6th Advanced In silico Drug Design workshop/challenge

30 January - 3 February 2023
Olomouc, Czech Republic

Univerzita Palackého v Olomouci

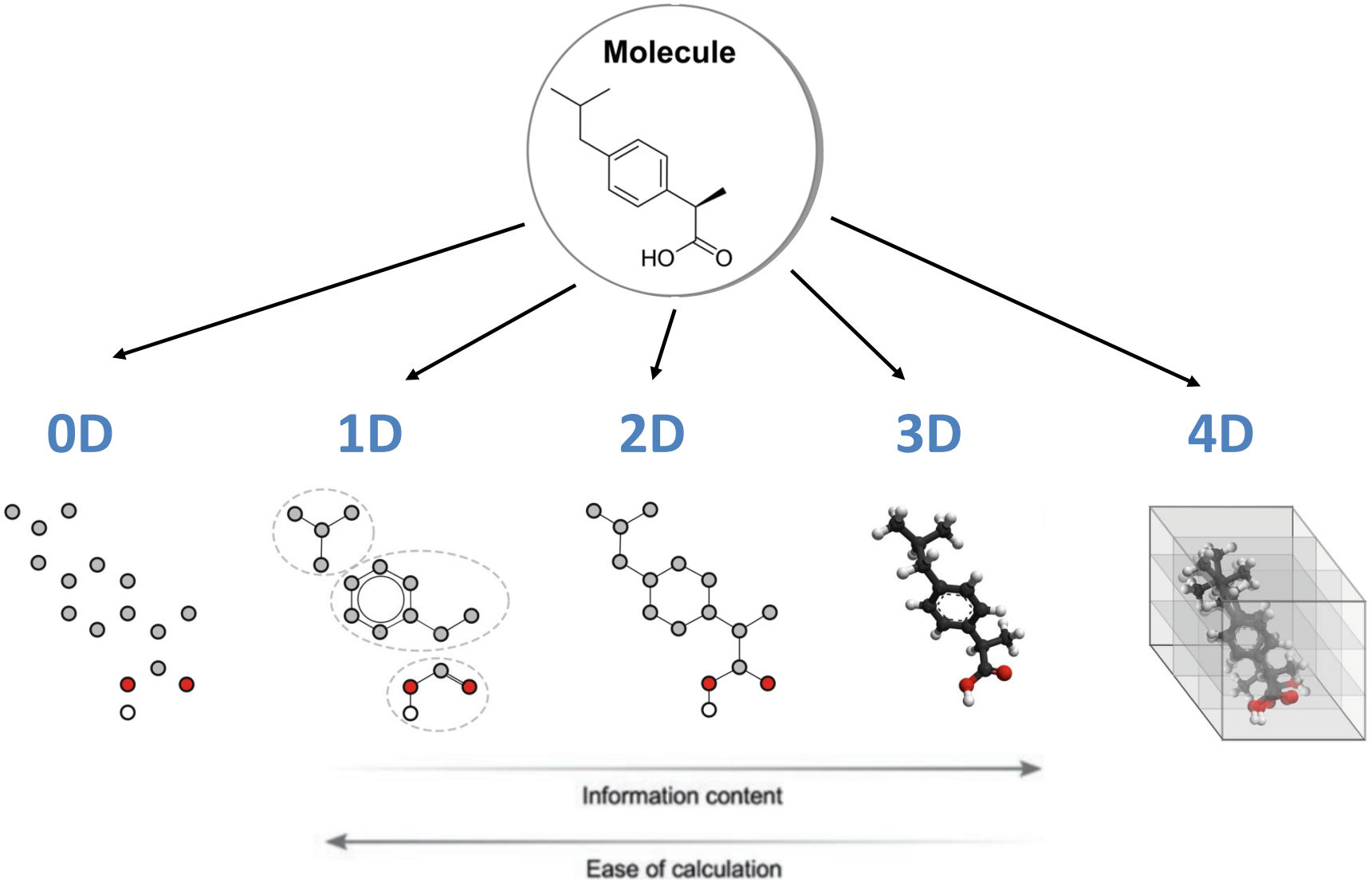
Multi-instance learning

Pavel Polishchuk

Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University

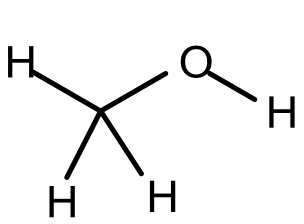
pavlo.polishchuk@upol.cz
qsar4u.com

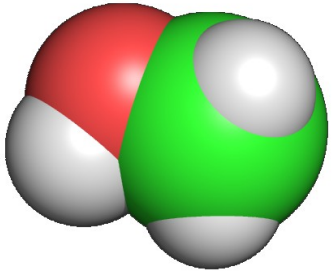
Molecular representation levels



Molecular representation levels

Molecular graph is only skeleton of a molecule
Real interactions between molecules go through their surface!



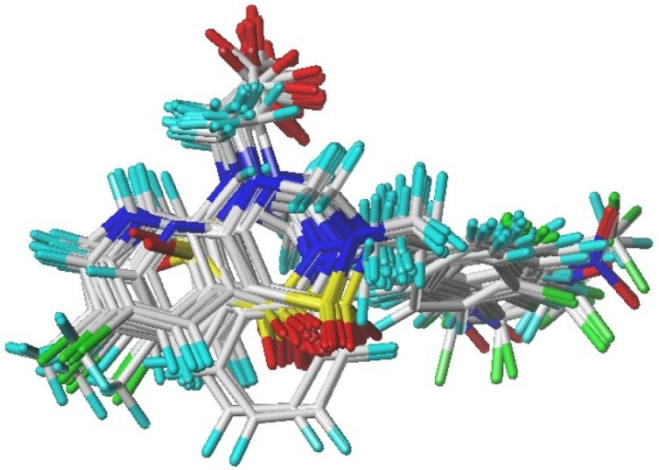


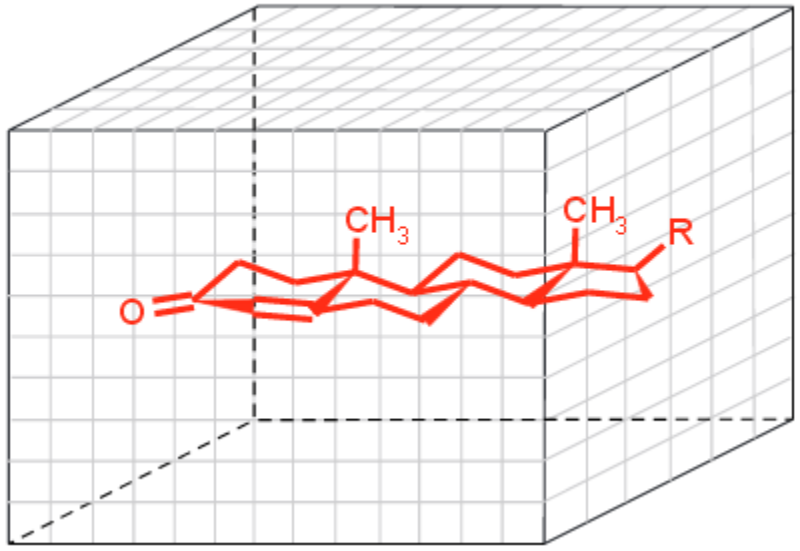
The skeleton:

How aliens would reconstruct the animal:

The animal:

Comparative Molecular Fields Analysis (CoMFA)





probe: electrostatic (H^+) steric (Csp^3)

PLS model

Y
2.4
3.7
...
8.1

Mol	Electrostatic field descriptors								Steric field descriptors							
	(0,0,0)	(0,0,1)	...	...	...	...	...	(G,G,G)	(0,0,0)	(0,0,1)	...	...	...	...	...	(G,G,G)
1																
2																
...																
N																

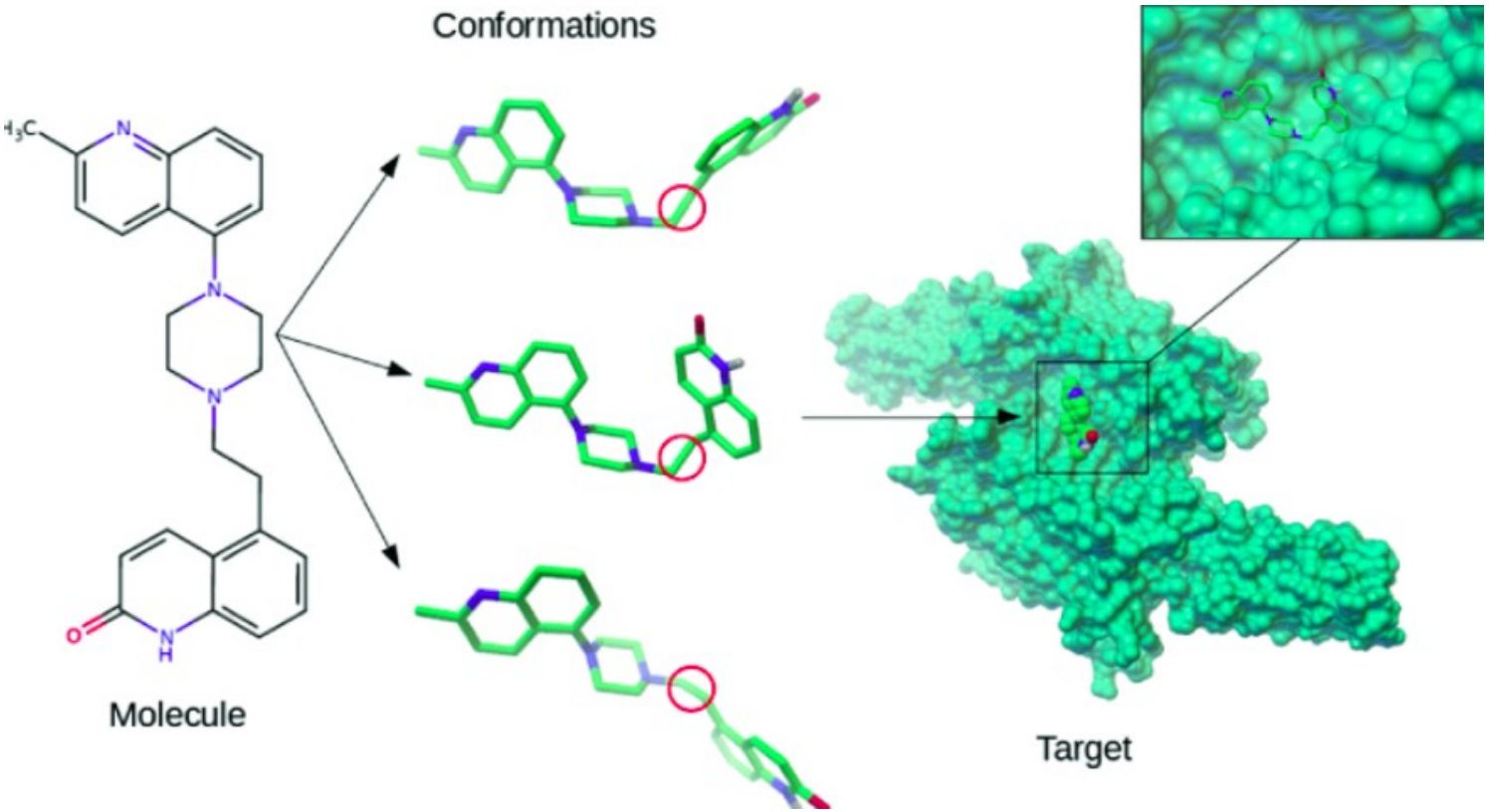
Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988,** *110*, 5959-5967

CoMFA: summary

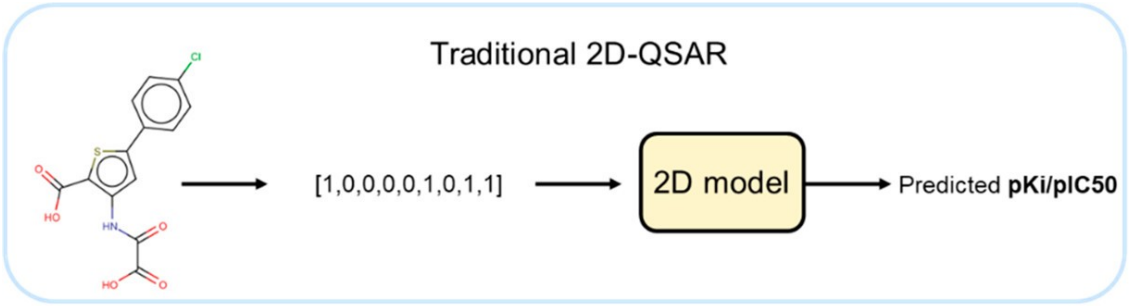
+ clear interpretation
+ explicit representation of stereochemistry

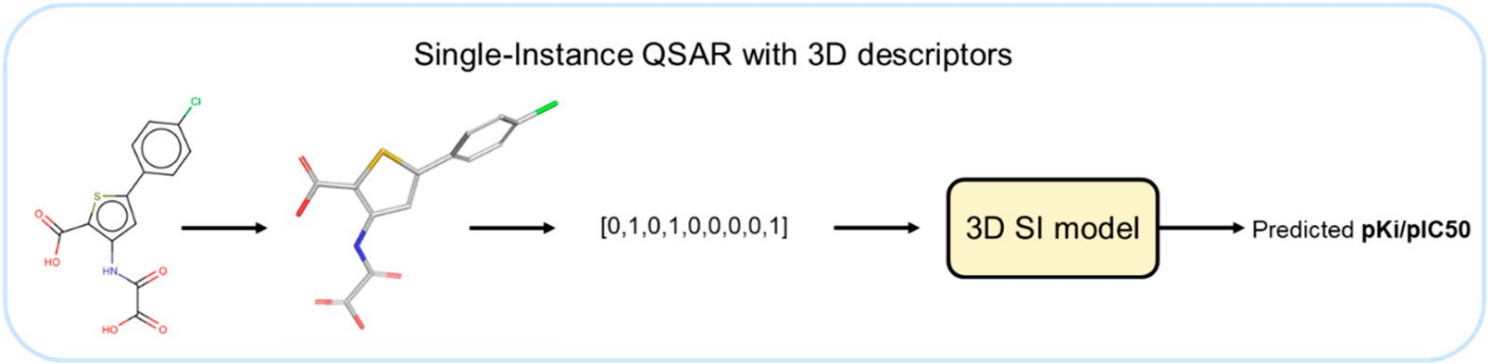
- dependence on chosen conformations
- dependence on alignment algorithm
- sensitivity to cell size
- sensitivity to scheme of partial atomic charges calculation
- doesn’t describe properly H-bonding and π-interactions

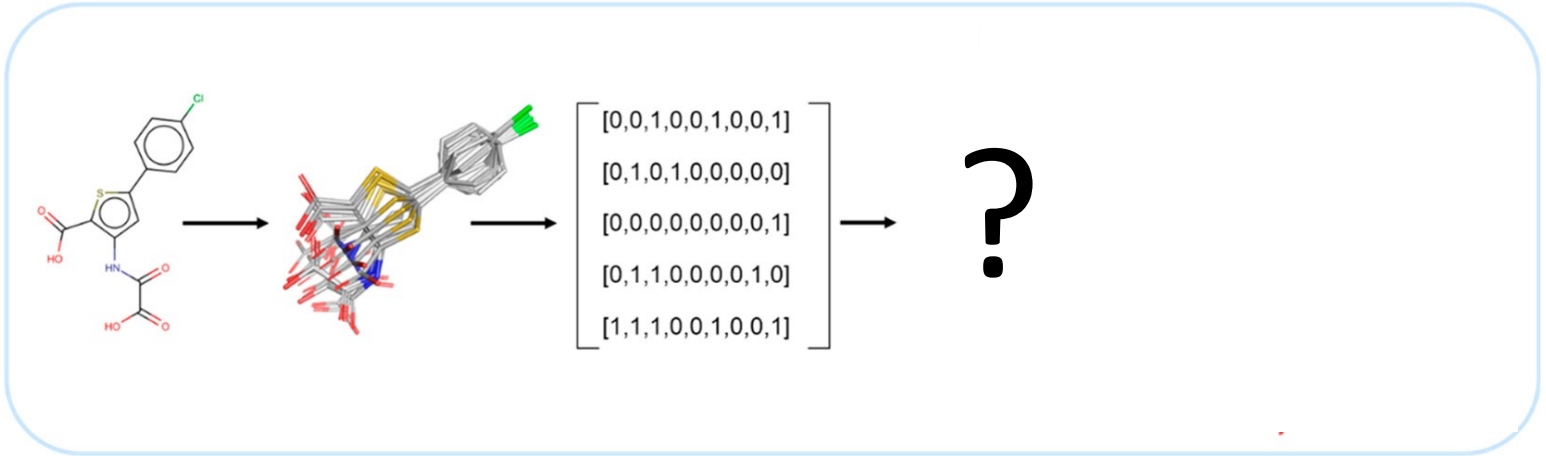
“Bioactive” conformation



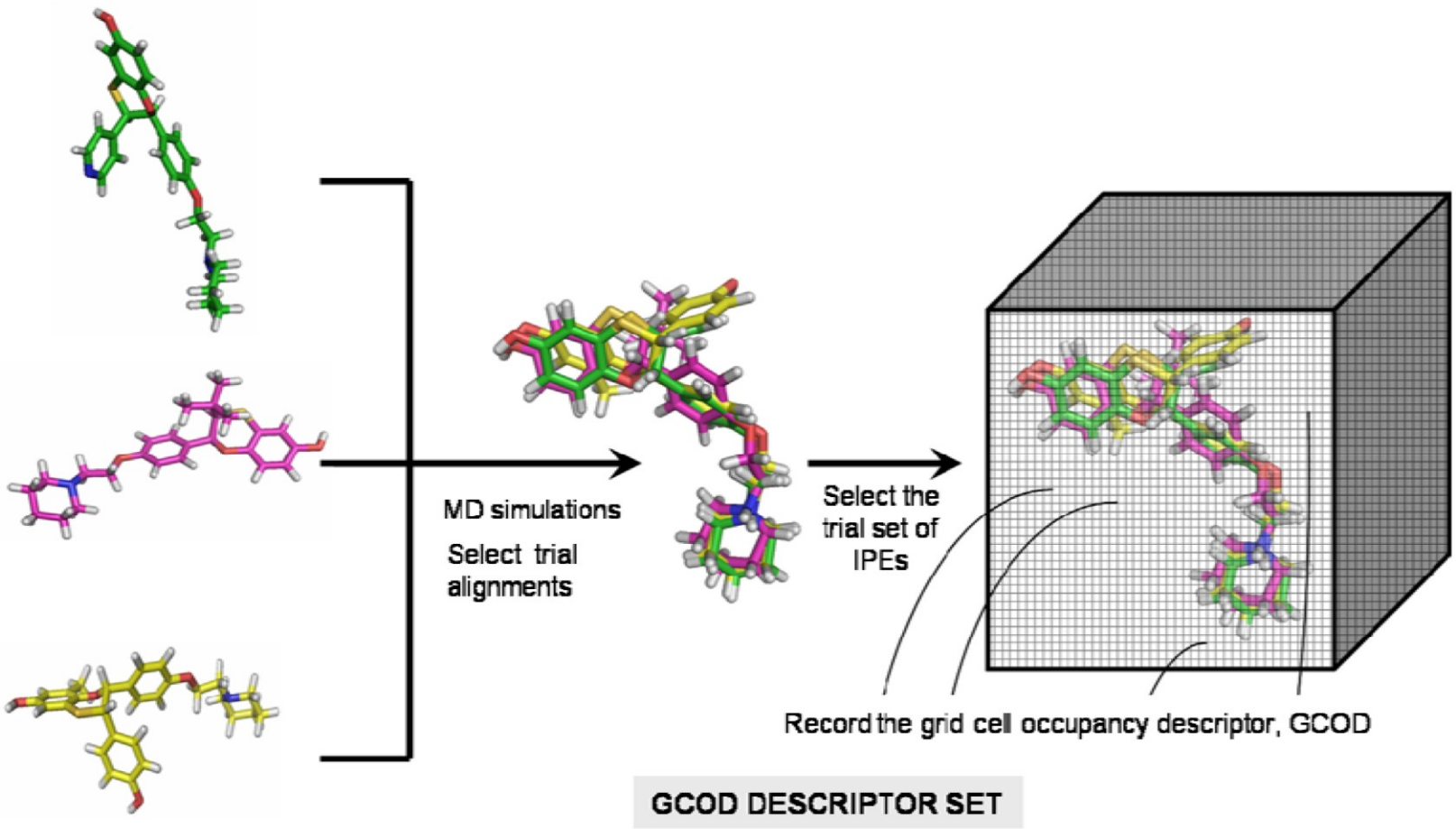
Multi-conformer challenge







4D QSAR



Andrade, C. H. et al, 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, 15, 3281-3294.

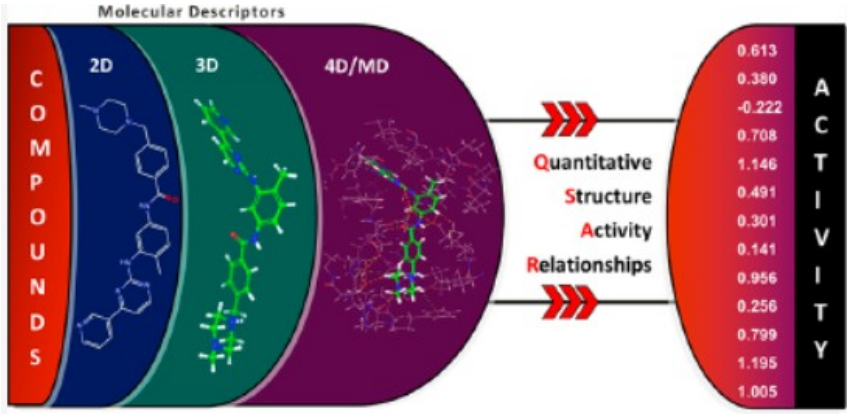
4D QSAR

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict?

Phyo Phyo Kyaw Zin, Alexandre Borrel, and Denis Fourches*



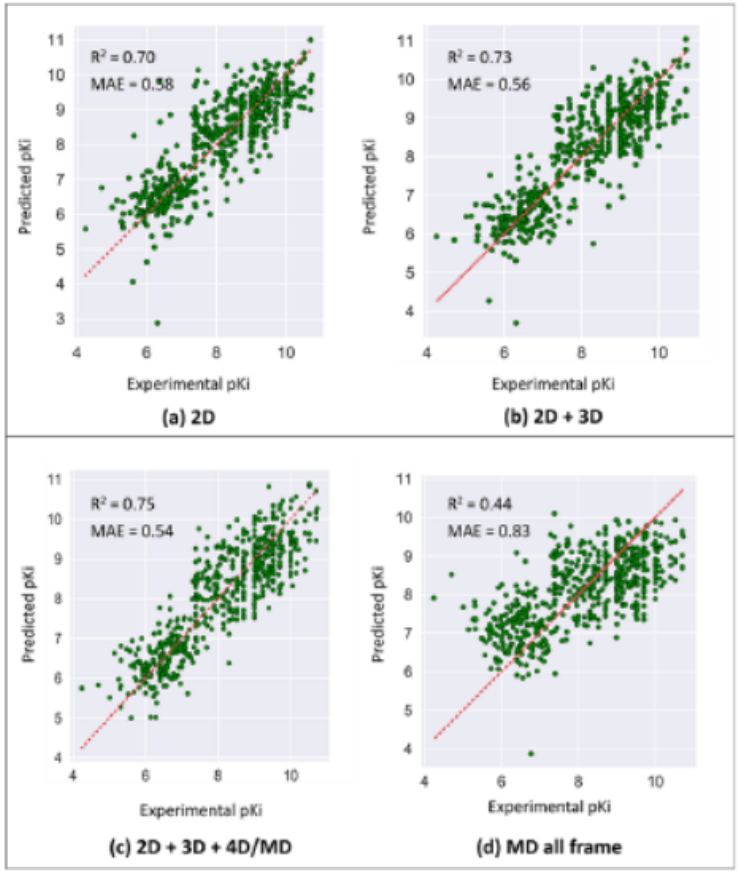


Figure 15. Native 10-fold cross-validation predictions of pKi-reported Imatinib analogues by DNN-based regression models using descriptors of (A) 2D, (B) 2D+3D, (C) 2D+3D+4D/MD, and (D) MD all frame. Red dotted line represents an ideal fit between experimental pKi and predicted pKi.

4D QSAR: summary

+ clear interpretation

+ explicit representation of stereochemistry

- dependence on alignment algorithm

- equal treating of all conformers (some weighted scheme exist to solve this)

Multi-instance learning

ELSEVIER

Artificial Intelligence 89 (1997) 31–71

Artificial Intelligence

Solving the multiple instance problem with axis-parallel rectangles

Thomas G. Dietterich [a,*], Richard H. Lathrop [b], Tomás Lozano-Pérez [c,d]

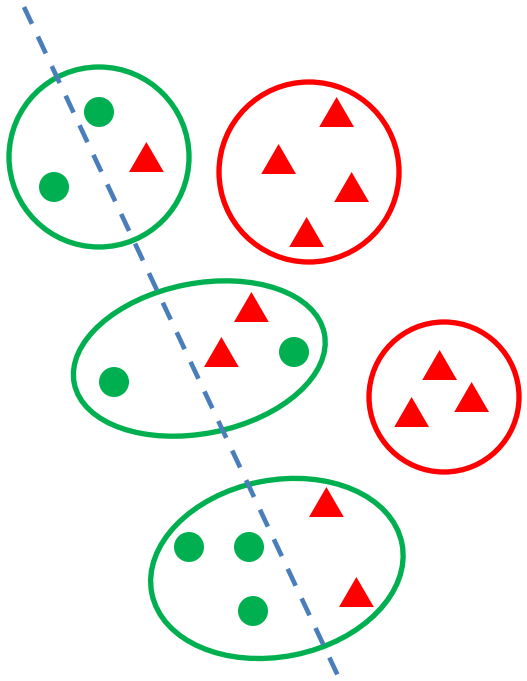
[a] *Department of Computer Science, Oregon State University, Dearborn Hall 303, Corvallis, OR 97331-3202, USA*

[b] *Department of Information and Computer Science, University of California, Irvine, CA 92697, USA*

[c] *Arris Pharmaceutical Corporation, 385 Oyster Pt. Blvd., South San Francisco, CA 94080, USA*

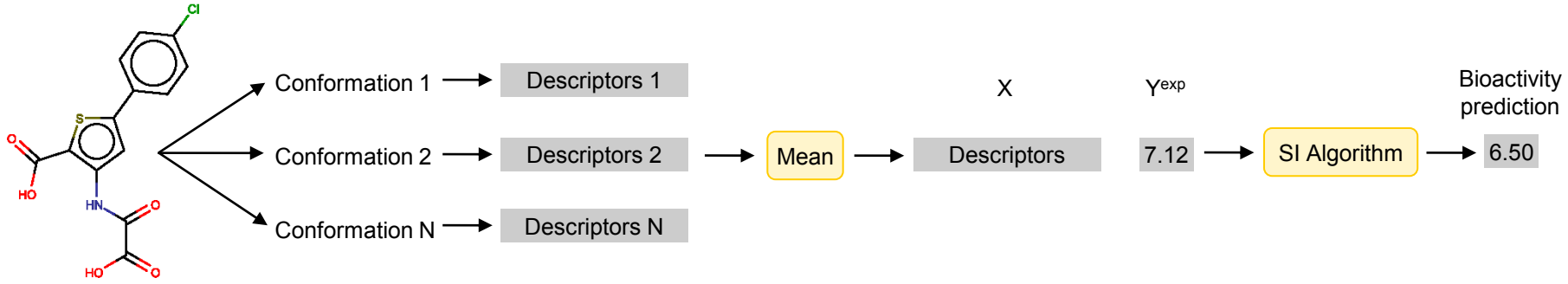
[d] *MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139, USA*

Received August 1994; revised July 1996

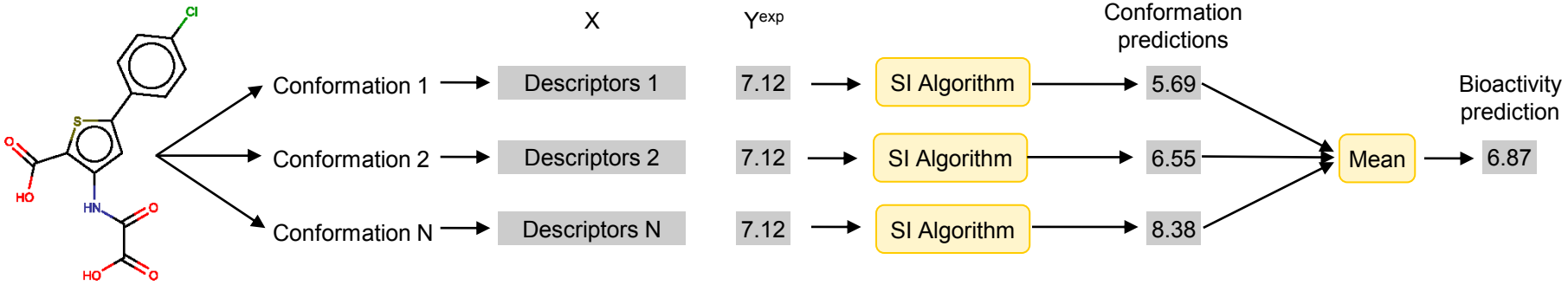


Multi-instance learning: naïve approaches (wrappers)

Bag-wrapper: averaging of conformation **descriptors**

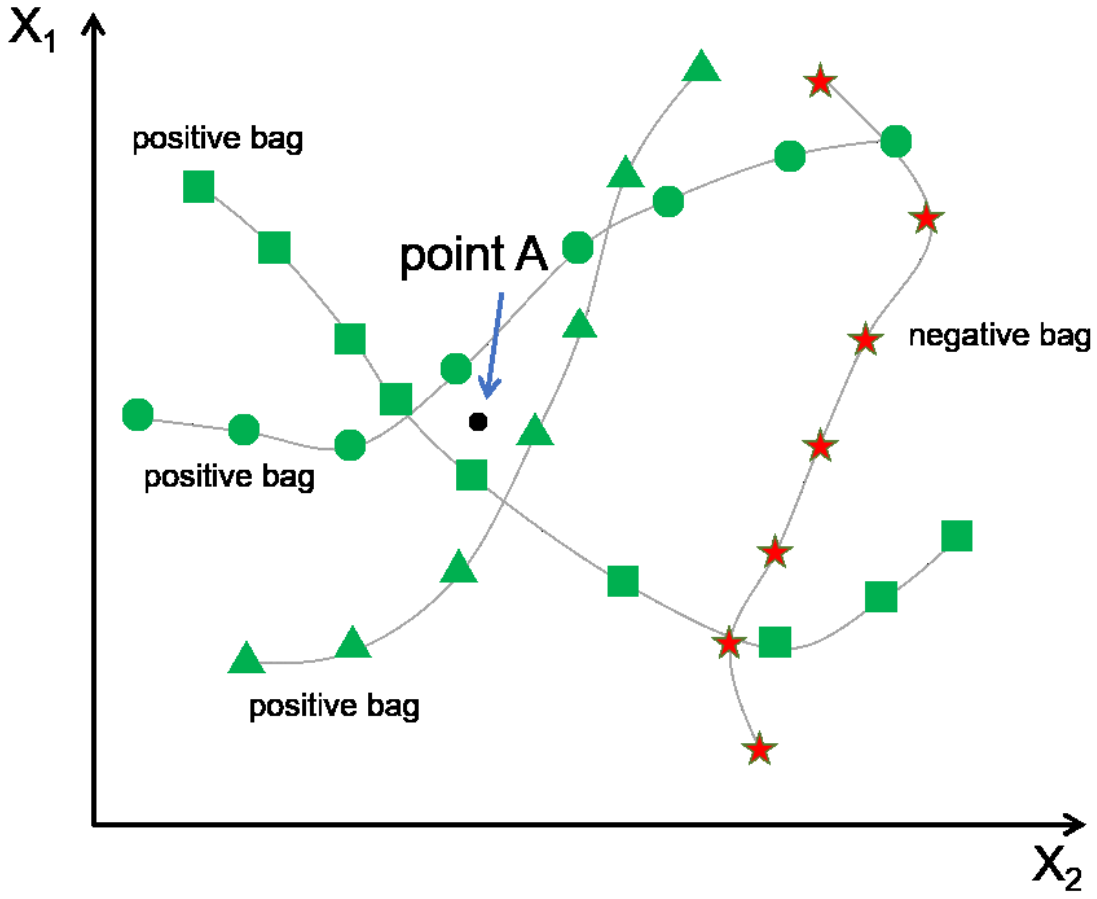


Instance-wrapper: averaging of conformation **predictions**



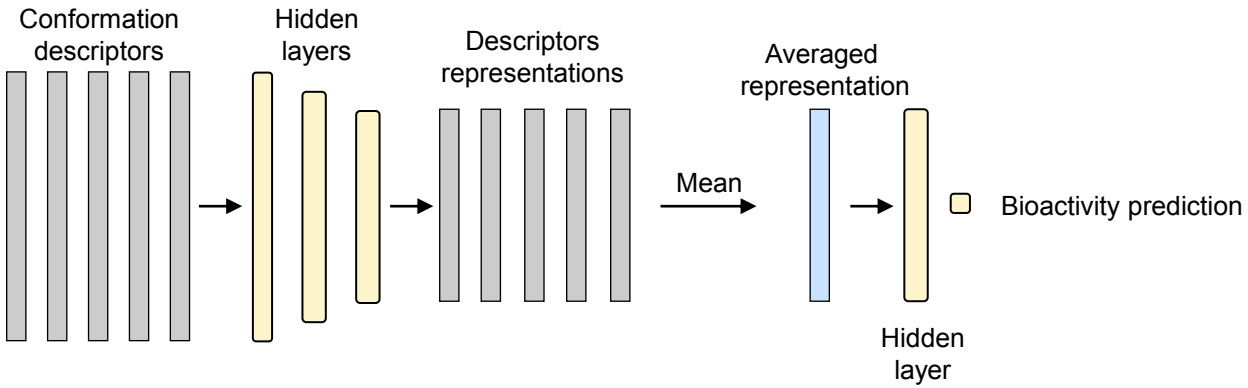
Multi-instance learning: conventional approaches

Diversity density

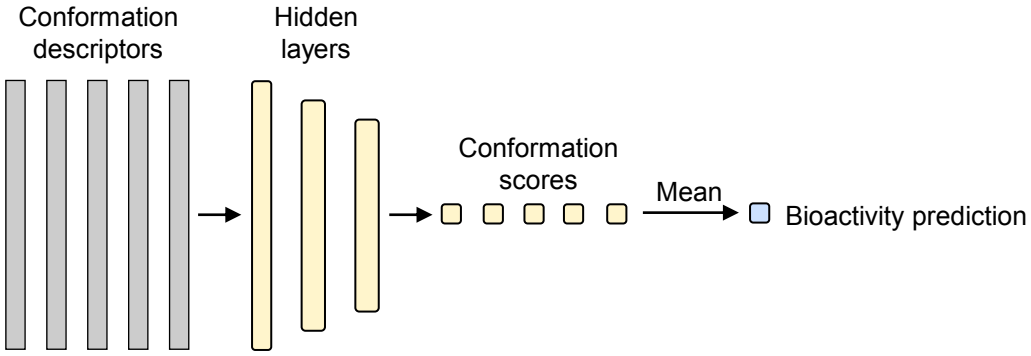


Multi-instance learning: NN approaches

Bag-Net: averaging of conformation **embeddings**

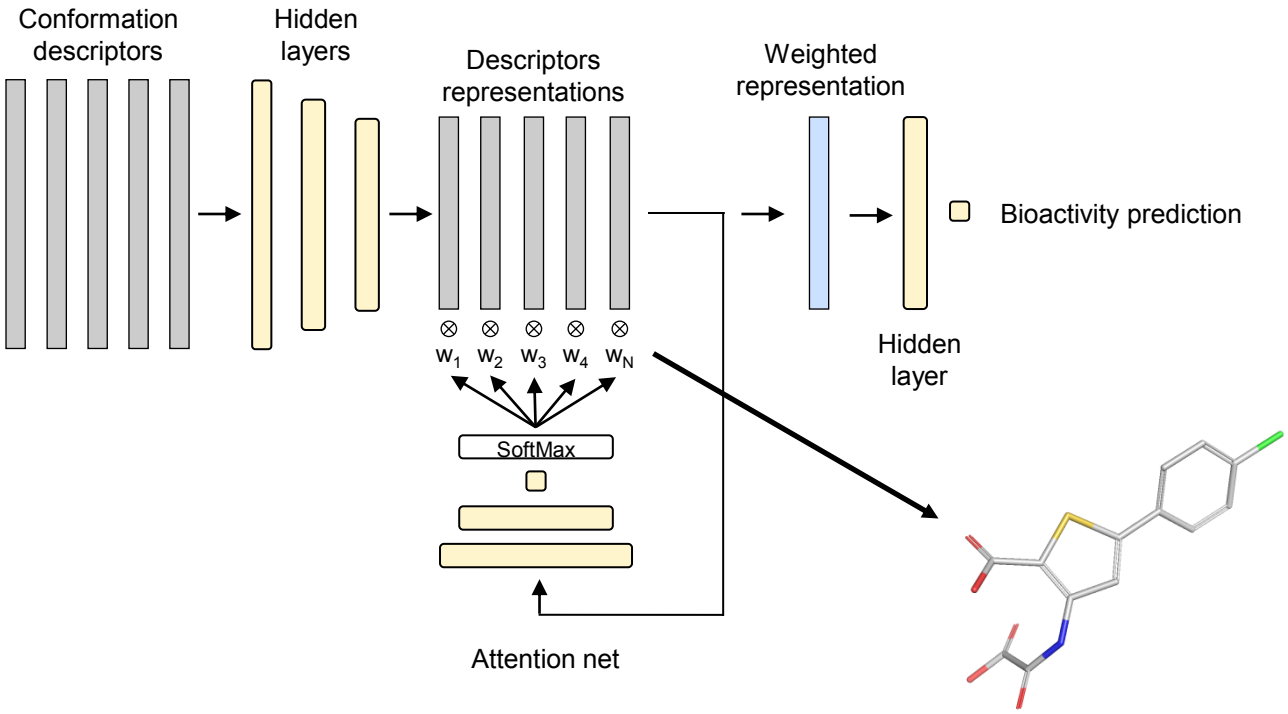


Instance-Net: averaging of conformation **scores**

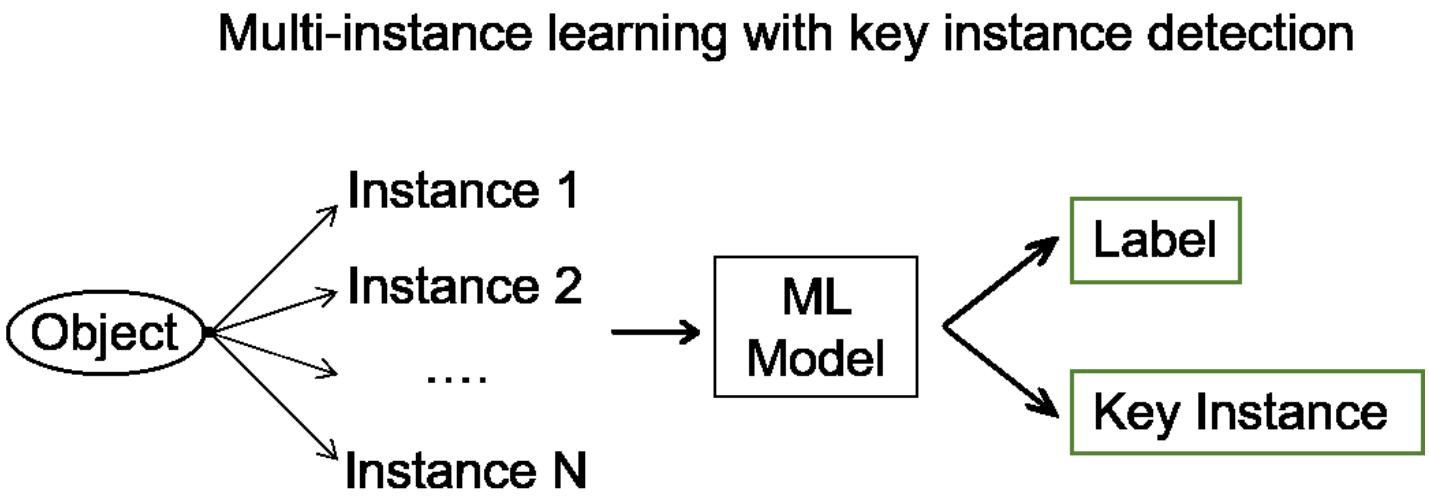


Multi-instance learning: attention-based NN approach

Bag-Attention Net: weighted averaging of conformation **embeddings**



Multi-instance learning: key instance detection (KID)



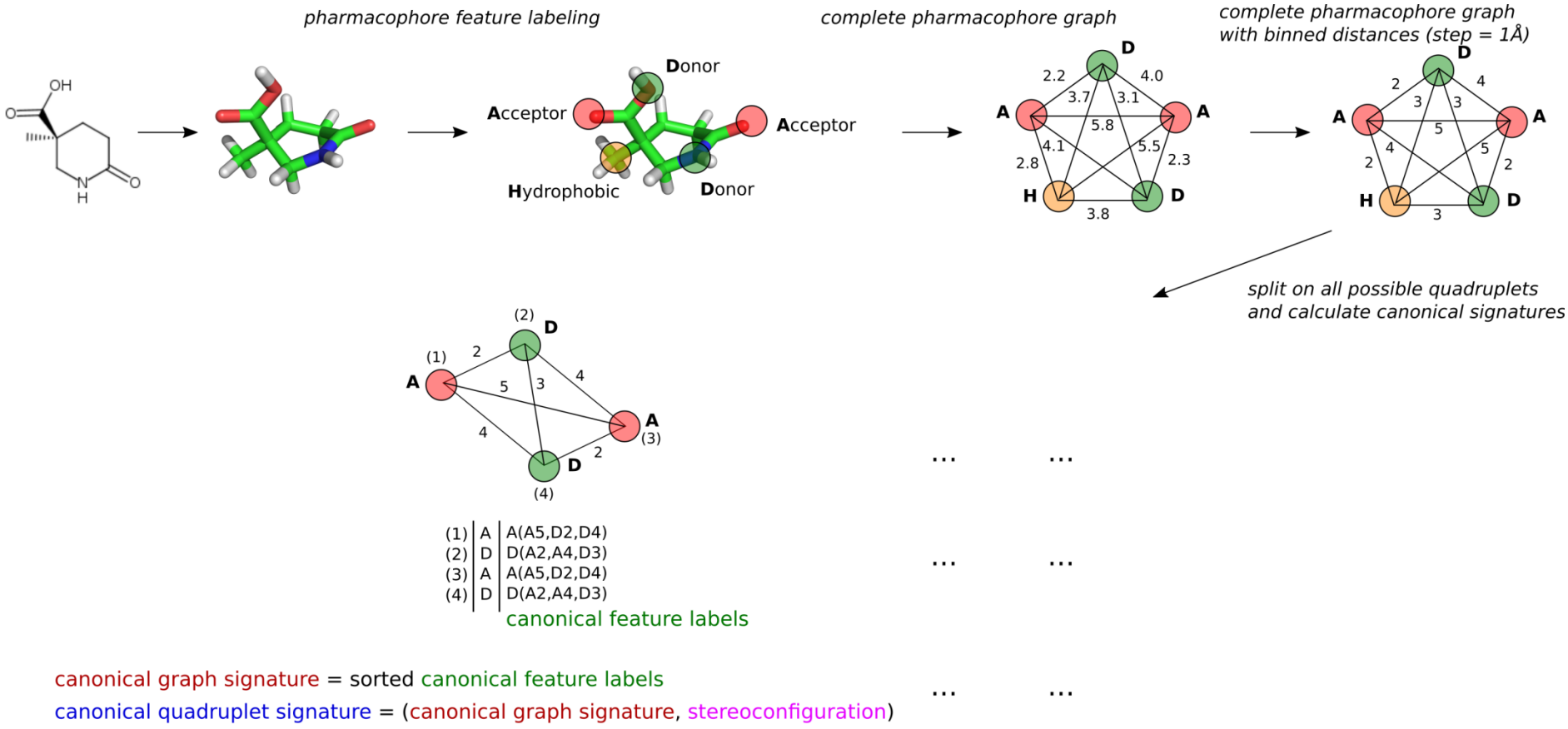
Model-specific:

- instance-based
- bag-based

Model-agnostic:

- single instance prediction
- single instance masking

Pmapper: 3D pharmacophore descriptors



https://github.com/DrrDom/pmapper

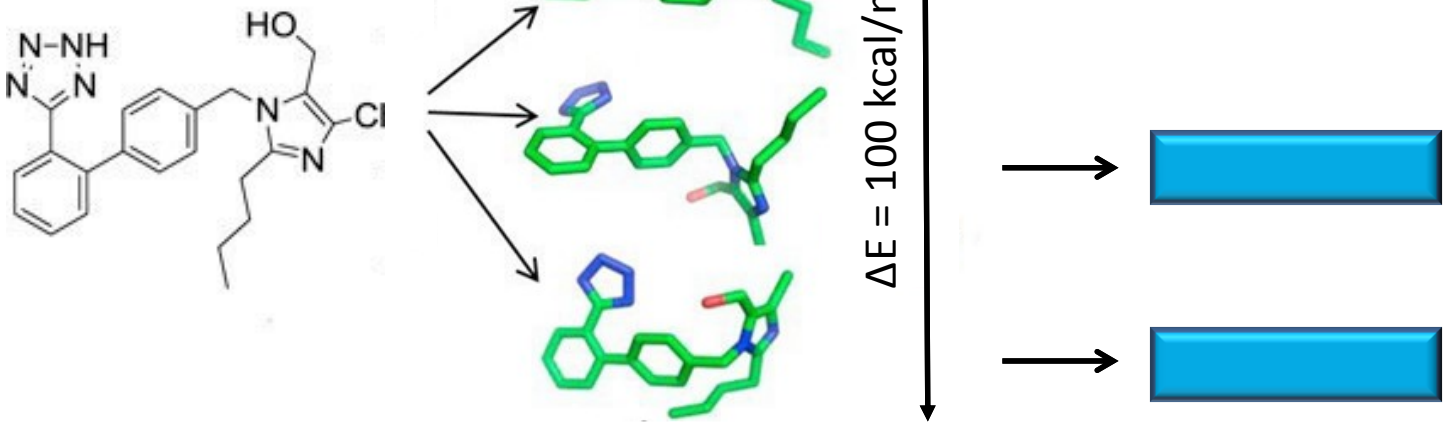
Kutlushina, A. et al., Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **2018**, 23, 3094.

MIL study: conformer and descriptor generation

175 data sets from ChEMBL

Generation of
up to **100**
conformations per
molecule
RDKit

Calculation of
Pmapper 3D descriptors

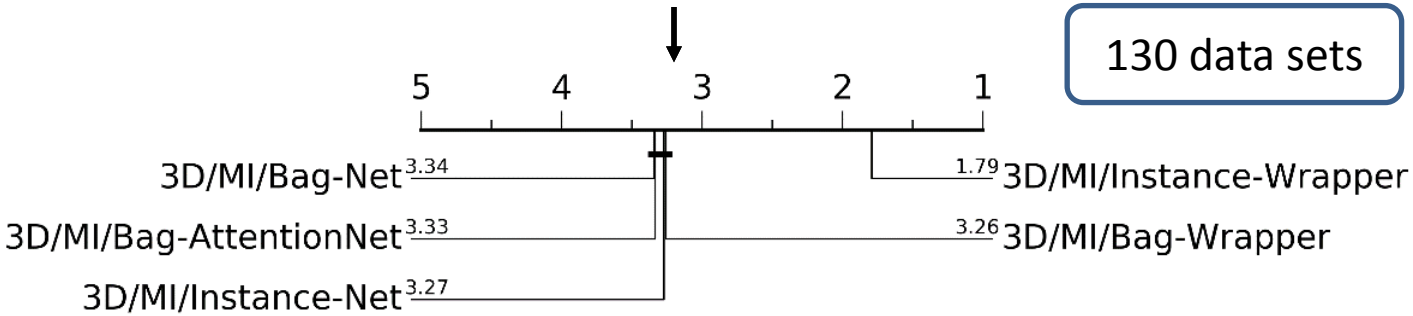


Kutlushina, A. et al. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **2018**, 23, 3094.

MIL study: comparison of MIL algorithms

	3D/MI/Bag-AttentionNet	3D/MI/Instance-Wrapper	3D/MI/Bag-Wrapper	3D/MI/Bag-Net	3D/MI/Instance-Net
~~CHEMBL1855~~	~~0.248~~	~~0.272~~	~~0.322~~	~~0.261~~	~~0.267~~
CHEMBL2034	0.624	0.664	0.645	0.644	0.638
CHEMBL261	0.313	0.404	0.326	0.322	0.321
CHEMBL322	0.348	0.432	0.357	0.377	0.375
CHEMBL2094122	0.254	0.310	0.420	0.422	0.430
...	...	...	...	...	...
~~CHEMBL238~~	~~0.395~~	~~0.397~~	~~0.367~~	~~0.374~~	~~0.373~~
CHEMBL210	0.535	0.636	0.546	0.548	0.558
CHEMBL3571	0.357	0.420	0.359	0.383	0.379
CHEMBL222	0.418	0.437	0.391	0.400	0.395
CHEMBL1899	0.255	0.660	0.578	0.584	0.582

130 data sets



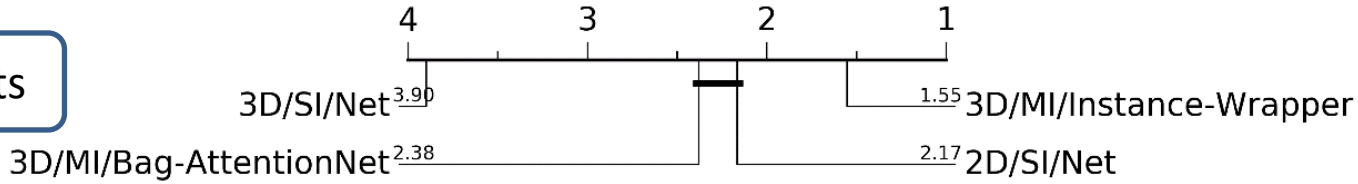
Groups of models that are not significantly different (at a confidence level of 0.05) are connected by the thick line

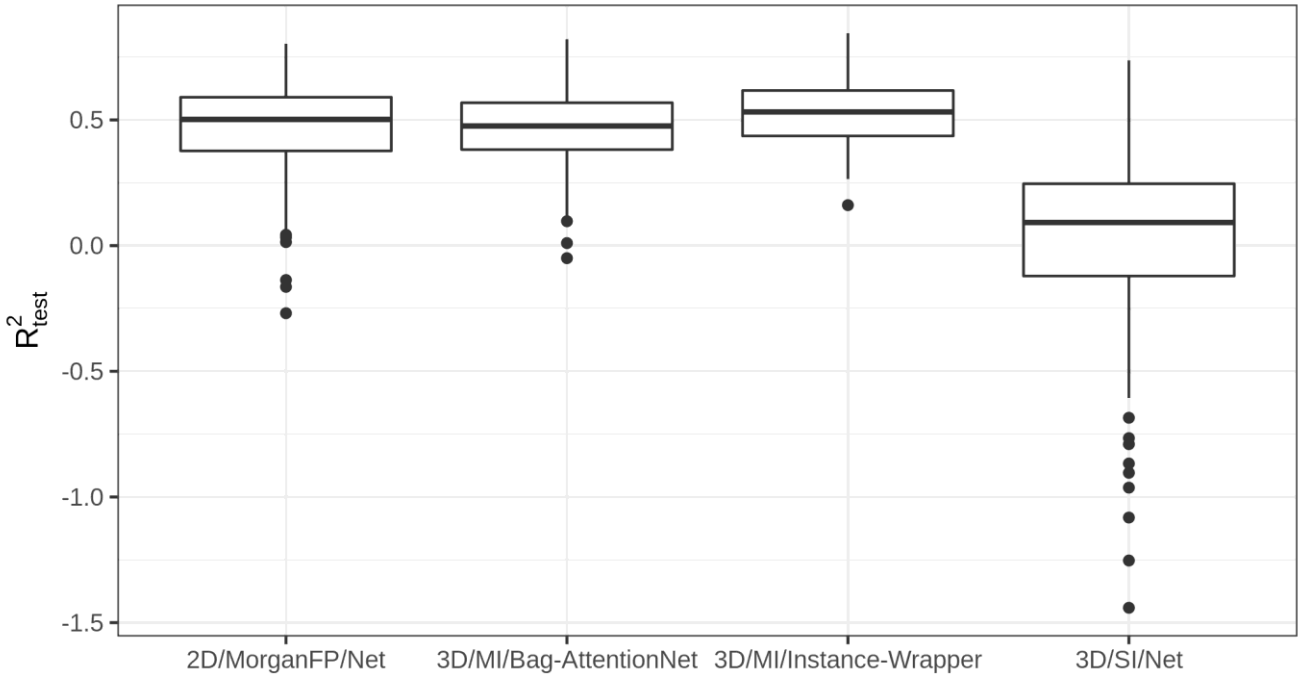
Zankov, D. V. et al, QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923

MIL study: comparison between 2D, 3D and MIL models

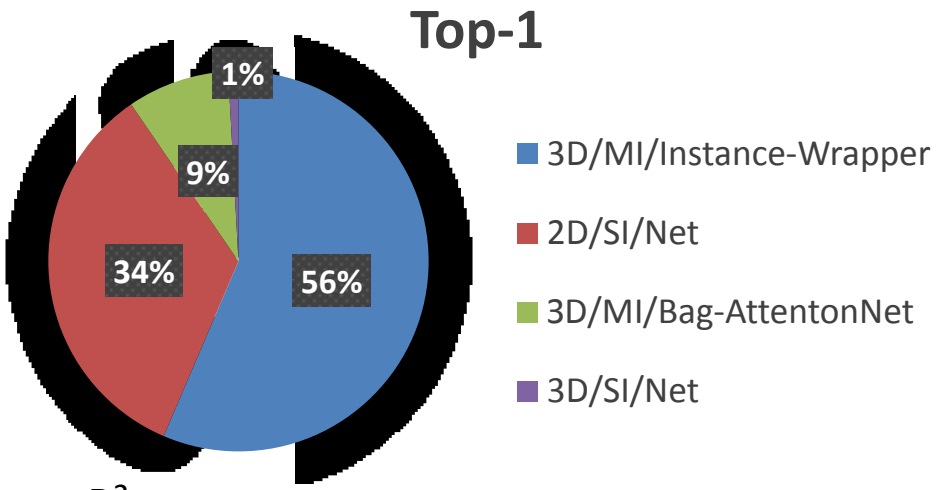
Model	Mean R^2_{test}	Median R^2_{test}
3D/MI/Instance-Wrapper	0.530 ± 0.123	0.532
2D/SI/Net	0.474 ± 0.189	0.503
3D/MI/Bag-Attention Net	0.474 ± 0.157	0.476
3D/SI/Net	0.027 ± 0.374	0.092

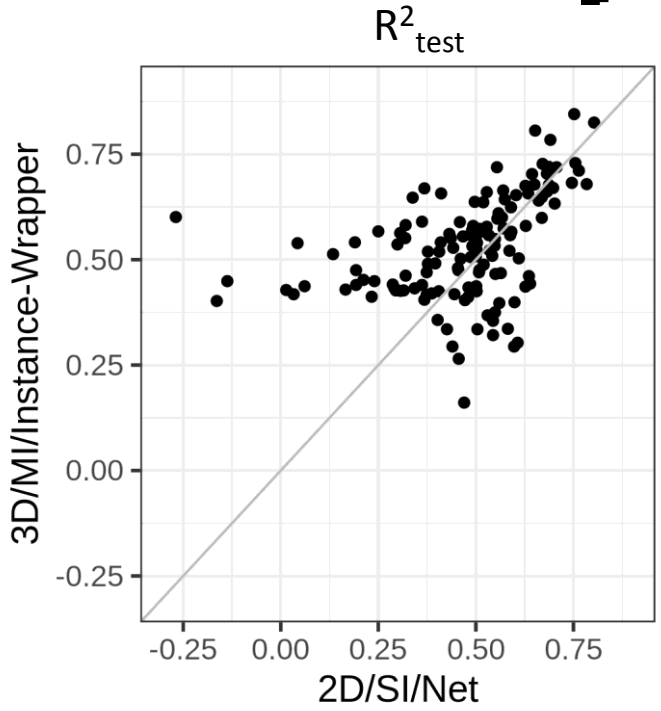
139 data sets

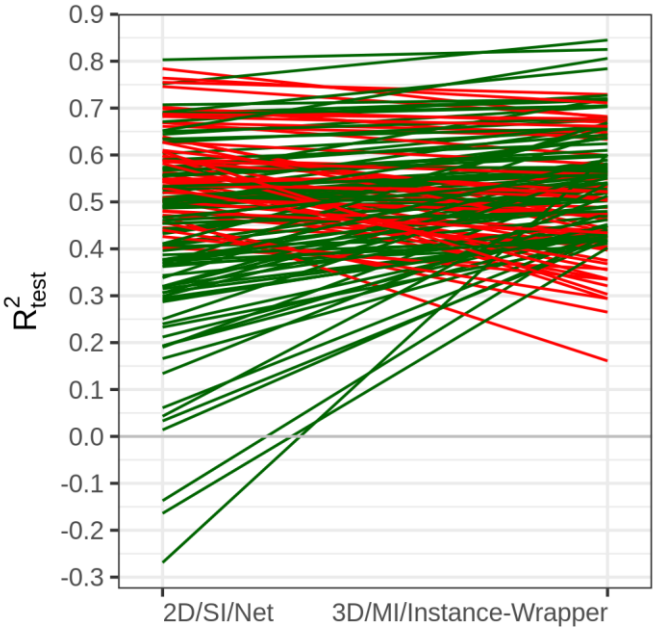




MIL study: comparison between 2D, 3D and MIL models



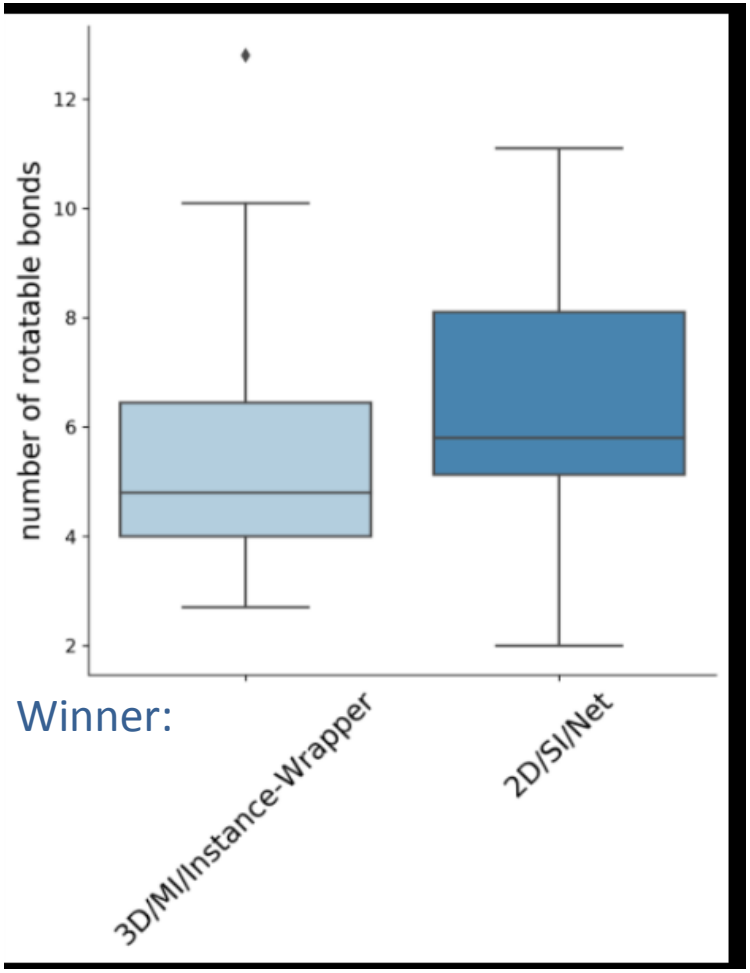




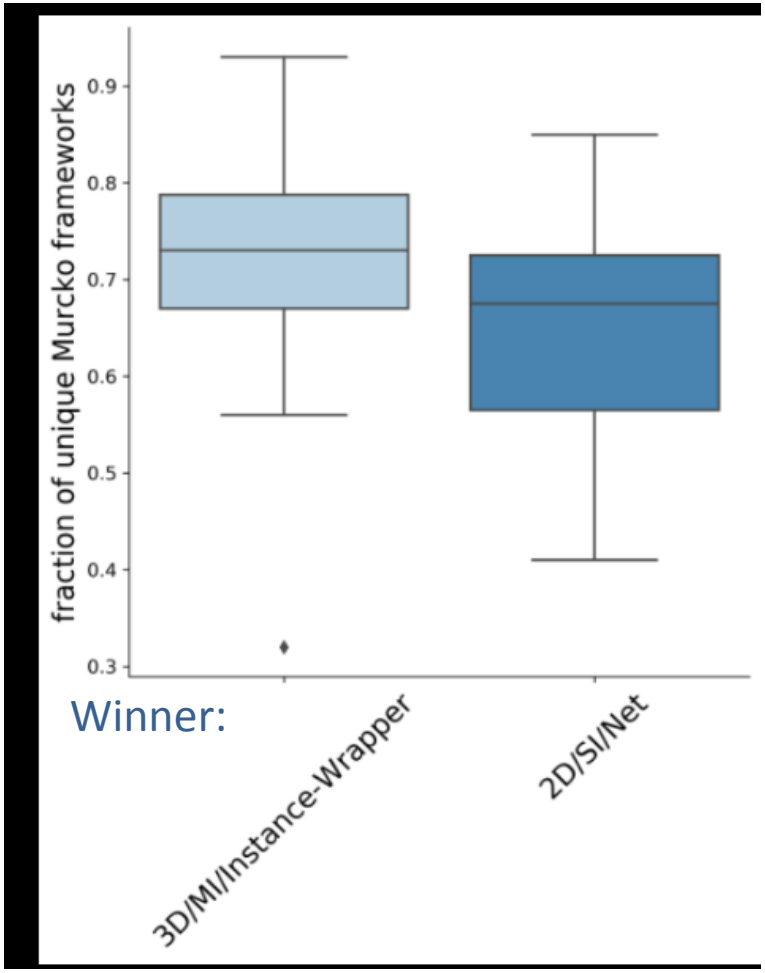
Zankov, D. V. et al, QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923

MIL study: comparison between 2D and MIL models

Number of rotatable bonds

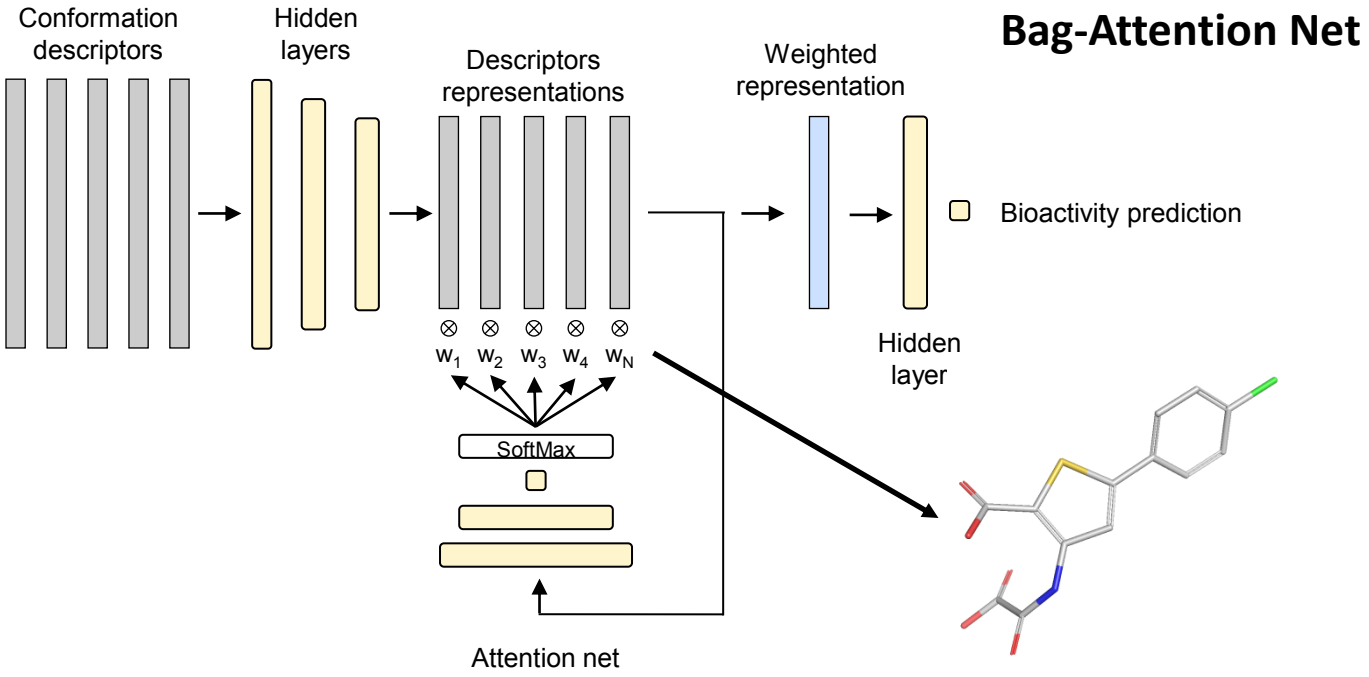


Fraction of distinct Bemis-Murcko scaffolds



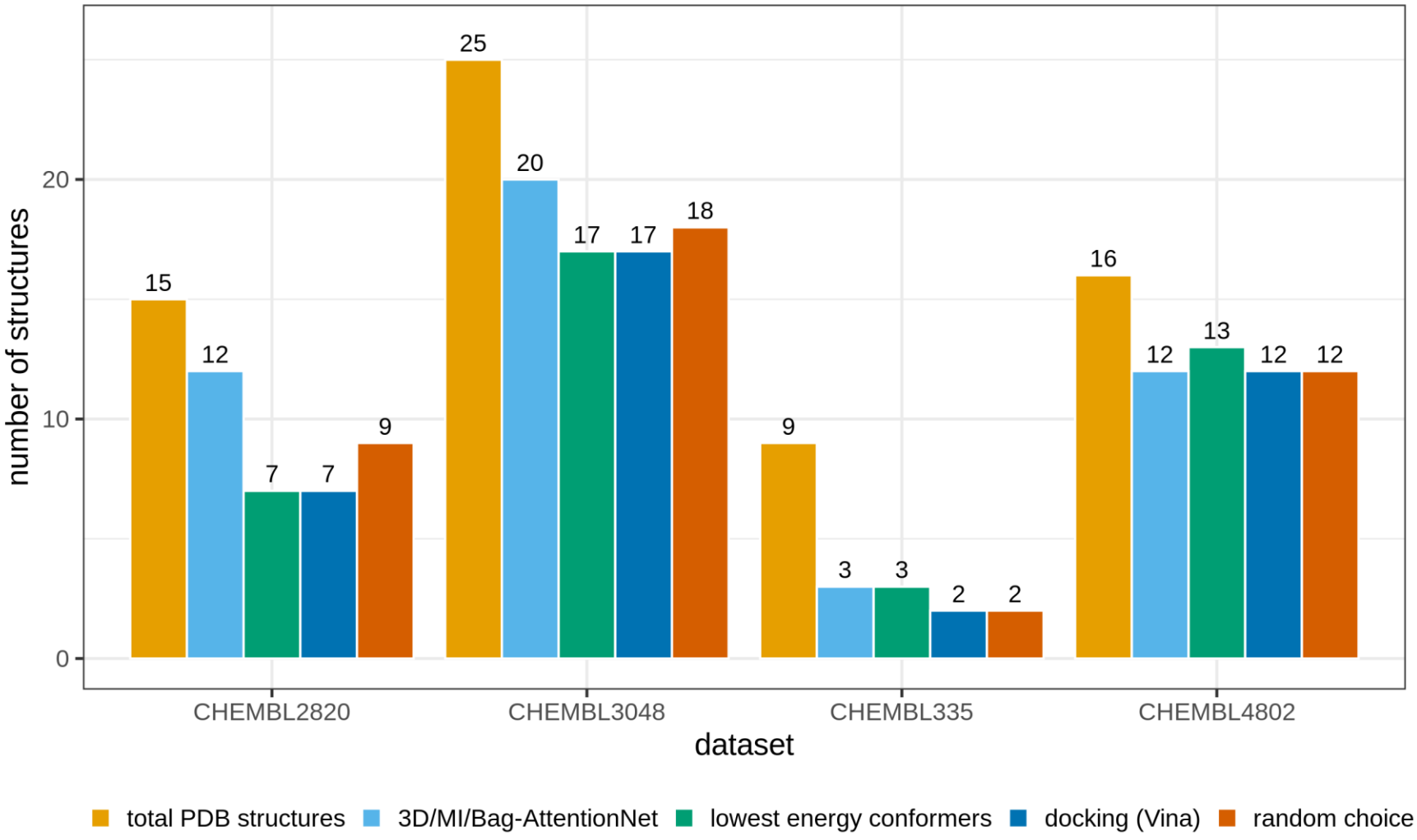
Zankov, D. V. et al, QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923

MIL study: identification of “bioactive” conformers



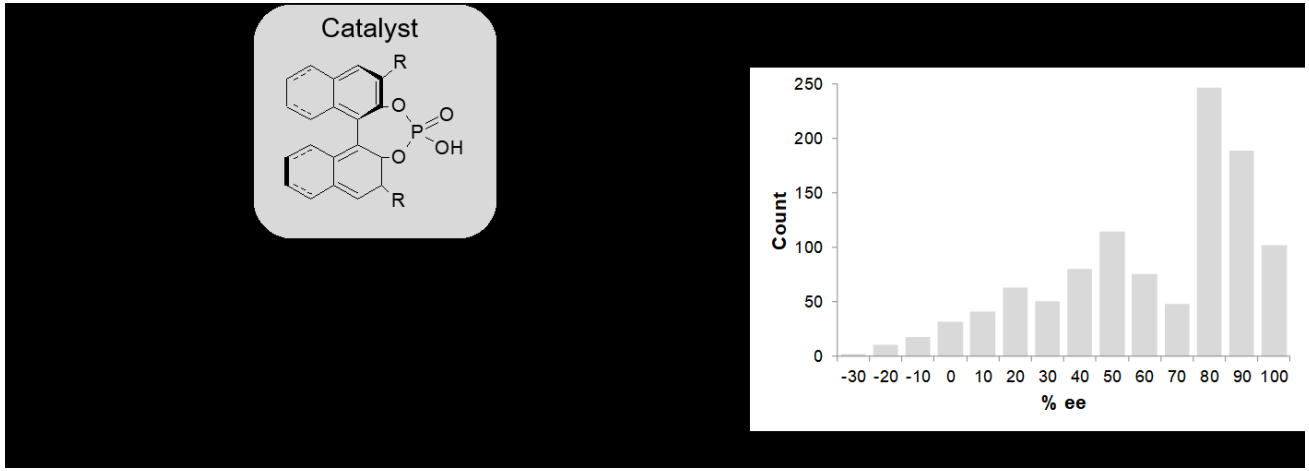
MIL study: identification of “bioactive” conformers

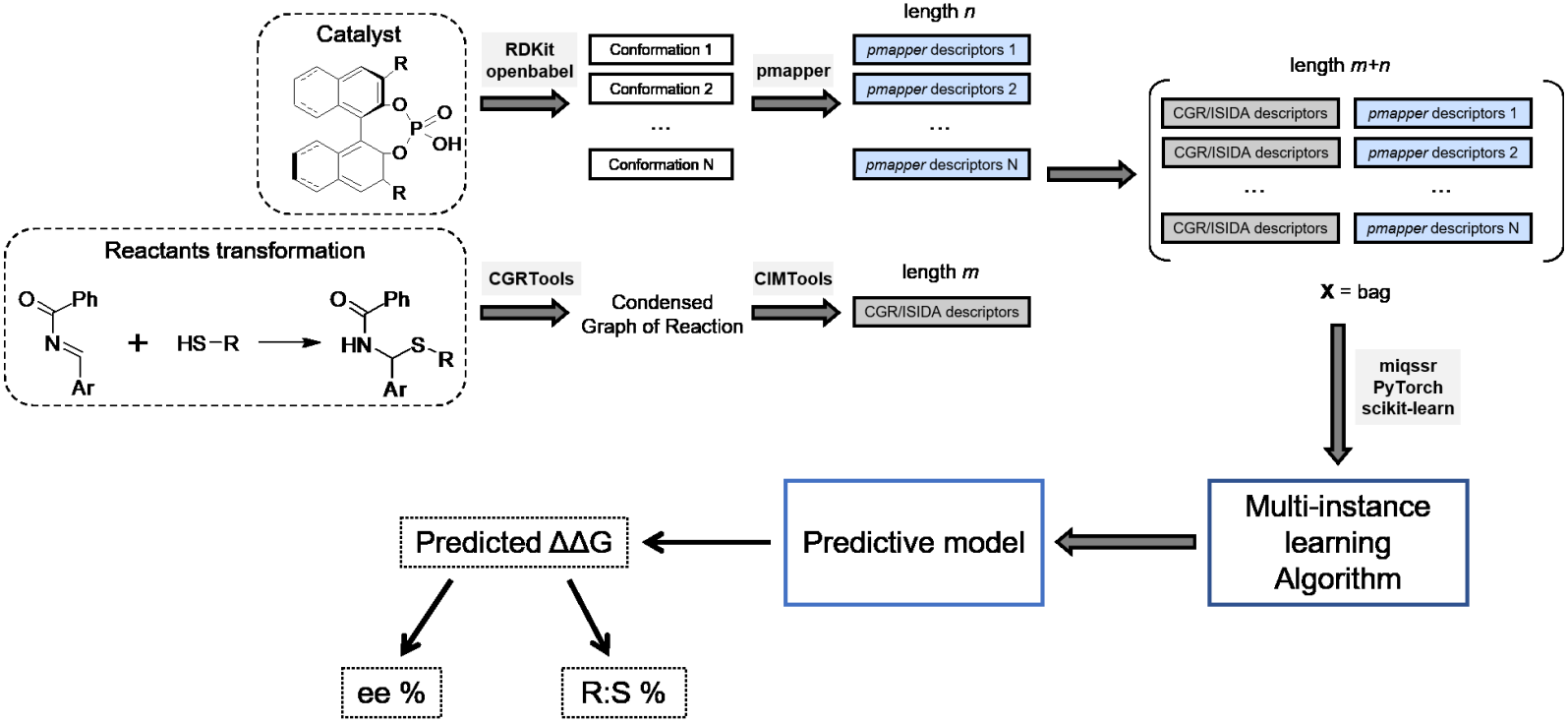
Selected compounds had average RMSD of generated conformers > 2A relative to PDB structure



Zankov, D. V. et al, QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *Journal of Chemical Information and Modeling* **2021**, 61, 4913-4923

Enantioselectivity prediction



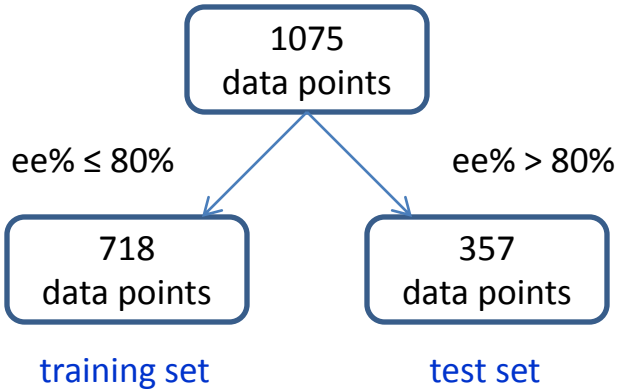


Enantioselectivity prediction

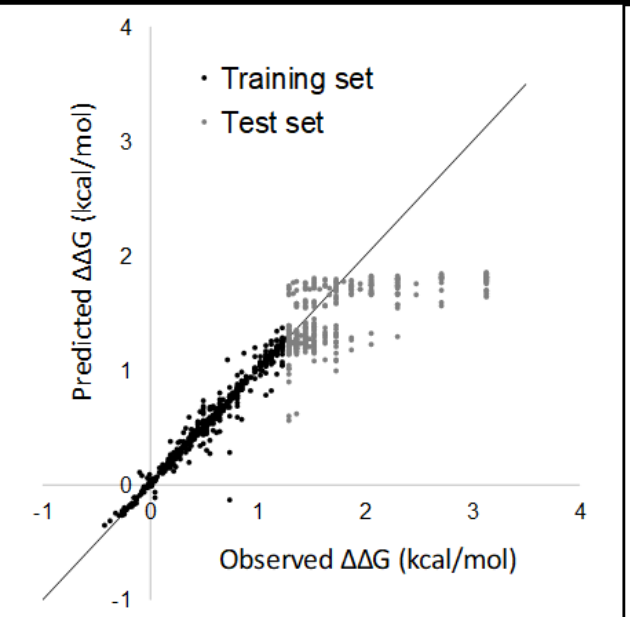
Model	MAE of ΔΔG predictions (kcal/mol)		
	Reaction-out set	Catalyst-out set	Both-out set
2D model	0.16	0.30	0.36
3D Single-conformation model	0.15	0.41	0.44
3D Multi-conformation model	0.13	0.22	0.26
Zahrt's model	0.16	0.21	0.24

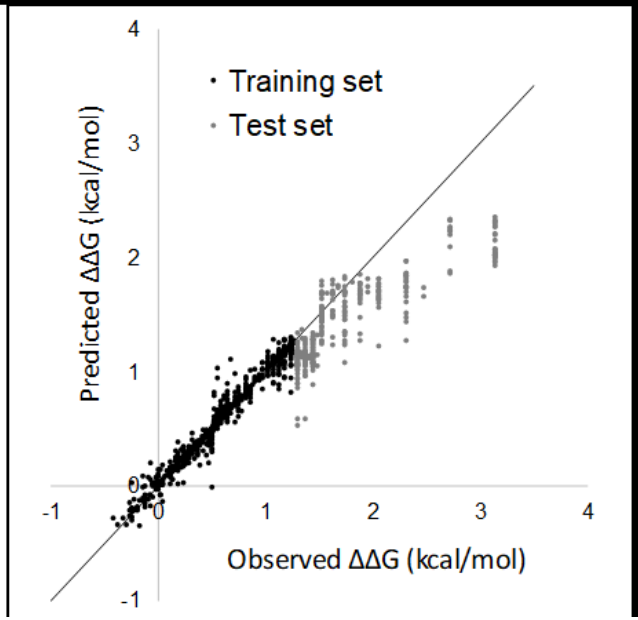
Zahrt, A.F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, 363, eaau5631

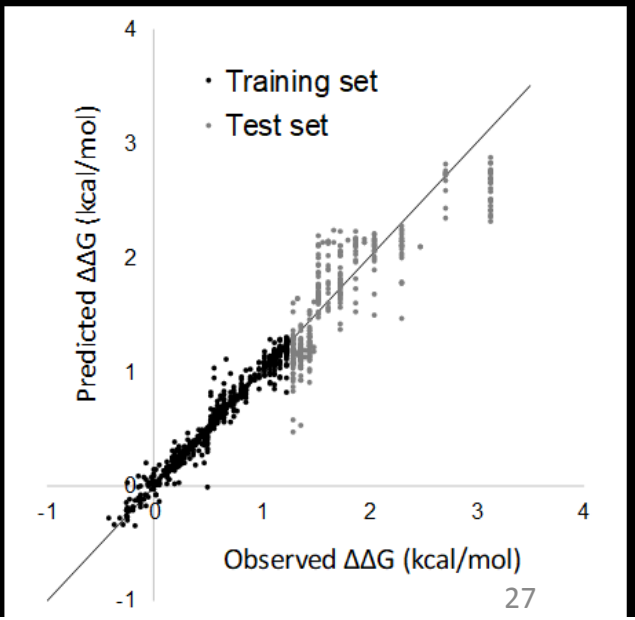
Enantioselectivity prediction: extrapolation scenario



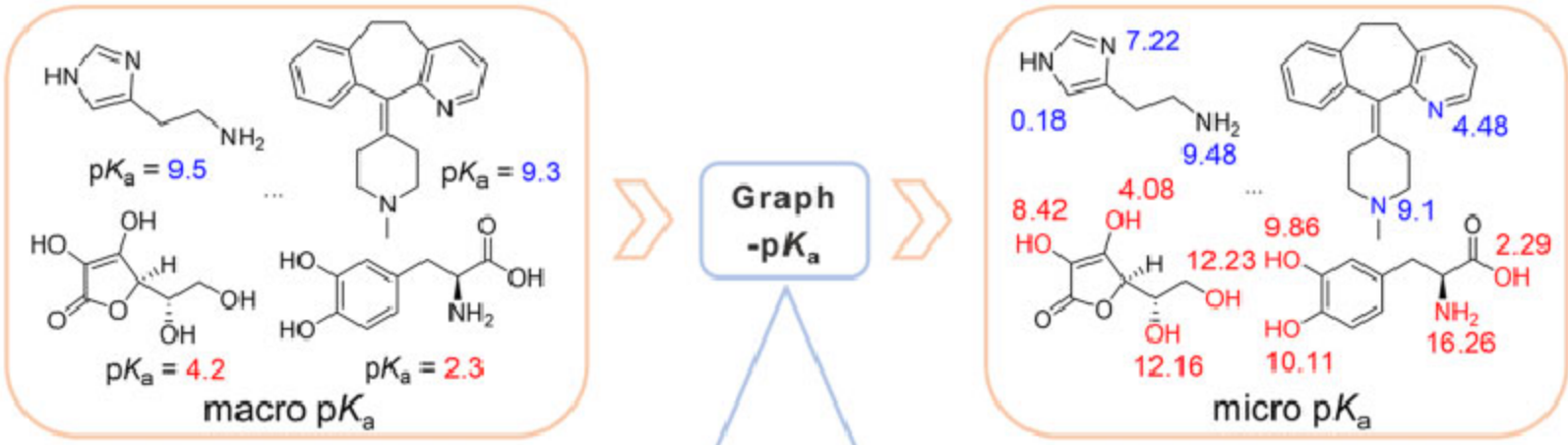
Model	R^2_{Test}	MAE_{Test}
2D model	0.10	0.34
3D Single-conformation model	0.64	0.24
3D Multi-conformation model	0.68	0.21
Zahrt's model	-	0.33

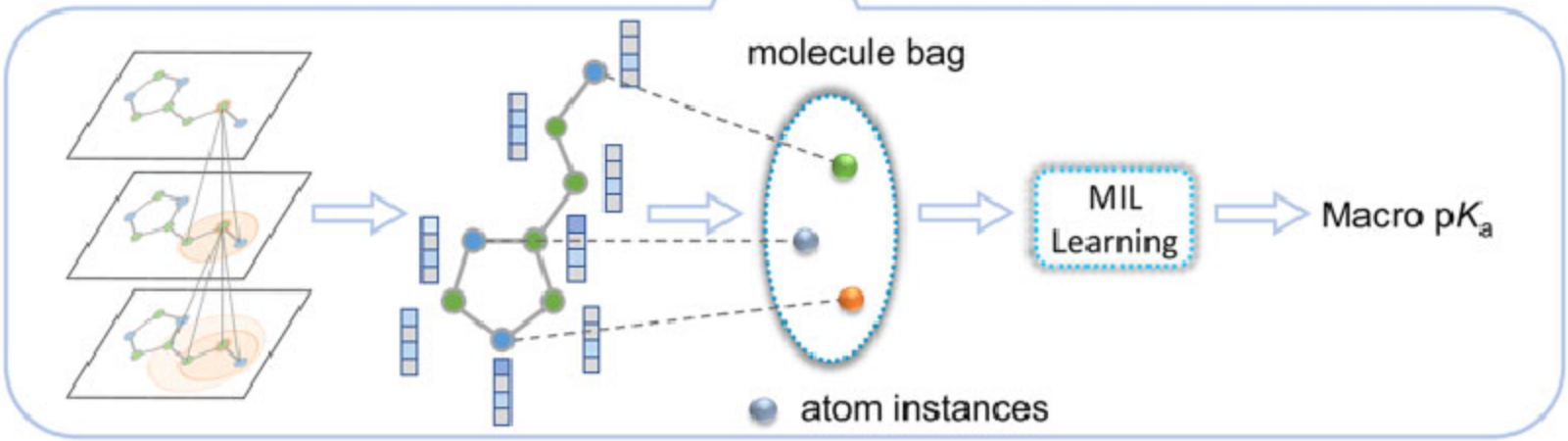






Atoms as instances

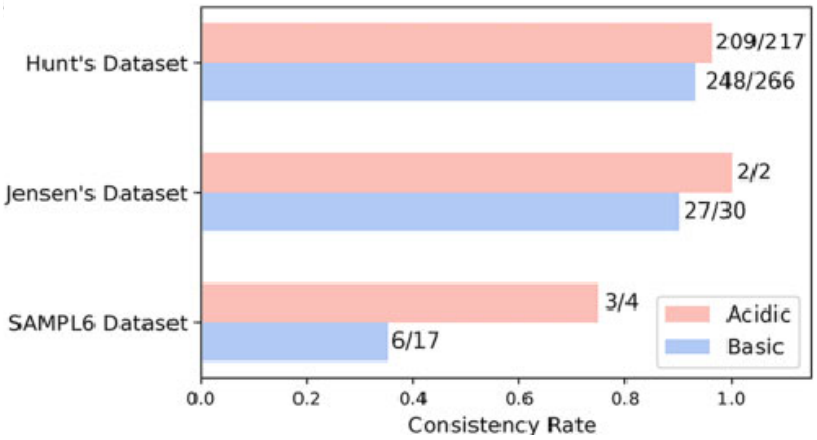


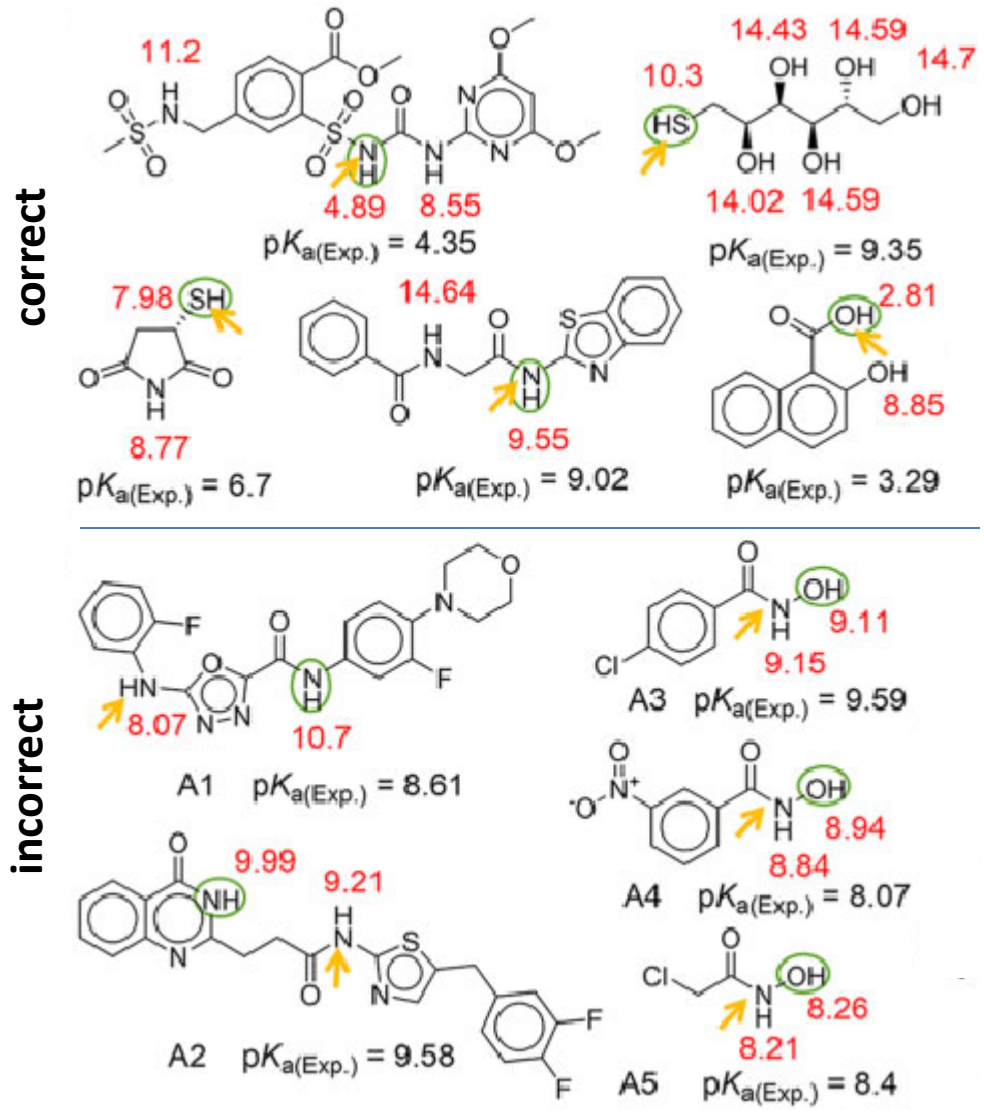


$$pK_{a(acidic)} = -\log\left(\sum_{i=1}^{N} 10^{-pK_{a(acidic)}^{i}}\right)$$

Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M., Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics* **2021**, 38, 792-798.

Atoms as instances

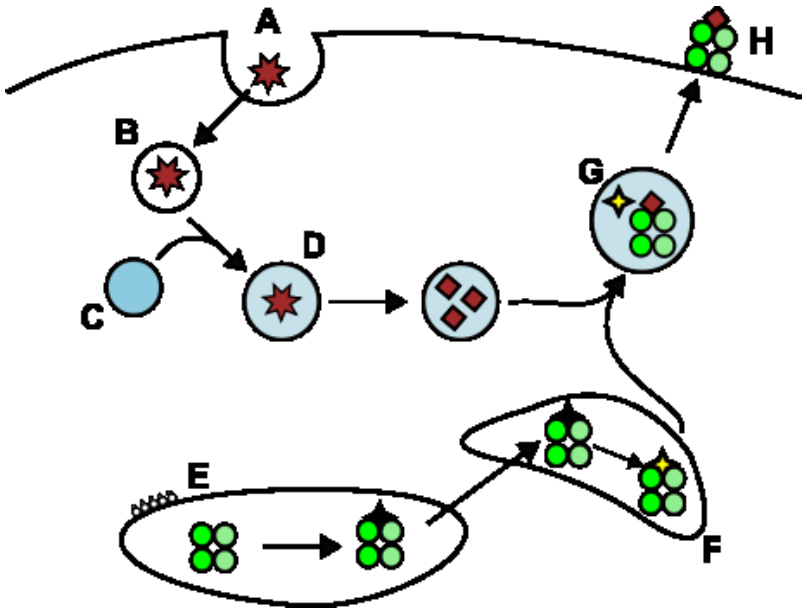




Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M., Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics* **2021**, 38, 792-798.

Peptide sequences as instances

major histocompatibility complex
class II molecules (MHC II)



Key segment

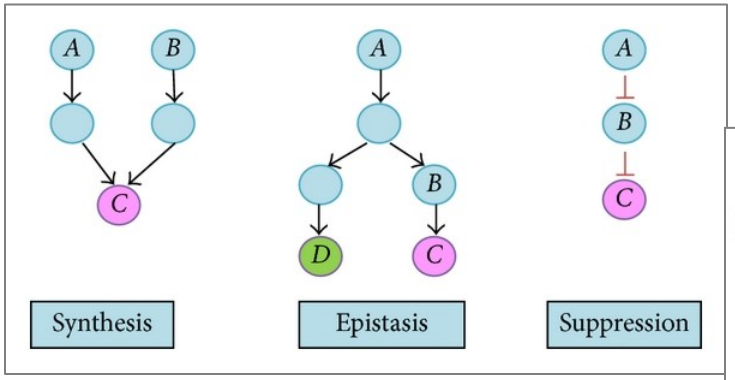
Protein sequence HNNNE**KNISYDKN**LVQKGND ...

9-mer sliding window

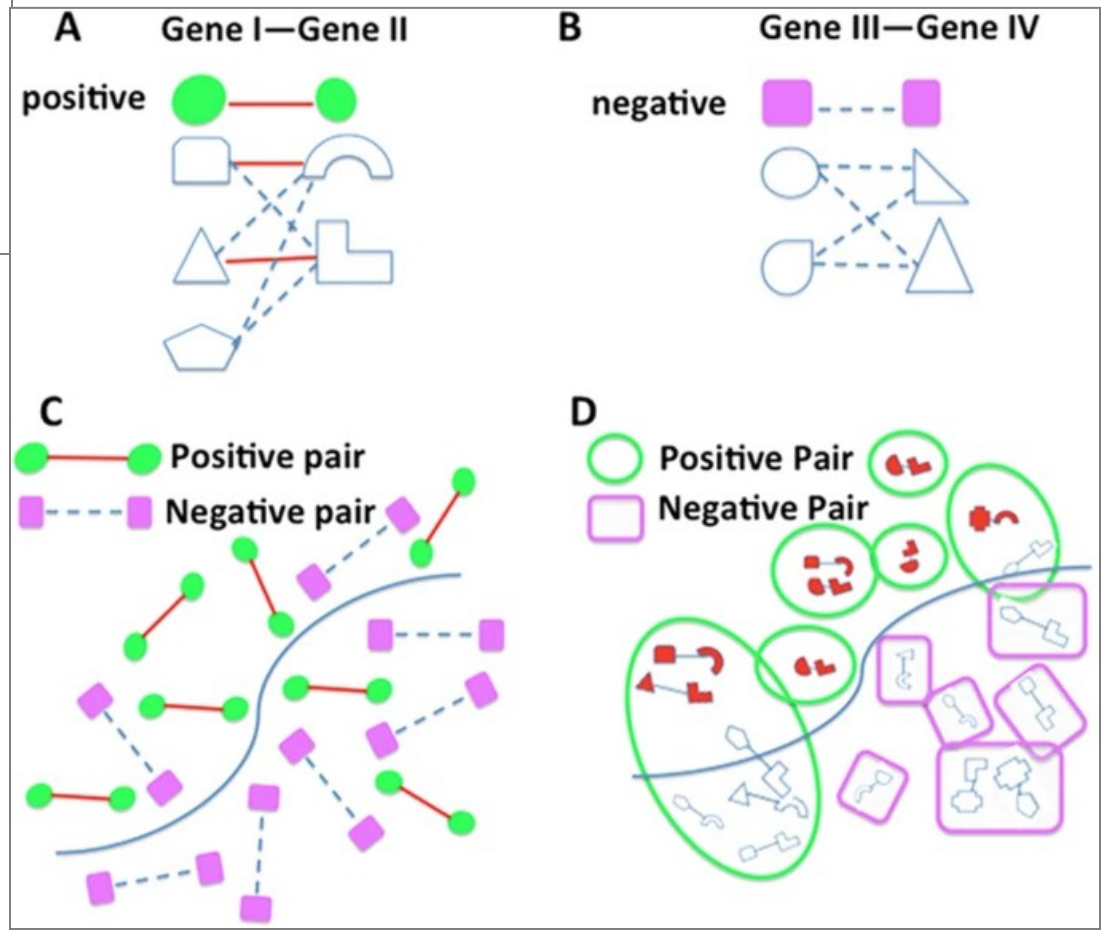
HNNNEKNI
NNNEKNISY
NEKNISYDK
EKNISYDKN
KNISYDKNL
NISYDKNLV
ISYDKNLVQ
SYDKNLVQK
YDKNLVQKG
DKNLVQKGN
KNLVQKGND
...

Bag of protein subsequences

Isoforms as instances



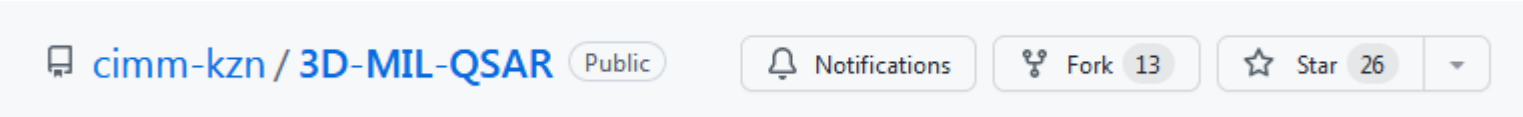
Gene-gene interaction



Li, HD., Menon, R., Eksi, R. et al. A Network of Splice Isoforms for the Mouse. Sci Rep 6, 24507 (**2016**). https://doi.org/10.1038/srep24507

Take-home message

- Multi-instance models outperform both single-instance 3D models and traditional QSAR models built on 2D descriptors in many cases
- Multi-instance models is good alternative to 2D modeling if the latter fails
- Multi-instance neural network with an attention mechanism can correctly identify a “bioactive” conformation close to the experimental structure of a ligand retrieved from PDB
- Atoms, tautomers, protomers, stereoisomers, etc can be considered as instances



https://github.com/cimm-kzn/3D-MIL-QSAR