8th Advanced In Silico Drug Design workshop

27 - 31 January 2025
Olomouc, Czech Republic

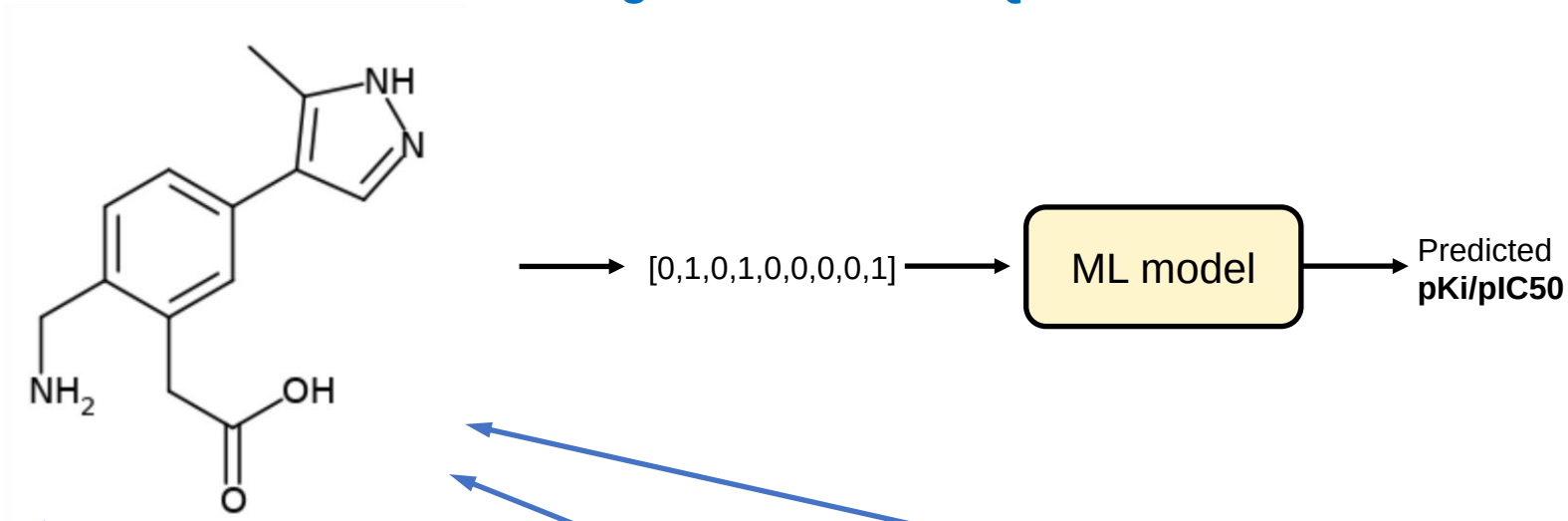
Univerzita Palackého v Olomouci

Multi-instance learning as a response to the complexity of molecular entities

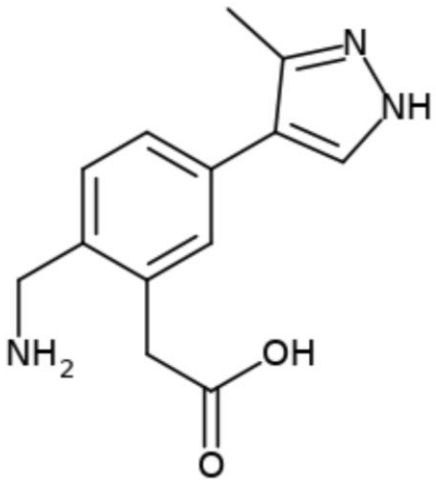
Pavel Polishchuk

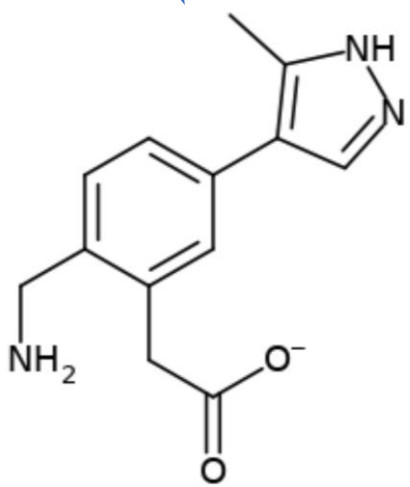
Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University
Czech Republic
pavlo.polishchuk@upol.cz
https://imtm.cz/chemoinformatics-and-drug-design

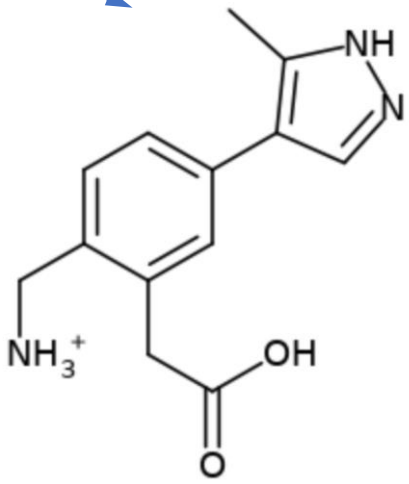
Single-Instance 2D-QSAR

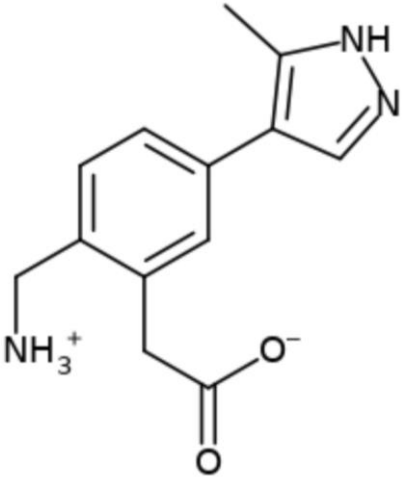


[0,1,0,1,0,0,0,0,1] → ML model → Predicted **pKi/pIC50**



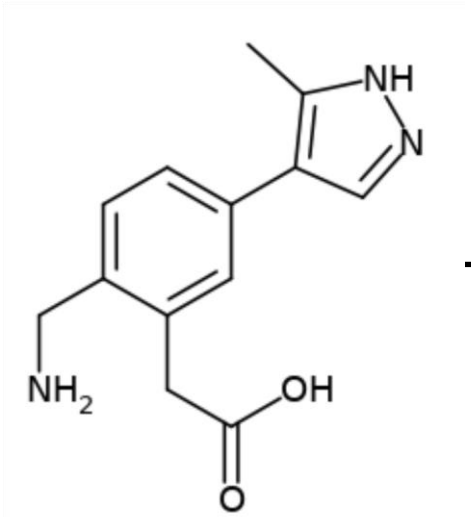


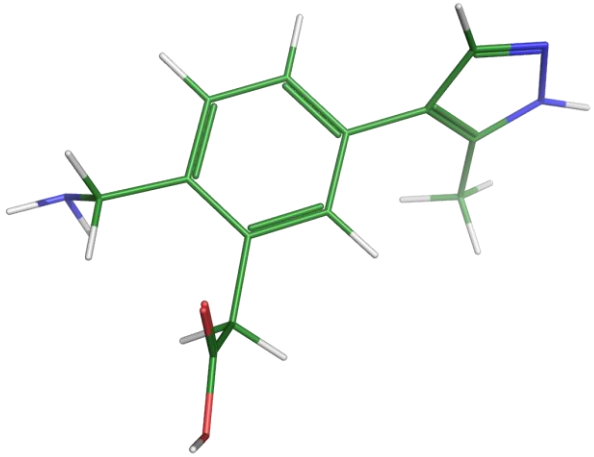




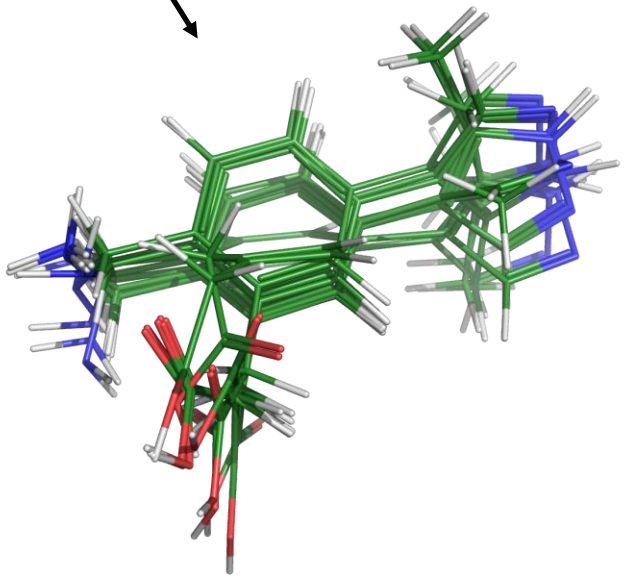
IMTM

Single-Instance 3D-QSAR





[0,1,0,1,0,0,0,0,1] → ML model → Predicted **pKi/pIC50**



[0,0,1,0,0,1,0,0,1]
[0,1,0,1,0,0,0,0,0]
[0,0,0,0,0,0,0,0,1]
[0,1,1,0,0,0,0,1,0]
[1,1,1,0,0,1,0,0,1]

→ ?

4D-QSAR

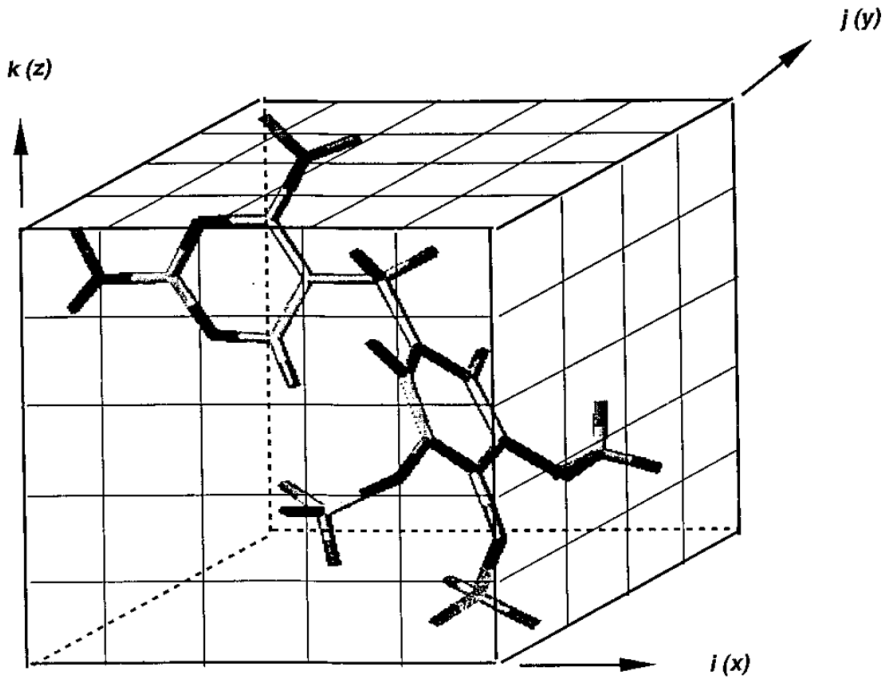


Table 1. The Ten Operational Steps in Performing a (RI) 4D-QSAR Analysis

step no.	description of the step operation
1.	Generate the reference grid and initial 3D models for all compounds in the training set.
2.	Select the trial set of interaction pharmacophores elements, IPEs.
3.	Perform a conformational ensemble sampling of each compound to generate its conformational ensemble profile, CEP.
4.	Select a trial alignment.
5.	Place each conformation of each compound in the reference grid cell space according to the alignment and record the grid cell occupancy profile, GCOP, for each IPE and choice in occupancy measure. The resulting composite set of grid cell properties constitute the set of grid cell occupancy descriptors, GCODs.
6.	Perform a PLS data reduction of the entire set of GCODs against the biological activity measures.
7.	Use the most highly weighted PLS GCODs and any other user-selected descriptors for the initial descriptor basis set in a GA analysis.
8.	Return to Step 4 and repeat Steps 4−7 unless all trial alignments have been included in the analysis.
9.	Select the optimum set of 3D-QSAR models with respect to alignment and any of the methodology parameters.
10.	Adopt the lowest-energy conformer state from the set sampled for each compound, which predicts the maximum activity using the optimum 3D-QSAR model as the “active” conformation (shape).

Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, 119 (43), 10509-10524.

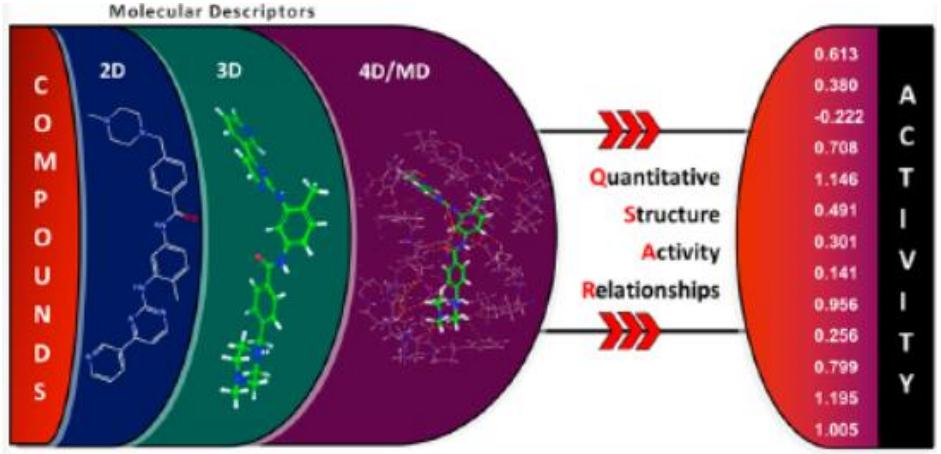
4D-QSAR

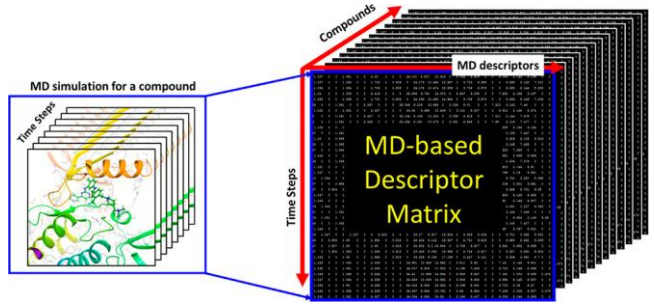
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

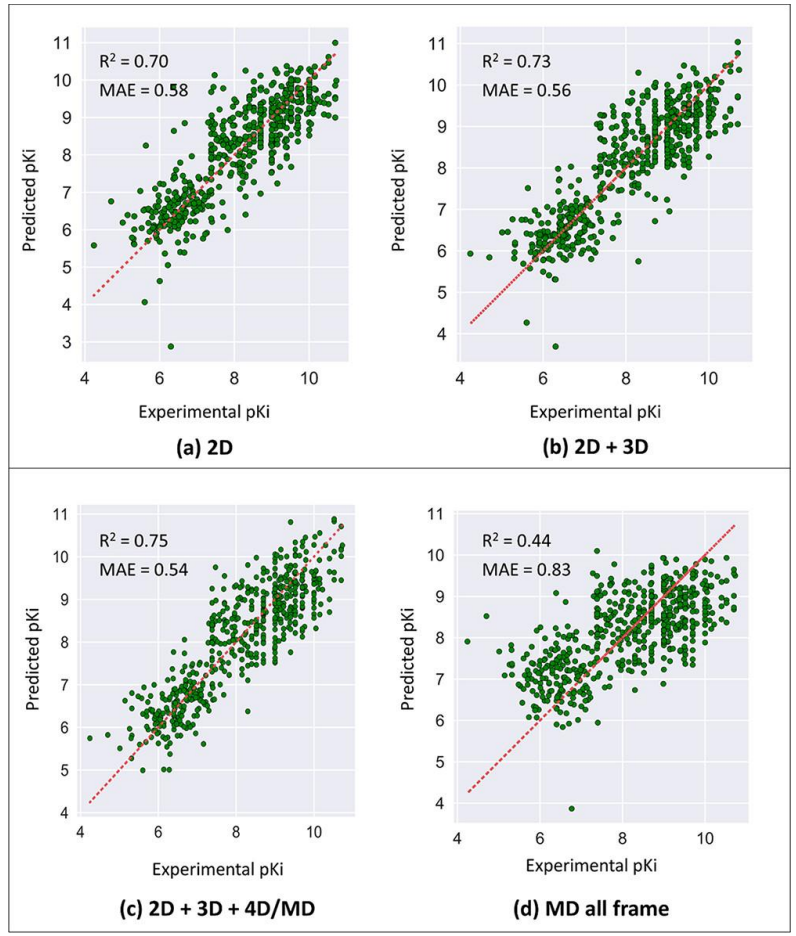
pubs.acs.org/jcim Article

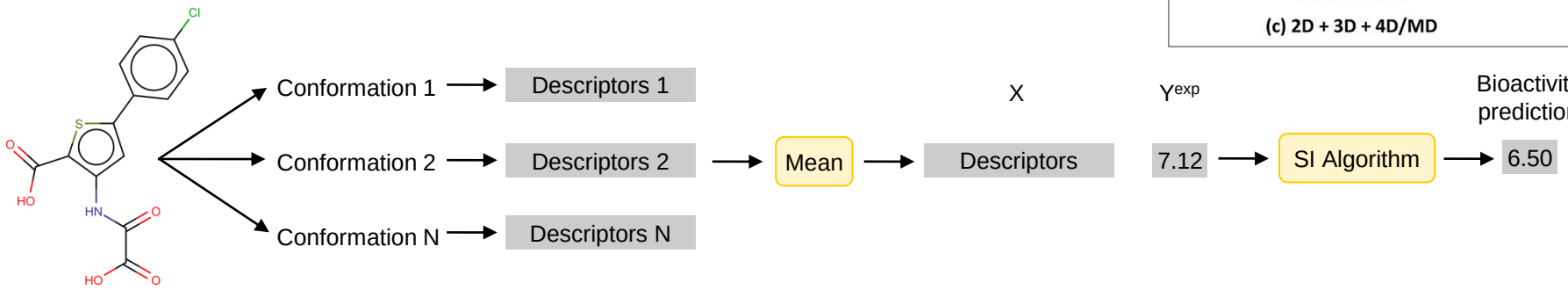
Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict?

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









4D-QSAR

International Journal of *Molecular Sciences*

MDPI

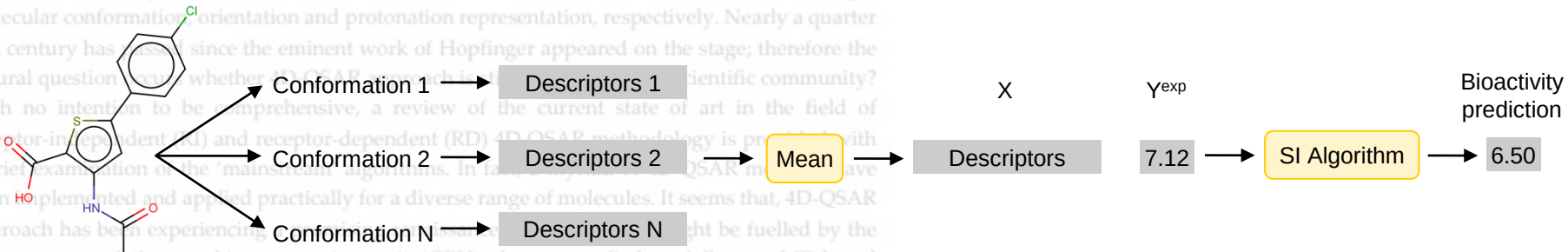
Review

Two Decades of 4D-QSAR: A Dying Art or Staging a Comeback?

Andrzej Bak

Department of Chemistry, University of Silesia, 40007 Katowice, Poland; andrzej.bak@us.edu.pl;
Tel.: +4816-359-11-97

Abstract: A key question confronting computational chemists concerns the preferable ligand geometry that fits complementarily into the receptor pocket. Typically, the postulated 'bioactive' 3D ligand conformation is constructed as a 'sophisticated guess' (unnecessarily geometry-optimized) mirroring the pharmacophore hypothesis—sometimes based on an erroneous prerequisite. Hence, 4D-QSAR scheme and its 'dialects' have been practically implemented as higher level of model abstraction that allows the examination of the multiple molecular conformation, orientation and protonation representation, respectively. Nearly a quarter of a century has [illegible] since the eminent work of Hopfinger appeared on the stage; therefore the natural question occurs whether 4D-QSAR approach is [illegible] scientific community? With no intention to be comprehensive, a review of the current state of art in the field of receptor-independent (RI) and receptor-dependent (RD) 4D-QSAR methodology is presented with a brief examination of the 'mainstream' applications. In fact, 4D-QSAR models have been implemented and applied practically for a diverse range of molecules. It seems that, 4D-QSAR approach has been experiencing [illegible] might be fuelled by the rising power of the graphics processing unit (GPU) clusters applied to full-atom MD-based simulations of the protein-ligand complexes.



Citation: Bak, A. Two Decades of 4D-QSAR: A Dying Art or Staging a Comeback? *Int. J. Mol. Sci.* **2021**, *22*, 5212. https://doi.org/10.3390/ijms22105212

Keywords: 4D-QSAR; structure-based SAR; receptor-dependent models; 4D-derived descriptors

IMTM

Multi-instance problem

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]

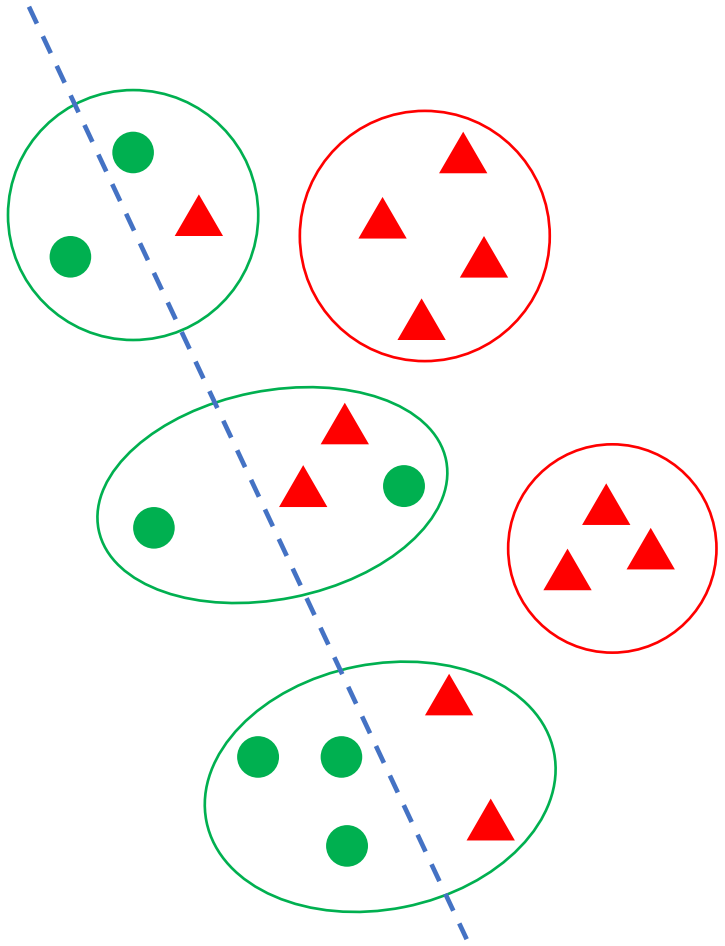
[a] *Department of Computer Science, Oregon State University, Dearborn Hall 303, Corvallis, OR 97331-3202, USA*

[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



Axis-parallel hyperrectangles (APR)

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]

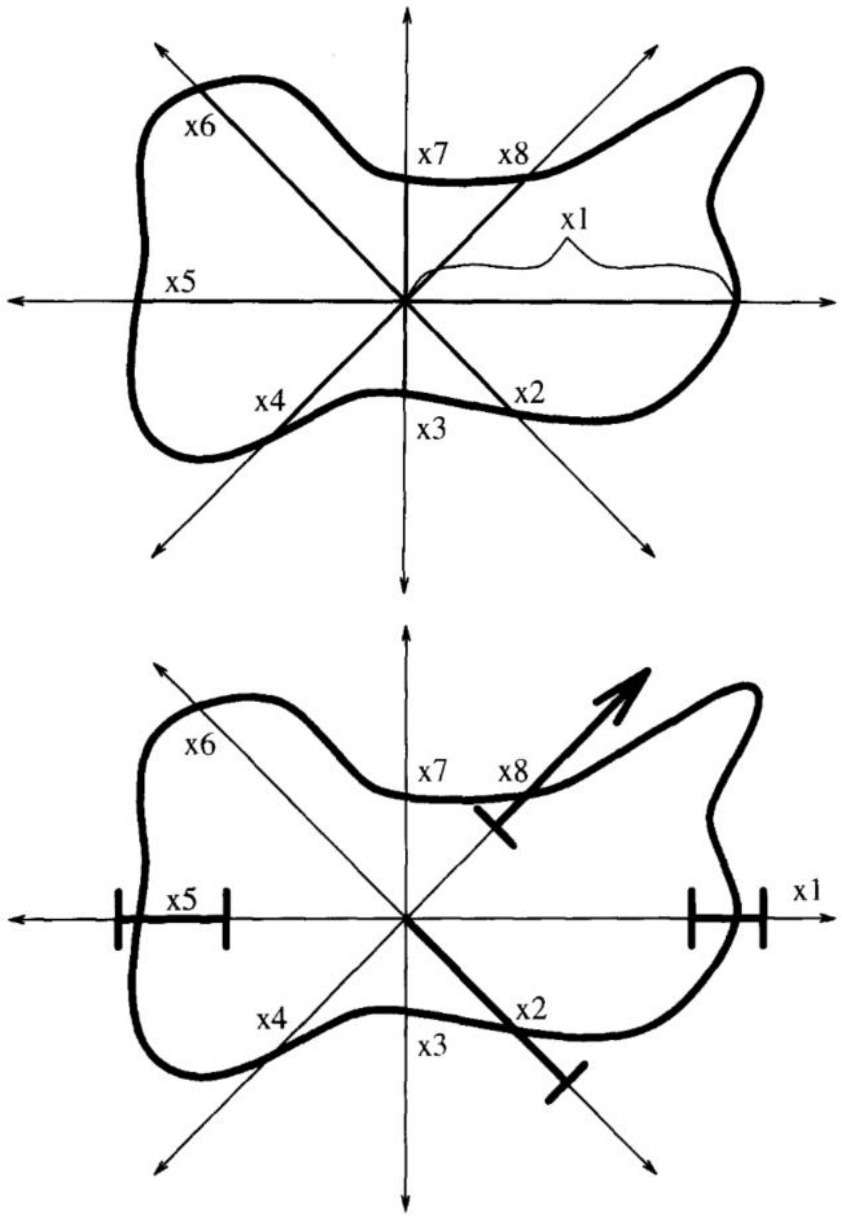
[a] *Department of Computer Science, Oregon State University, Dearborn Hall 303, Corvallis, OR 97331-3202, USA*

[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



Dietterich, T. G.; Lathrop, R. H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* **1997**, 89 (1), 31-71.

IMTM

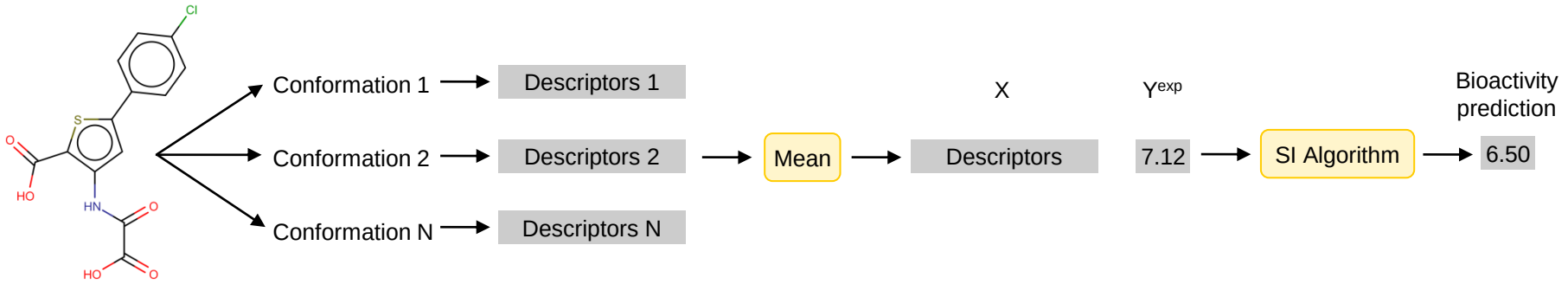
MIL papers in chemoinformatics

Paper	Year	Task	Datasets
Jain et al	1994	Bioactivity of molecules (musk strength)	MUSK1 (102 molecules)
Dietterich et al	1997	Bioactivity of molecules (musk strength)	MUSK1 (92 molecules) and MUSK2 (102 molecules)
Davis et al	2007	Binding affinity of molecules	Dopamine agonists, thermolysin inhibitors, and thrombin inhibitors
Bergeron et al	2008	Identification of metabolic sites of molecules	227 compounds metabolized by cytochrome CYP3A4
Bergeron et al	2012	Identification of metabolic sites of molecules	10 CYP datasets
Fu et al	2012	Inhibitory activities of molecules	Inhibitors against GSK-3, P-gp, and CBrs receptors
Nikonenko et al	2021	Bioactivity of molecules	162 ChEMBL datasets
Zankov et al	2021	Bioactivity of molecules	175 ChEMBL datasets
Zankov et al	2021	Enantioselectivity of organic catalysts	Phosphoric acid catalysts
Xiong et al	2022	macro- and micro-pK_a of molecules	16595 compounds associated with 17489 pKa values
Zankov et al	2023	Enantioselectivity of organic catalysts	Phosphoric acid catalysts; two datasets on phase-transfer catalysts; disulfonimides
Feeney et al	2023	Mutagenicity	6512 compounds tested in the Ames test

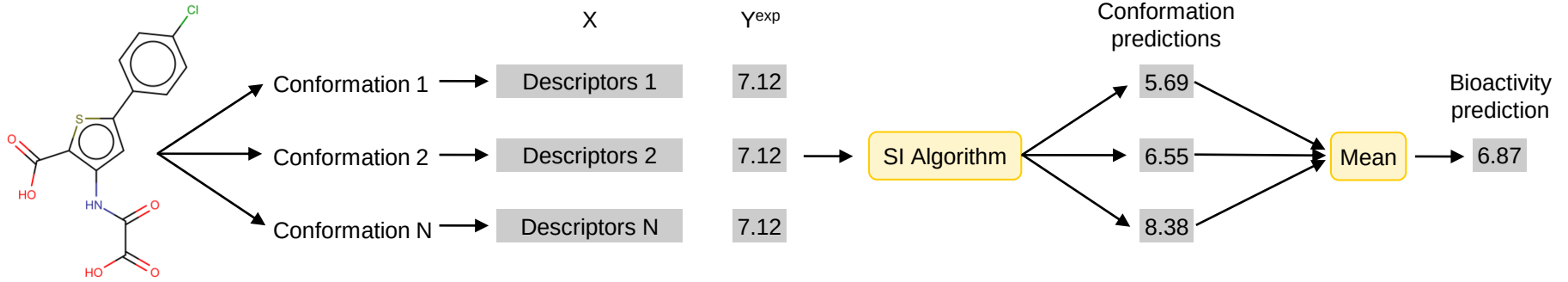
Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. Chemical complexity challenge: Is multi-instance machine learning a solution? *WIREs Computational Molecular Science* **2024**, 14 (1), e1698.

Multi-instance learning approaches

Bag-wrapper: averaging of conformation descriptors

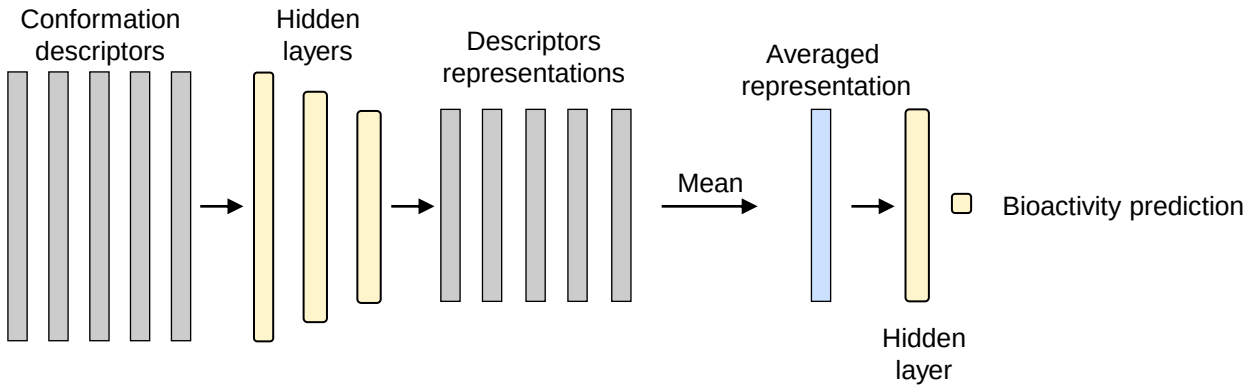


Instance-wrapper: averaging of conformation predictions

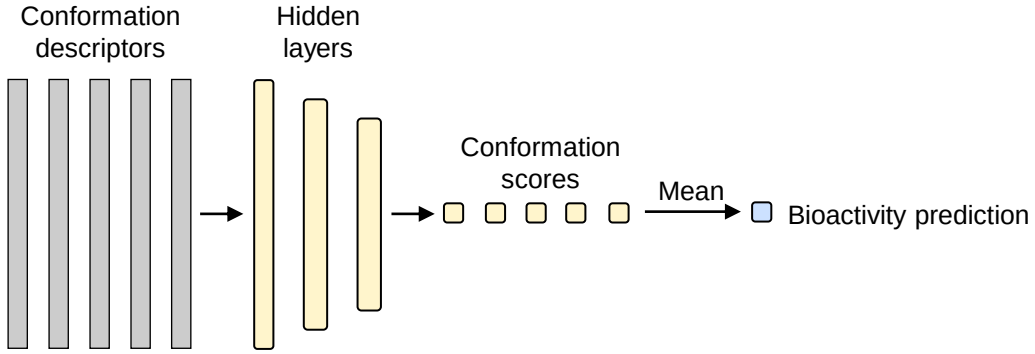


Multi-instance learning approaches

Bag-Net: averaging of conformation embeddings

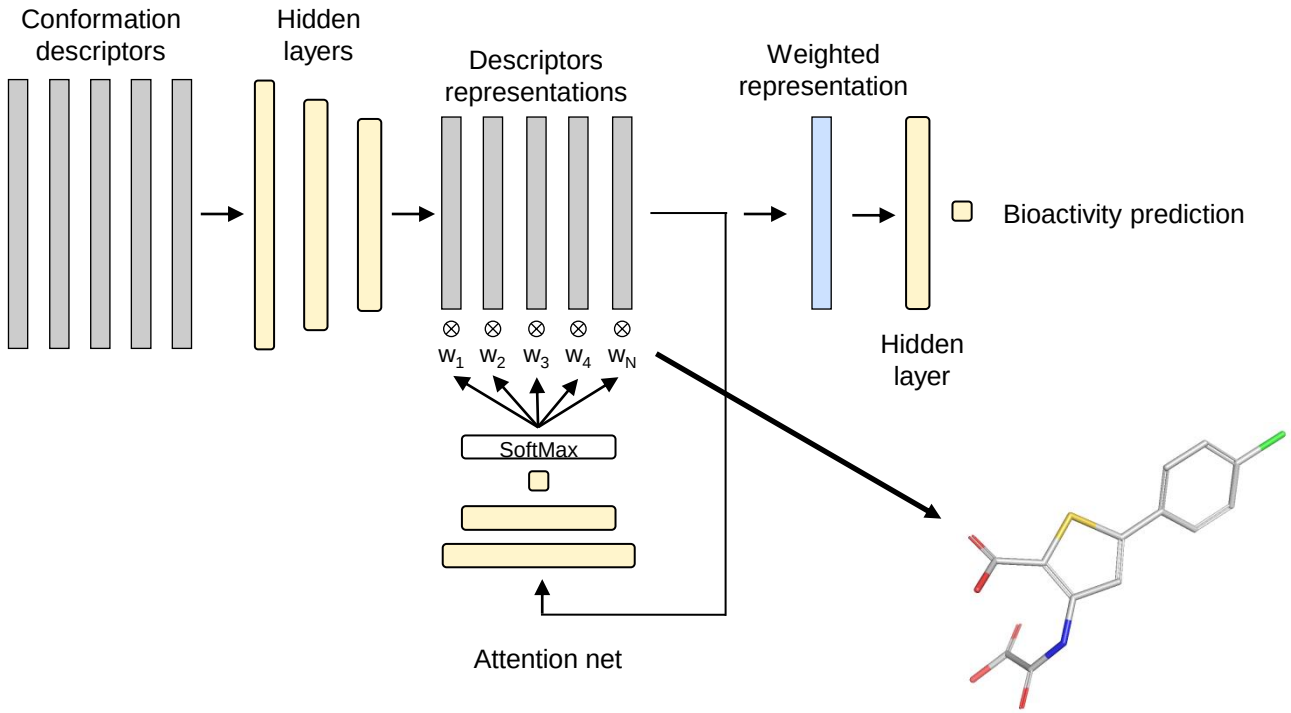


Instance-Net: averaging of conformation scores



Multi-instance learning approaches

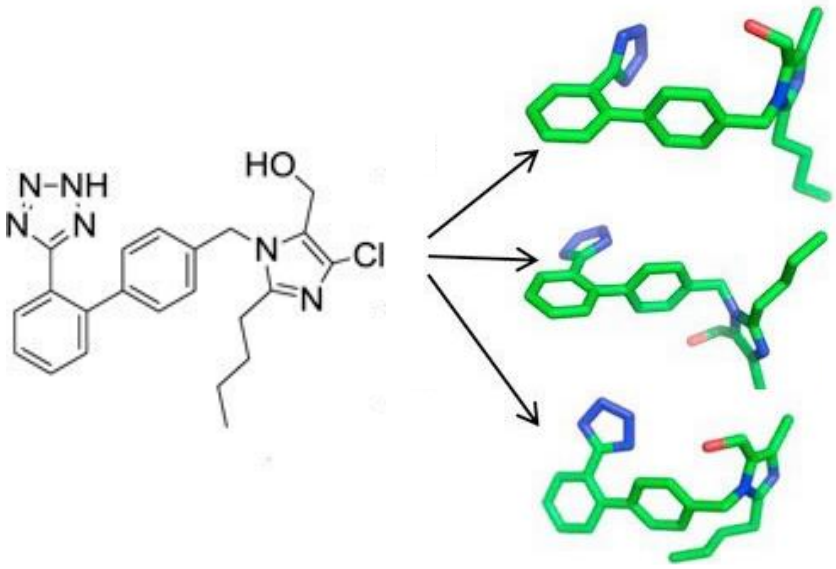
Bag-Attention Net: weighted averaging of conformation embeddings



Multi-instance learning case study

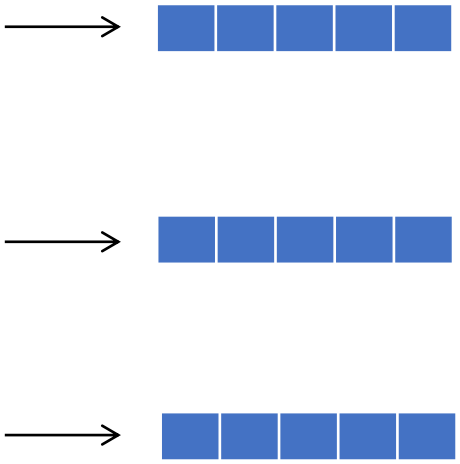
175 data sets from ChEMBL

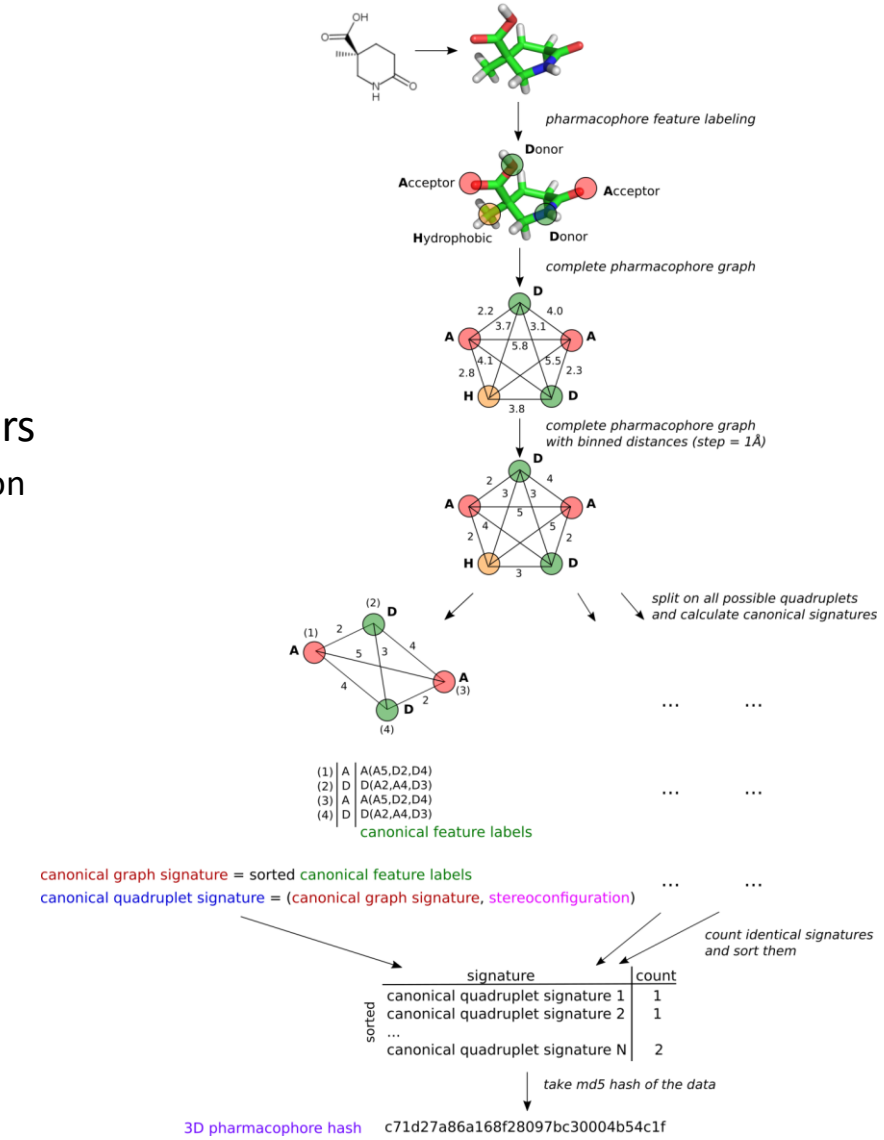
Generation of
up to **100**
conformations per
molecule
RDKit



ΔE = 100 kcal/m

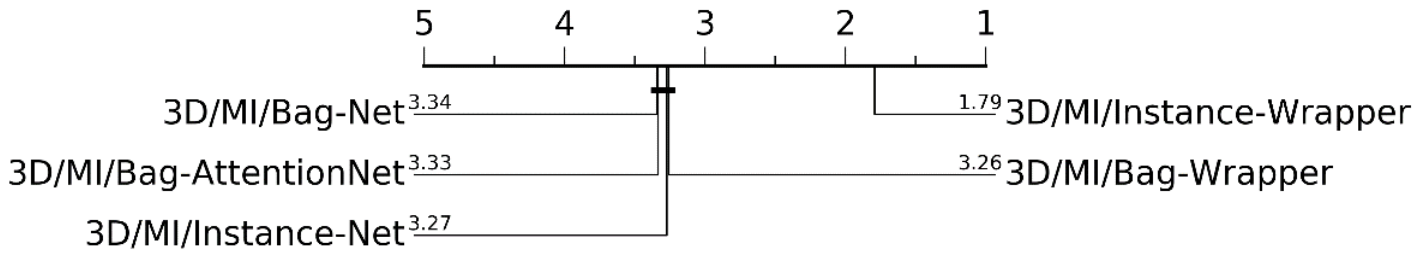
Calculation of
Pmapper 3D descriptors
alignment-free representation





Kutlushina, A.; Khakimova, A.; Madzhidov, T.; Polishchuk, P. Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. *Molecules* **2018**, 23 (12), 3094.

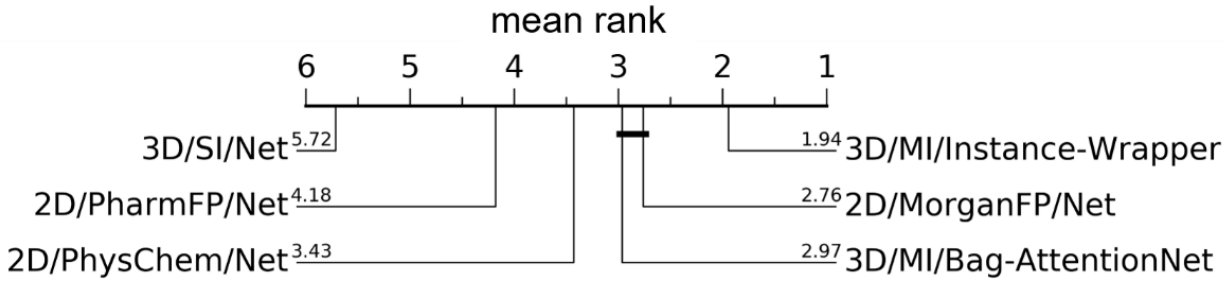
Comparison of 3D MIL approaches



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

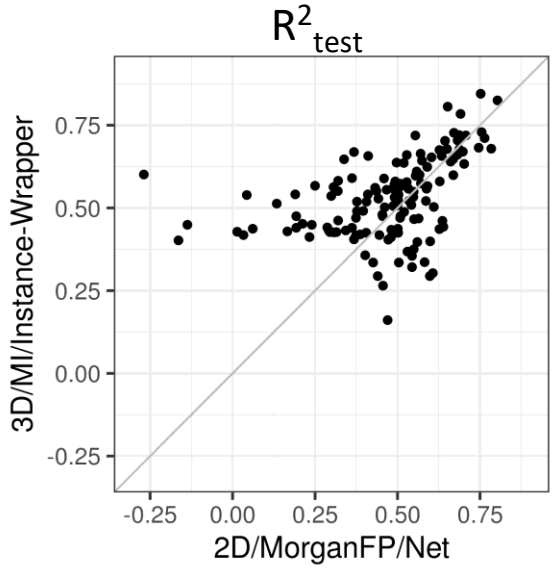
Comparison of 3D MIL, 3D and 2D single instance approaches

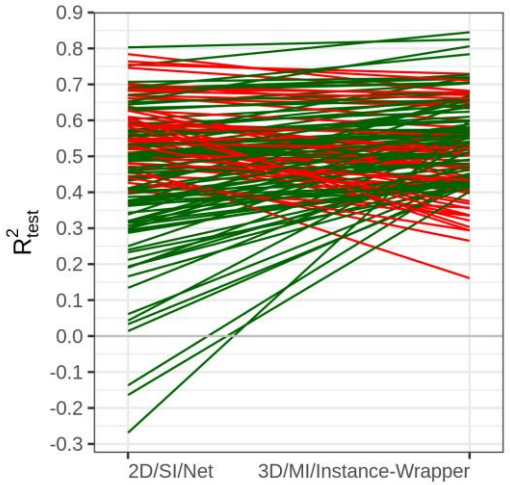
Model	Mean R^2_{test}	Median R^2_{test}	Top-1
3D/MI/Instance-Wrapper	0.530 ± 0.123	0.532	71
2D/MorganFP/Net	0.474 ± 0.189	0.503	41
3D/MI/Bag-Attention Net	0.474 ± 0.157	0.476	14
2D/PhysChem/Net	0.436 ± 0.165	0.443	17
2D/PharmFP/Net	0.357 ± 0.275	0.383	3
3D/SI/Net	0.027 ± 0.374	0.092	0

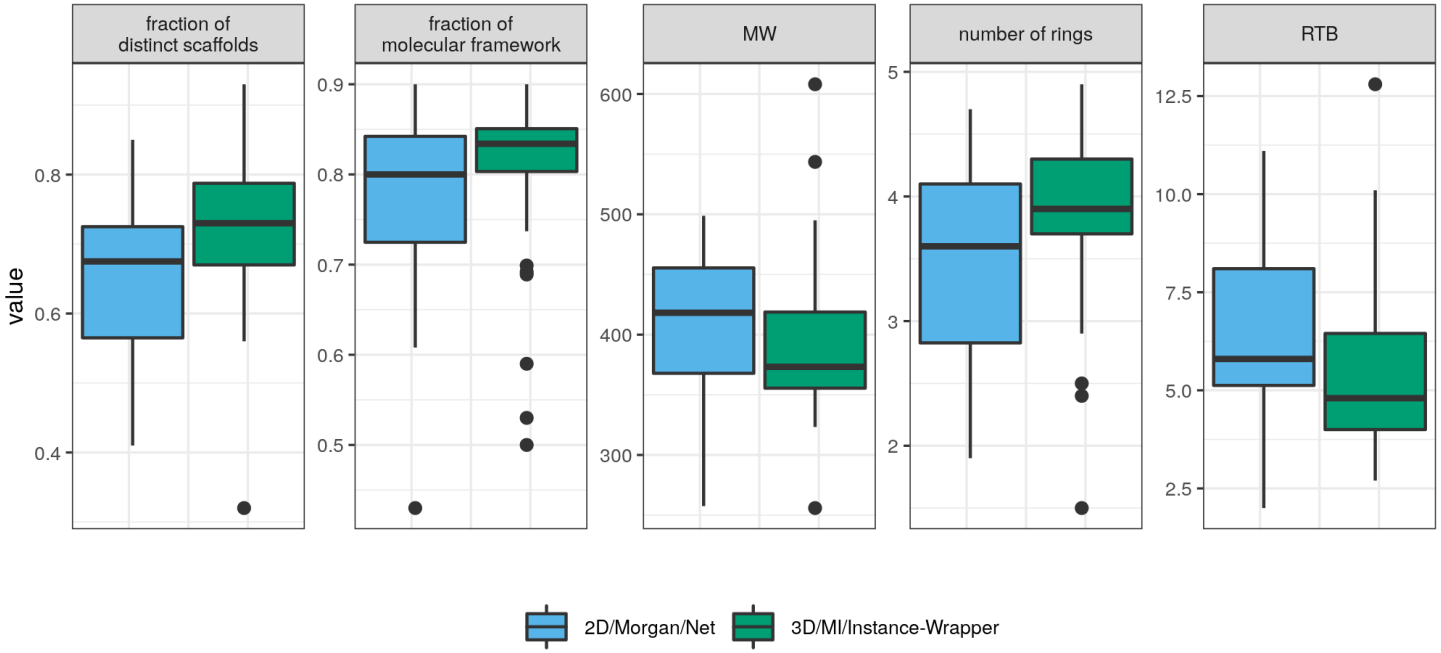


Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I.
QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *J. Chem. Inf. Model.* **2021**, 61 (10), 4913-4923.

3D Instance-wrapper vs 2D Morgan models

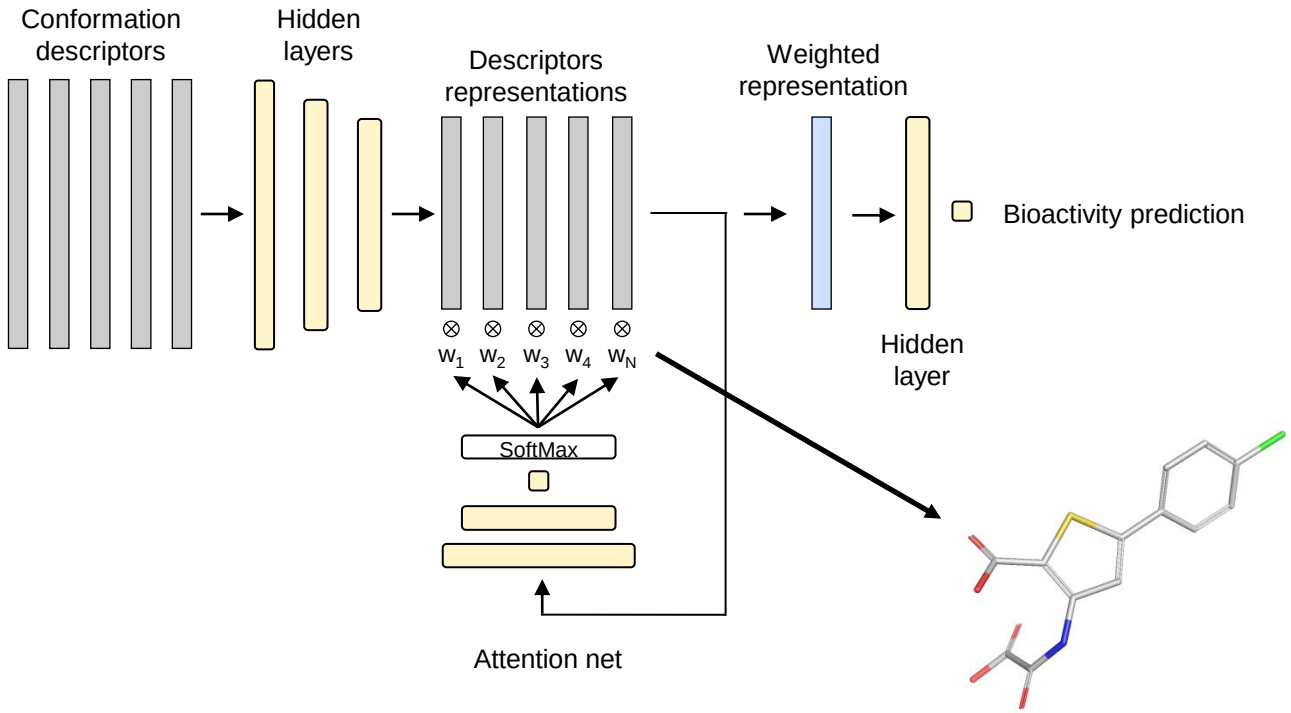






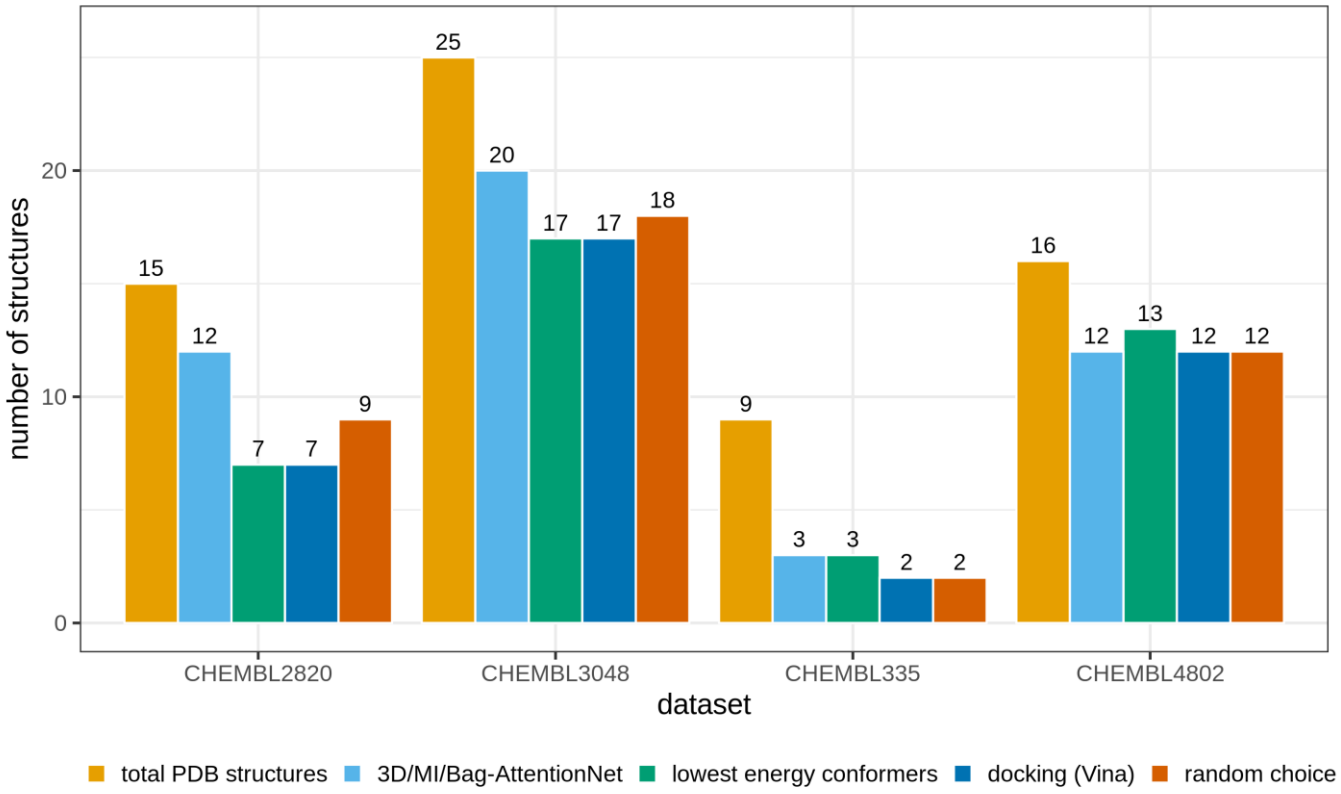
Multi-instance learning approaches

Bag-Attention Net: weighted averaging of conformation embeddings

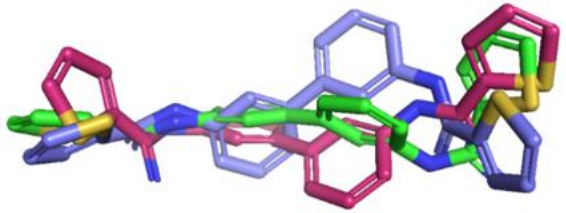


Identification of biologically relevant conformers

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



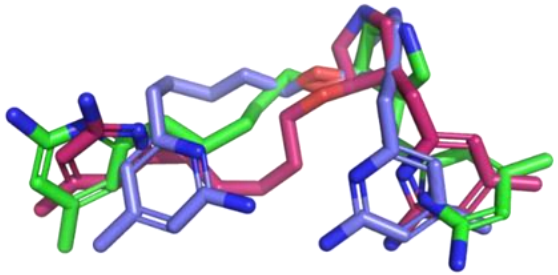
coagulation factor XI (CHEMBL2820)



(a)

Experimental pK_i: 6.10
3D/SI/Net pKi: 6.48, RMSD = 2.42 Å
3D/MI/Bag-AttentionNet pK_i: 6.31
(attention weight: 0.83), RMSD = 1.70 Å

protein-tyrosine phosphatase 1B (CHEMBL335)

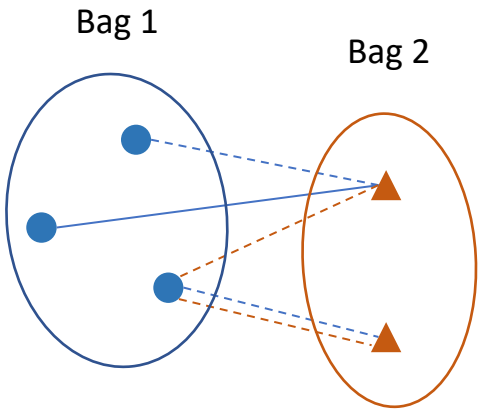


(b)

Experimental pK_i: 7.42
3D/SI/Net pKi: 7.86, RMSD = 2.78 Å
3D/MI/Bag-AttentionNet pK_i: 7.41
(attention weight: 0.59), RMSD = 1.55 Å

Conventional MIL approaches

Citation-kNN

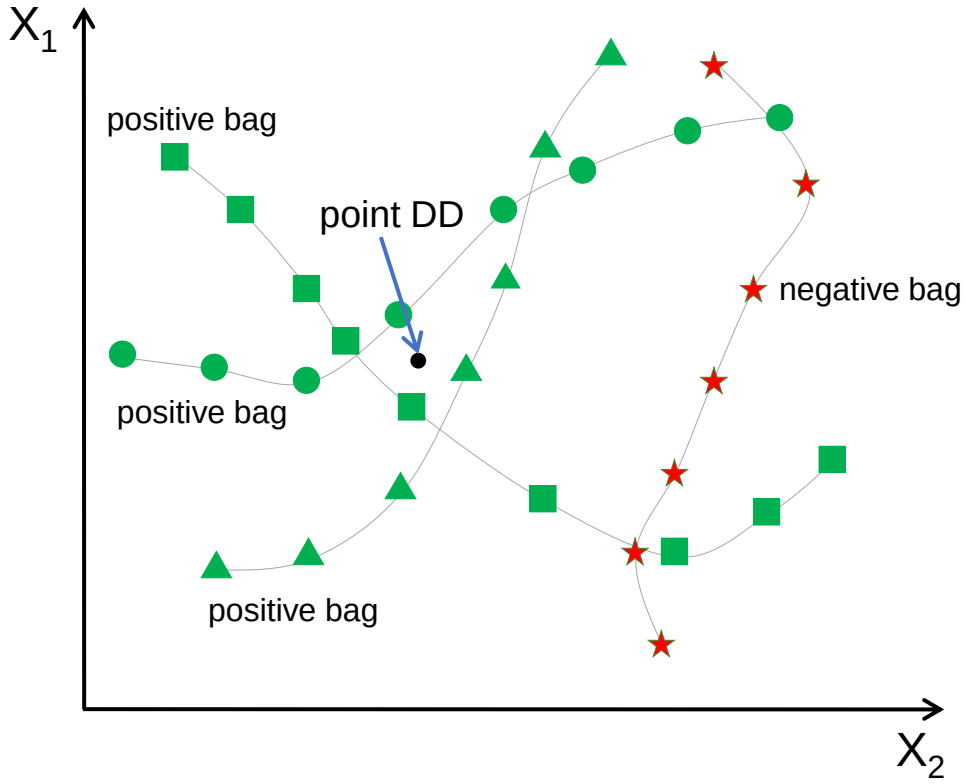


Hausdorff distance - the greatest of all the distances from a point in one set to the closest point in the other set

Wang J, Zucker JD. Solving multiple-instance problem: a lazy learning approach. In: Proceedings 17th IEEE international conference on machine learning. **2000**:1119–1125.

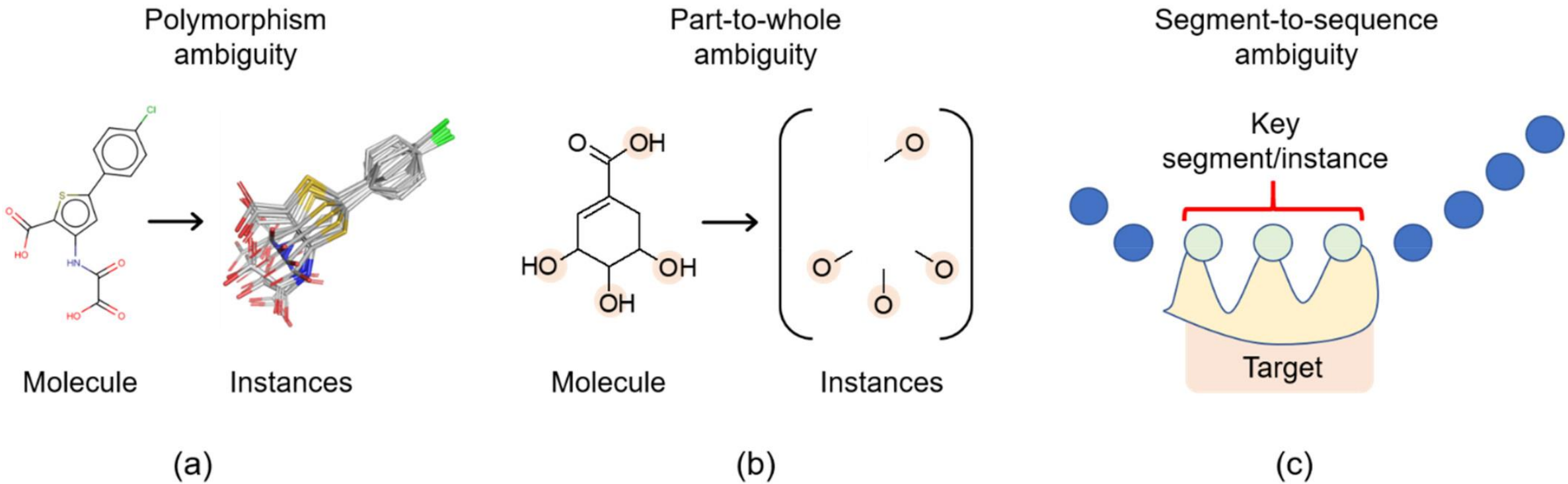
Diversity density (DD)

searching for points where many instances of positive bags are close and instances of negative bags are far

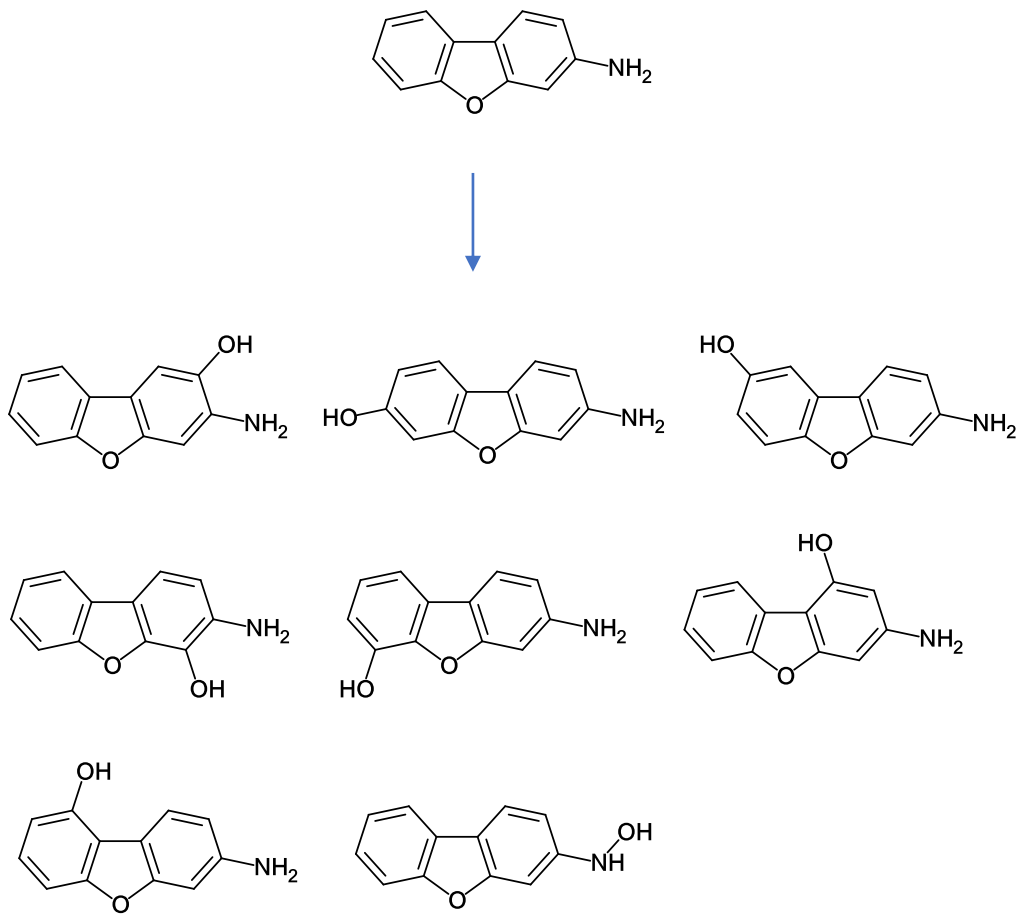


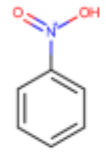
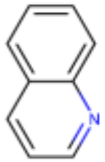
Maron O, Lozano-Pérez T. A framework for multiple-instance learning. In: Advances in neural information processing systems. **1998**: 570–576.

Types of ambiguity in tasks related to modelling molecular properties and functions



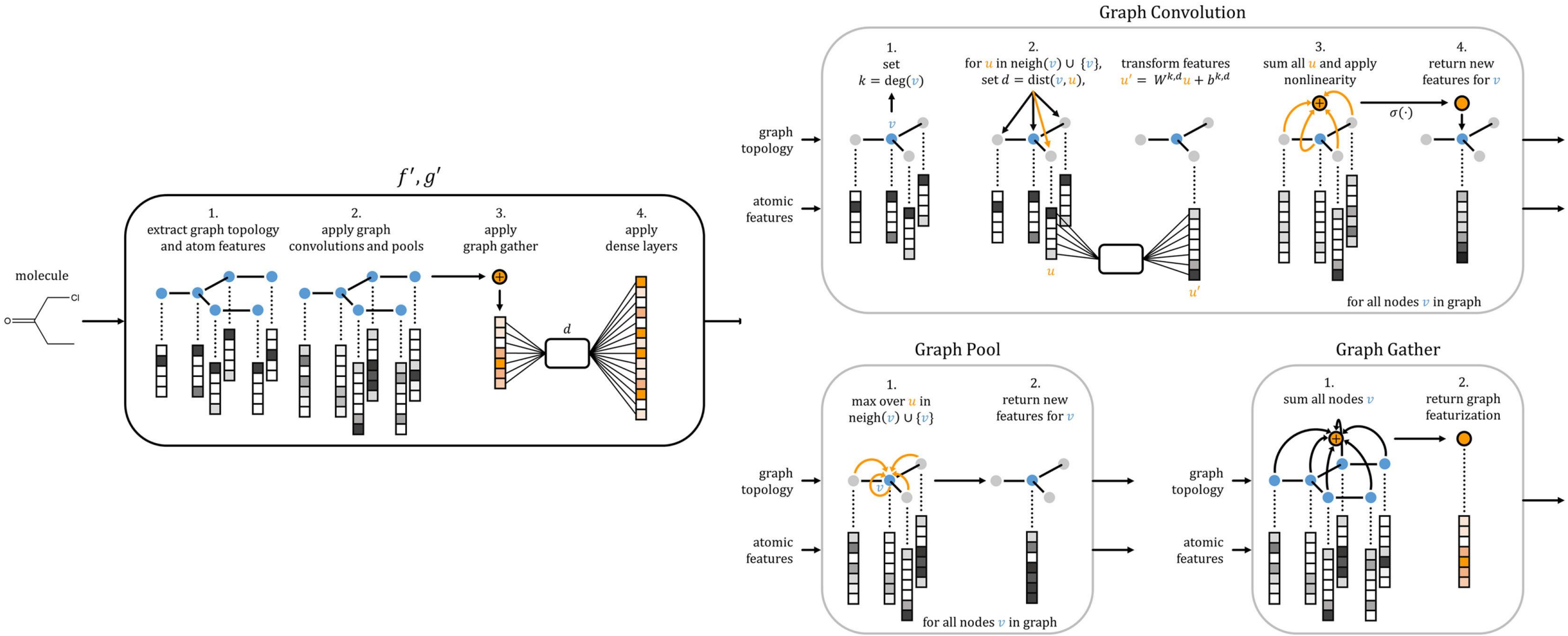
Metabolites as instances



Structural class	Count (%)	Ames + (%)	Excl. (%)	Model	Sens.	Spec.	Acc.	Bal. Acc.
Nitro aromatic compounds 	771 / 6505 (11.8%)	631 / 771 (81.8%)	0	Morgan NSK polynomial	0.913	0.507	0.839	0.710
	1018 / 9681 (10.5%)	841 / 1018 (82.6%)	4 / 1018 (0.39%)	DEREK	0.994	0.131	0.845	0.562
			434 / 1018 (42.63%)	LSMA	0.948	0.087	0.762	0.518
			1 / 1018 (0.10%)	MC4PC	0.964	0.379	0.989	0.612
Quinolines 	140 / 6505 (2.15%)	66 / 140 (47.1%)	0	Morgan NSK polynomial	0.833	0.730	0.779	0.782
	174 / 9681 (1.80%)	67/174 (38.5%)	3 / 174 (1.72%)	DEREK	0.828	0.523	0.637	0.676
			54 / 174 (31.03%)	LSMA	0.931	0.297	0.450	0.614
			88 / 174 (50.57%)	MC4PC	0.438	0.759	0.640	0.599
Furan w/o nitro 	38 / 6505 (0.58 %)	9 / 38 (23.7%)	0	Morgan NSK polynomial	0.444	0.690	0.632	0.567
	64 / 9681 (0.66%)	9 / 64 (14.1%)	1 / 64 (1.56%)	DEREK	0.556	0.704	0.683	0.630
			8 / 64 (12.00%)	LSMA	0.857	0.327	0.393	0.592
			28 / 64 (43.75%)	MC4PC	0.167	0.833	0.722	0.500

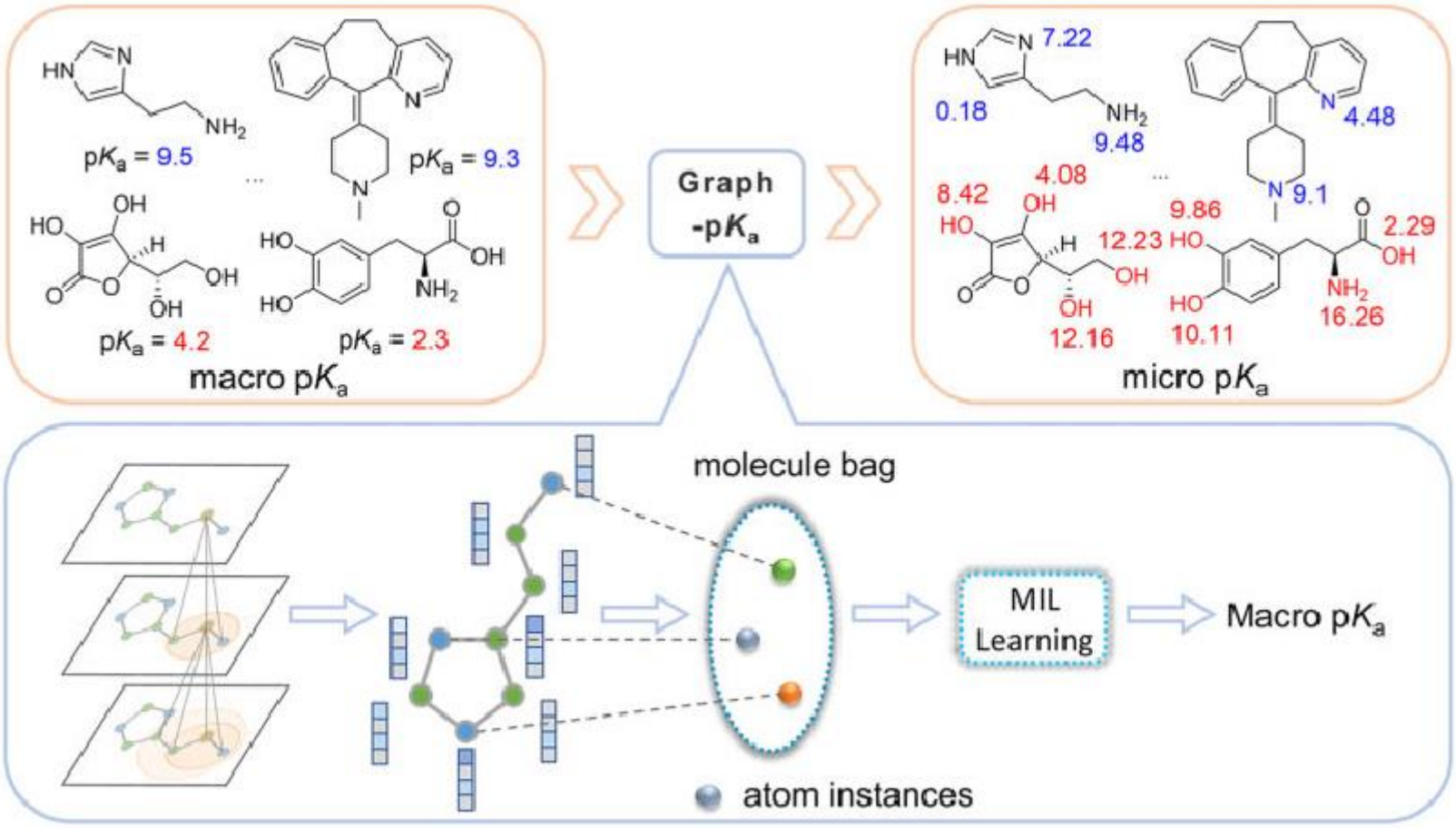
Feeney, S. V.; Lui, R.; Guan, D.; Matthews, S. Multiple Instance Learning Improves Ames Mutagenicity Prediction for Problematic Molecular Species. *Chemical Research in Toxicology* **2023**, 36 (8), 1227-1237.

Atoms as instances



Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, 3 (4), 283-293.

Atoms as instances

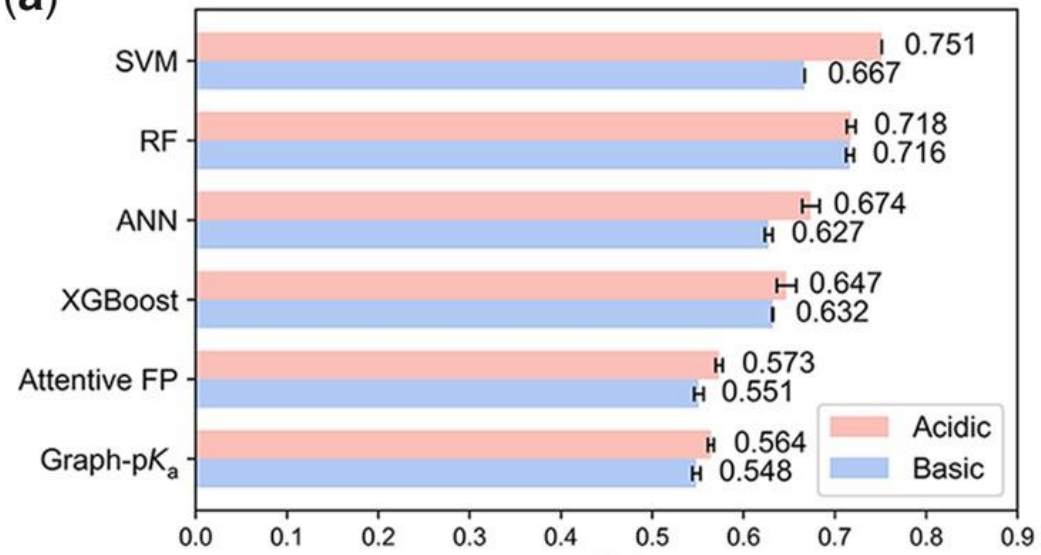


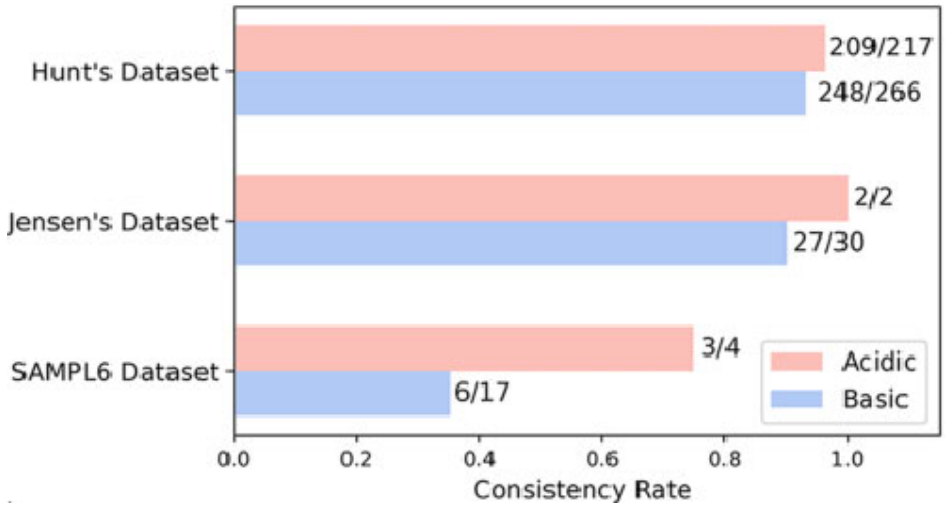
$$pK_{a(acidic)} = -\log\left(\sum_{i=1}^{N} 10^{-pK_{a(acidic)}{}^{i}}\right)$$

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.

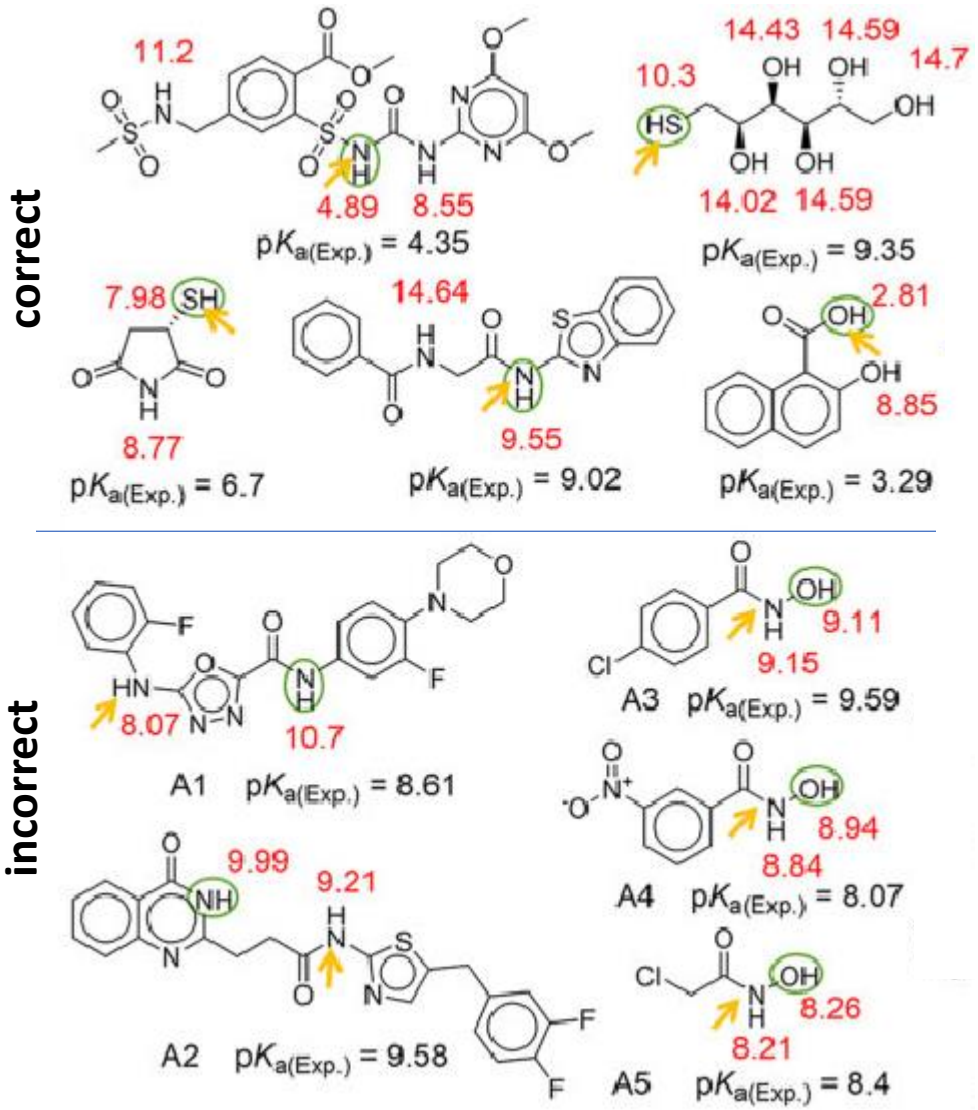
IMTM

(a)





Atoms as instances



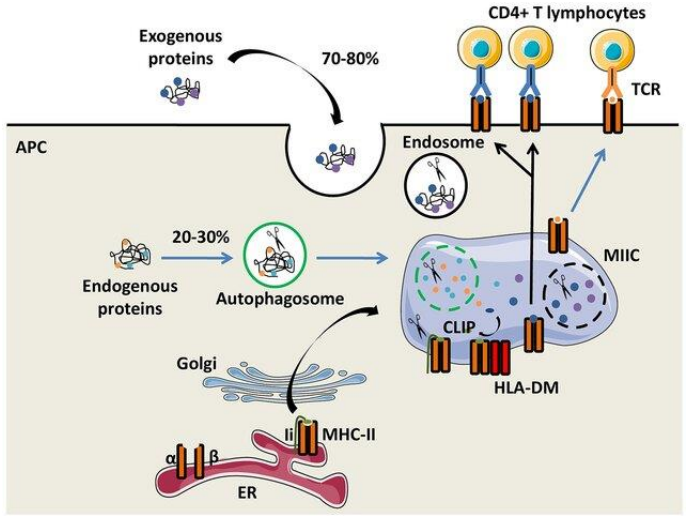
model

human experts

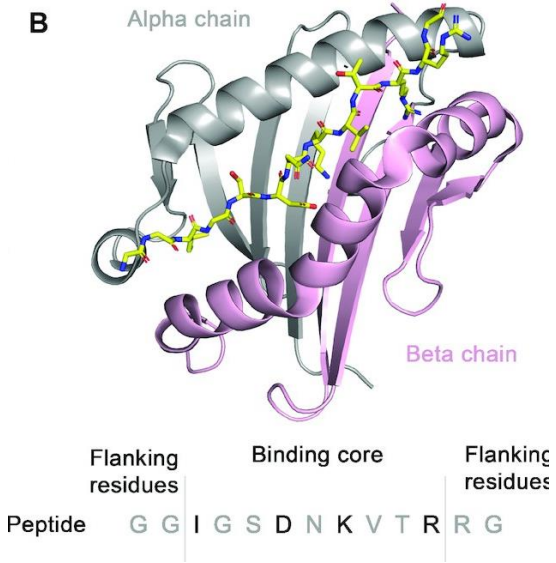
Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M., Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics* **2021**, 38, 792-798.

Peptide sequences as instances

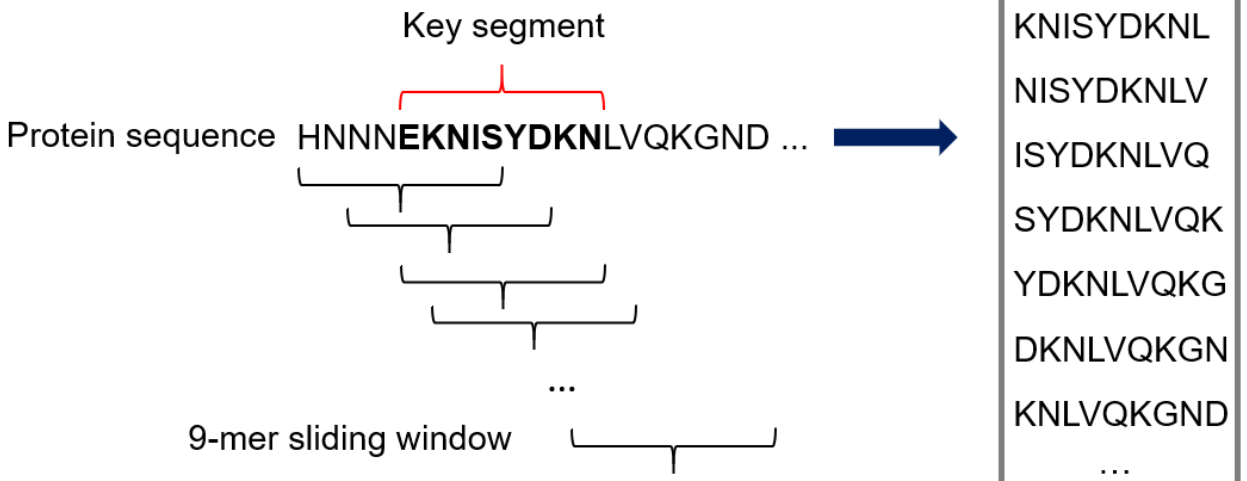
major histocompatibility complex
class II molecules (MHC II)



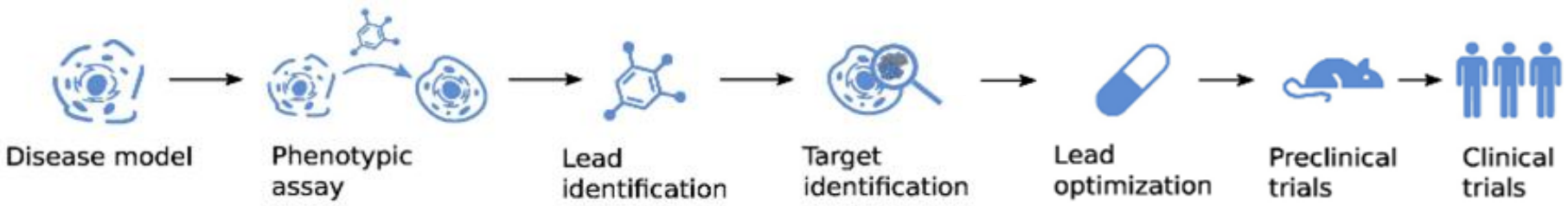
http://dx.doi.org/10.3389/fimmu.2019.01081



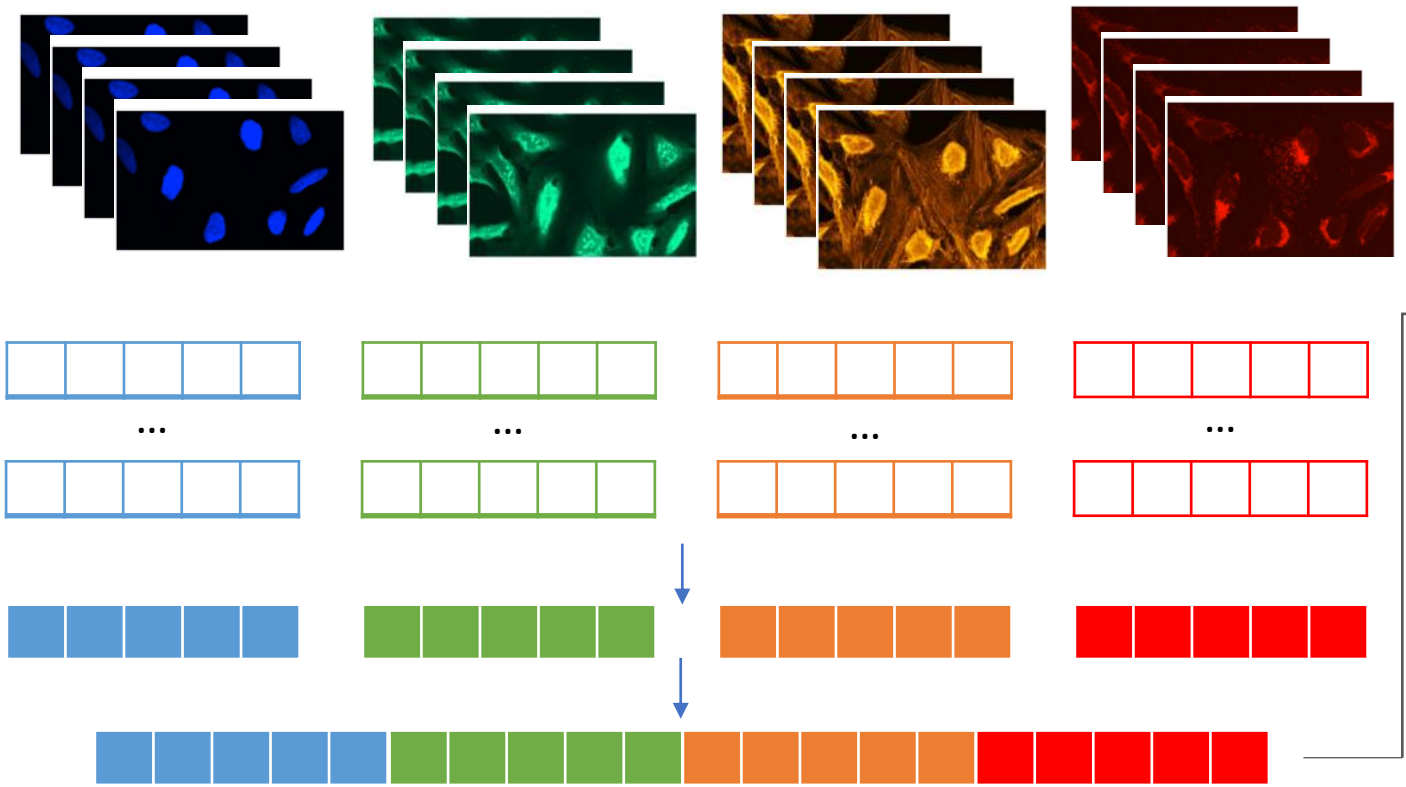
https://doi.org/10.1093/nar/gkac965

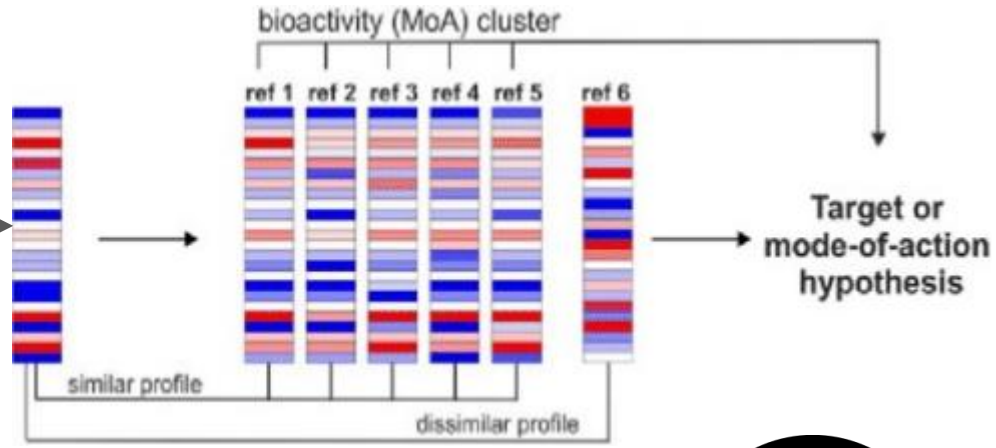


Morphological profiling



(Krentzel et al. 2023)





IMTM

Received: 8 January 2023 | Revised: 1 November 2023 | Accepted: 7 November 2023

DOI: 10.1002/wcms.1698

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

ADVANCED REVIEW

Chemical complexity challenge: Is multi-instance machine learning a solution?

Dmitry Zankov[1] | **Timur Madzhidov**[2] | **Alexandre Varnek**[1,3] | **Pavel Polishchuk**[4]

[1]ICReDD, Hokkaido University, Sapporo, Japan

[2]Chemistry Solutions, Elsevier, Oxford, United Kingdom

[3]Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France

[4]Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Olomouc, Czech Republic

Correspondence
Pavel Polishchuk, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Olomouc, Czech Republic.
Email: pavlo.polishchuk@upol.cz

Abstract

Molecules are complex dynamic objects that can exist in different molecular forms (conformations, tautomers, stereoisomers, protonation states, etc.) and often it is not known which molecular form is responsible for observed physicochemical and biological properties of a given molecule. This raises the problem of the selection of the correct molecular form for machine learning modeling of target properties. The same problem is common to biological molecules (RNA, DNA, proteins)—long sequences where only key segments, which often cannot be located precisely, are involved in biological functions. Multi-instance machine learning (MIL) is an efficient approach for solving problems where objects under study cannot be uniquely represented by a single instance, but rather by a set of multiple alternative instances. Multi-instance learning was formalized in 1997 and motivated by the problem of conformation selection in drug activity prediction tasks. Since then MIL has

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

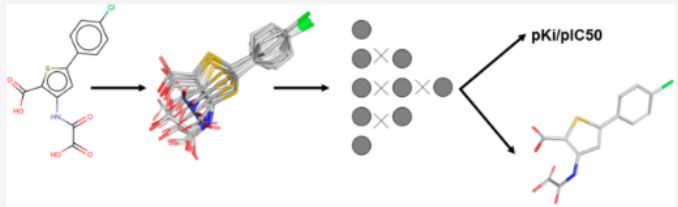
pubs.acs.org/jcim Article

QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach

Dmitry V. Zankov, Mariia Matveieva, Aleksandra V. Nikonenko, Ramil I. Nugmanov, Igor I. Baskin, Alexandre Varnek,* Pavel Polishchuk,* and Timur I. Madzhidov*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 4913–4923 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



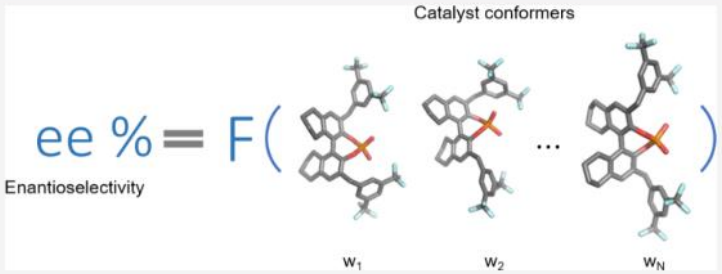
pubs.acs.org/jcim Article

Multi-Instance Learning Approach to the Modeling of Enantioselectivity of Conformationally Flexible Organic Catalysts

Dmitry Zankov, Timur Madzhidov, Pavel Polishchuk, Pavel Sidorov, and Alexandre Varnek*

Cite This: https://doi.org/10.1021/acs.jcim.3c00393 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



IMTM

MIL toolkits and software

Tool	Programming language	Link	Description
WEKA	Java	https://waikato.github.io/weka-wiki/multi_instance_classification/	Contains a module of multi-instance classification algorithms (at least 14 algorithms) as part of the WEKA tool
KEEL	Java	https://sci2s.ugr.es/keel/category.php?cat=mul	Contains some multi-instance learning classification algorithms (APR, CitationKNN, DD, etc.)
Multiple Instance Learning Matlab toolbox	Matlab	https://github.com/DMJTax/mil	Multi-instance learning classification algorithms
Multiple-Instance Learning Python Toolbox	Python	https://github.com/jmarrietar/MILpy	Multi-instance learning classification algorithms
MILL	Matlab	https://www.cs.cmu.edu/~juny/MILL/	Contains some multi-instance learning classification algorithms (APR, DD, Citation-kNN, etc.)
MISVM	Python	https://github.com/garydoranjr/misvm	Python implementation of numerous support vector machine (SVM) algorithms for the multiple-instance (MI) learning framework
AttentionDeepMIL	Python	https://github.com/AMLab-Amsterdam/AttentionDeepMIL	PyTorch implementation of attention-based deep multiple Instance learning neural network
Set Transformer	Python	https://github.com/juho-lee/set_transformer	PyTorch implementation of the paper Set Transformer
Graph neural networks	Python	https://github.com/KostiukIvan/Multiple-instance-learning-with-graph-neural-networks	Multi-instance learning with graph neural networks
3D-MIL-QSAR	Python	https://github.com/cimm-kzn/3D-MIL-QSAR	QSAR modeling based on conformation ensembles using a multi-instance learning approach

IMTM

Conclusions

- Multi-instance models outperformed both single-instance 3D models and conventional QSAR models built on 2D descriptors in many cases
- Multi-instance models is a 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
- Mixtures can also be modeled within MIL framework, where instances are individual components

Future directions

- The ability to identify relevant instances, e,g conformers, is very intriguing and underexplored. There are different approaches to identify relevant instances and which one is more suitable for chemical problems is unknown.
- How different 3D representation will work on different chemical problems.

Acknowledgements

Strasbourg University (France):

Prof. Alexandre Varnek

Strasbourg University (France)
ICReDD, Hokkaido University (Japan):

Dmitry Zankov

Palacky University (Czech Republic):

Dr. Mariia Matveieva

Aleksandra Ivanova (Nikonenko)

Kazan Federal University (Russia)
Elsevier Ltd (UK):

Dr. Timur Madzhidov

Janssen (Belgium):

Dr. Ramil Nougmanov

Technion (Israel):

Prof. Igor Baskin