic Topology → Persistent Homology

Topology is a field of Mathematics concerned with studying characteristics of shapes in terms of connectivity and closeness, using a combinatorial process known as ***simplicial complexes.***

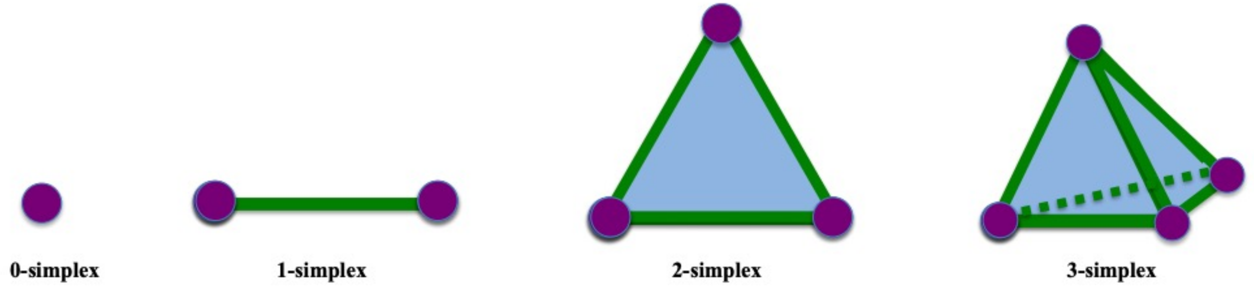
The main tool in algebraic topology we use is called **homology**.

Homology uses simplicial complexes to measure topological properties of spaces such as number of **connected pieces, number of holes/loops, number of cavities** ...etc.

Simplicial Complexes → (SC) Homology

Consider $V = \{v_0, v_1, \dots, v_n\}$ to be the set of vertices. A SC with a vertex set V is a collection $\mathbb{S}$ of subsets of V whereby the following two conditions satisfied:

- The singleton $\{v\} \in \mathbb{S}$, where $v \in V$.
- Let $\tau \in \mathbb{S}$ and $\sigma \subset \tau$, then $\sigma \in \mathbb{S}$.



For each integer $k \geq 0$, the boundary operator defines a linear transformation $\partial_k : \mathcal{C}_k(\mathbb{S}) \to \mathcal{C}_{k-1}(\mathbb{S})$

Then, we can define a sequence of homeomorphism of abelian groups (i.e. chain complex) as follows:

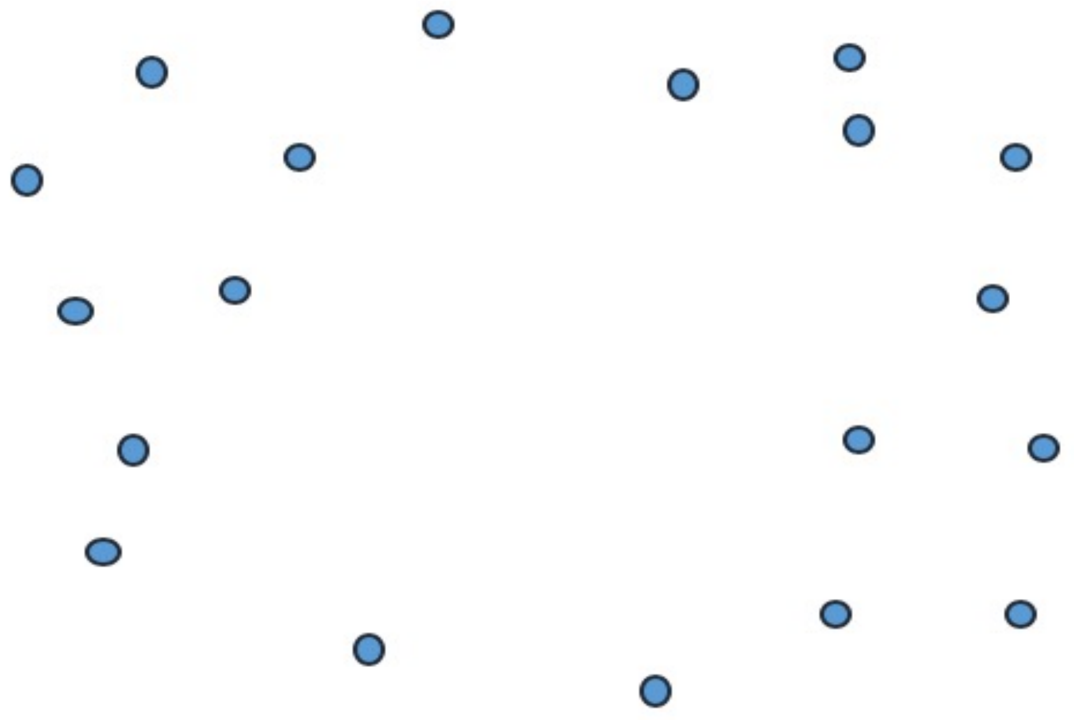
$$\dots \to \mathcal{C}_{k+1}(\mathbb{S}) \xrightarrow{\partial_{k+1}} \mathcal{C}_k(\mathbb{S}) \xrightarrow{\partial_k} \mathcal{C}_{k-1}(\mathbb{S}) \xrightarrow{\partial_{k-1}} \dots \xrightarrow{\partial_2} \mathcal{C}_1(\mathbb{S}) \xrightarrow{\partial_1} \mathcal{C}_0(\mathbb{S}) \xrightarrow{\partial_0} 0.$$

Finally, we can define the k-th homology group of $\mathbb{S}$ by the quotient vector space as follows: $H_k(\mathbb{S}) = \ker(\partial_k) / Im(\partial_{k+1})$

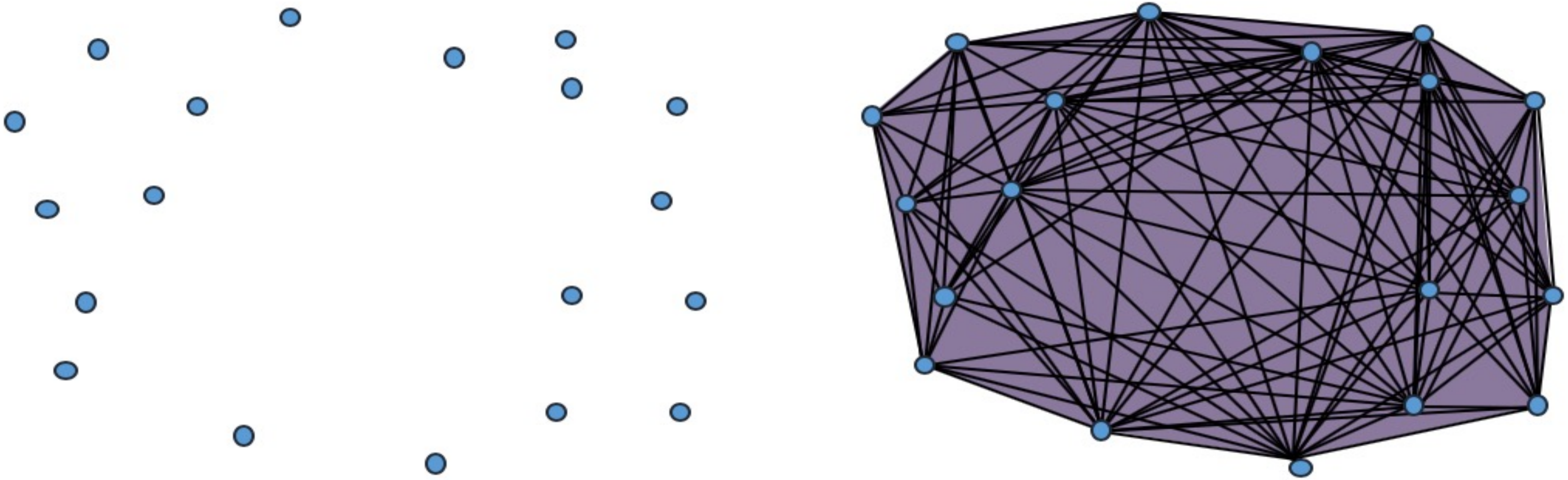
Dimensions of the homology groups are known as Betti numbers : $B_k(\mathbb{S}) \coloneqq \dim(H_k(\mathbb{S}) = \dim\big(Ker(\partial_k)\big) - \dim\big(Im(\partial_{k+1})\big)$

Betti numbers of dimension zero = B_0 = connected Components
Betti numbers of dimension one = B_1 = loops / holes.
Betti numbers of dimension two = B_2 = Cavities/Voids

How to build SCs from Data?



How to build SCs from Data?

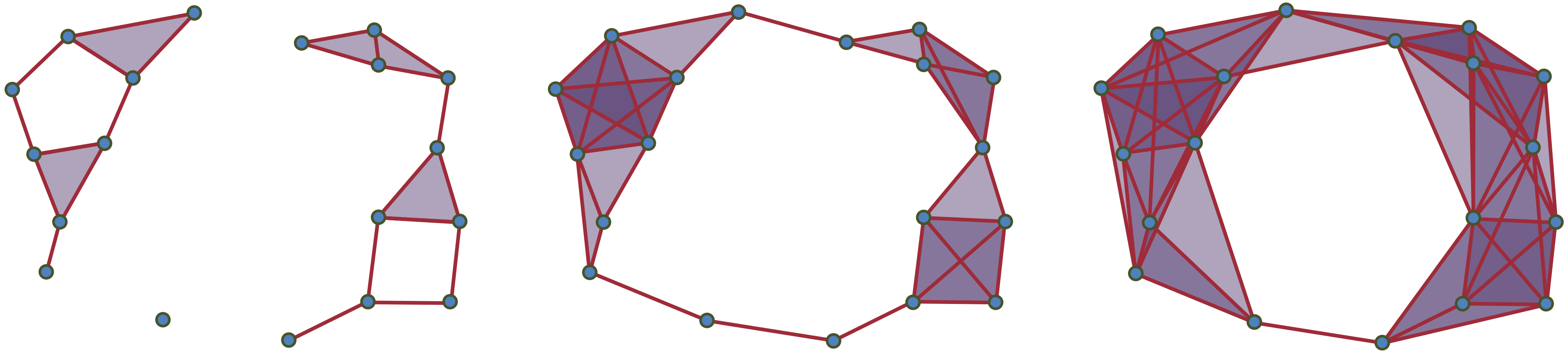


Idea: Consider *a series of* distances *thresholds* and analyse pattern of change in the topology of the corresponding SCs as thresholds increase, known as **Persistent Homology**.

Persistent Homology

Consider the sequence (SC_i) of simplicial complexes associated to a point cloud for a sequence of distance values:

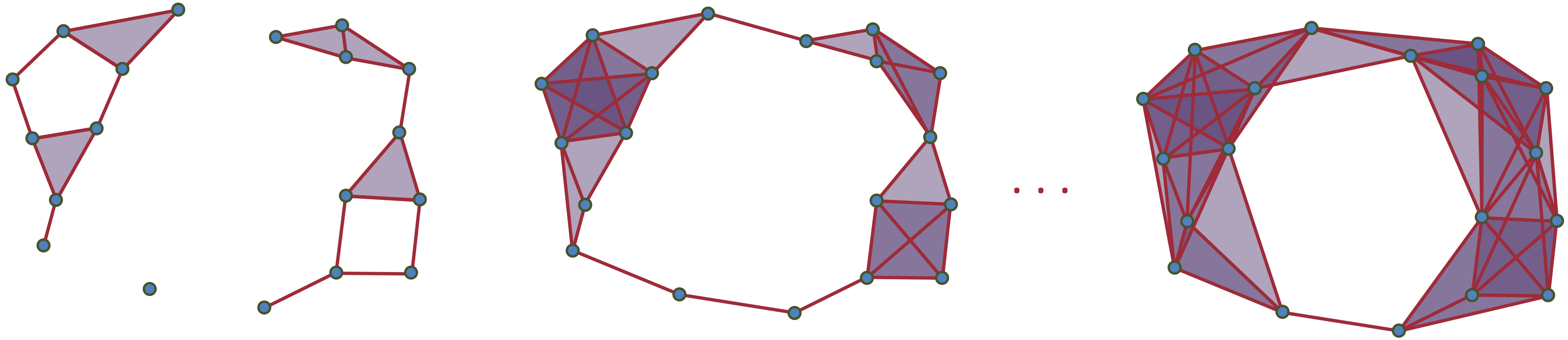
$$\begin{tikzcd} SC_1 \arrow[r, hook, "\delta"] & SC_2 \arrow[r, hook, "\delta"] & SC_3 \end{tikzcd}$$



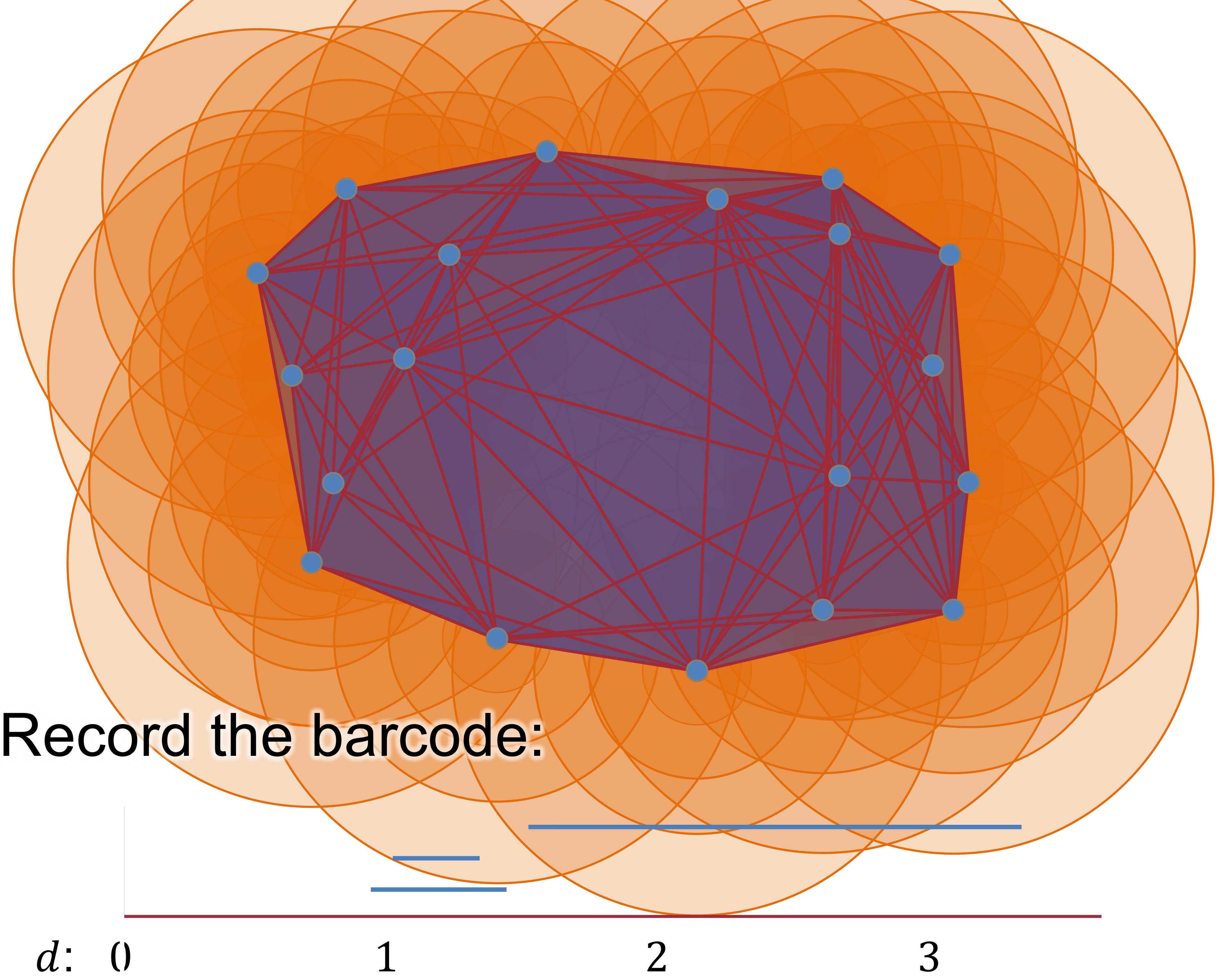
Persistent Homology

Consider a sequence of nested simplicial complexes (SC_i) associated to a point cloud for a sequence of distance values:

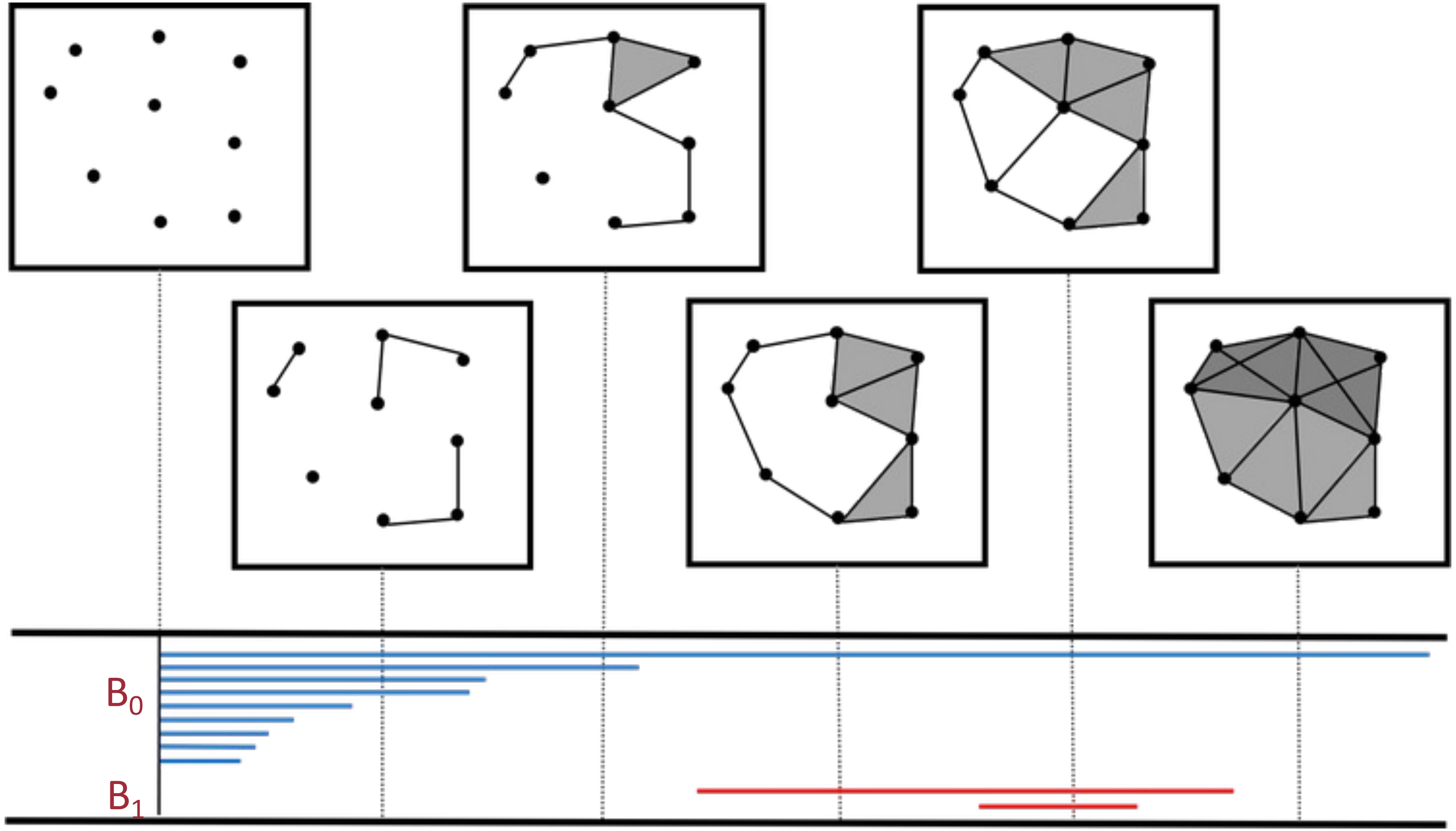
$$\cdots \hookrightarrow SC_1 \hookrightarrow SC_2 \hookrightarrow SC_3 \hookrightarrow SC_4 \hookrightarrow SC_5 \hookrightarrow SC_6 \hookrightarrow \mathrm{SC}_7 \hookrightarrow \cdots$$

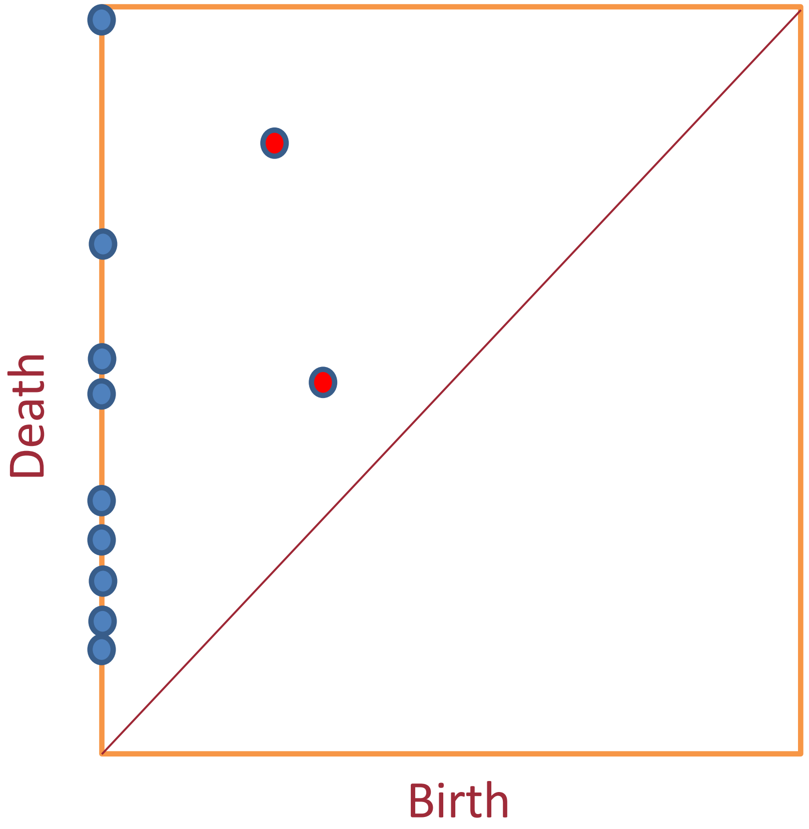


This sequence of complexes, with maps, is a **filtration** of the final SC.
Note the change of connected components & holes (***Top. Invariants***)


Record the barcode:

Persistent Homology: Barcodes & Persistent Diagram Representations

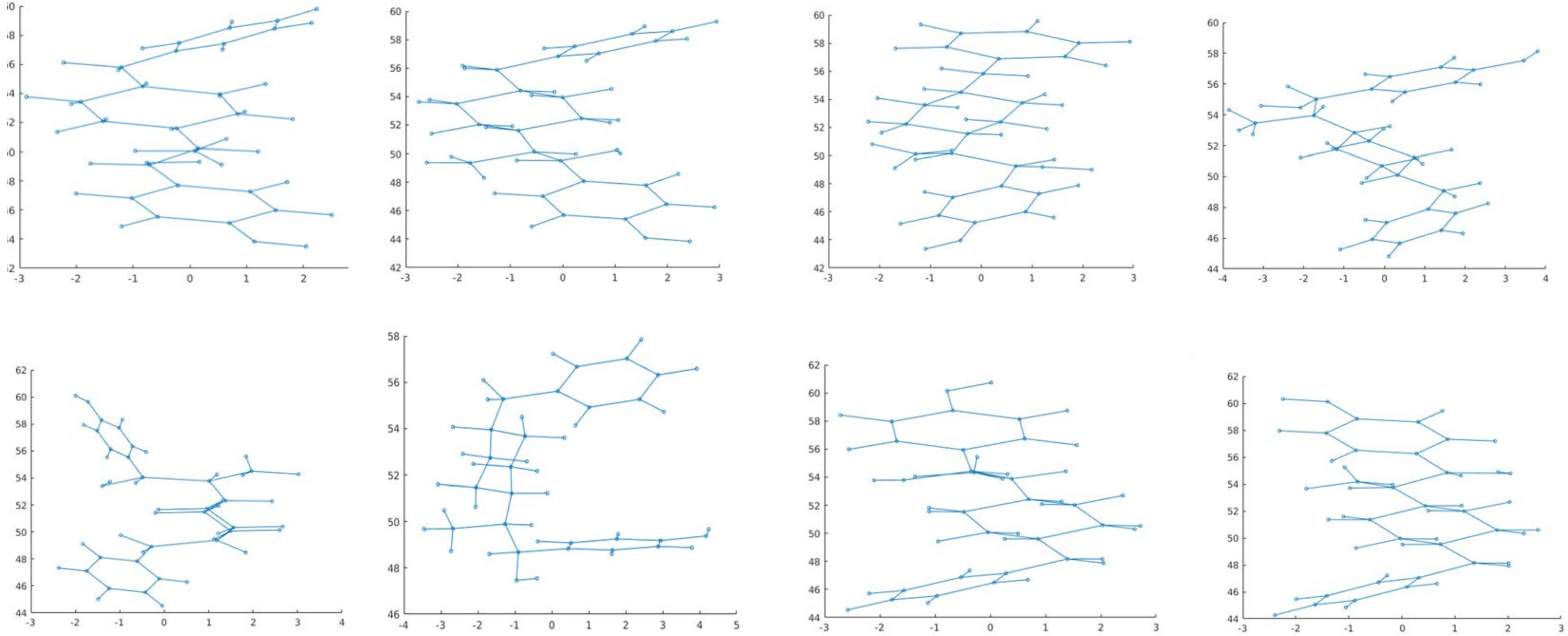




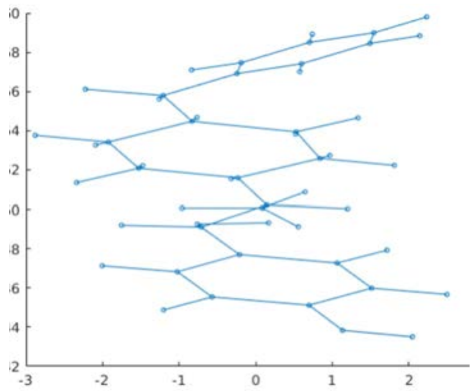
Building Persistent Homological Features from Molecules

ilfenprodil (52 atoms,(21C, 1N,2O,28H)).

Vertices are atoms and the line segments constructed based on the distance between co-ordinates of atoms. Here, the distance threshold T=1.6.

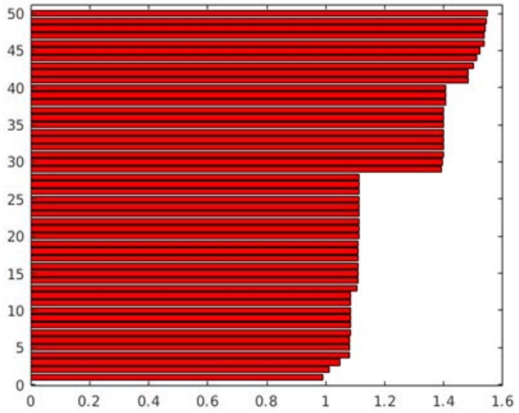


Building Persistent Topological Features from Molecules

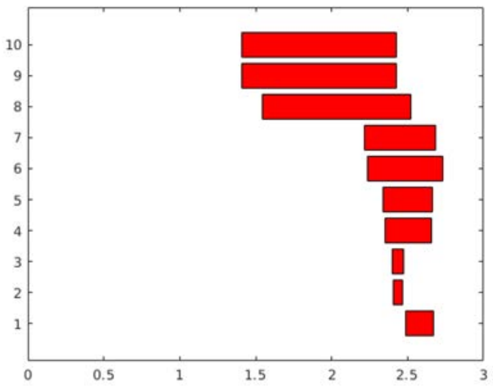


Input Point Cloud

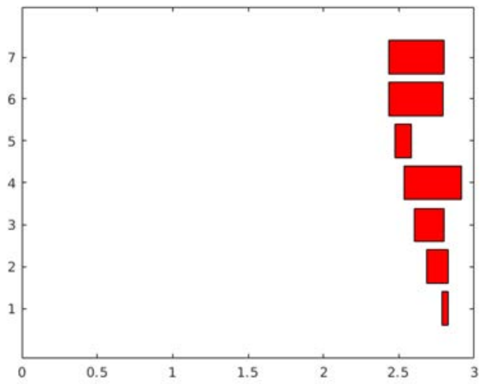
- ✓ 3D atom coordinates,
- ✓ 4D (3D + Partial Charges)
- ✓ 5D (4D + lipophilicity)



Persistent 0-D homology Features
i.e. Connected Components

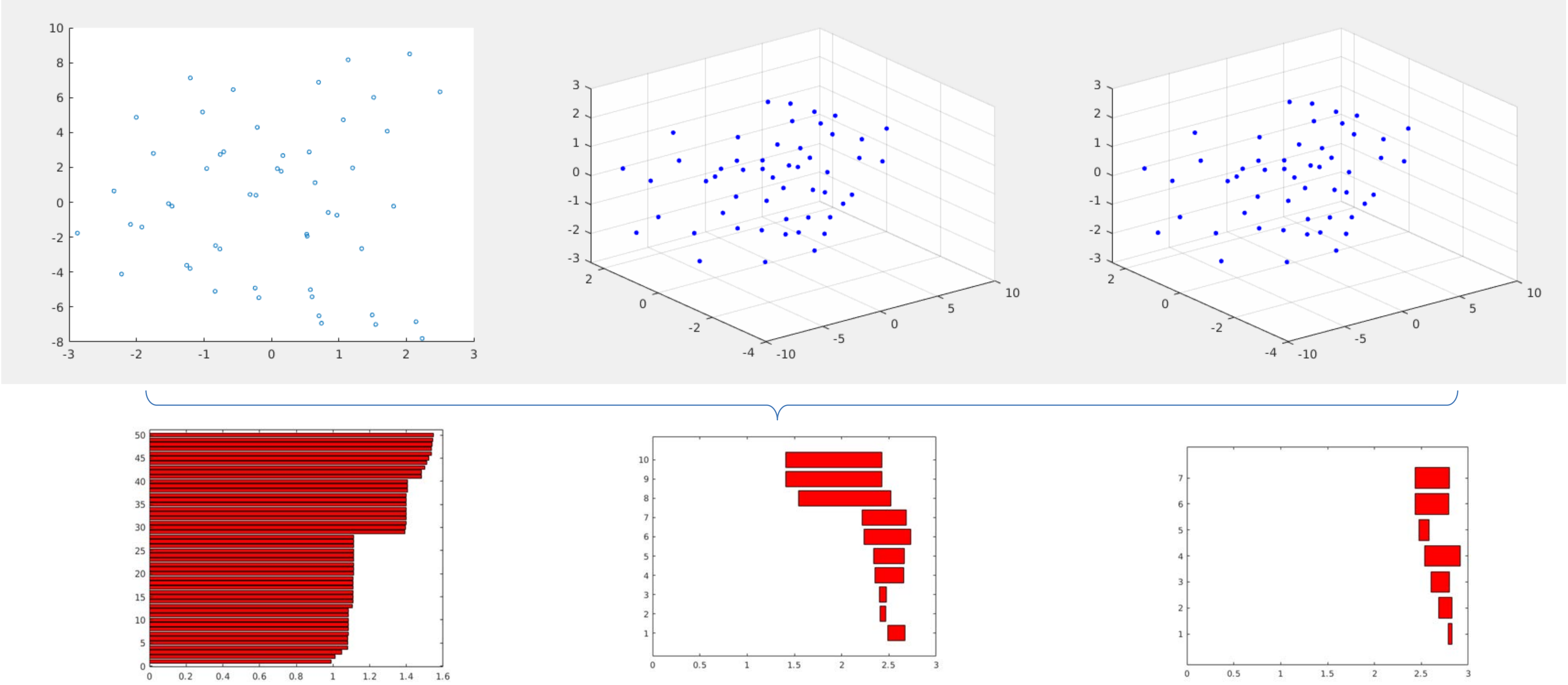


Persistent 1-D homology Features
i.e. 1-D loops/holes

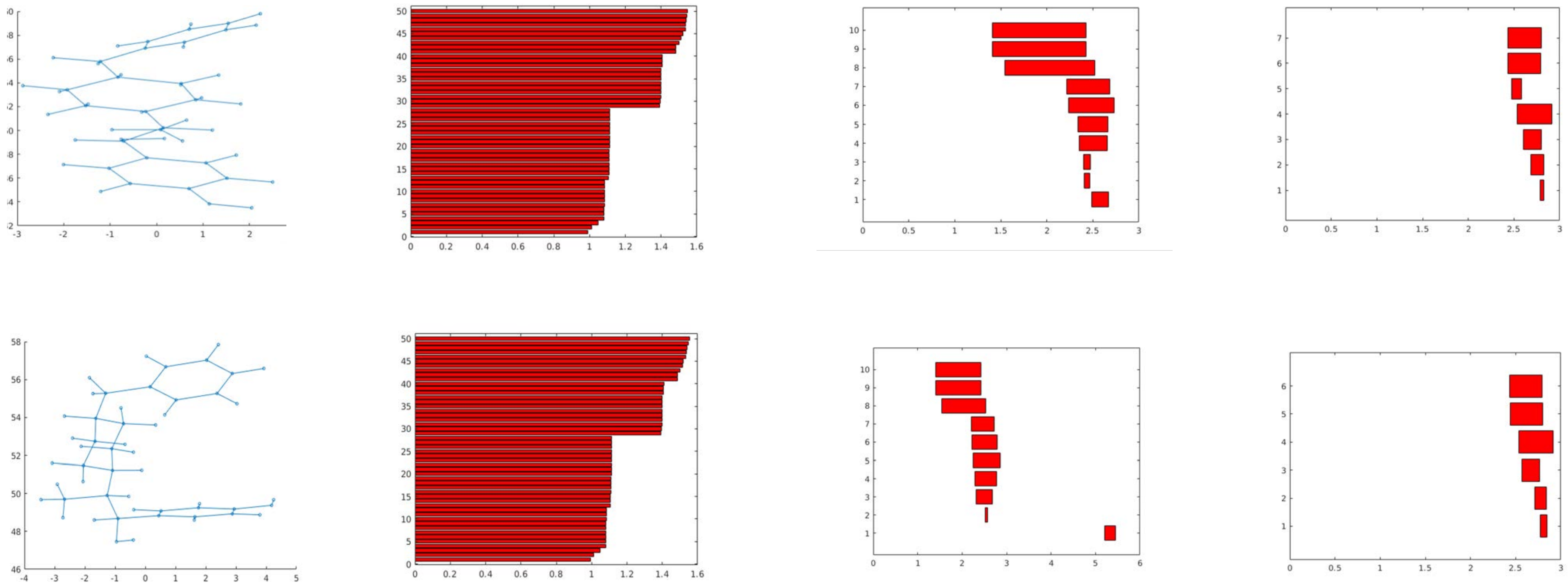


Persistent 2-D homology Features
i.e. 2-D Cavities/Voids

Building Persistent Topological/Homological Features from Molecules

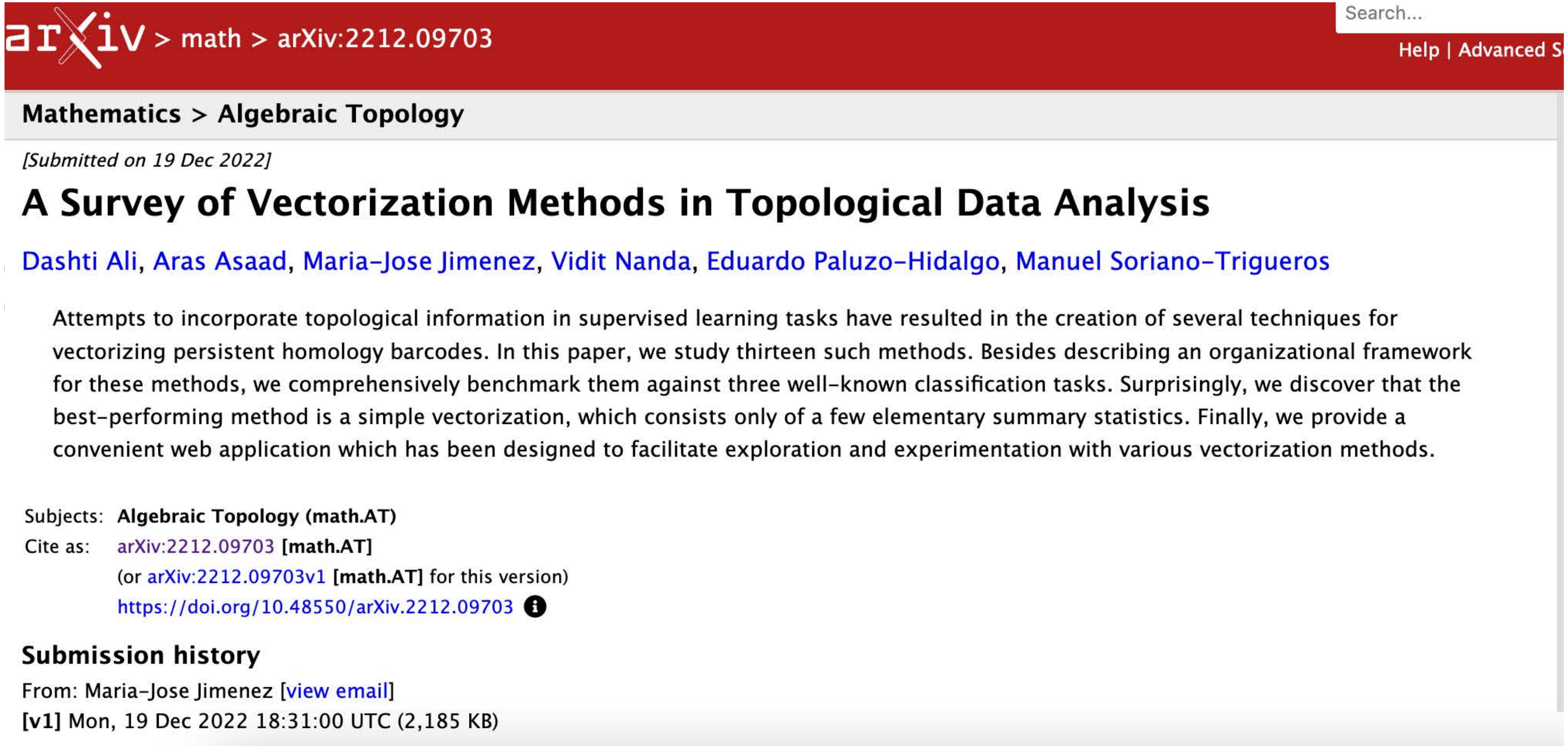


Molecule-Conformer: Persistent Barcode Visualizations



How to Use Persistence Barcodes to differentiate Ligands from Decoys?

How to Vectorise the Space of Persistence Barcodes?



Mathematics > Algebraic Topology

[Submitted on 19 Dec 2022]

A Survey of Vectorization Methods in Topological Data Analysis

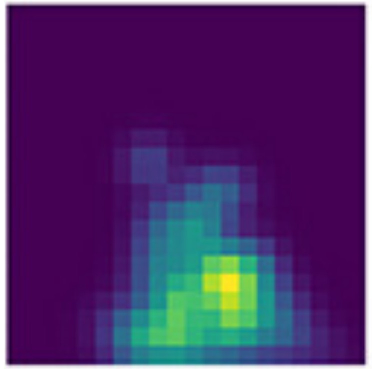
Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, Manuel Soriano-Trigueros

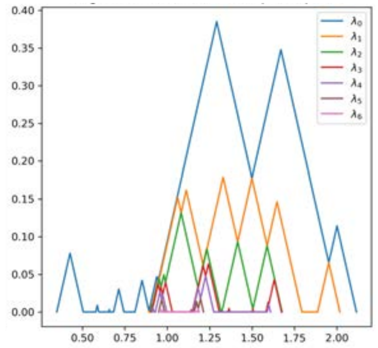
Attempts to incorporate topological information in supervised learning tasks have resulted in the creation of several techniques for vectorizing persistent homology barcodes. In this paper, we study thirteen such methods. Besides describing an organizational framework for these methods, we comprehensively benchmark them against three well-known classification tasks. Surprisingly, we discover that the best-performing method is a simple vectorization, which consists only of a few elementary summary statistics. Finally, we provide a convenient web application which has been designed to facilitate exploration and experimentation with various vectorization methods.

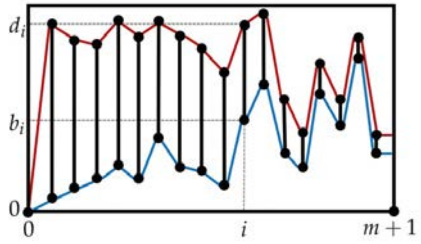
Subjects: **Algebraic Topology (math.AT)**
Cite as: arXiv:2212.09703 **[math.AT]**
(or arXiv:2212.09703v1 **[math.AT]** for this version)
https://doi.org/10.48550/arXiv.2212.09703

Submission history

From: Maria-Jose Jimenez [view email]
[v1] Mon, 19 Dec 2022 18:31:00 UTC (2,185 KB)







Statistical summaries:

Average of Birth of Bars
STDEV of Birth of Bars
Median of Birth of Bars
IQR of a Birth of Bars

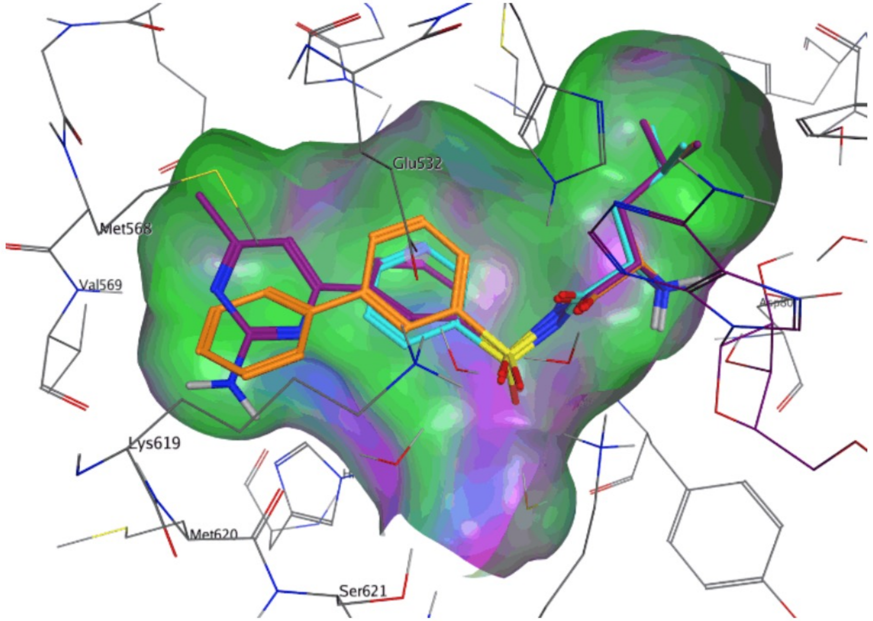
Average of Death of Bars
STDEV of Death of Bars
Median of Death of Bars
IQR of a Death of Bars

Average of Life-Span of Bars
STDEV of Life-Span of Bars
Median of Life-Span of Bars

Range, entropy and 10th ,25th ,90th IQR of midpoints ,life-spans

Dataset, Performance Metrics and ML

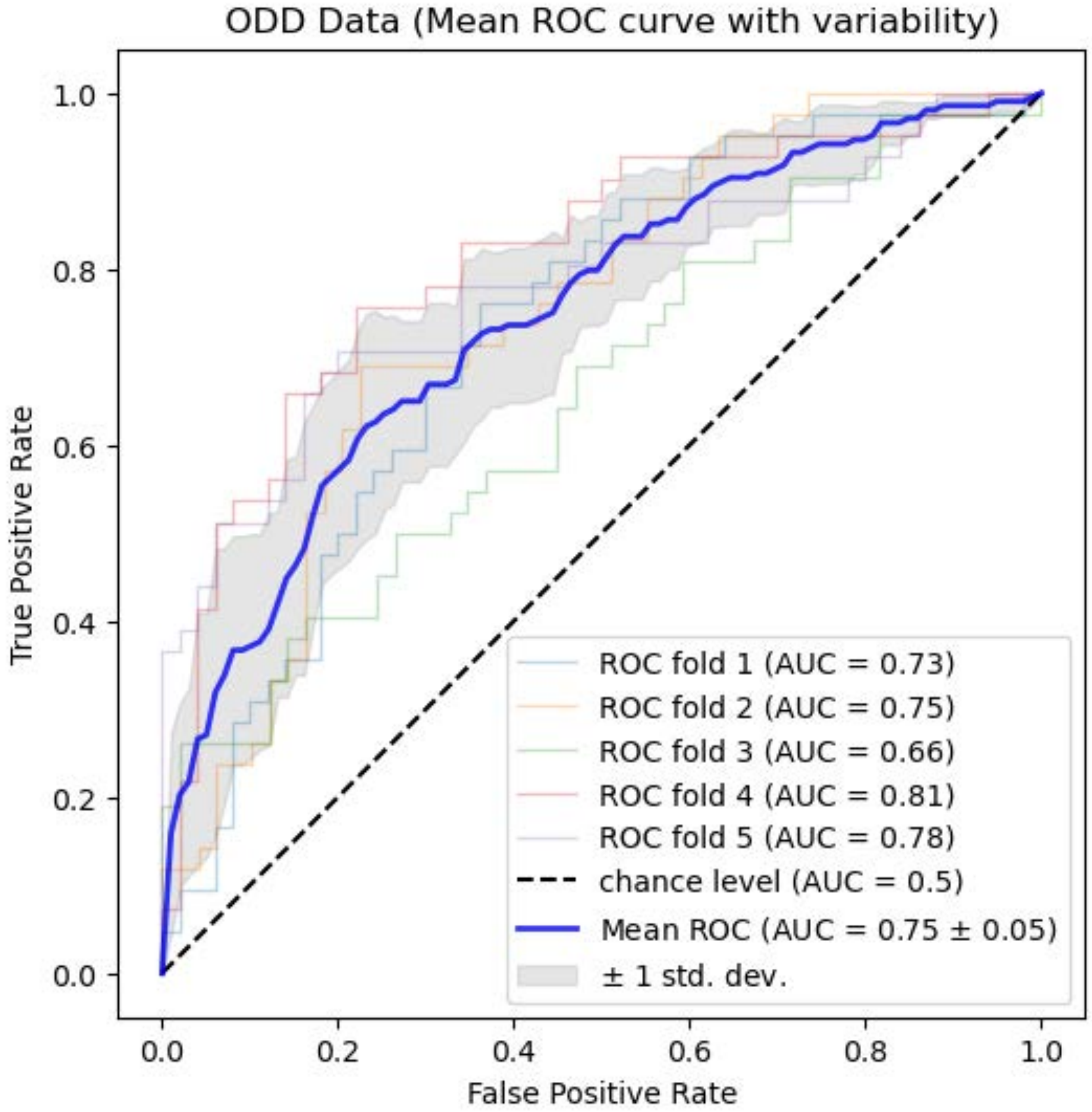
- DUD-E dataset:
  - We used all 102 DUD-E targets.
  - Conformers generated using an internal ODD pipeline.
  - Minimum Energy conformer used in our experiments.
- Internal Dataset:
  - Protein target: leucyl-tRNA synthetase
  - Number of Active molecules: 208
  - Number of Inactive Molecules: 248
- Performance Metrics:
  - Enrichment Factor at 1%.
  - Hit-Rate (also known as relative Enrichment Factor at 1%)
  - Area under the ROC-curve (AUC).
- Machine Learning:
  - Light-GBM classifier with optimizing hyperparameters
  - Stratified 5 fold cross validation to partition the training and Testing.



Results From Internal Dataset

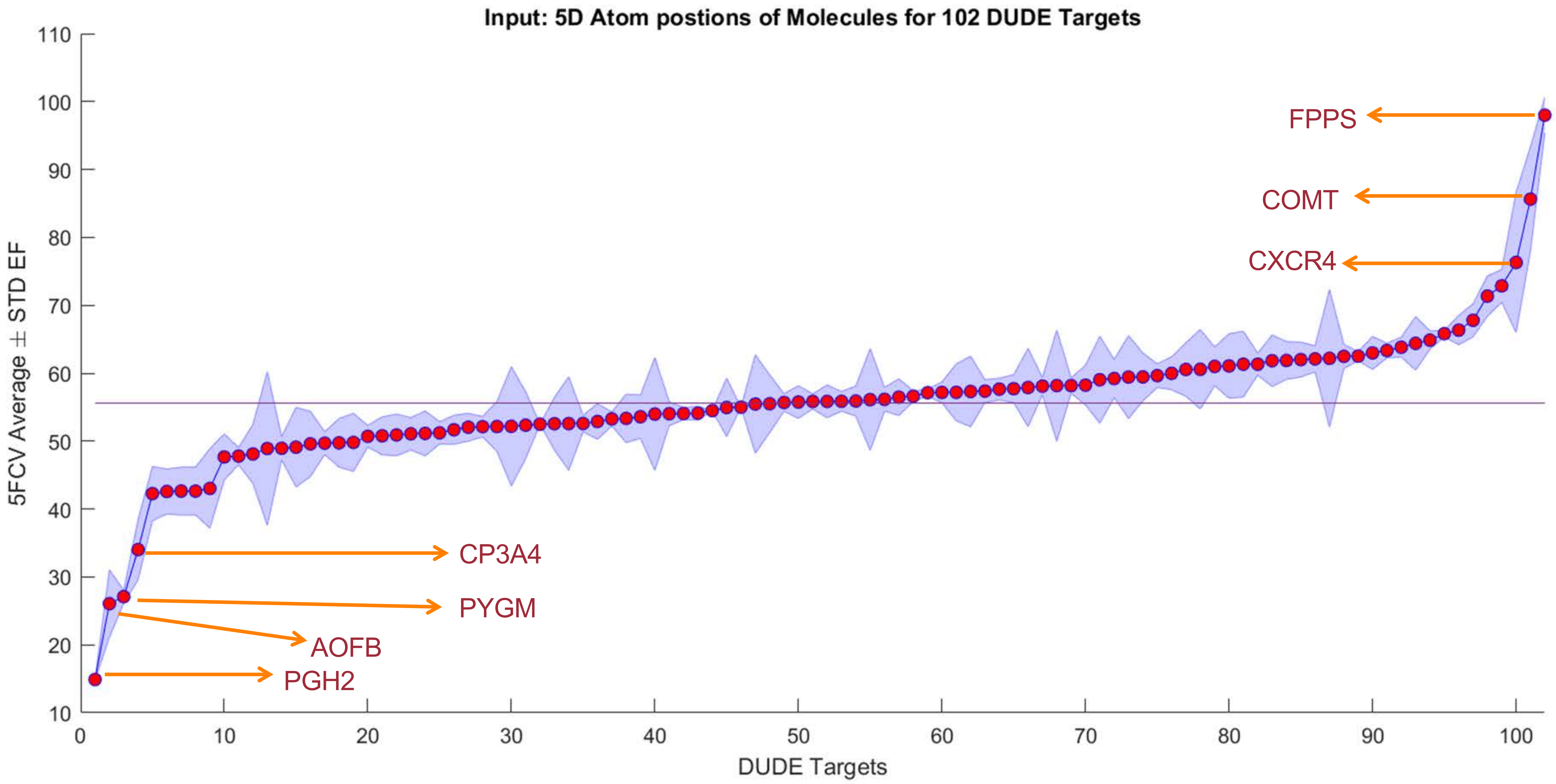
ODD Internal Results

	AUC	1% EF	Hit-Rate %
Fold 1	0.73	2.19	100
Fold 2	0.75	2.17	100
Fold 3	0.66	2.17	100
Fold 4	0.81	2.22	100
Fold 5	0.78	2.21	100
Average	**0.75**	2.12	100
STDEV	**0.05**	0.024	0

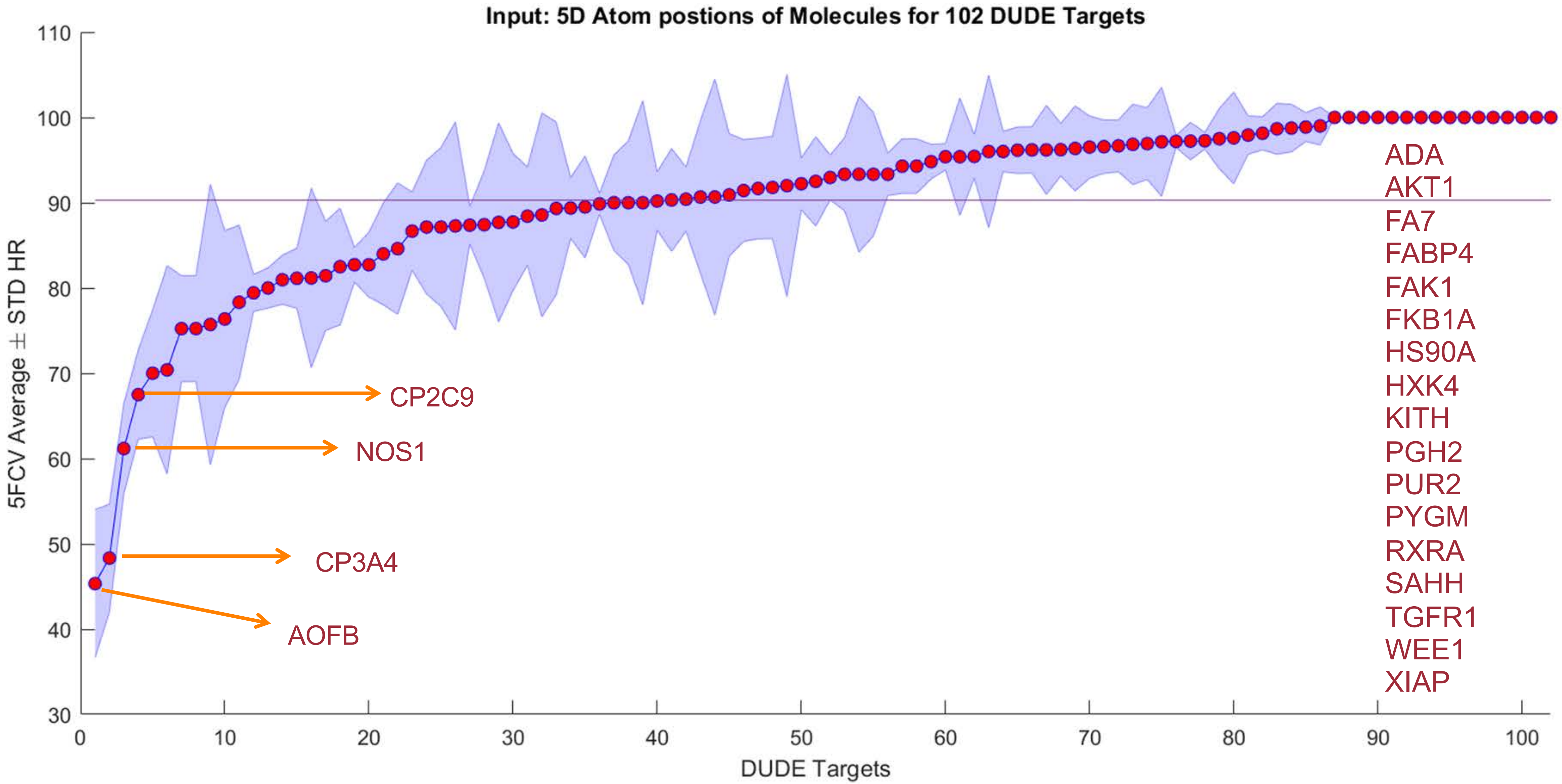


Results From DUD-E dataset

The Input: 5D atom Positions, Metric: 1% Enrichment Factor

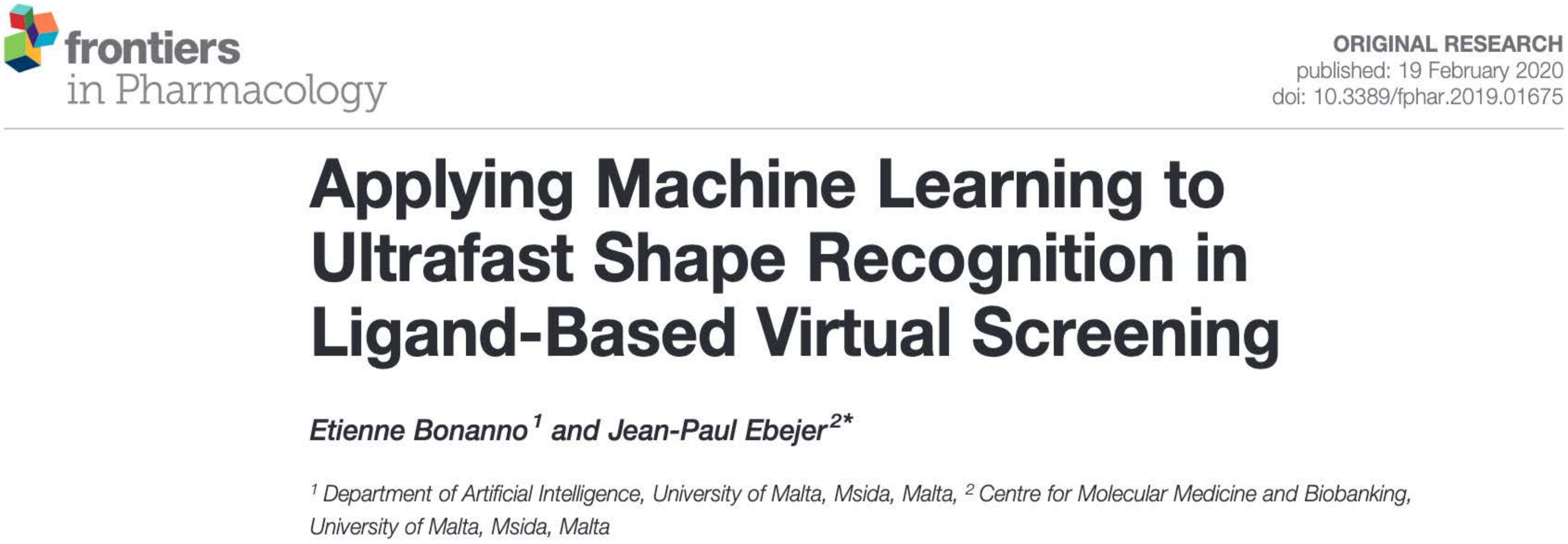


The Input: 5D atom Positions, Metric: Hit-Rate (HR)



ADA
AKT1
FA7
FABP4
FAK1
FKB1A
HS90A
HXK4
KITH
PGH2
PUR2
PYGM
RXRA
SAHH
TGFR1
WEE1
XIAP

Comparing with Literature Results



frontiers in Pharmacology

ORIGINAL RESEARCH
published: 19 February 2020
doi: 10.3389/fphar.2019.01675

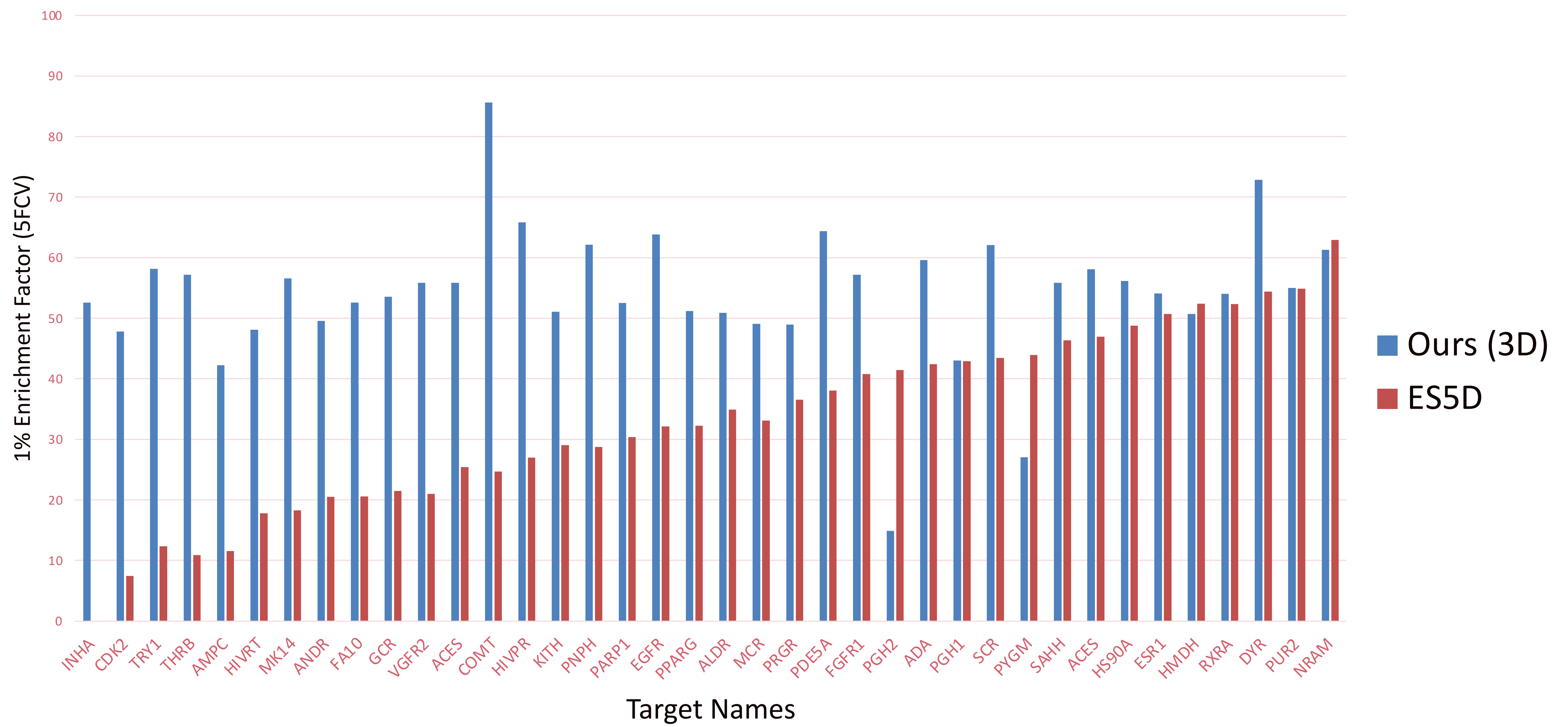
Applying Machine Learning to Ultrafast Shape Recognition in Ligand-Based Virtual Screening

Etienne Bonanno[1] *and Jean-Paul Ebejer*[2*]

[1] Department of Artificial Intelligence, University of Malta, Msida, Malta, [2] Centre for Molecular Medicine and Biobanking, University of Malta, Msida, Malta

- 38 DUD-E targets used.
- Three Machine Learning classifiers used: Gaussian Mixture Models, Isolation-Forest and Neural Networks (ANN).
- 5 fold cross validation used (4 folds to optimize hyperparameters and used for the testing partition in each round)
- 1% Enrichment Factor used as a performance metric (as well as ROC-AUC).

Comparing TDA with Literature (Bonanno & Ebejer paper)



Conclusion and Future Research Directions

✓ Persistent Homology is a novel method to represent molecules in the form of persistence barcodes which encodes both global topological features as well as geometrical features.

✓ Persistence Homological Statistical Summary is an effective featurisation approach to use Persistence barcodes with machine learning.

✓ Persistence Statistics is a state-of-the-art ligand based virtual screening method tested on DUD-E, MUV and also validated on in-house antimicrobial drug design project.

Future Research direction:

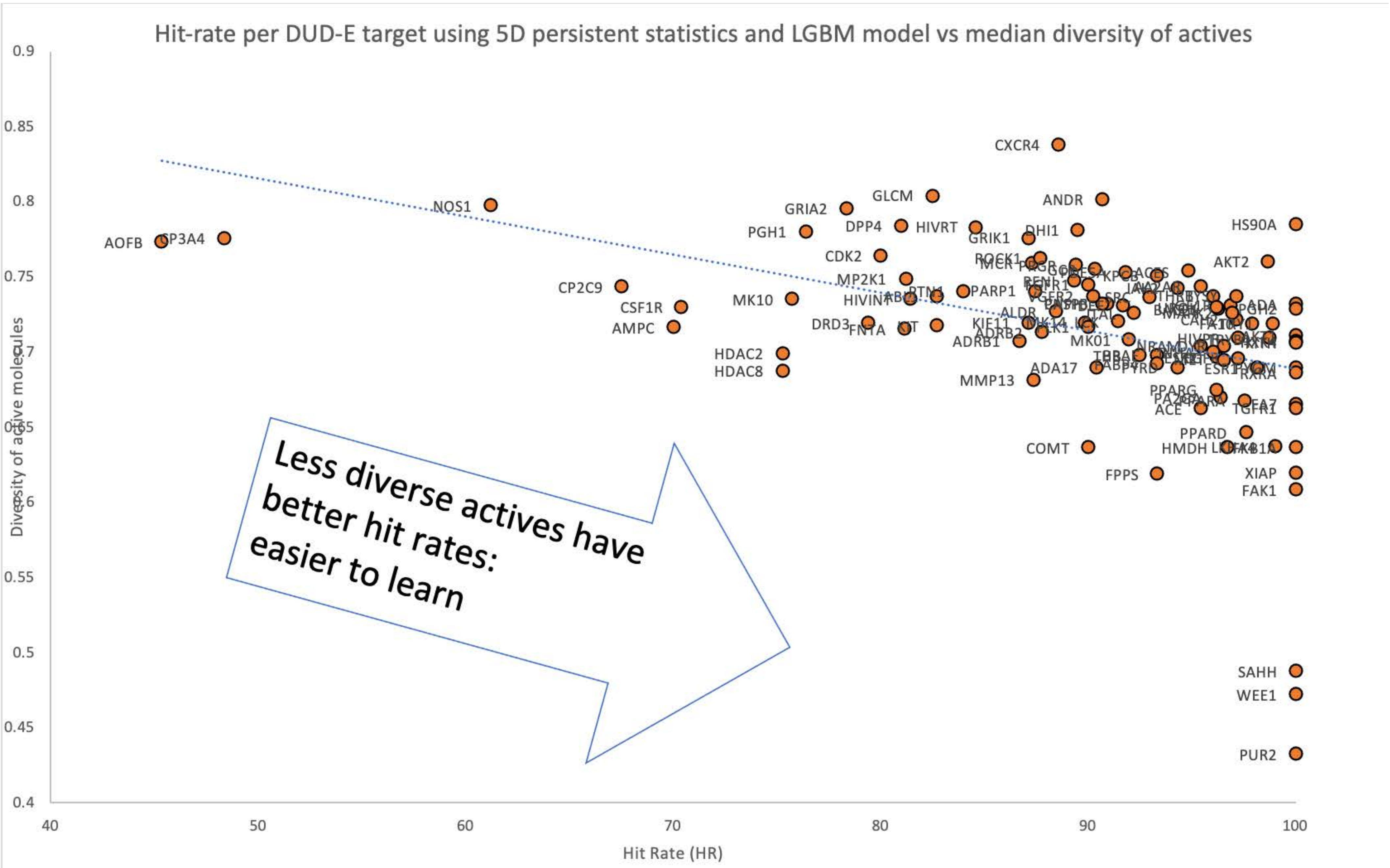
1- **Multi-parameter persistent Homology**.

2- **Incorporating protein information together with ligand topological features to improve the activity prediction**.

Thank You for Your Attention!

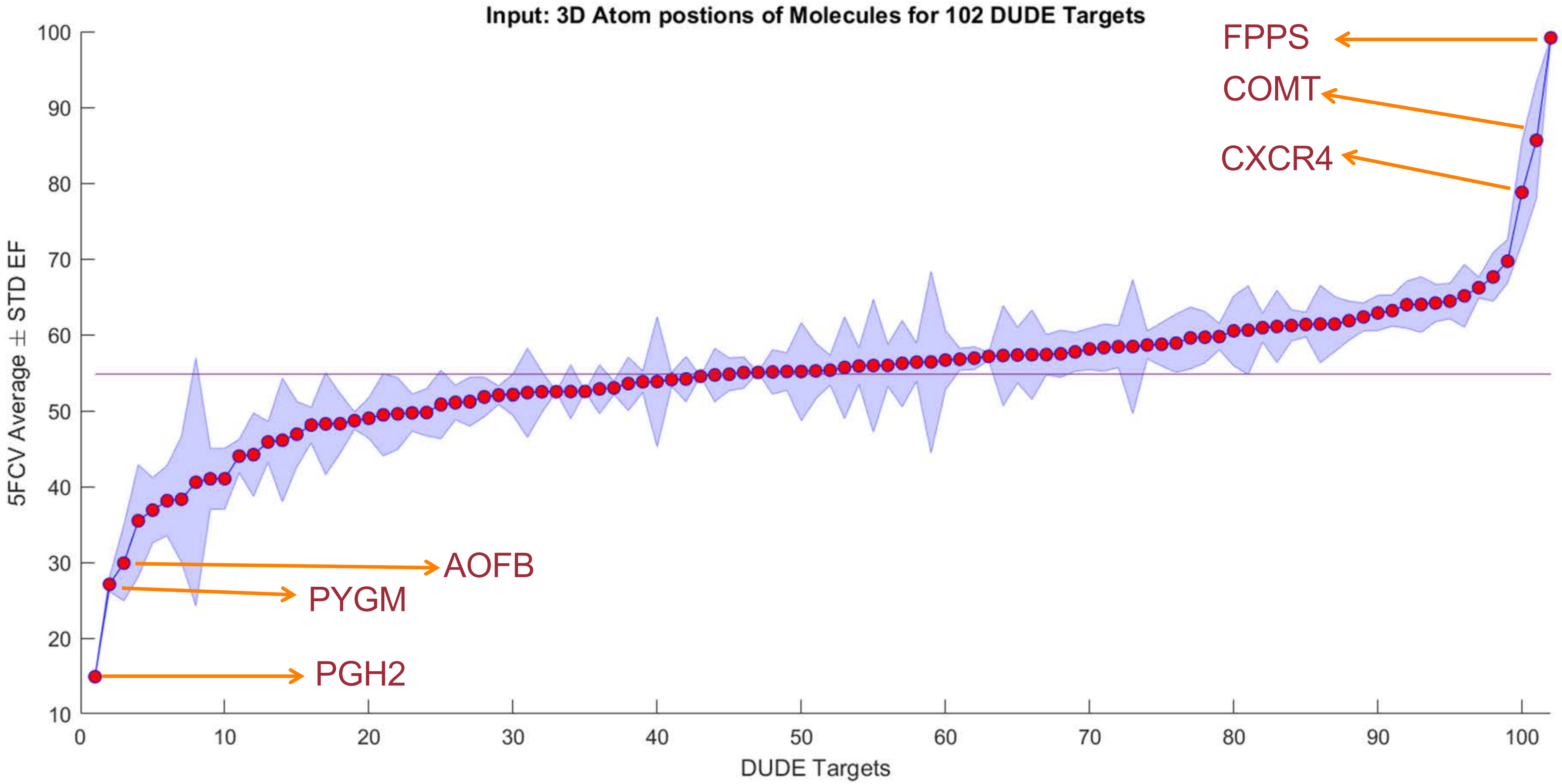
Diversity Analysis for DUD-E Targets

Diversity= (1-DiceSimilarity) of the Morgan fingerprints (radius 3)

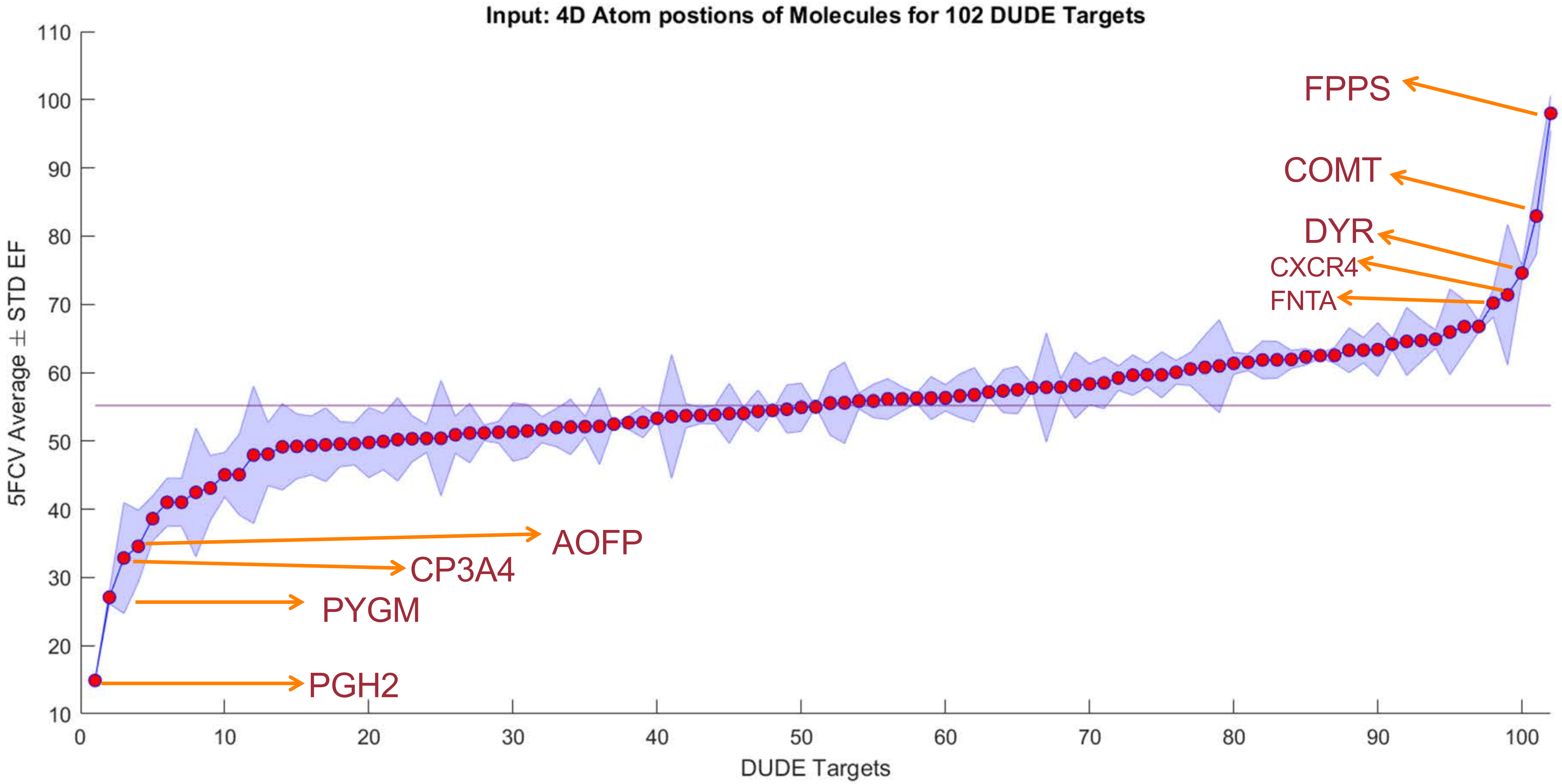


Less diverse actives have better hit rates: easier to learn

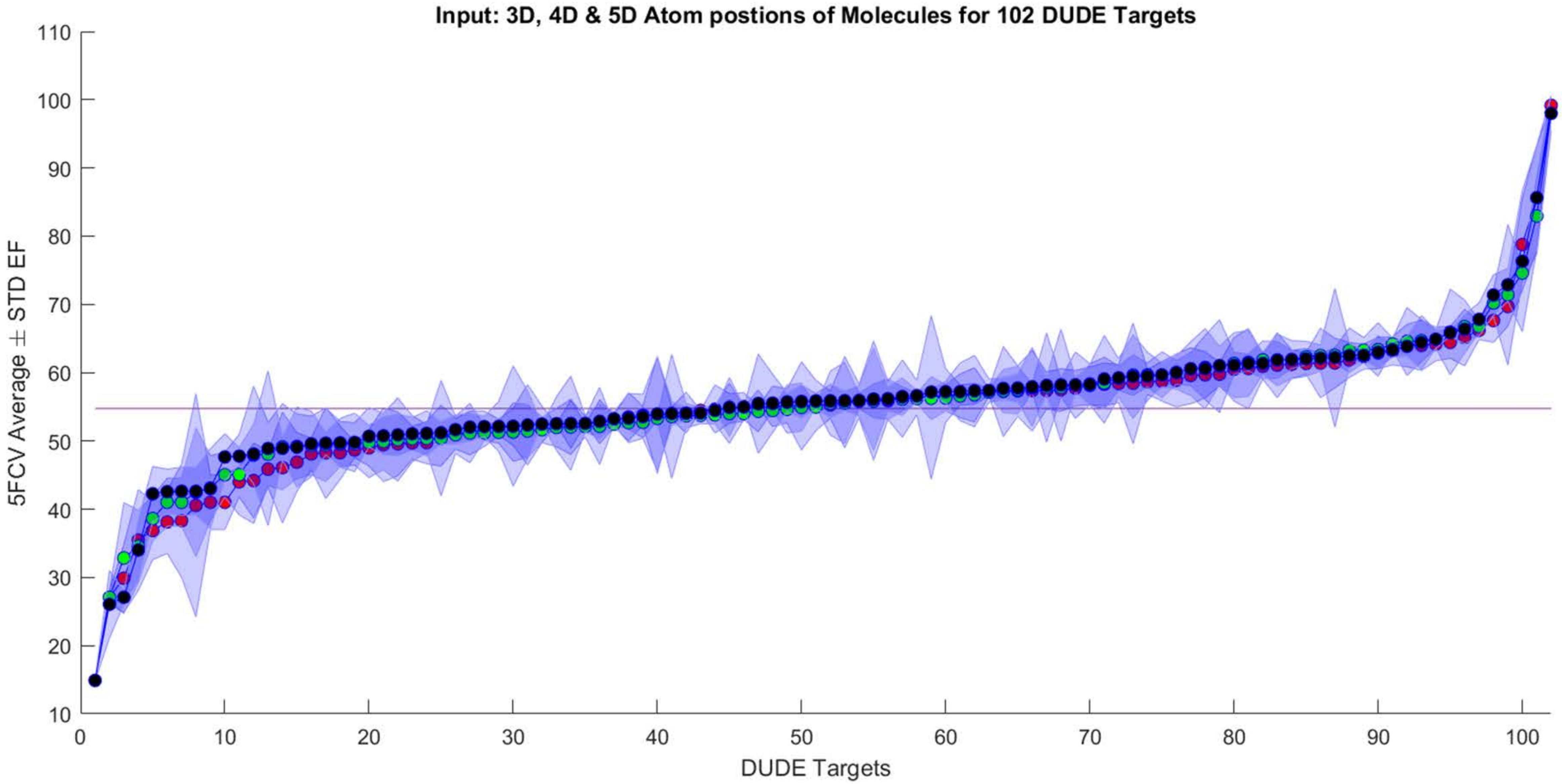
The Input: 3D atom Positions, Metric: 1% Enrichment Factor



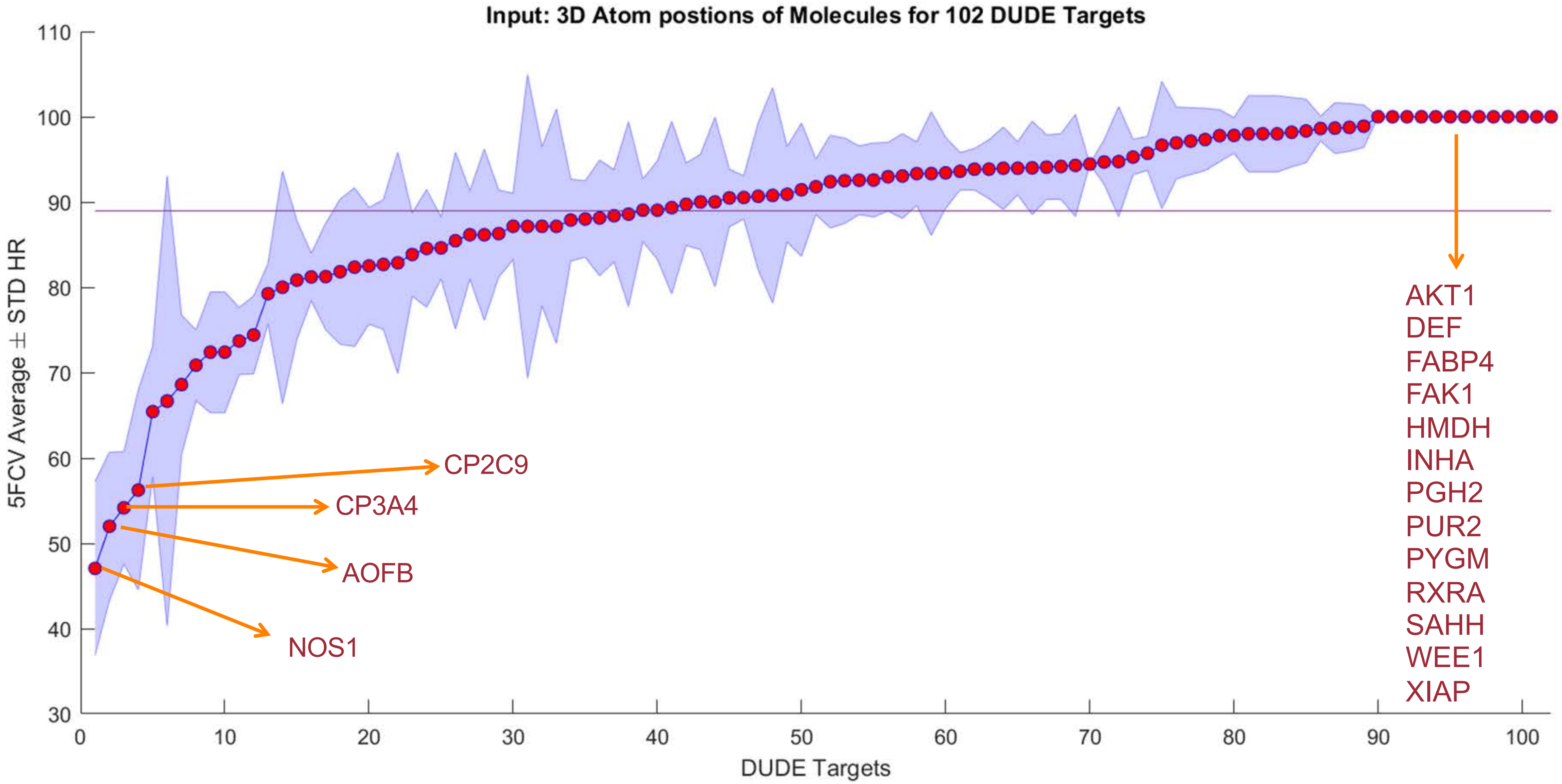
The Input: 4D atom Positions, Metric: 1% Enrichment Factor



The Input: 5D atom Positions, Metric: 1% Enrichment Factor

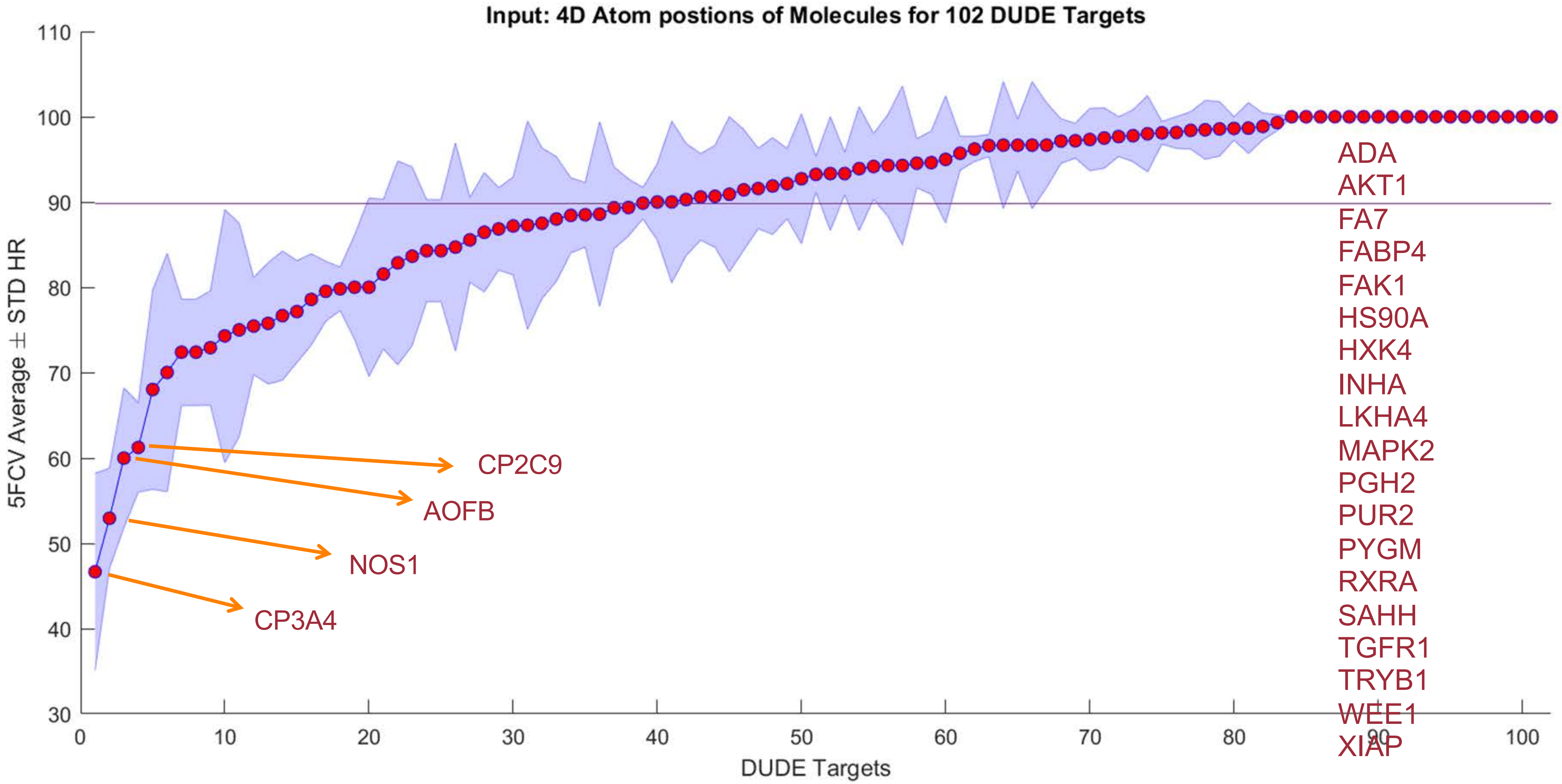


The Input: 3D atom Positions, Metric: Hit-Rate (HR)

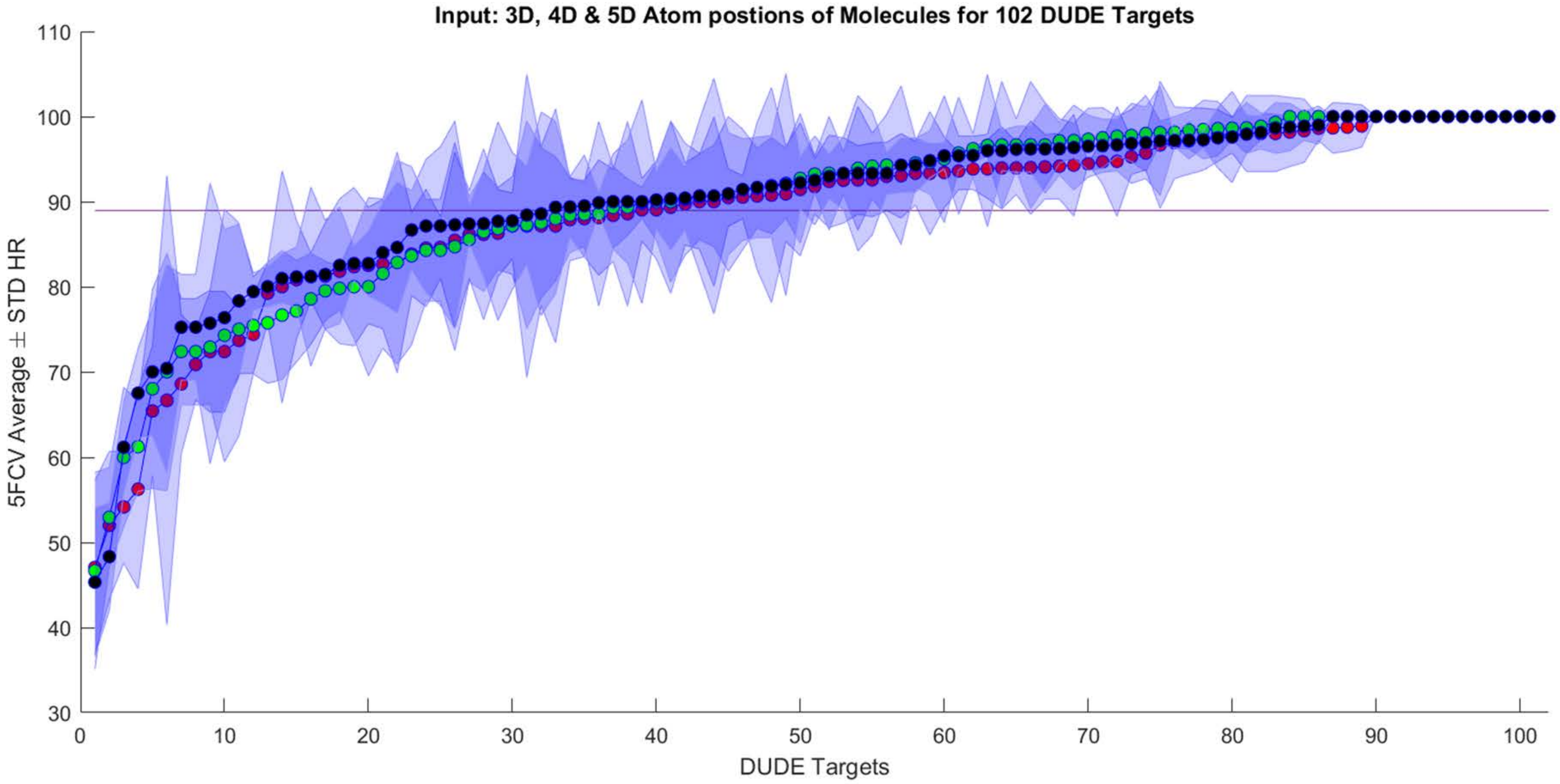


AKT1
DEF
FABP4
FAK1
HMDH
INHA
PGH2
PUR2
PYGM
RXRA
SAHH
WEE1
XIAP

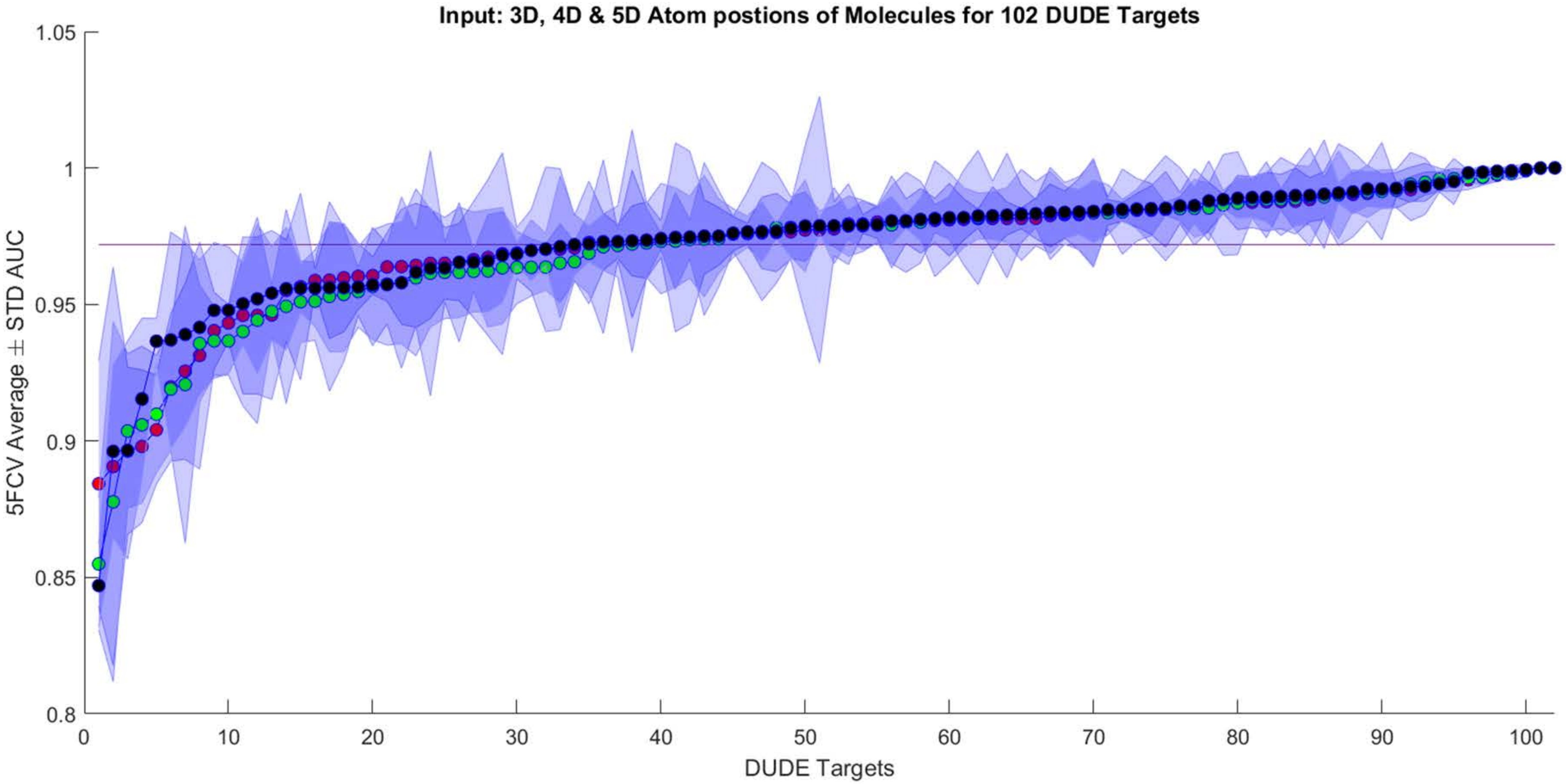
The Input: 4D atom Positions, Metric: Hit-Rate (HR)



The Input: 3D,4D,5D atom Positions, Metric: Hit-Rate (HR)



The Input: 3D, 4D, 5D   atom Positions, Metric: AUC



Results on Testing on MUV Dataset

Comparing to the work of Tiikkainen et al. (J. Chem. Inf. Model. 2009, 49, 2168–2178)

MUV-AUC	PersStats	ROCS	BRUTUS	EON
Average	0.6839	0.5771	0.5129	0.542
STDEV	0.10	0.08	0.08	0.07

MUV Dataset:
The data set comprises confirmed active molecules and decoys for 17 target classes.
For each target class there are 30 active molecules and 15,000 decoys.

*Authors of the MUV data set had chosen the active molecules so that they occupy different areas of chemical space as defined with simple chemical properties such as heavy atom count and hydrogen bond donors and acceptors. In contrast, decoys that resemble the active molecules with respect to these simple properties were chosen. This process led to data sets where active molecules cannot be separated from decoys using simple properties