SIMILARITY ASSESSMENTS IN DRUG DISCOVERY

D. Filimonov, V. Poroikov

Institute of Biomedical Chemistry, Moscow, Russia

E-mail: dmitry.filimonov@ibmc.msk.ru vladimir.poroikov@ibmc.msk.ru

Acknowledgements

IBMC Institute of Biomedical Chemistry

Laboratory of Structure-Function Based Drug Design

News

The XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

From 24 to 26 May 2022 Institute of Biomedical Chemistry (IBMC) performed the International XXVIII Symposium "Bioinformatics and Computer-Aided Drug Discovery" chaired by the Corresponding Member of the Russian Academy of Sciences Vladimir Poroikov (IBMC) and Professor Roman Efremov (IBC RAS). The Symposium was held in the framework of the World-Class Scientific Center (NCMU) "Digital Biodesign and Personalized Health Care", within the framework of the National Science Project.

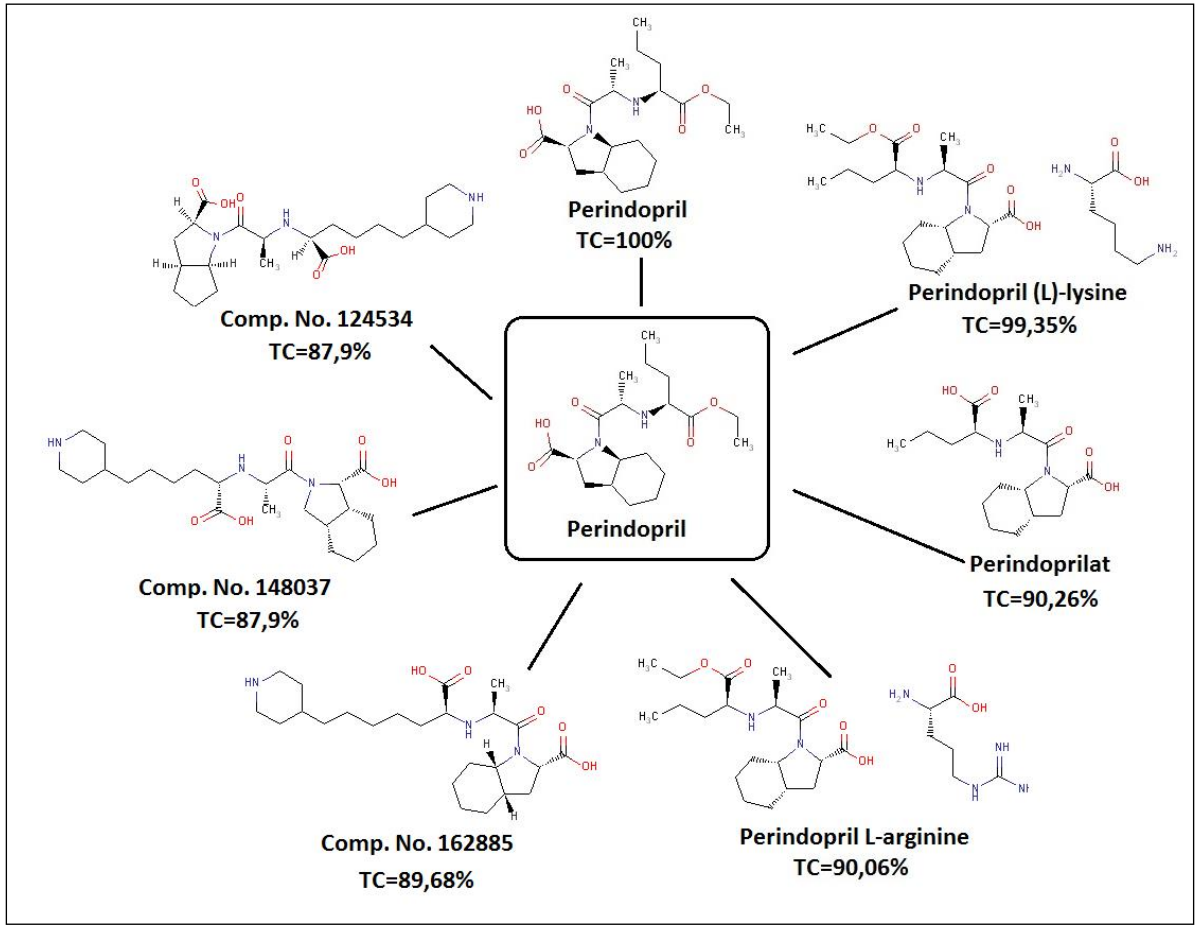
Alexey Lagunin,
Oleg Gomazkov,
Alexander Dmitriev,
Anastasia Rudik,
Boris Sobolev,
Dmitry Druzhilovskiy,
Olga Tarasova,
Pavel Pogodin,
Sergey Ivanov,
Tatyana Gloriozova,
Leonid Stolbov,
Dmitry Karasev,
Polina Savosina,
Nikita Ionov,
Nadezhda Biziukova,
Vladislav Sukhachev

25 YEARS RFBR
RUSSIAN FOUNDATION FOR BASIC RESEARCH

Financial Support

The study was supported in the framework of the Russian Foundation for Basic Research grant No. 20-04-60285.

Cortellis Drug Discovery Intelligence (**CDDI**) database, structural formula of **Perindopril** as a query. Structural similarity search results, only the most similar of 15 "similar" structures are presented.



Hierarchical clustering using similarity

Let's start with *N* clusters – each object is a separate cluster.

Next, we consistently combine clusters for which the distance between them is minimal.

The result of clustering depends on the selected method of determining the distance between clusters.

The distance between the "centers of mass":

$$D_{ij} = d\left(\frac{1}{n_i}\sum_{x \in G_i} \boldsymbol{x}, \frac{1}{n_j}\sum_{x \in G_j} \boldsymbol{x}\right)$$

where D_{ij} is the distance between two clusters, G_i и G_j.

Hausdorff distance:

$$D_{ij} = \min_{i,j} \max_{x',x''} \left(d(\boldsymbol{x}', \boldsymbol{x}'') \middle| \boldsymbol{x}' \in G_i, \boldsymbol{x}'' \in G_j\right)$$

- two clusters for which the distance between the most distant points is minimal are combined into one cluster.

Generative Topographic Mapping (GTM):
Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison

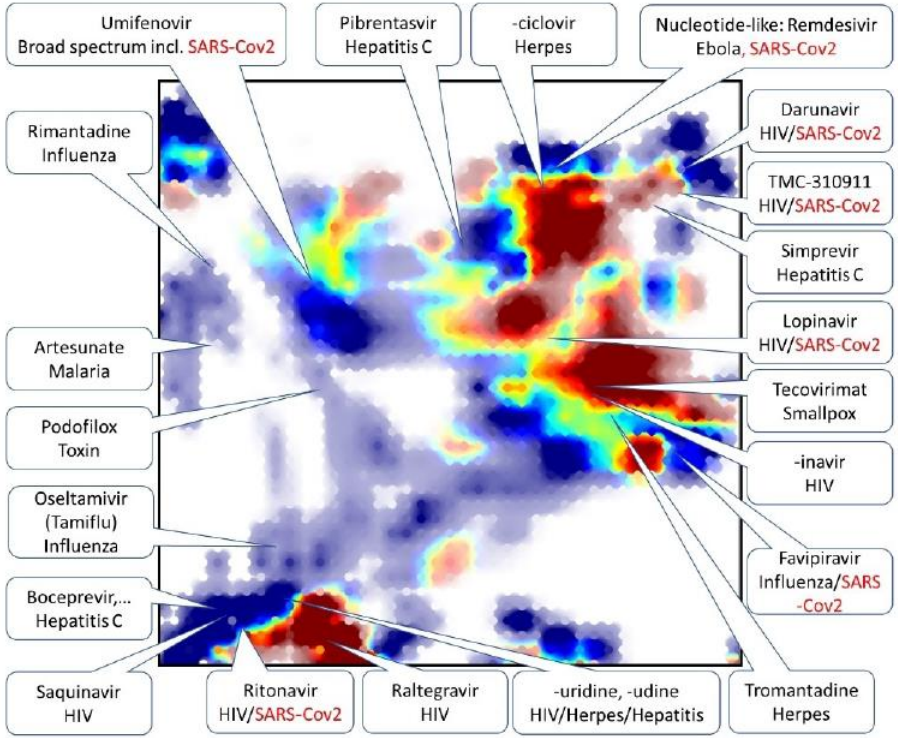
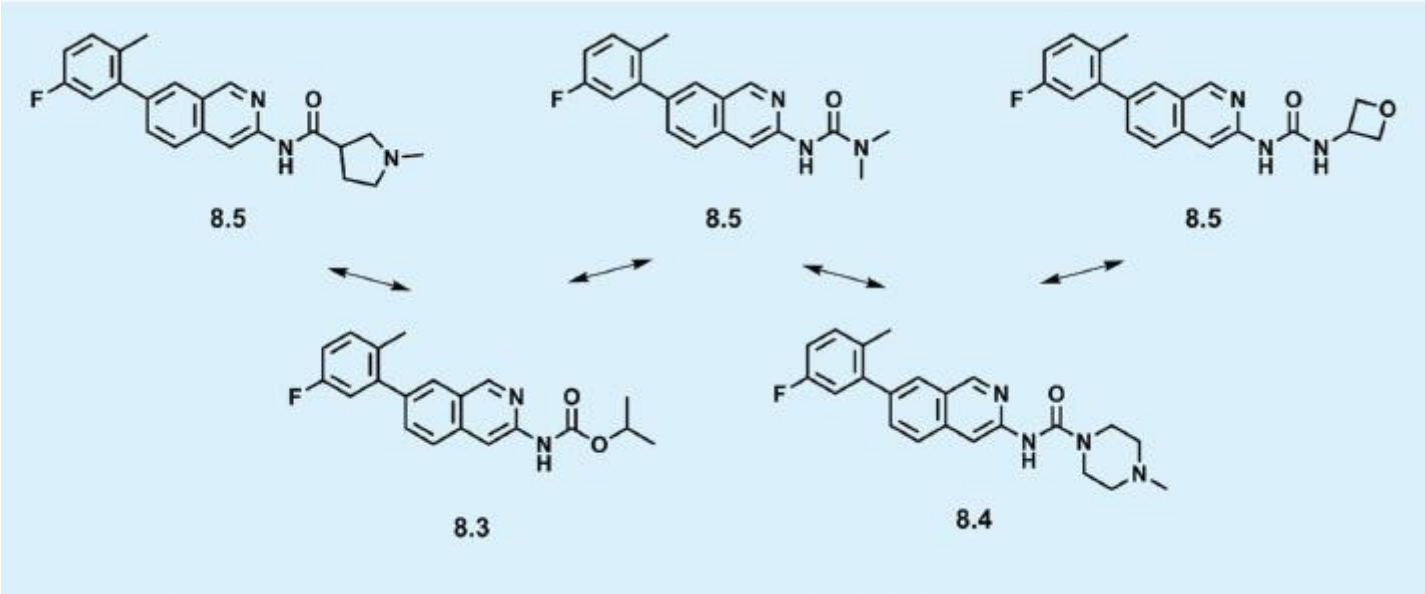


Figure 5. Pool of 1000 compounds predicted to inhibit the 3CL proteinase of the novel SARS-CoV-2, (red) mapped against the SARS-CoV compounds (blue), within the DrugBank reference frame.

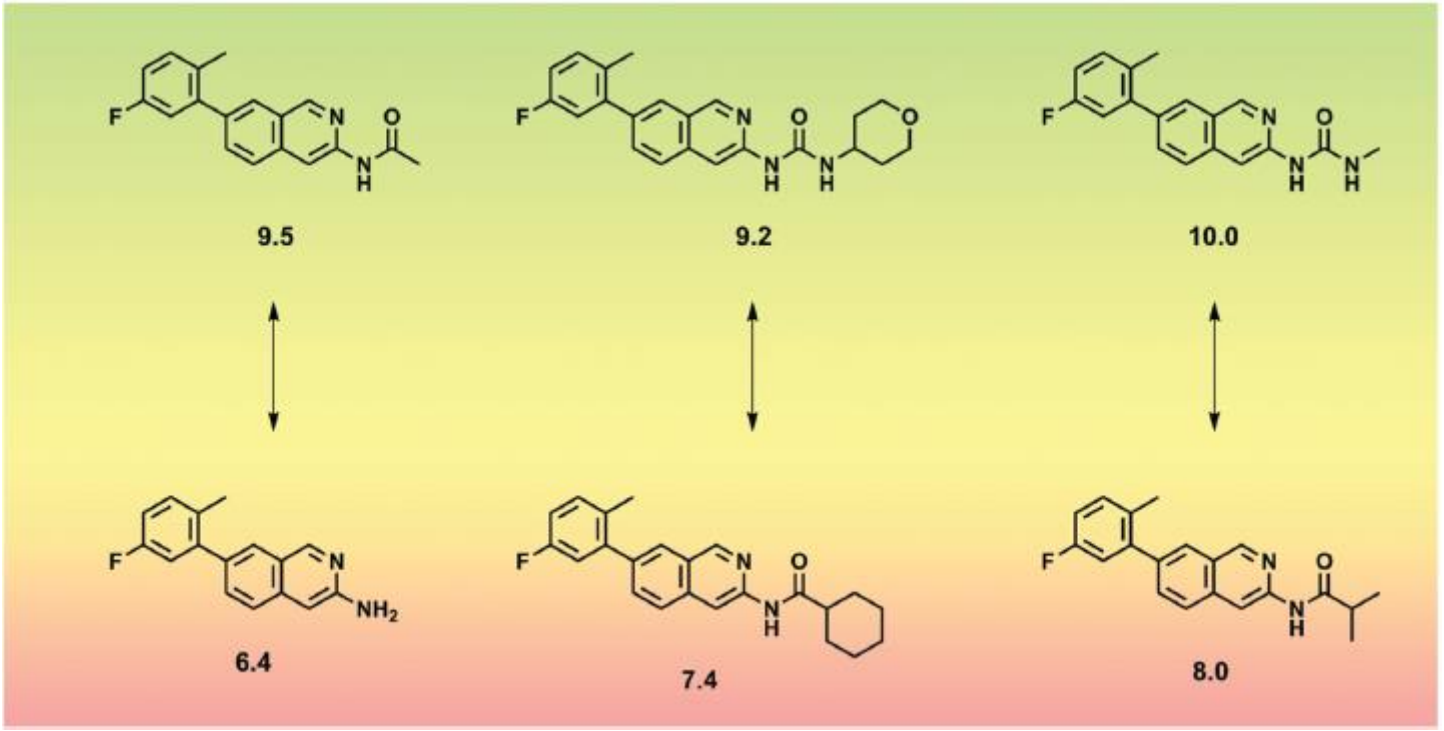
Horvath D. et al. Molecular Informatics, 2020, 39, 12. https://doi.org/10.1002/minf.202000080
Gaspar H.A. et al. J. Chem. Inf. Model. 2015, 55, 11, 2403–2410. https://doi.org/10.1021/acs.jcim.5b0039
Kireeva N. et al. Molecular Informatics, 2012, 31(3-4), 301-312. https://doi.org/10.1002/minf.201100163

The hypothesis that structurally similar compounds exhibit similar biological effects or some other properties is taken as an axiom.



Tyrosine kinase ABL inhibitors, reported pKi values.
SAR continuity is observed where gradually changes in compound structure (tracked by horizontal arrows) are accompanied by moderate activity alterations.

Muratov EN et al. (2020) QSAR without borders, Chemical Society Reviews, 49, 3525-3564. https://doi.org/10.1039/d0cs00098a



Tyrosine kinase ABL inhibitors, reported pKi values.
The inhibitors display SAR discontinuity - small structural modifications lead to large changes in activity. Vertical arrows indicate the formation of pairwise activity cliffs.

Muratov EN et al. (2020) QSAR without borders, Chemical Society Reviews, 49, 3525-3564. https://doi.org/10.1039/d0cs00098a

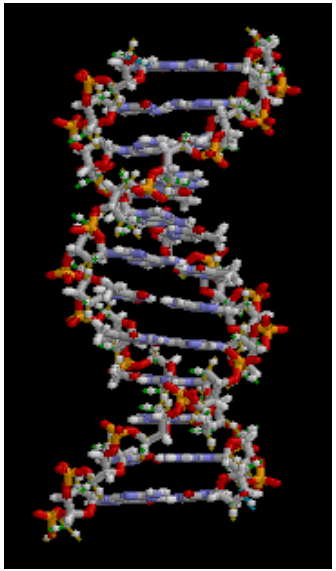
Molecular similarity, as a paradigm, contains many implicit and explicit assumptions.

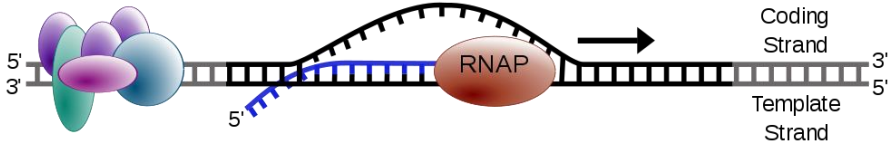
One does not know a priory which properties of the molecular structure are essential for its biological activity; therefore, the description of the structure can be only heuristic.

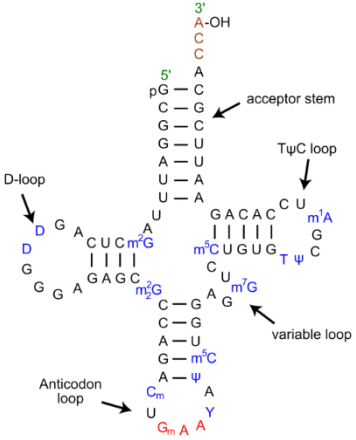
The selection of molecular descriptors and the estimation of molecular similarity based on this selection crucially determine the final result of the study.

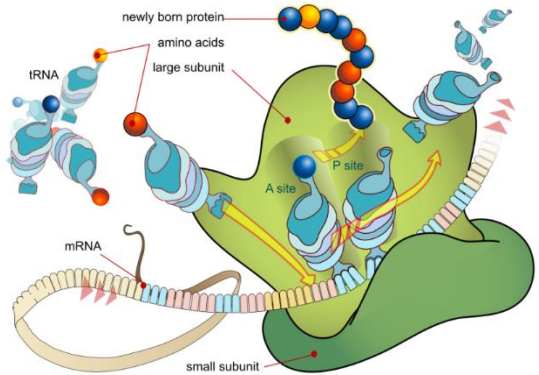
However, for novel pharmacological targets (like SARS-CoV-2 coronavirus proteins), when only limited number of antiviral agents that may be used as a "query" are known, similarity assessment is the method-of-the-choice.

Local Correspondence Concept









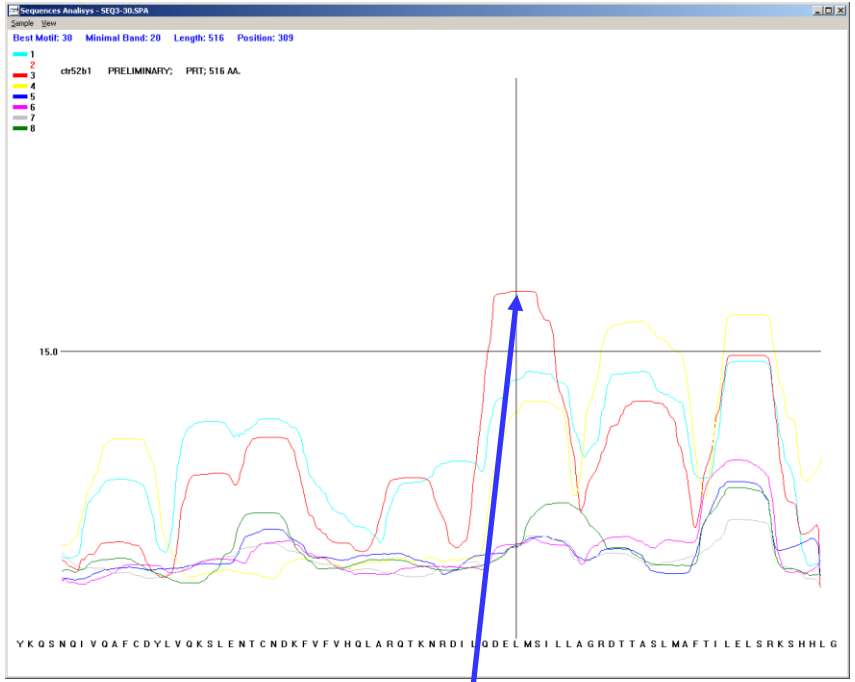
Sequence Local Similarity. Frame 20, shift from 0 to 17

```
AANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVA  2
ANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVAL  1
NRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALR  1
RDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRA  0
DPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRAL  1
PSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALF  2
SQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFG  1
QFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGR  1
FPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRF  2
PDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFP  0
DPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPA  1
PHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPAL  0
HRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALS  9   The best match
RFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSL  0
FDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLG  3
DVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGI  1
VTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGID  1
TRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGIDA  2

GTAINKPLSEKMMLFGMGKRRCIGEVLAKWEIFLFLAILLQQLEFSV  9   Query sequence
                 |
                Ri = 9
```

Sobolev B., Filimonov D., Lagunin A. et al. (2010) BMC Bioinformatics, 11:313.
http://www.biomedcentral.com/1471-2105/11/313

Sequence Local Similarity. It is descriptor itself!



Descriptor is defined as the similarity value S_{ik}
for position i of sequence under study
and experimentally annotated sequence k.

Local Correspondence Concept. Neighborhoods of atoms descriptors

The most biological activities of organic compounds are the result of molecular recognition, which in turn depends on the correspondence between particular atoms of the ligand and the target.

MOLECULAR BIOLOGY

QUANTUM CHEMISTRY

QUANTUM FIELD THEORY

$$M_i = V_i + V_i g M = V_i + V_i g(M_1 + M_2 + \ldots + M_m)$$

All descriptors are based on the concept of atoms' of molecule description subject to the neighborhood of them:

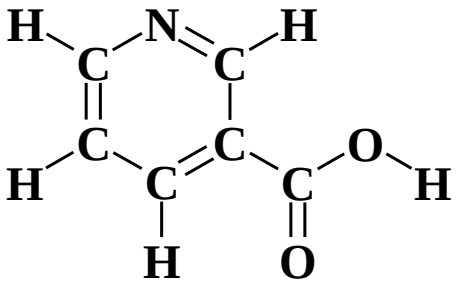
MNA - Multilevel Neighborhoods of Atoms

LMNA - Labeled Multilevel Neighborhoods of Atoms

QNA - Quantitative Neighborhoods of Atoms

Filimonov D.A., Poroikov V.V. (2008) In: Chemoinformatics Approaches to Virtual Screening. Eds. Alexandre Varnek and Alexander Tropsha. Cambridge (UK): RSC Publishing, 182-216.

Multilevel neighborhoods of atoms descriptors – MNA



MNA/2

C(C(CC-H)C(CC-C)-H(C))
C(C(CC-H)C(CN-H)-H(C))
C(C(CC-H)C(CN-H)-C(C-O-O))
C(C(CC-H)N(CC)-H(C))
C(C(CC-C)N(CC)-H(C))
N(C(CN-H)C(CN-H))
-H(C(CC-H))
-H(C(CN-H))
-H(-O(-H-C))
-C(C(CC-C)-O(-H-C)-O(-C))
-O(-H(-O)-C(C-O-O))
-O(-C(C-O-O))

Filimonov D.A., Poroikov V.V. (2008) In: Chemoinformatics Approaches to Virtual Screening. Eds. Alexandre Varnek and Alexander Tropsha. Cambridge (UK): RSC Publishing, 182-216.

PASS (Prediction of Activity Spectra for Substances) software

Theoretical basis:
The Local Correspondence Concept
The original Bayesian Classifier
High-quality training dataset
Knowledge base "structure-mechanism-effect"

1992
The first version of PASS:
9,314 BAS, 114 A, 0.85 APA

1998
New version of PASS:
31,000 BAS, 435 A, 0.94 APA

1999
The first PASS Online program on the Internet

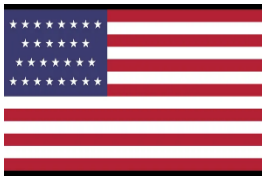
2009
New version of PASS:
206,000 BAS, 3,750 A, 0.95 APA

2022
New version of PASS:
1,614,066 BAS, 8,387 A, 0.94 APA

2021: 20 web services
PASS Online Statistics:
1 137 046 predictions;
>1 200 publications;
30 758 users
from 104 countries of the world


Russia (27,6%)


India (12,4%)


USA (8%)


China (5,3%)

1990 2000 2010 2021

BAS is the number of substances in the training set; A is the number of predicted activity types; APA is the average prediction accuracy

Quantitative neighborhoods of atoms descriptors – QNA

In fact, interatomic and intermolecular forces are electrical in nature according to the Hellman-Feynman theorem.

Feynman R. Ph. Phys. Rev., 1939, 56, 340-343.

$P_i = B_i\sum_k(Exp(-½C))_{ik}B_k$

$Q_i = B_i\sum_k(Exp(-½C))_{ik}B_kA_k$

$A = ½(IP + EA)$

$B = (IP - EA)^{-½}$

C is the connectivity matrix of a molecule,

IP is the first ionization potential,

EA is the electron affinity.

Normalization:

$P' = (P - E(P))/D(P)$

$Q'' = (Q - E(Q))/D(Q)$

$Q' = (\mu P' - Q'')/D(PQ)$

$E(P') = 0,\ D(P') = 1$

$E(Q') = 0,\ D(Q') = 1$

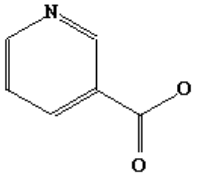
$Cov(P'Q') = 0$

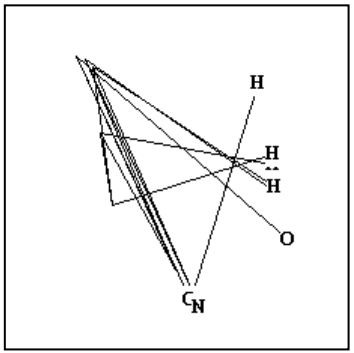
Robert G. Parr et al. J. Chem. Phys., 1978, 68(8), 3801-3807.
Gasteiger J, Marsili M. Tetrahedron, 1980, 36, 3219-3228.
Rappe A K and W A Goddard III. J. Ph. Ch., 1991, 95, 3358-3363.

Filimonov D.A. et al. (2009) SAR and QSAR Environ. Res., 20 (7-8), 679-709.

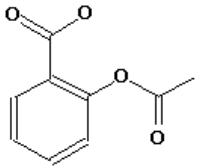
QNA descriptors' space

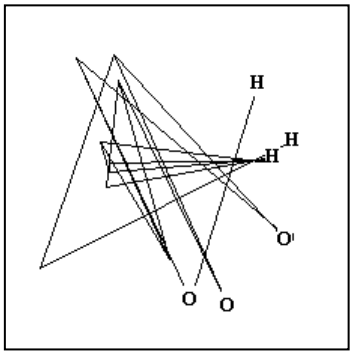
Nicotinic Acid



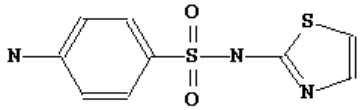


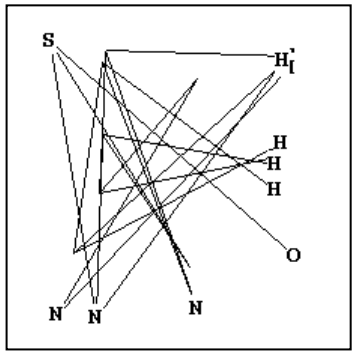
Acetylsalicylic acid





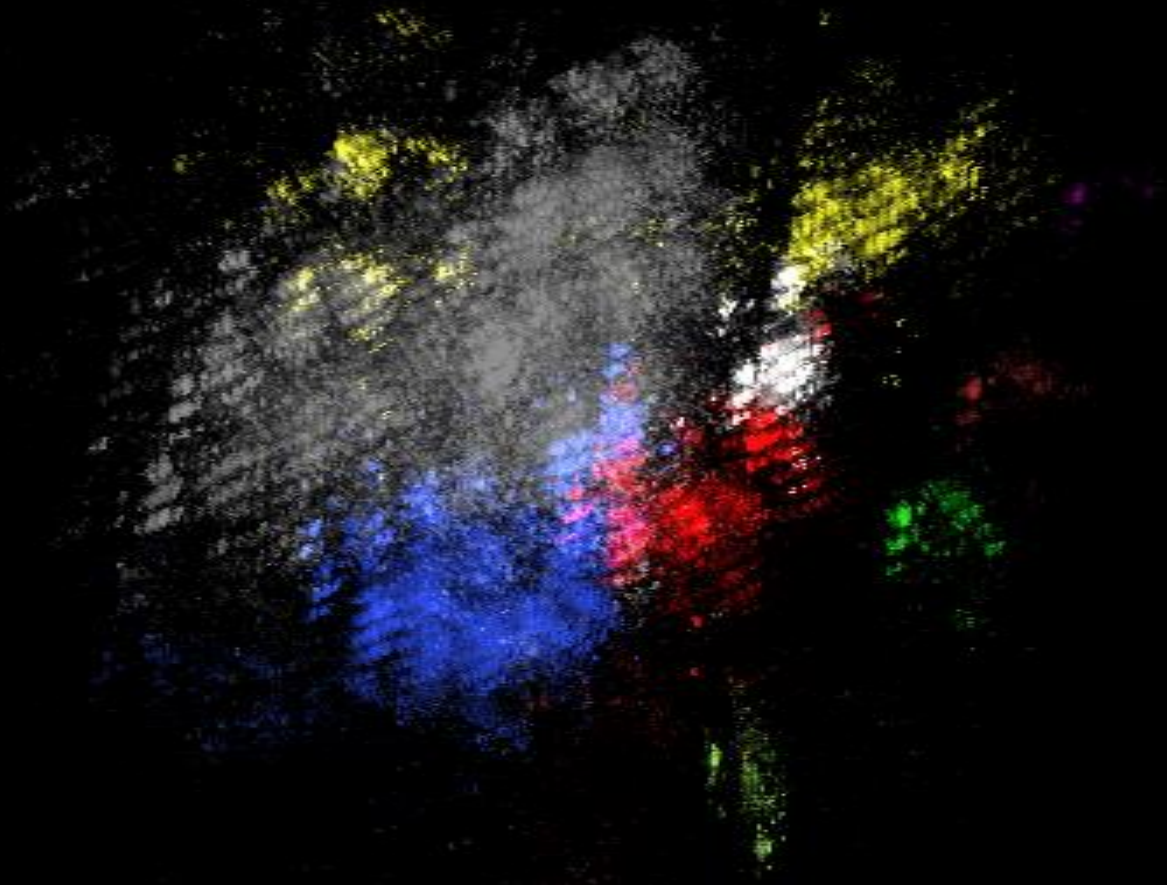
Sulfathiazole





Filimonov D.A. et al. (2009) SAR and QSAR Environ. Res., 20 (7-8), 679-709.

QNA descriptors' space



Similarity estimation using the QNA descriptors:

$$F(A,B) = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)}$$

$$n(A \cap B) = \frac{1}{2}\left(\sum_A \max_{b \in B}[s_{ab}] + \sum_B \max_{a \in A}[s_{ba}]\right)$$

where s_{ab} and s_{ba} are the pairwise similarities of QNA descriptor of atom ***a*** in a molecule *A* and QNA descriptor of atom *b* in a molecule *B*:

$$s_{ab} = Exp\left(-4N_B\left((P_a - P_b)^2 + (Q_a - Q_b)^2\right)\right)$$

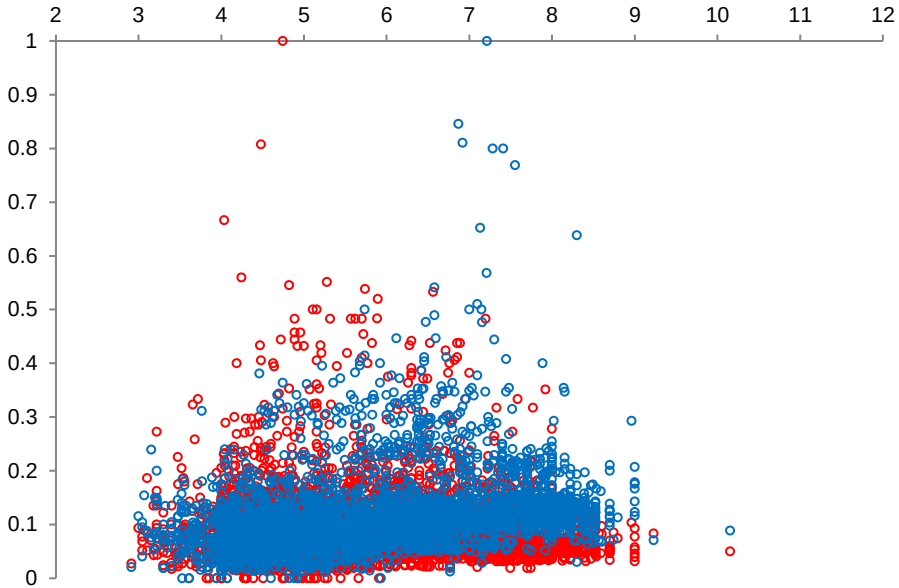
$$s_{ba} = Exp\left(-4N_A\left((P_a - P_b)^2 + (Q_a - Q_b)^2\right)\right)$$

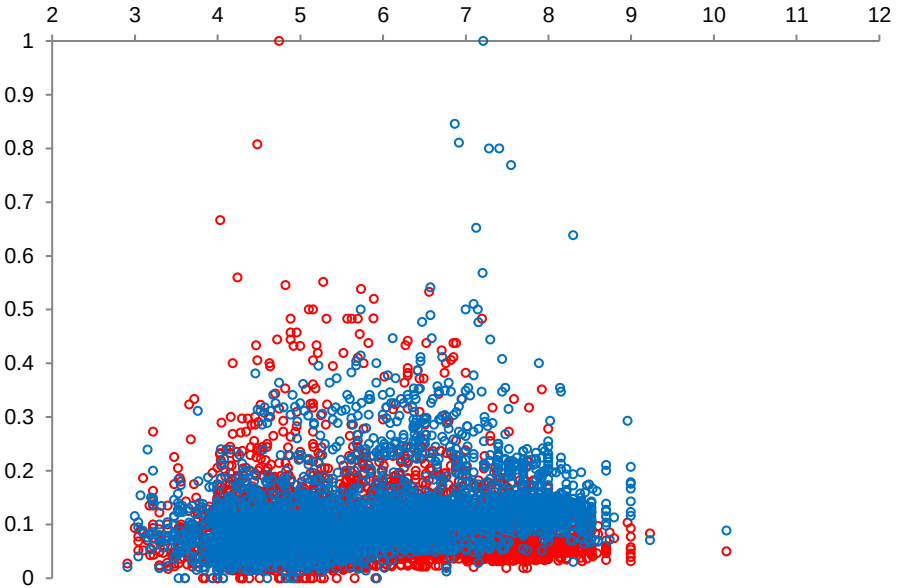
where P_a and Q_a are values of QNA descriptor of atom ***a*** in a molecule *A*, P_b and Q_b are values of QNA descriptor of atom *b* in a molecule *B*.

Similarity of compounds – HIV-1 integrase inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.

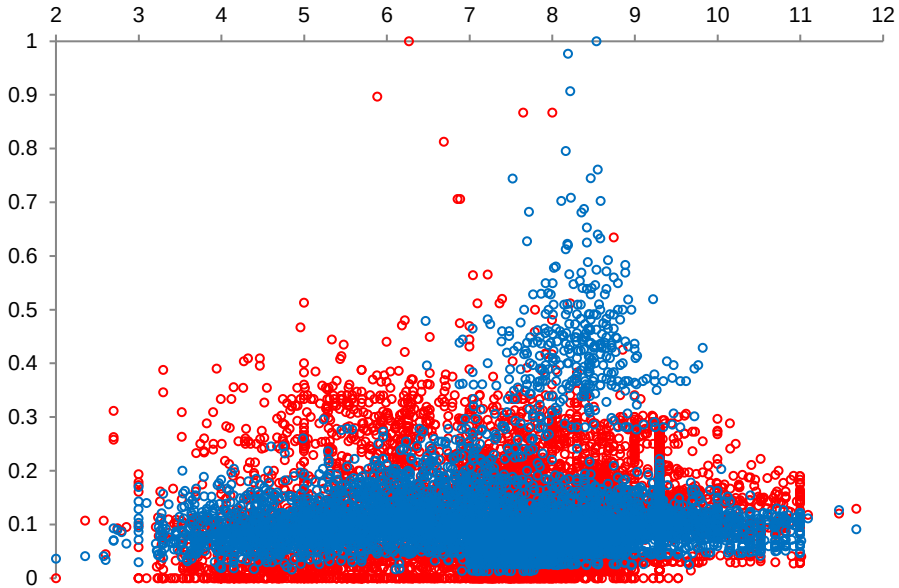


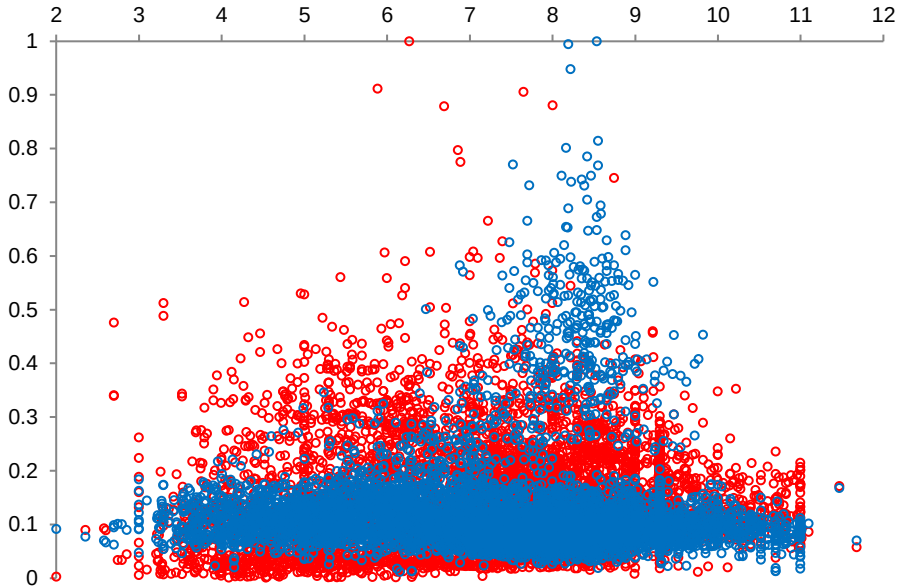


Similarity of compounds – HIV-1 protease inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.

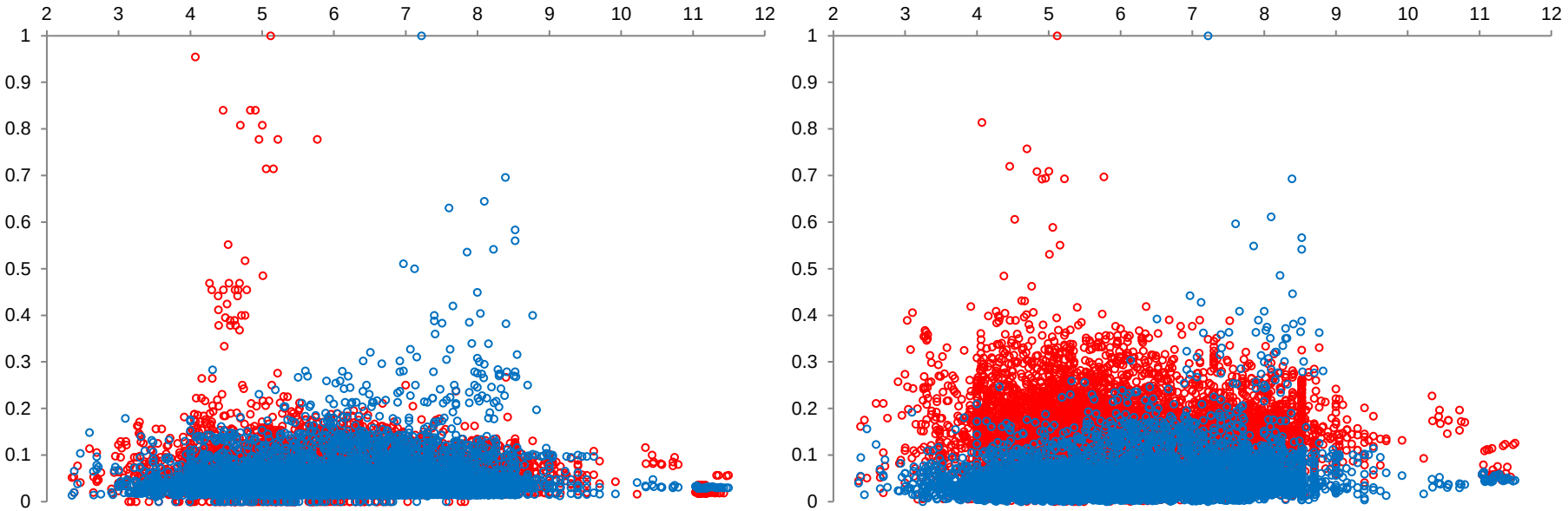




Similarity of compounds – HIV-1 reverse transcriptase inhibitors.

Dependencies between similarity (on the Y axis) and pIC50 (on the X axis).

On the left – T(A,B) for MNA descriptors, on the right – F(A,B) for QNA descriptors.



Big Chemical Data

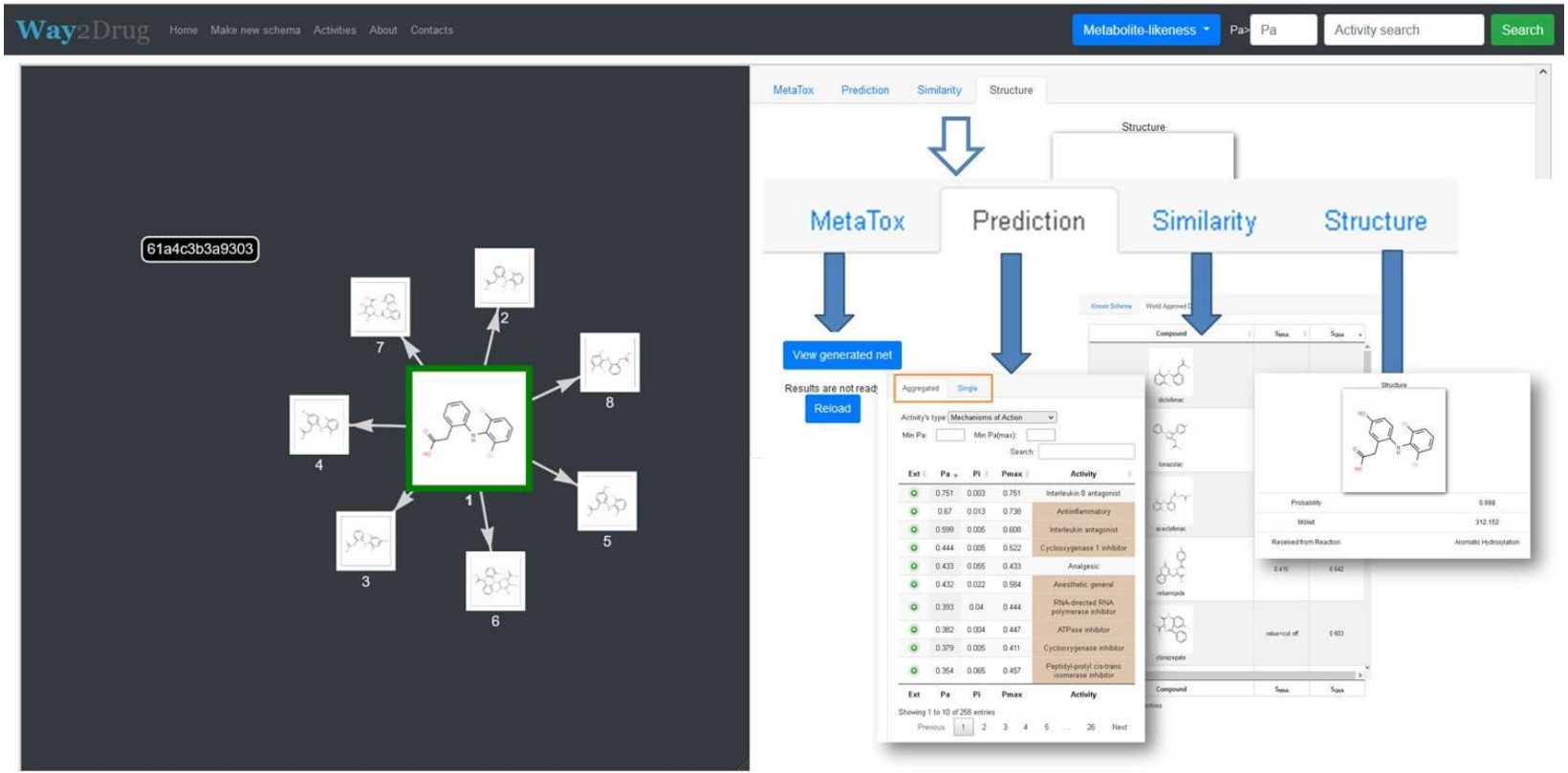
Many different sources are used – thus, there is a problem of duplicating the molecule structures.

When the number of structures in data set, N, is many millions, the pairwise comparison of structures requires an inaccessible resources to perform N(N - 1)/2 comparisons.

To solve such tasks, we use the Q index which is the sum of the Q values of QNA descriptors.

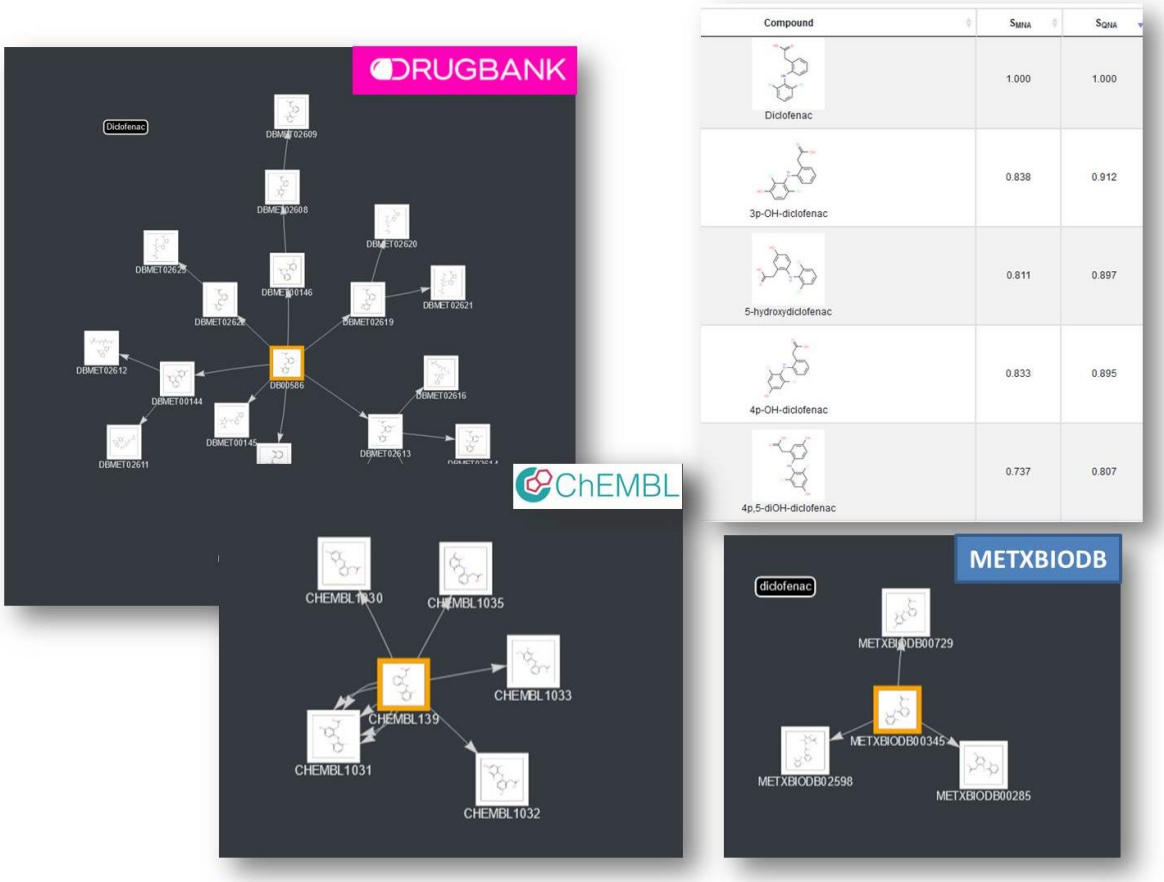
The resulting array of N real values can be easily sorted, and then you need to compare only those structures whose Q indexes differ by less than 1E-9.

The MetaTox 2.0 web portal provides an opportunity for a comprehensive analysis of the biological activity profiles of existing and developing drugs, taking into account their metabolism in the human.



http://www.way2drug.com/metatox

Three schemes of Diclofenac metabolism from different sources and structures of Diclofenac-like compounds with similarity estimates for MNA and QNA descriptors (top right).



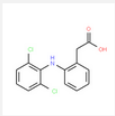
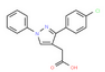
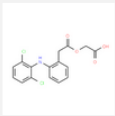
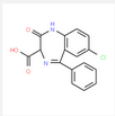
Compound	S_{MNA}	S_{QNA}
Diclofenac	1.000	1.000
3p-OH-diclofenac	0.838	0.912
5-hydroxydiclofenac	0.811	0.897
4p-OH-diclofenac	0.833	0.895
4p,5-diOH-diclofenac	0.737	0.807

More than 2000 biologically active compounds with known metabolic pathways, which were extracted from DrugBank, MetXBIODB and ChEMBL databases

http://www.way2drug.com/metatox

MetaTox | Prediction | Similarity | Structure

Known Schema | World Approved Drugs

Compound	S_{MNA}	S_{QNA}
 diclofenac	1.000	1.000
 lonazolac	0.467	0.677
 aceclofenac	0.789	0.676
 rebamipide	0.415	0.642
 clorazepate	.value<cut off	0.603
Compound	S_{MNA}	S_{QNA}

Showing 1 to 60 of 60 entries

Structures of Diclofenac-like drugs from the WWAD sample with similarity estimates for MNA and QNA descriptors.

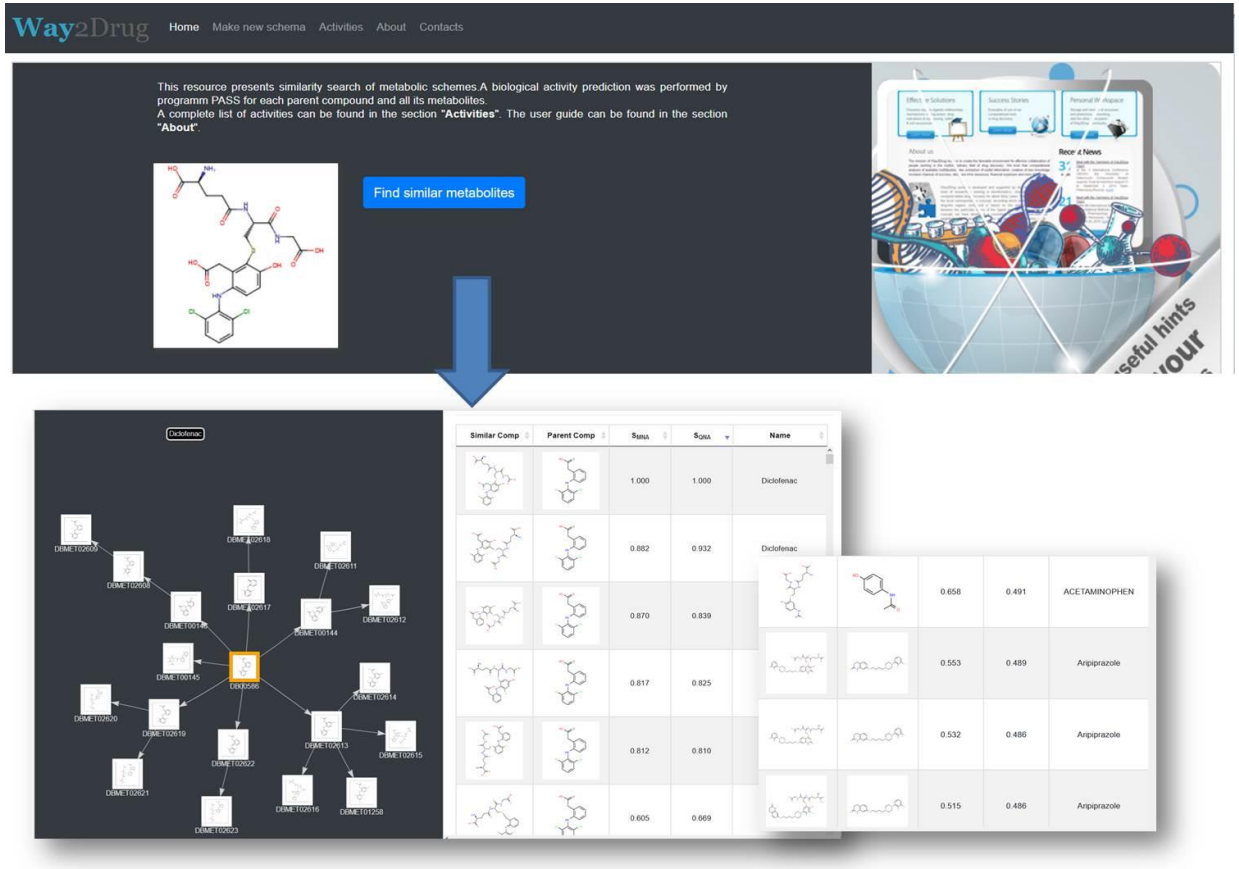
World Wide Approved Drugs (WWAD)
contains information on more than 4,000 medicines approved by regulatory authorities in various countries.

For the initial compound, a table is given containing the structure of the drug and two similarity values calculated using the MNA and QNA descriptors.

http://www.way2drug.com/metatox

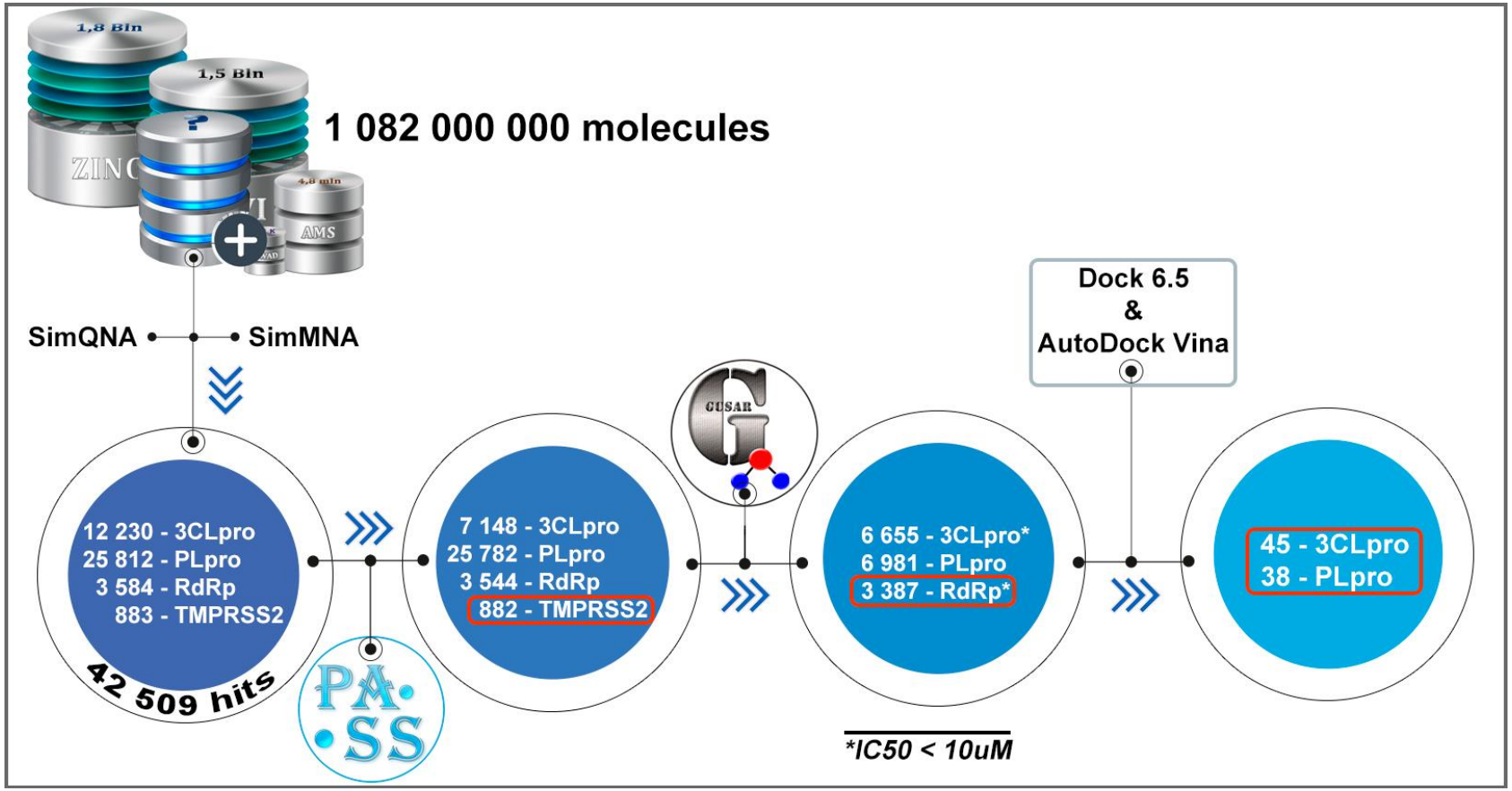
The search for precursor compounds by the similarity method for Diclofenac acyl glucuronide.

The method of evaluating possible precursors of compounds is to search for similar metabolites among the known ones.



http://www.way2drug.com/retro

Our selection of “hits” by virtual screening in the JEDI Grand Challenge



https://discord.com/channels/694851986042126366/694851987208011818

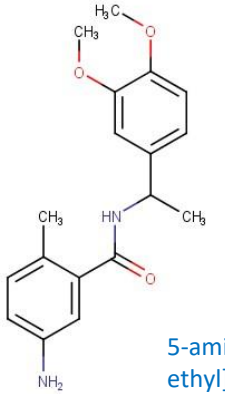
Results of synthesis and biological testing of 820 compounds (36 from us) in the framework of JEDI

PowerPoint Slide Show - [JEDI_Symposium_20220223.pptx]

Details of hits

Name	SMILES	Teams	List position	Predicted Target	Exp Target
ace2_24791097	CC(C(=O)NC(Cc1c[nH]c2ccccc12)C(O)=O)c1ccc(Cl)cc1	kyuken	9247	Ace2	S
S_dee55631224	CCc1cc(NC(=O)C2(CC2)c2cccc(F)c2)[nH]n1	deeplab	56	S	S
ace2_84865828	COC(=O)C(Cc1c[nH]c2cc(F)ccc12)NC(=O)C(C)Oc1ccccc1	kyuken	6848	Ace2	S
nsp5_15734065	CC(N(C)C(=O)Cn1nnc2ccccc12)c1cccc(Cl)c1	JKU	212 & 1157	Nsp5	Nsp5
nsp12s4614385	O=C(Nc1nc2CCN(Cc3ccccc3)Cc2s1)c1ccc2C(=O)N3CCCCCC3=Nc2c1	covid19ddc	46	Nsp12	Nsp5
ace2_53484105	Fc1ccc2c(CCNC(=O)c3cccc(Nc4ccc(cc4)C#N)c3)c[nH]c2c1	kyuken	34	Ace2	Nsp5
nsp12s8452882	O=C(Nc1nc2cc3OCCOc3cc2s1)C1CCN(CC1)S(=O)(=O)c1ccc2CCCc2c1	virtualflow	84 & 84	Nsp12, TMPRSS2	Nsp5
tmprss9288982	O=C(Nc1nc2CN(Cc3ccccc3)CCc2s1)c1ccc2C(=O)N3CCCCCC3=Nc2c1	cermn	92	TMPRSS2	Nsp5
tmprss5675826	Cc1ccc(Nc2cccc(c2)C(=O)Nc2cc([nH]n2)C(=O)OCc2ccccc2)nn1	ai4science	56 & 167	TMPRSS2	Nsp5
nsp5_10432016	CC(C)N(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	JKU	104	Nsp5	Nsp5
nsp5_39432161	CN(Cc1cccc(F)c1)C(=O)Cn1nnc2ccccc12	JKU	394	Nsp5	Nsp5
nsp5_12421240	O=C(Cn1nnc2ccccc12)N1CCCCCC1c1ccccc1	JKU	93 & 124	Nsp5	Nsp5
nsp5_54148358	CCN(C(c1ccccc1)c1ccccc1)C(=O)Cn1nnc2ccccc12	JKU	41	Nsp5	Nsp5
nsp5_84931882	CC(C)CN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	JKU	849	Nsp5	Nsp5
nsp5_53469020	CC(N(C1CC1)C(=O)Cn1nnc2ccccc12)c1ccccc1	JKU	34	Nsp5	Nsp5
nsp5_55279240	CC(N(C)C(=O)Cn1nnc2ccccc12)c1ccccc1	JKU	52	Nsp5	Nsp5
nsp5_43858913	CC(NC(=O)Cn1nnc2ccccc12)c1cccc(Cl)c1	JKU, aiwinter	1011 & 2438	Nsp5	Nsp5
nsp5_35872038	COc1ccc(Cl)cc1CN(C)C(=O)Cn1nnc2ccccc12	JKU	4358	Nsp5	Nsp5
nsp5_55727679	CCN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	JKU	57	Nsp5	Nsp5
nsp5_54641236	CN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	JKU	46	Nsp5	Nsp5
nsp5_2648068	CC(N(C1CC1)C(=O)Cn1nnc2ccccc12)c1cccc(c1)C(F)(F)F	JKU	264	Nsp5	Nsp5
nsp5_5489691	CCC(N(CC)C(=O)Cn1nnc2ccccc12)c1ccccc1	JKU	45 & 48	Nsp5	Nsp5
nsp3_56422252	COc1ccc(cc1OC)C(C)NC(=O)c1cc(N)ccc1C	way2drug	63 & 64	Nsp3	Nsp3
nsp3_51533607	Fc1ccc2[nH]c(CNC(=O)c3ccc(F)nc3)nc2c1	kyuken	515	Nsp3	Nsp3 macrodomain
nsp3_OS110511	Fc1ccc(cn1)C(=O)NCc1nc2ccc(Cl)cc2[nH]1	kyuken	1	Nsp3	Nsp3 macrodomain
nsp12S5746081	O=C(CC1=NNC(=O)c2ccccc12)Nc1ccc(NS(=O)(=O)c2ccc3OCCOc3c2)cc1	covid19ddc	57	Nsp12	Nsp12
nsp1295277234	Oc1nc(no1)-c1ccc(NC(=O)C2CCCS2)cc1	mlinch	4952	Nsp12	Nsp12

63



5-amino-N-[1-(3,4-dimethoxyphenyl) ethyl]-2-methylbenzamide

PASS_MNA_COUNT: 40

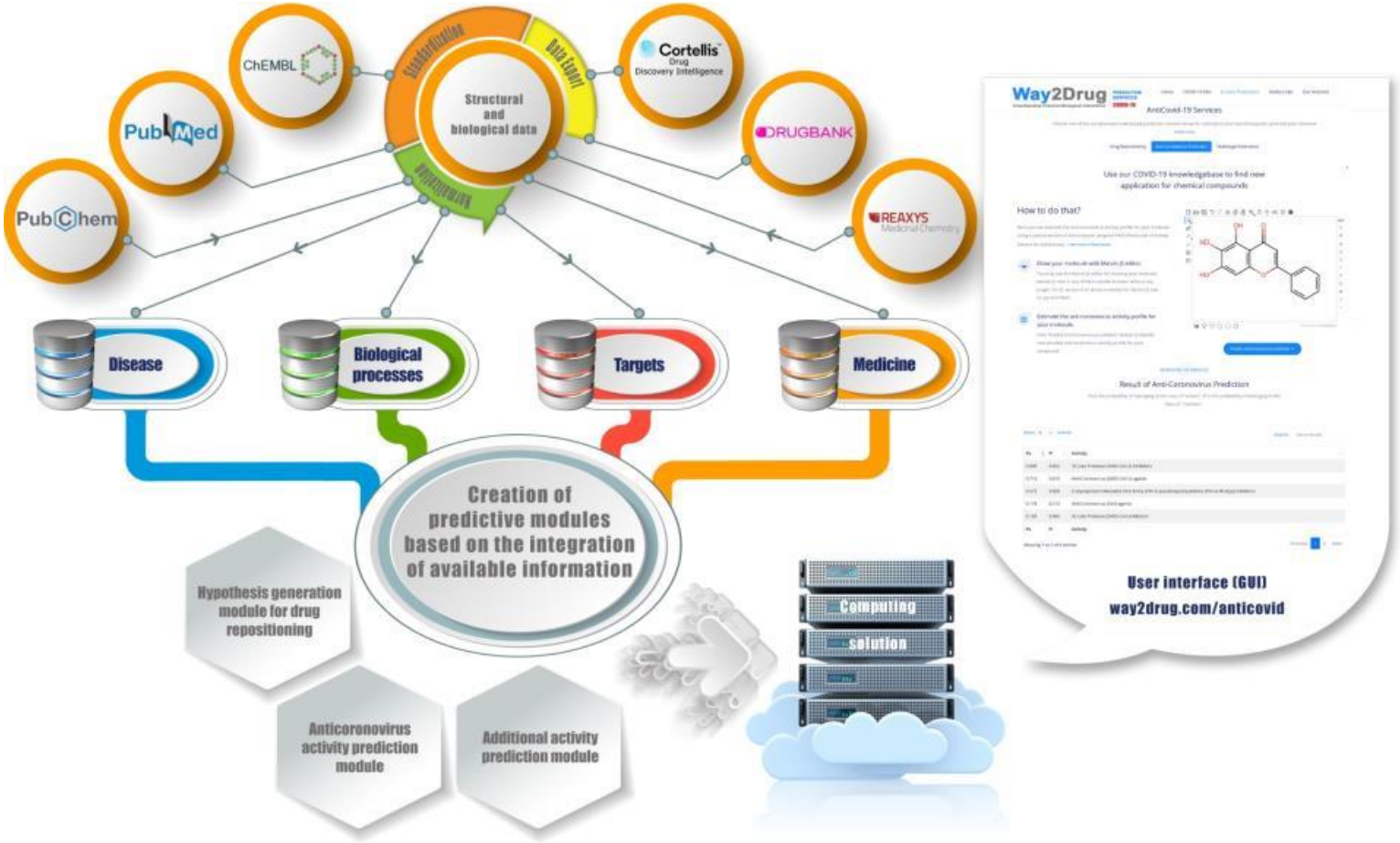
PASS_MNA_NEW_COUNT: 0

PASS_RESULT_COUNT: 4 of 8 Possible Activities at Pa>Pi

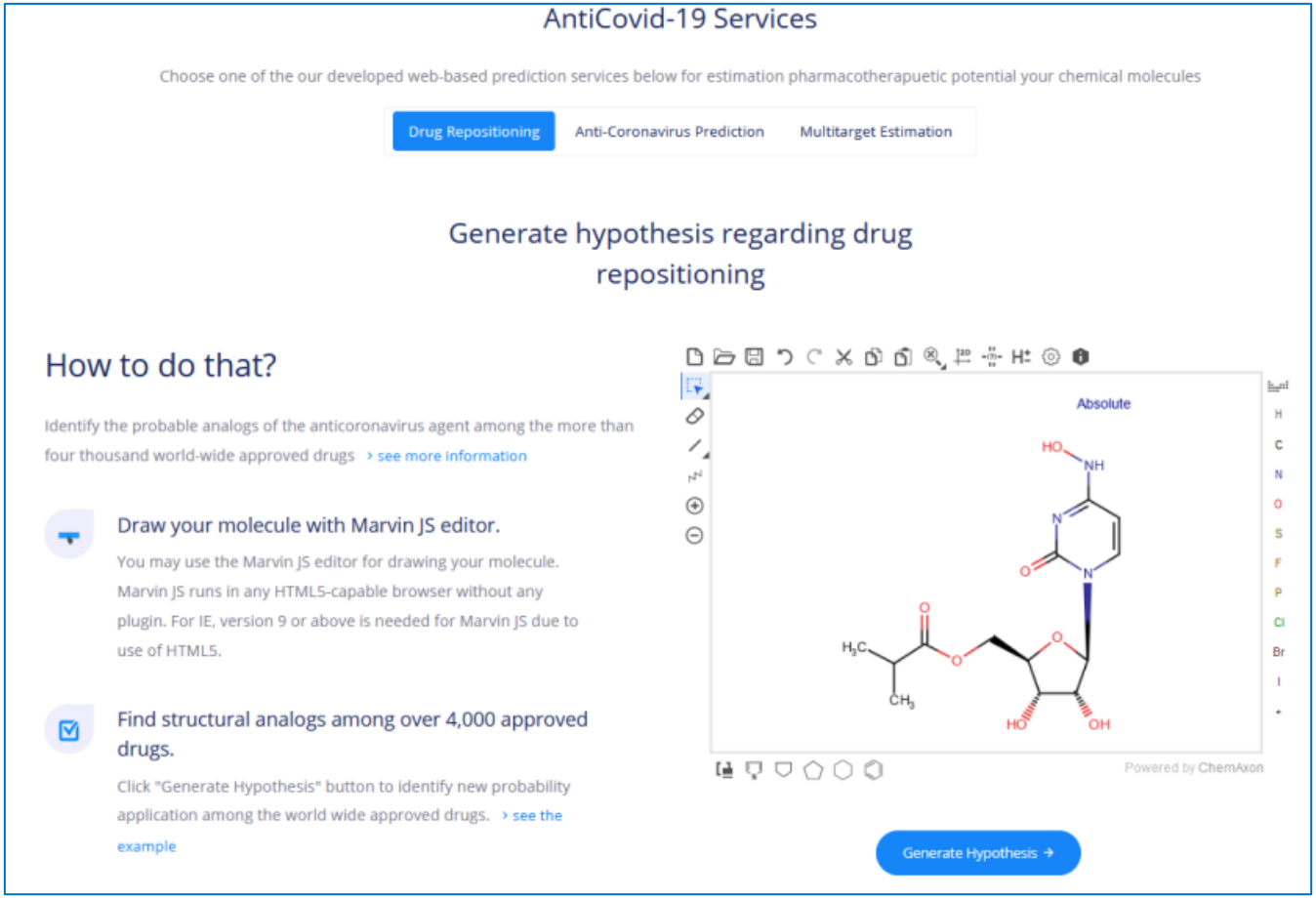
PASS_ACTIVITY_SPECTRUM

Pa	Pi	Activity
0.908	0.001	Papain-like Protease (SARS-CoV-2) Inhibitors
0.341	0.188	Spike Glycoprotein (S) (SARS-CoV-2)/ACE2 Interaction inhibitors
0.365	0.249	SARS-CoV-2 Infection Reduction in Cell-Based Assay
0.136	0.090	3C-Like Protease (SARS-CoV) Inhibitors

Informational-Computational System AntiCOVID-19



Drug repurposing: Molnupriravir as a query



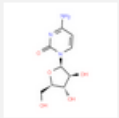
Structural analogs of Molnupriravir among the launched drugs

ANTICOVID-19 SERVICES

Result of similarity estimation

Show 5 entries

Search: Search Results

Structure	Name	Indication	Activity	Approval	Target	Similarity MNA	Similarity QNA
	cytarabine	› Myeloid Leukemia › Lymphoid leukemia Read more	› Antineoplastic (my... › CYP3A4 substrate Read more	1969-06-17 FDA	› DNA polymerase ... › Cytidine deaminas... Read more	0.576	0.602
	enocitabine	› Acute leukemia	› Antileukemic	2009-09-01 PMDA		0.529	0.452

Thank you for your attention! We are open for collaboration.