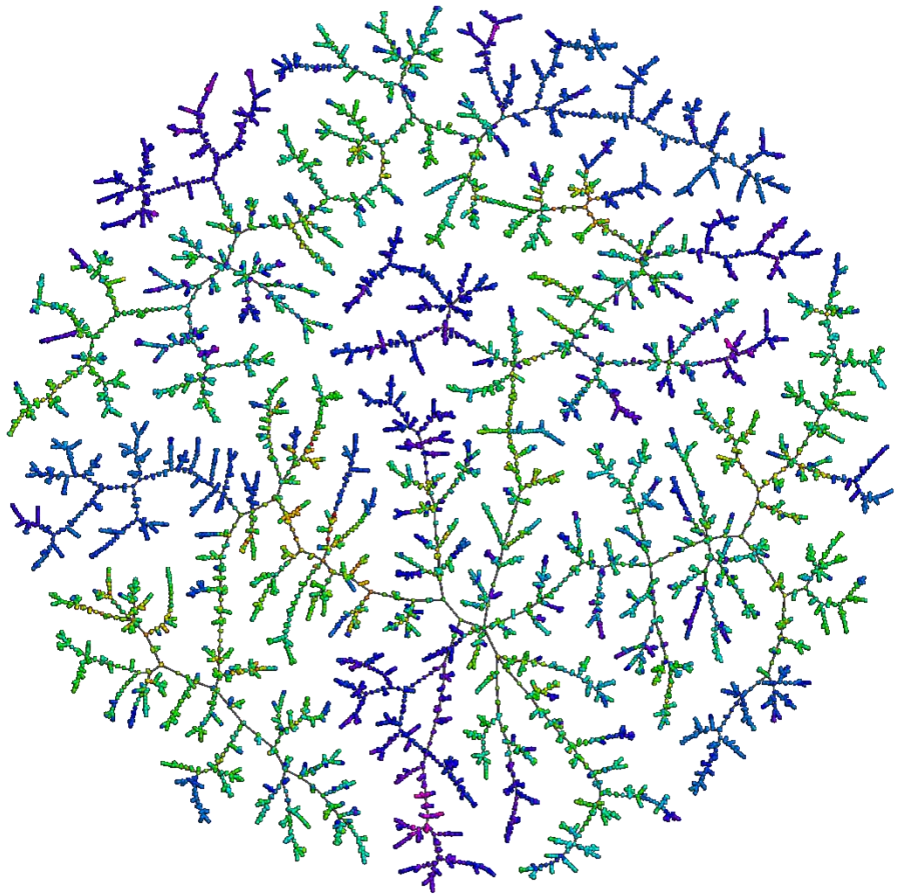
The GDB databases and their use for drug discovery

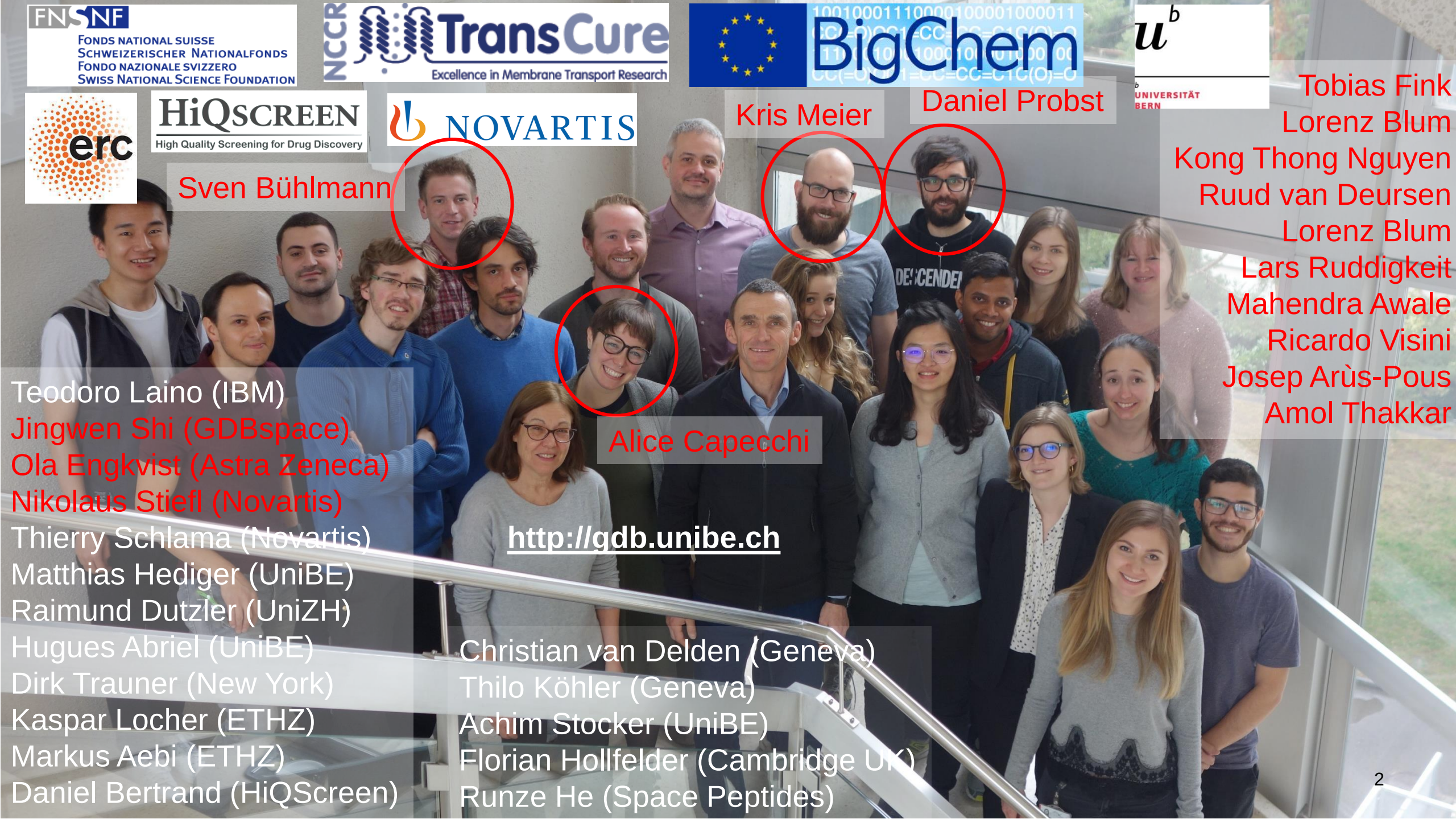
http://gdb.unibe.ch

Jean-Louis Reymond

December 2nd, 2020, NIH Virtual Workshop on Ultra-Large Chemistry Databases

1. Constructing GDB
2. Visualizing and Searching GDB
3. Synthesizing GDB

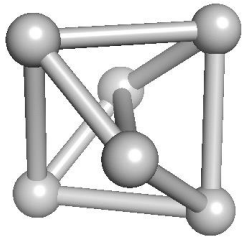



FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation
NCCR TransCure
Excellence in Membrane Transport Research
BigChem
UNIVERSITÄT BERN
erc
HiQSCREEN
High Quality Screening for Drug Discovery
NOVARTIS
Kris Meier
Daniel Probst
Sven Bühlmann
Alice Capecchi
Tobias Fink
Lorenz Blum
Kong Thong Nguyen
Ruud van Deursen
Lorenz Blum
Lars Ruddigkeit
Mahendra Awale
Ricardo Visini
Josep Arùs-Pous
Amol Thakkar
Teodoro Laino (IBM)
Jingwen Shi (GDBspace)
Ola Engkvist (Astra Zeneca)
Nikolaus Stiefl (Novartis)
Thierry Schlama (Novartis)
Matthias Hediger (UniBE)
Raimund Dutzler (UniZH)
Hugues Abriel (UniBE)
Dirk Trauner (New York)
Kaspar Locher (ETHZ)
Markus Aebi (ETHZ)
Daniel Bertrand (HiQScreen)
http://gdb.unibe.ch
Christian van Delden (Geneva)
Thilo Köhler (Geneva)
Achim Stocker (UniBE)
Florian Hollfelder (Cambridge UK)
Runze He (Space Peptides)

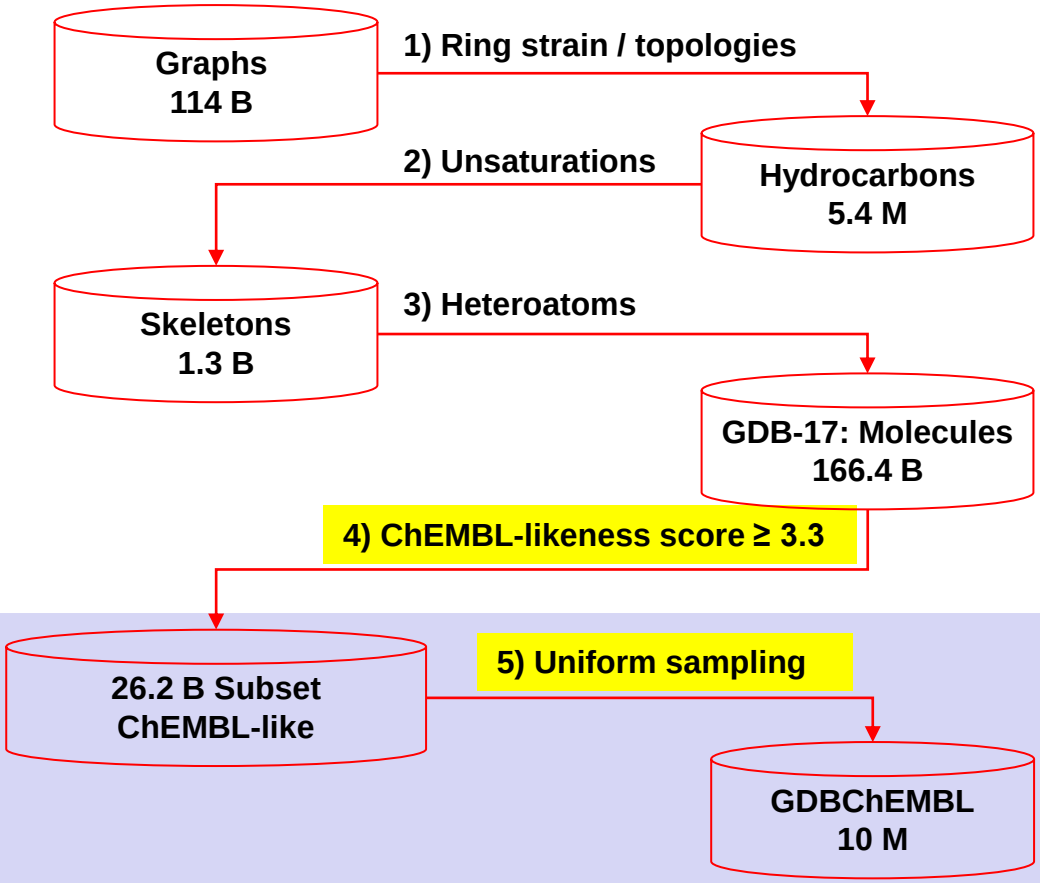
The Lipinski-graben ?

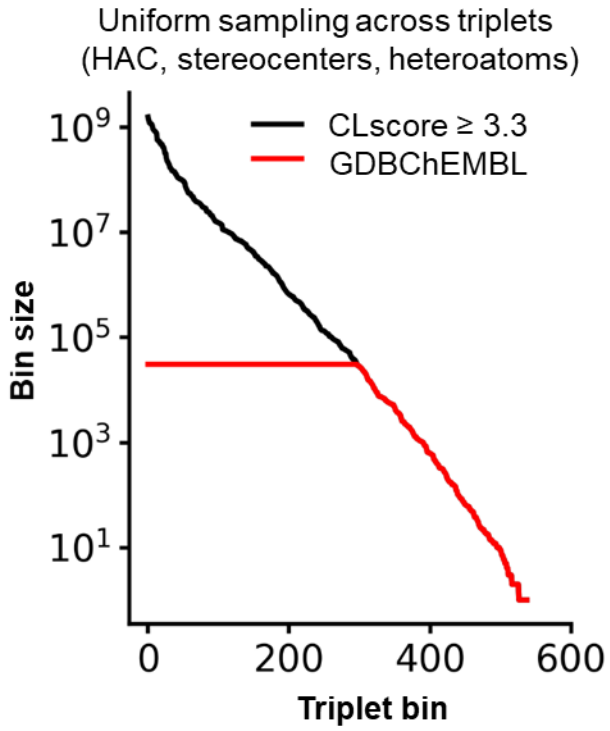
Academic molecules	Pharma molecules
Large	Small
Complex	Simple
Unlimited options	Limited options?
Useless?	Useful

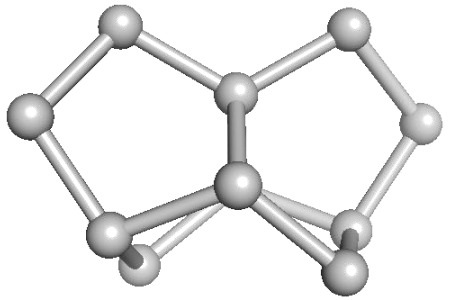
1. Constructing GDB



Claus Benzol
(1867)



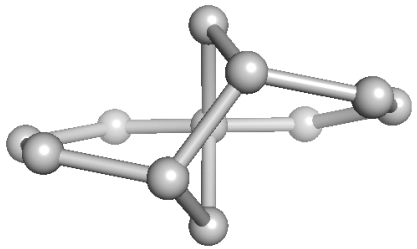


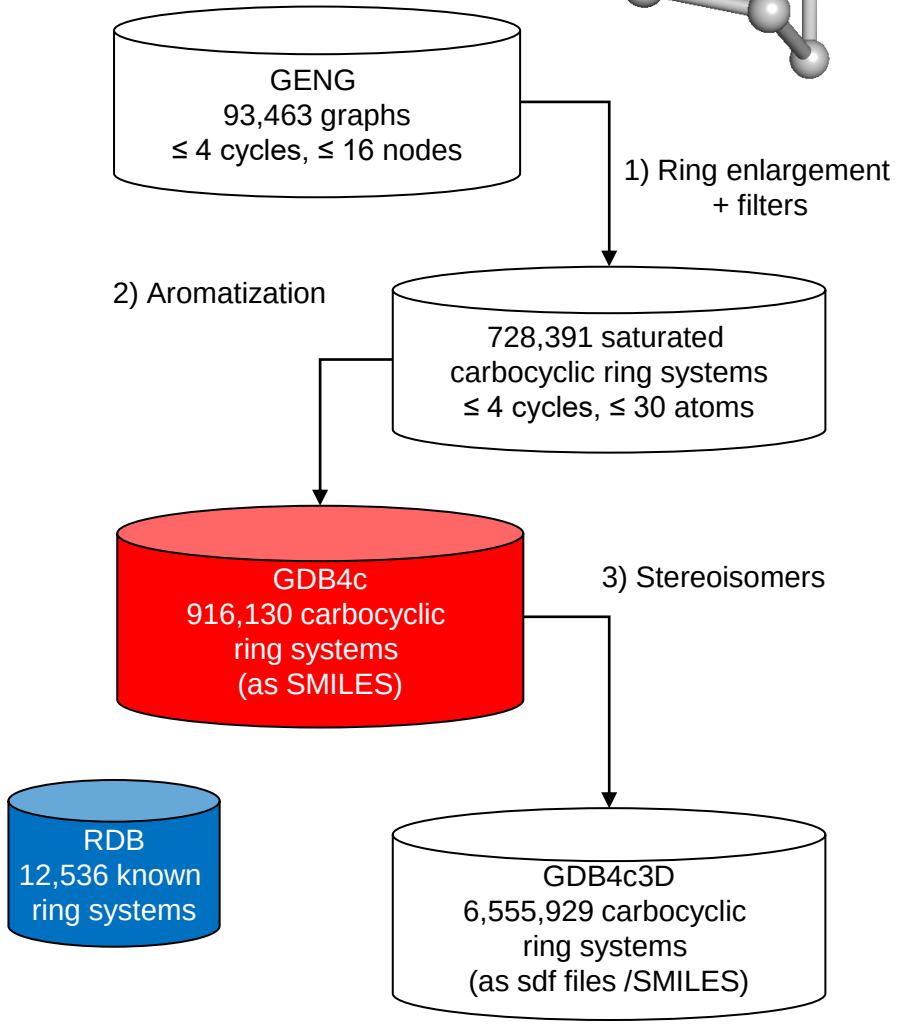


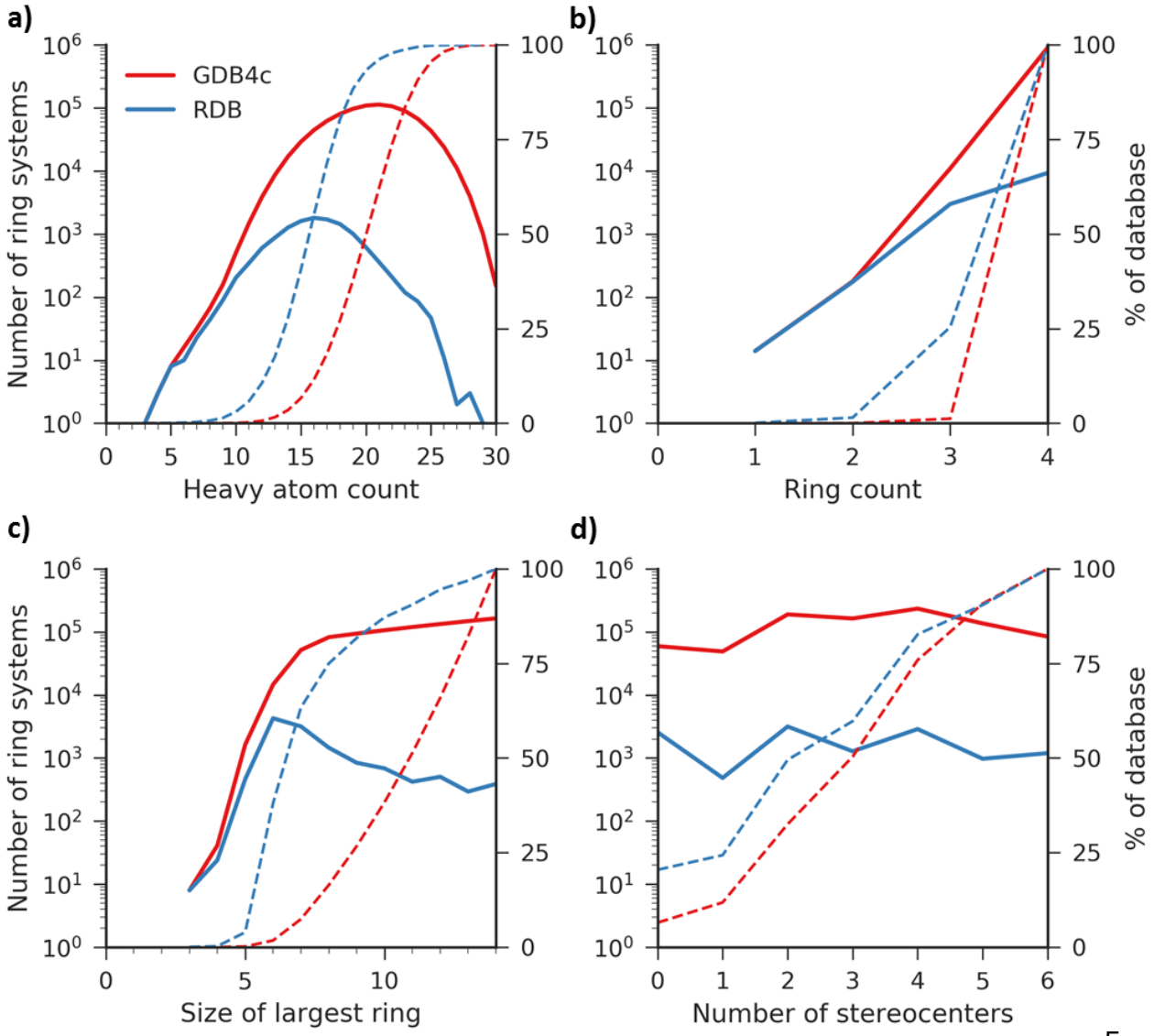
Trinorbornane
(2007)

Tobias Fink et al. *Angew. Chem. Int. Ed.* **2005**, *44*, 1504-1508, *J. Chem. Inf. Model.* **2007**, *47*, 342-353 (GDB-11)
Lorenz C. Blum et al., *J. Am. Chem. Soc.* **2009**, *131*, 8732-3 (GDB-13);
Lars Ruddigkeit et al.**,** *J. Chem. Inf. Model.* **2012,** *52*, 2864-2875 (GDB-17)
Ricardo Visini et al., *J. Chem. Inf. Model.* **2017,** *57*, 700-709 (FDB17), *J. Chem. Inf. Model.* **2017**, *57*, 2707-2718 (GDB4c)
Mahendra Awale et al., *Mol. Inf.* **2019**, *38*, 1900031 (GDBMedChem)
Sven Bühlmann et al., *Front. Chem.* **2020**, doi:10.3389/fchem.2020.00046

Ring Systems

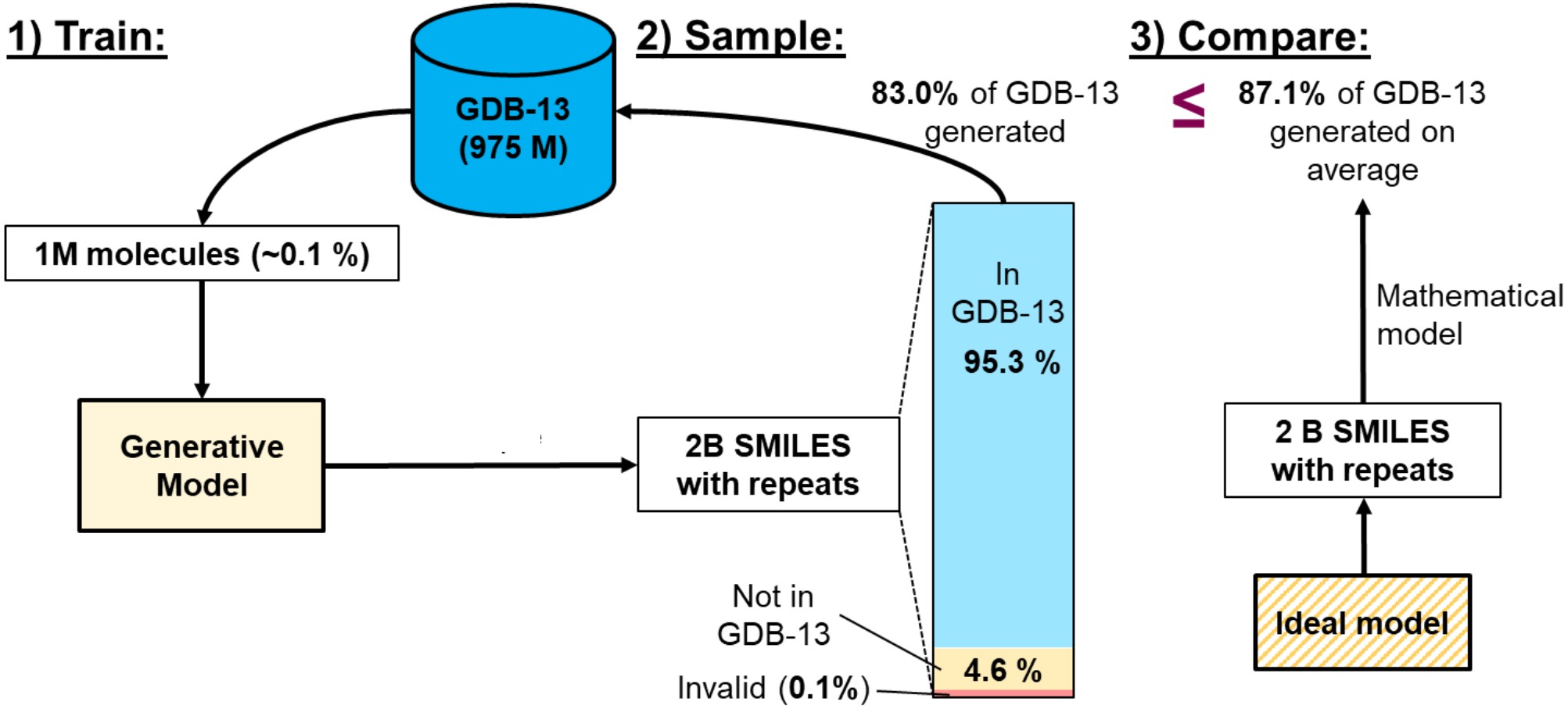






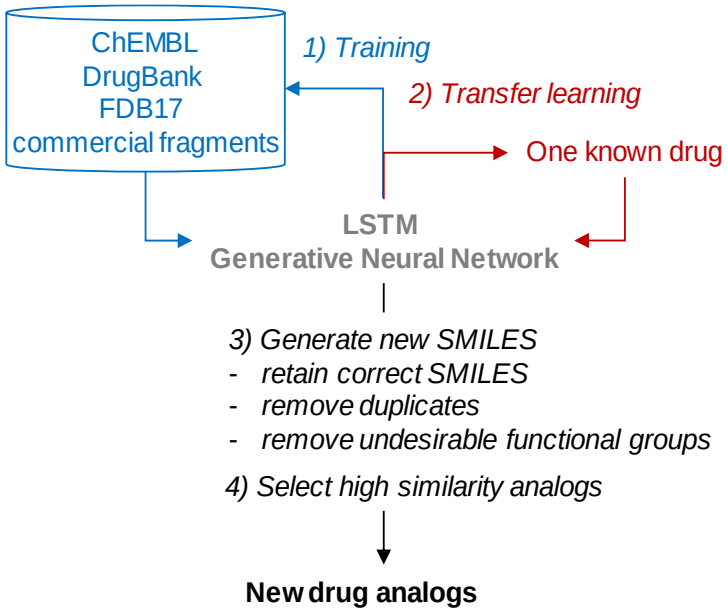
Ricardo Visini, Josep Arùs-Pous et al., *J. Chem. Inf. Model.* **2017,** *57*, 2707-2718

→ sampling chemical space using AI



Josep Arus-Pous et al., *J. Cheminf.* **2019**, *11*, 20

→ transfer learning



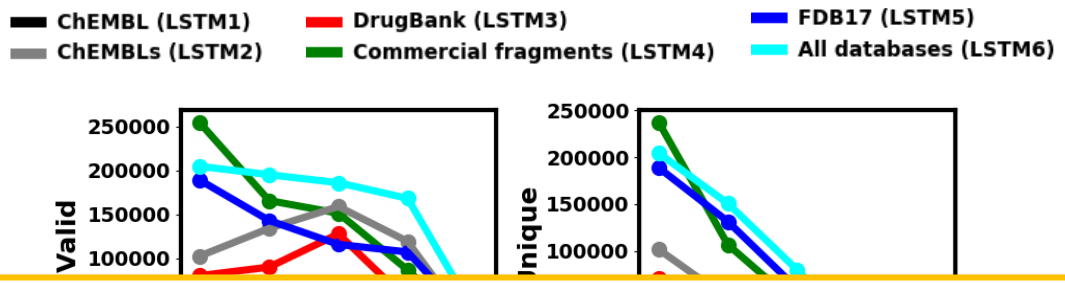
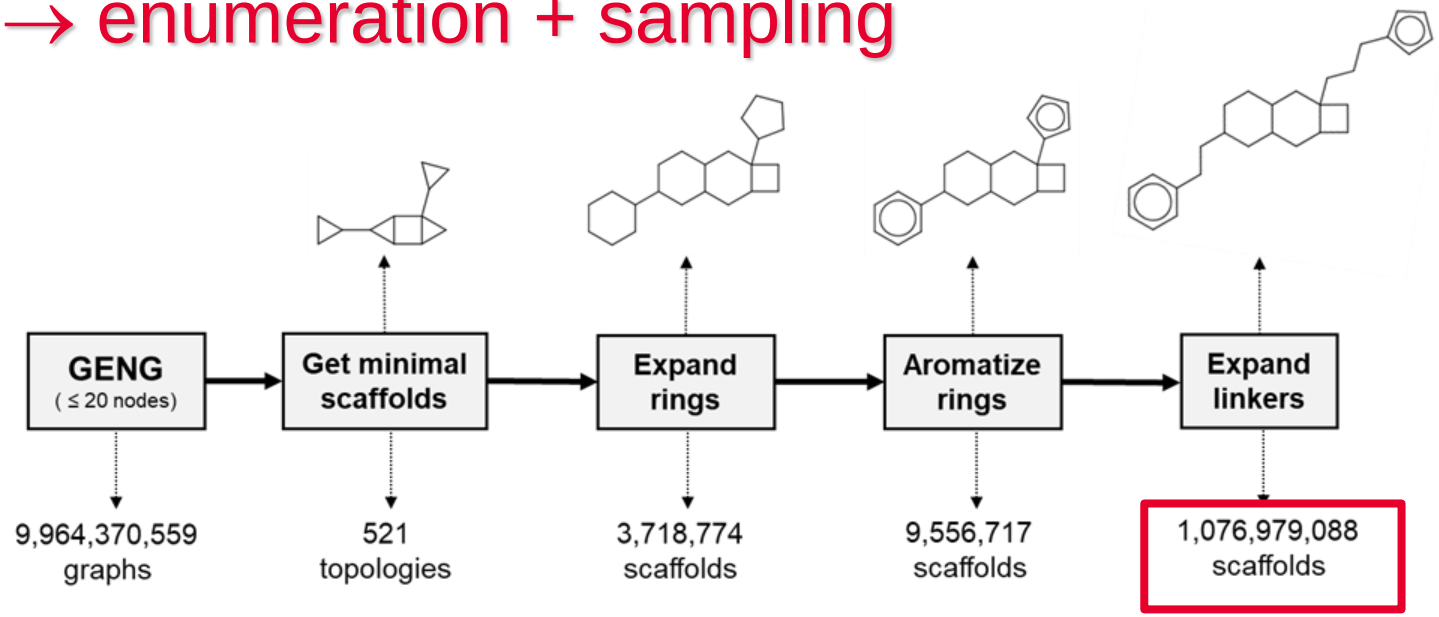


Table 2. Number of unique/total high similarity drug analogs produced by the different LSTM neural networks.

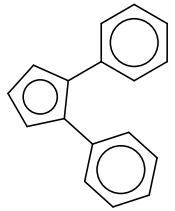
Neural Network	**LSTM1**	**LSTM2**	**LSTM3**	**LSTM4**	**LSTM5**	**LSTM6**	
Source database	ChEMBL	ChEMBLs	DrugBank	Commercial Fragments	FDB17	All databases	Unique across LSTMs
training cpds.	344,319	40,000	5,104	40,986	500,000	890,409	
Nicotine	0/23	32/82	1/32	32/93	9/47	16/67	166
Fencamfamine	15/42	126/218	40/96	130/231	92/164	41/114	580
Aminophenazone	5/26	34/96	23/71	38/99	22/66	19/65	223
Sulfadiazine	6/27	19/59	11/37	28/74	8/30	2/25	124
Miconazole	2/10	301/500	268/438	174/336	0/0	153/256	1134
Roflumilast	8/15	319/557	117/283	351/585	0/0	45/166	1126
Lovastatin	0/1	631/986	460/757	352/625	487/728	289/530	2729
Epothilone D	0/1	911/1301	561/831	807/1160	1595/2039	1163/1511	5707
Nilotinib	0/1	506/666	180/321	218/381	0/0	243/355	1362
Erythromycin	0/2	832/1042	174/243	524/709	1243/1444	1105/1288	4190

Mahendra Awale et al., *J. Chem. Inf. Model.* **2019**, *59*, 1347-1356

→ enumeration + sampling

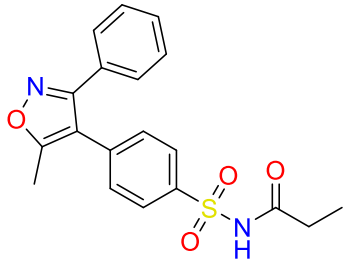


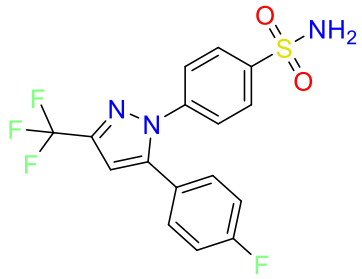
GDBscaffolds



c1(c2c(c3ccccc3)ccc2)ccccc1

LSTM or Transformer



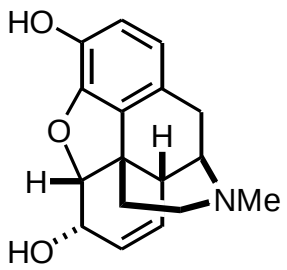


O=S(c1ccc(n2c(c3ccc(F)cc3)cc(C(F)(F)F)n2)cc1)(N)=O

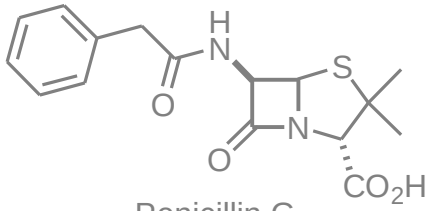
Training set	Architecture	% Invalid	% Correct	% Recovered	% Unique
ChEMBL	LSTM	8.66 %	61.22 %	16.65 %	54.03 %
	Transformer	1.73 %	87.87 %	36.25 %	58.33 %
ZINC	LSTM	2.62 %	88.10 %	25.94 %	59.56 %
	Transformer	0.52 %	96.67 %	31.17 %	51.94 %

Josep Arus-Pous, PhD thesis, **2020**

2. Visualizing and searching GDB



Morphine



Penicillin G

Atoms

Carbon	**17**	16
Fluorine	**0**	0
Chlorine	**0**	0
Bromine	**0**	0
Iodine	**0**	0
Sulphur	**0**	1
Phosphor	**0**	0
Acyclic nitrogen	**0**	1
Cyclic nitrogen	**1**	1
Acyclic oxygen	**2**	4
Cyclic oxygen	**1**	0
Heavy atom count	**21**	23

Bonds

Acyclic single bonds	**3**	8
Acyclic double bonds	**0**	3
Acyclic triple bonds	**0**	0
Cyclic single bonds	**18**	11
Cyclic double bonds	**4**	3
Cyclic triple bonds	**0**	0
Rotatable bonds	**0**	4

Polar groups

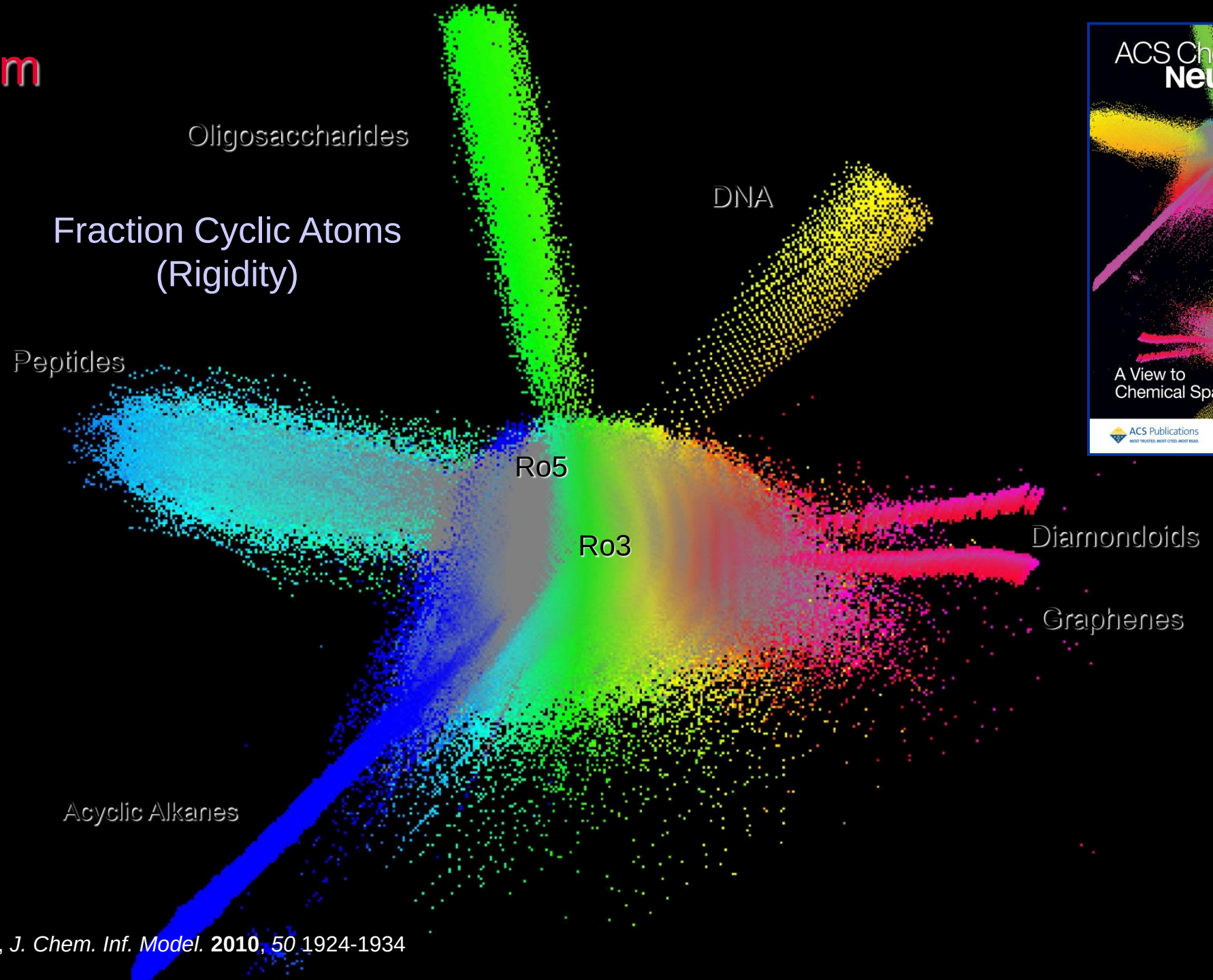
H-Bond donor atoms	**3**	1
H-Bond donor sites	**3**	1
H-Bond acceptor atoms	**3**	4
H-Bond acceptor sites	**3**	7
Positive charges	**1**	0
Negative charges	**0**	1

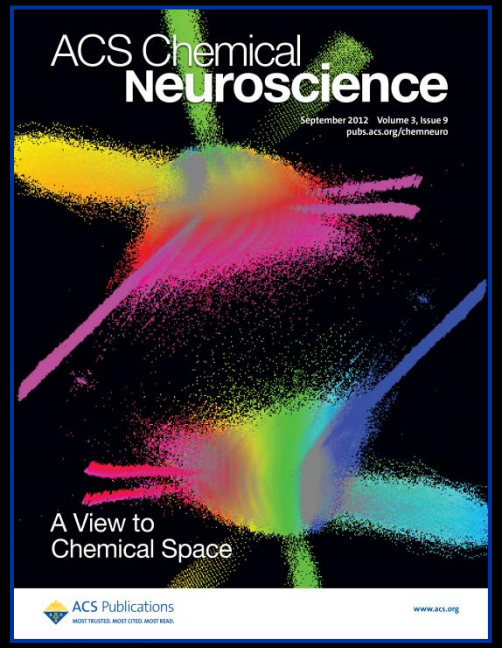
Topology

Acyclic monovalent nodes	**3**	6
Acyclic divalent nodes	**0**	2
Acyclic trivalent nodes	**0**	2
Acyclic tetravalent nodes	**0**	0
Cyclic divalent nodes	**8**	6
Cyclic trivalent nodes	**9**	6
Cyclic tetravalent nodes	**1**	1
3-Membered rings	**0**	0
4-Membered rings	**0**	1
5-Membered rings	**1**	1
6-Membered rings	**4**	1
7-Membered rings	**0**	0
8-Membered rings	**0**	0
9-Membered rings	**0**	0
≥ 10 membered rings	**0**	0
Atoms shared by fused rings	**7**	2
Bonds shared by fused rings	**6**	1

Kong Thong Nguyen et al., *ChemMedChem* **2009**, *4*, 1803-1805

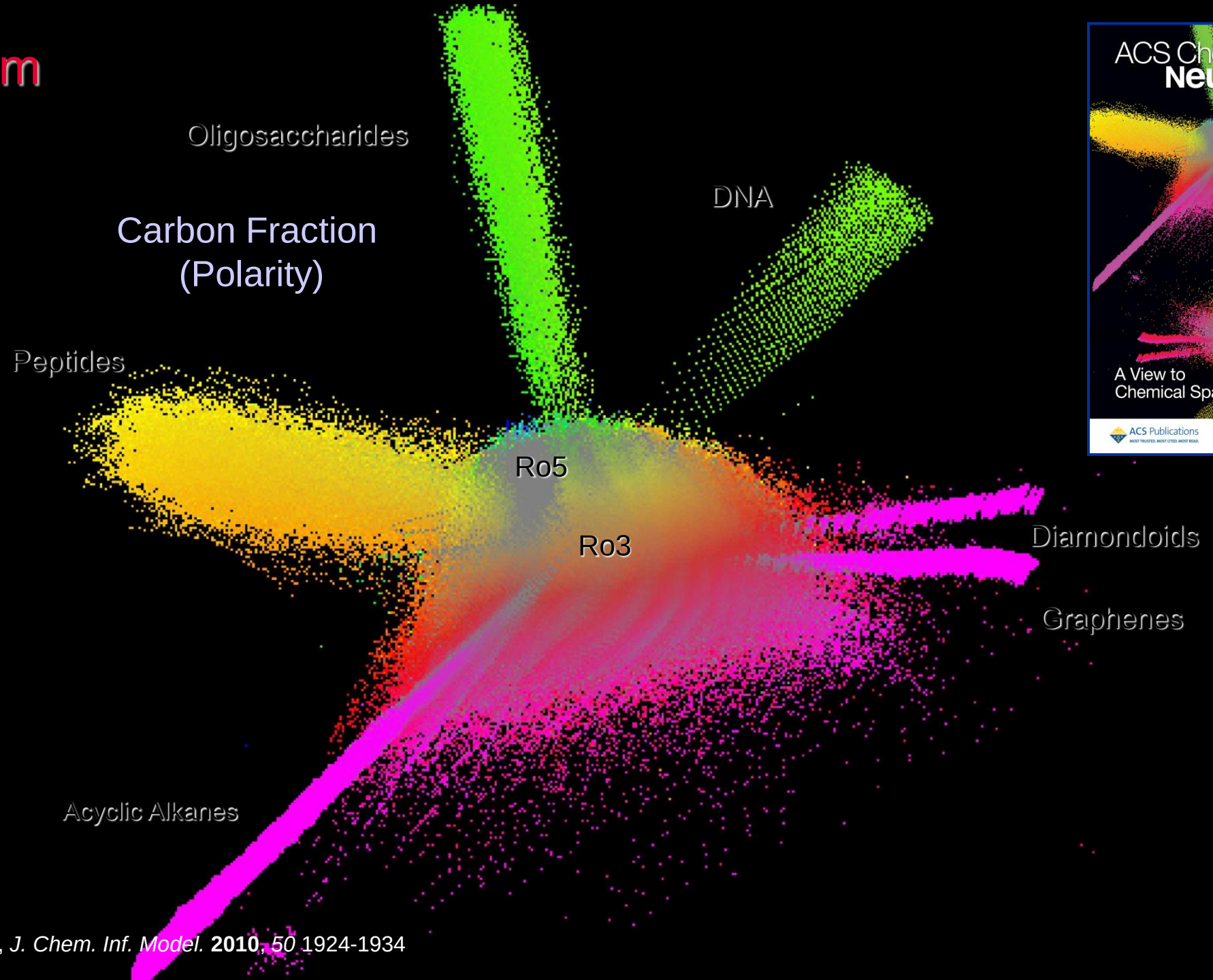
PubChem

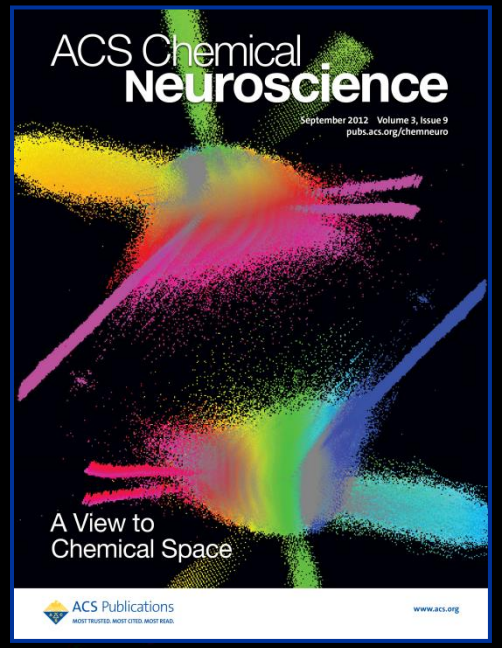




Ruud van Deursen et al., *J. Chem. Inf. Model.* **2010**, *50* 1924-1934

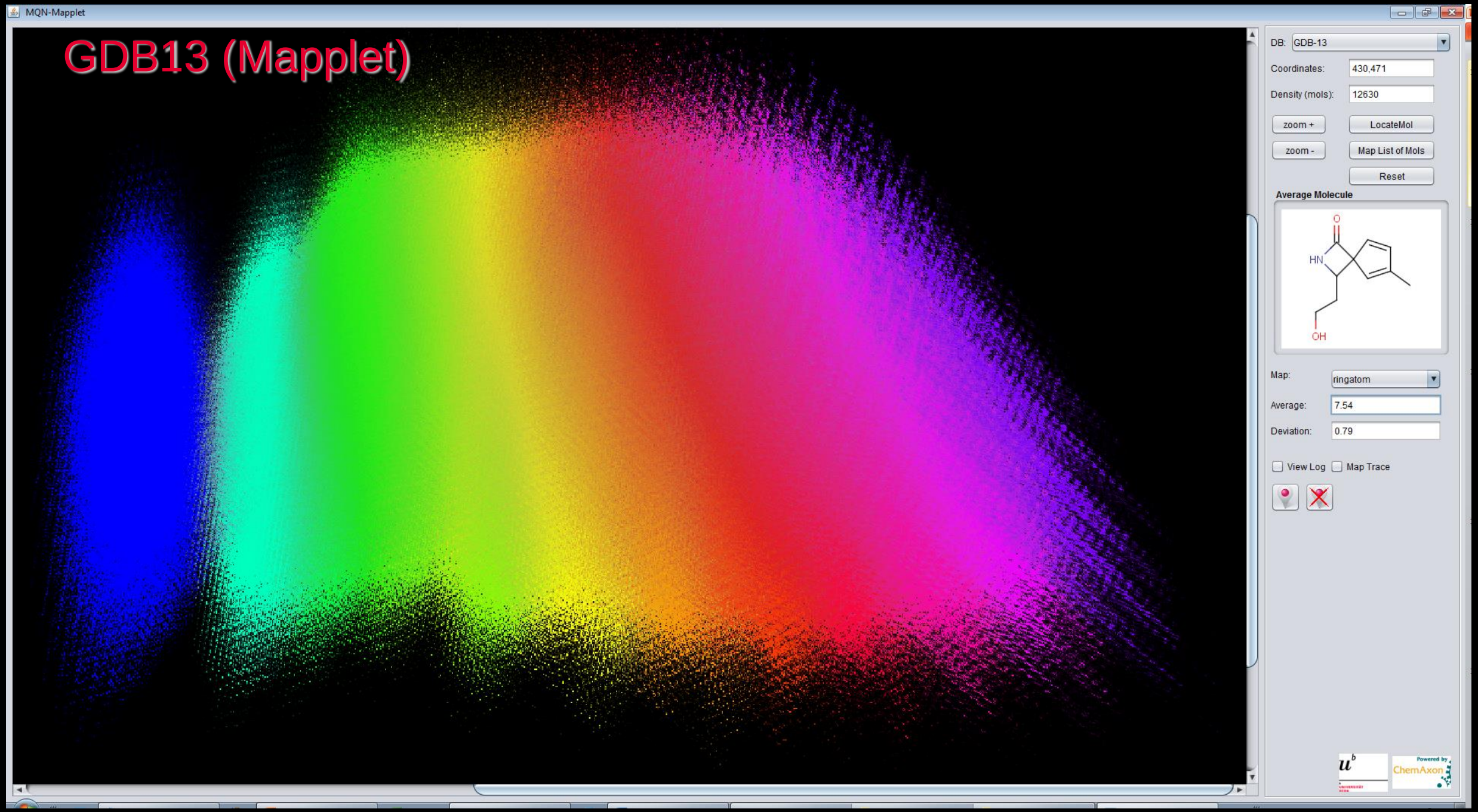
PubChem





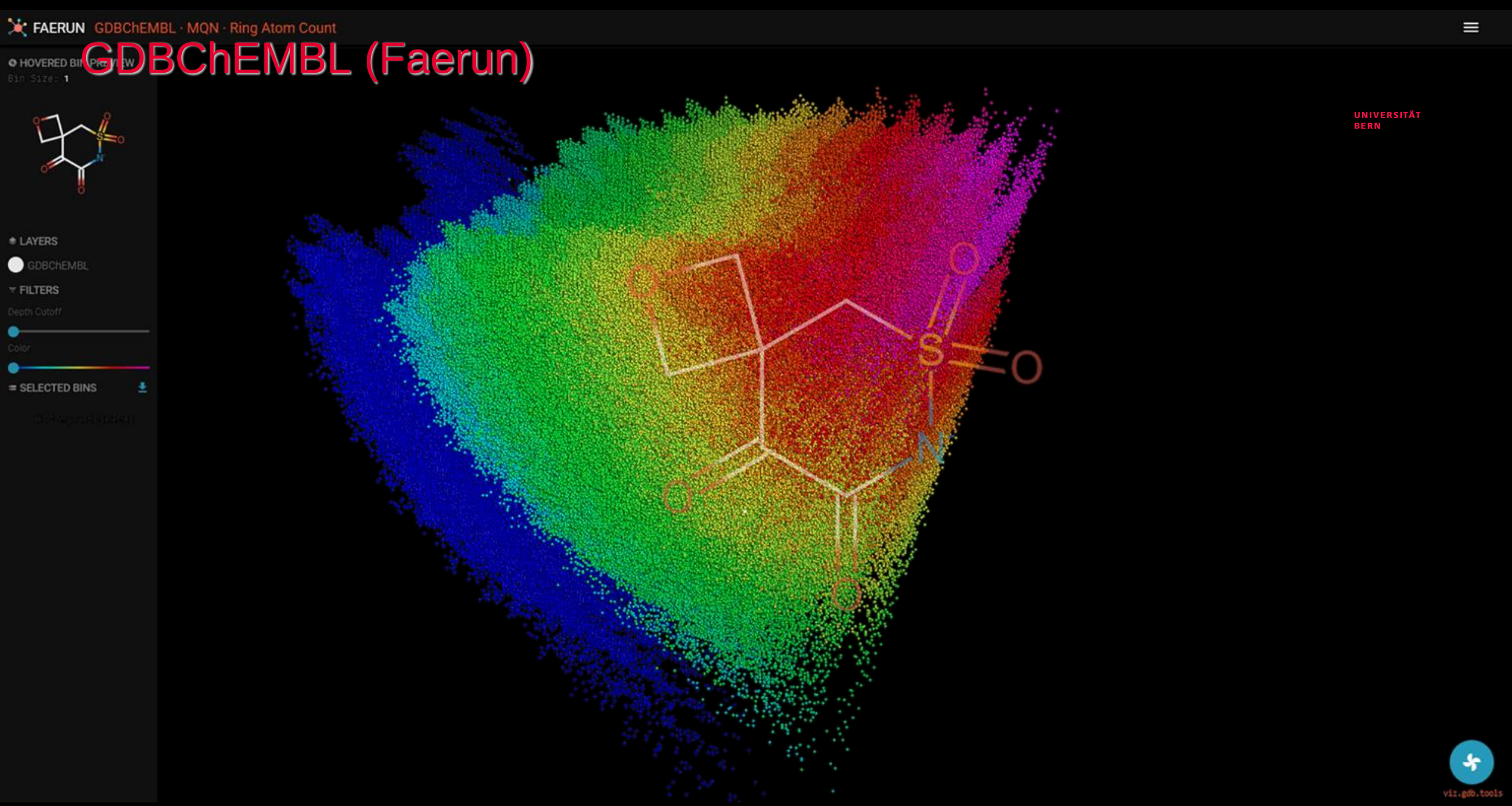
Ruud van Deursen et al., *J. Chem. Inf. Model.* **2010**, *50* 1924-1934

GDB13 (Mapplet)



Mahendra Awale, Ruud van Deursen et al, *J. Chem. Inf. Model.* **2013**, *53*, 509-518

GDBChEMBL (Faerun)



Daniel Probst *et al.*, *Bioinformatics*, **2018**, *34*, 1433-1435, *J. Chem. Inf. Model.* **2018**, *58*, 1–7

NN-searches in MQN-space

a) X-MQN string

MQN-sum | rest MQN to last non-zero value | Mask

A90 211/0A14 0A107206204A1303005010000300A16; C13H14N3 255 C1C2C[NH2+]CC1C1=CC3=C(C=C21)N=CC=N3

MQN 1-3/ MQN 4-5 | Formula | SMILES

b) Hash table file entry

File A90:

Mask 1 | Mask 2

211/0A14 0A107206204A1303005010000300A16; C13H14N3 239 111 255 116

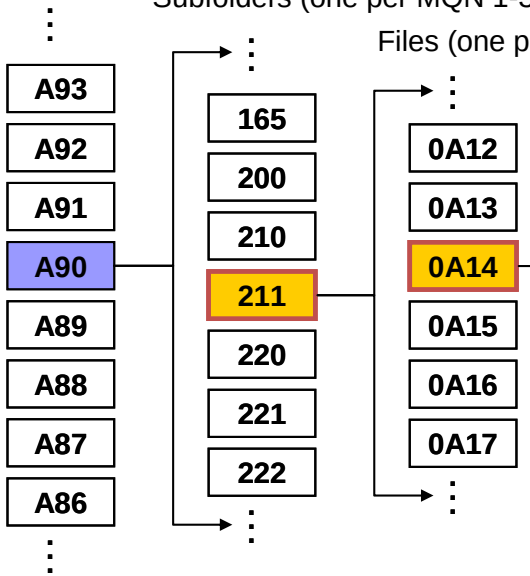
number of molecules with Mask values

c) MOLTREE

Folders (one per MQN sum, 120 MQN sums)

Subfolders (one per MQN 1-3 combination, 264 subfolders in folder A90)

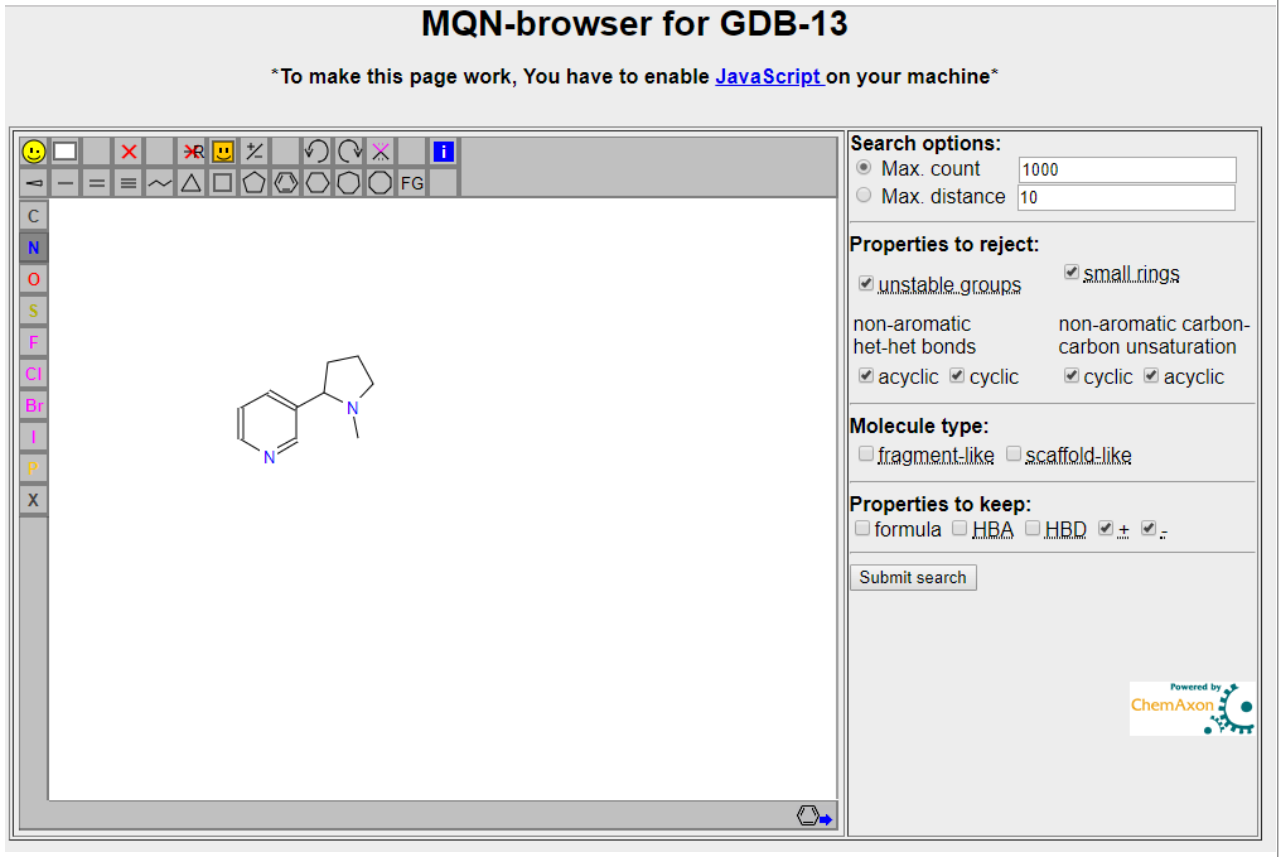
Files (one per MQN 4-5 combination, 45 files in subfolder 211)



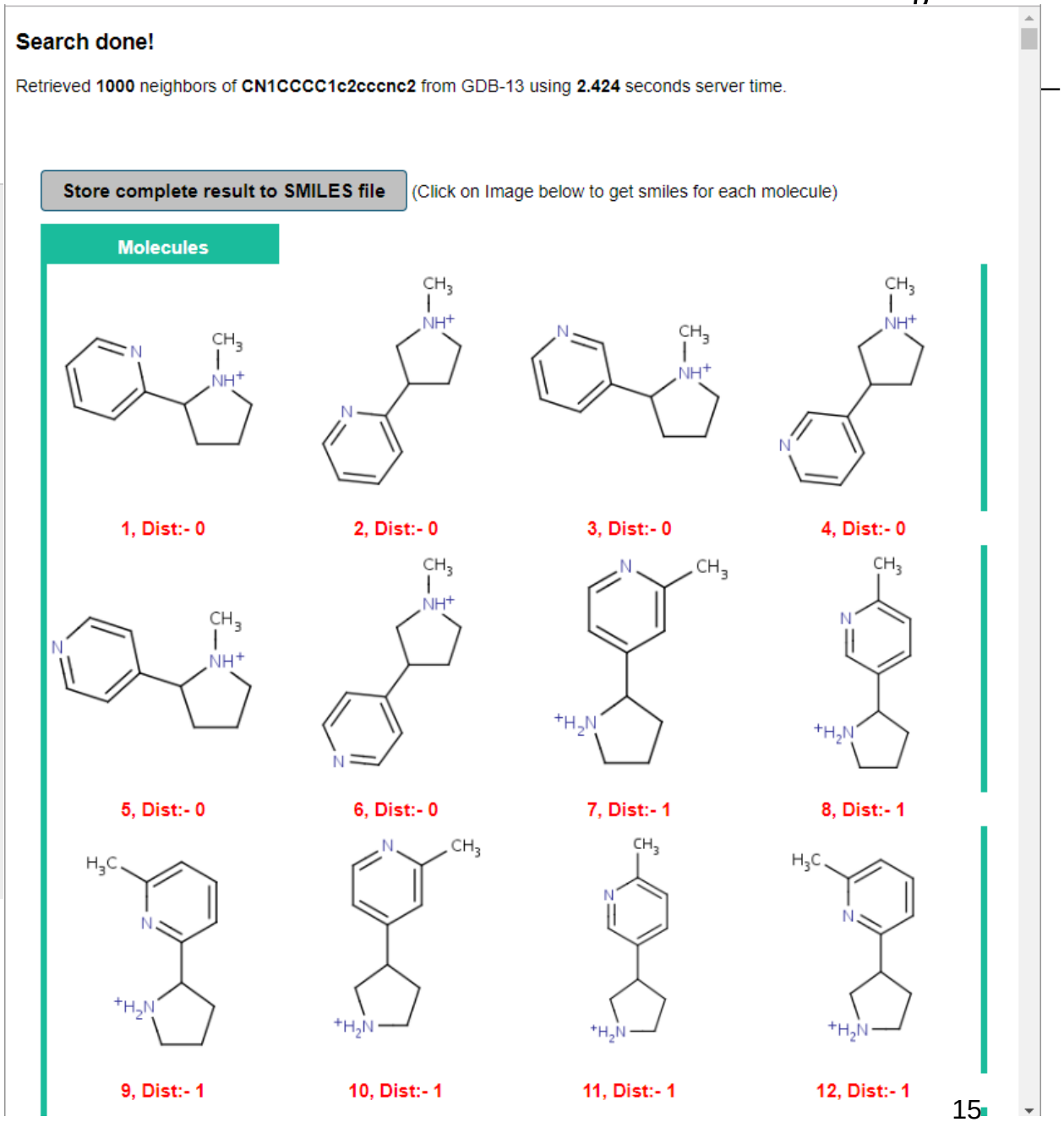
File 0A14 (83,799 MQN-bins, 41,761,344 SMILES):

....0A107206204A1303005010000300A16;
C13H14N3....255 .....C1C2C[NH2+]CC1C1
=CC3=C(C=C21)N=CC=N3;....

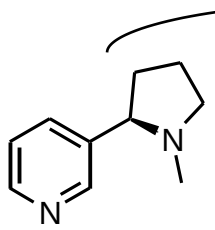
Lars Ruddigkeit, Lorenz C. Blum et al. *J. Chem. Inf. Model.* **2013**, *53*, 56-65



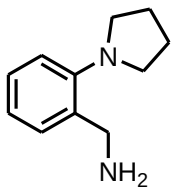
b



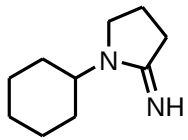
Lorenz C. Blum et al., *J. Chem. Inf. Model.* **2011,** *51*, 3105-3112



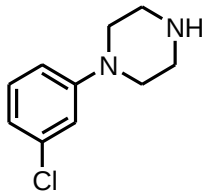
31504 MQN-neighbours
60 purchased and tested



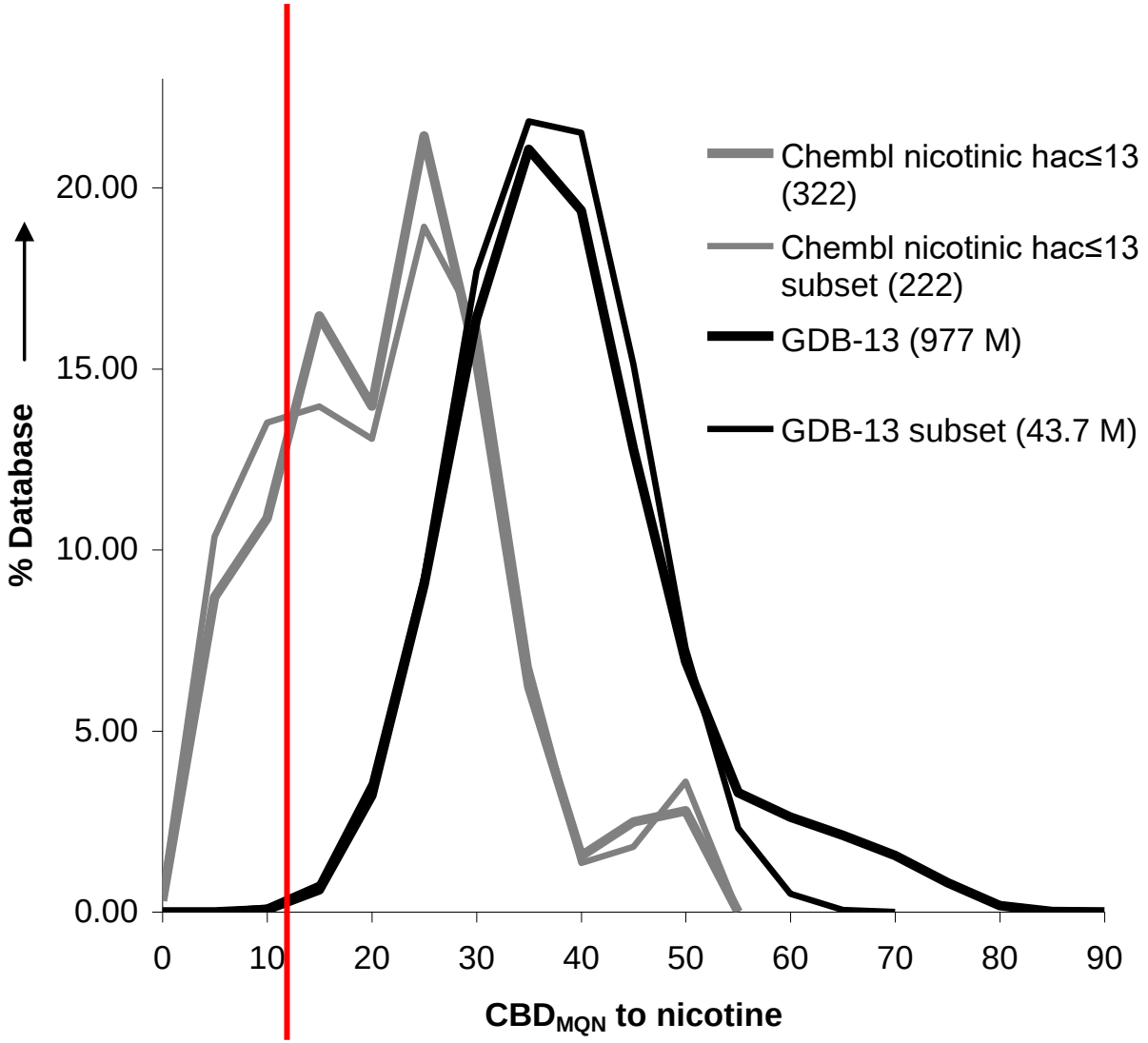
11
CBD_{MQN} = 9
antagonist
IC_{50} = 7.0 μM



18
CBD_{MQN} = 11
antagonist
IC_{50} = 9.2 μM

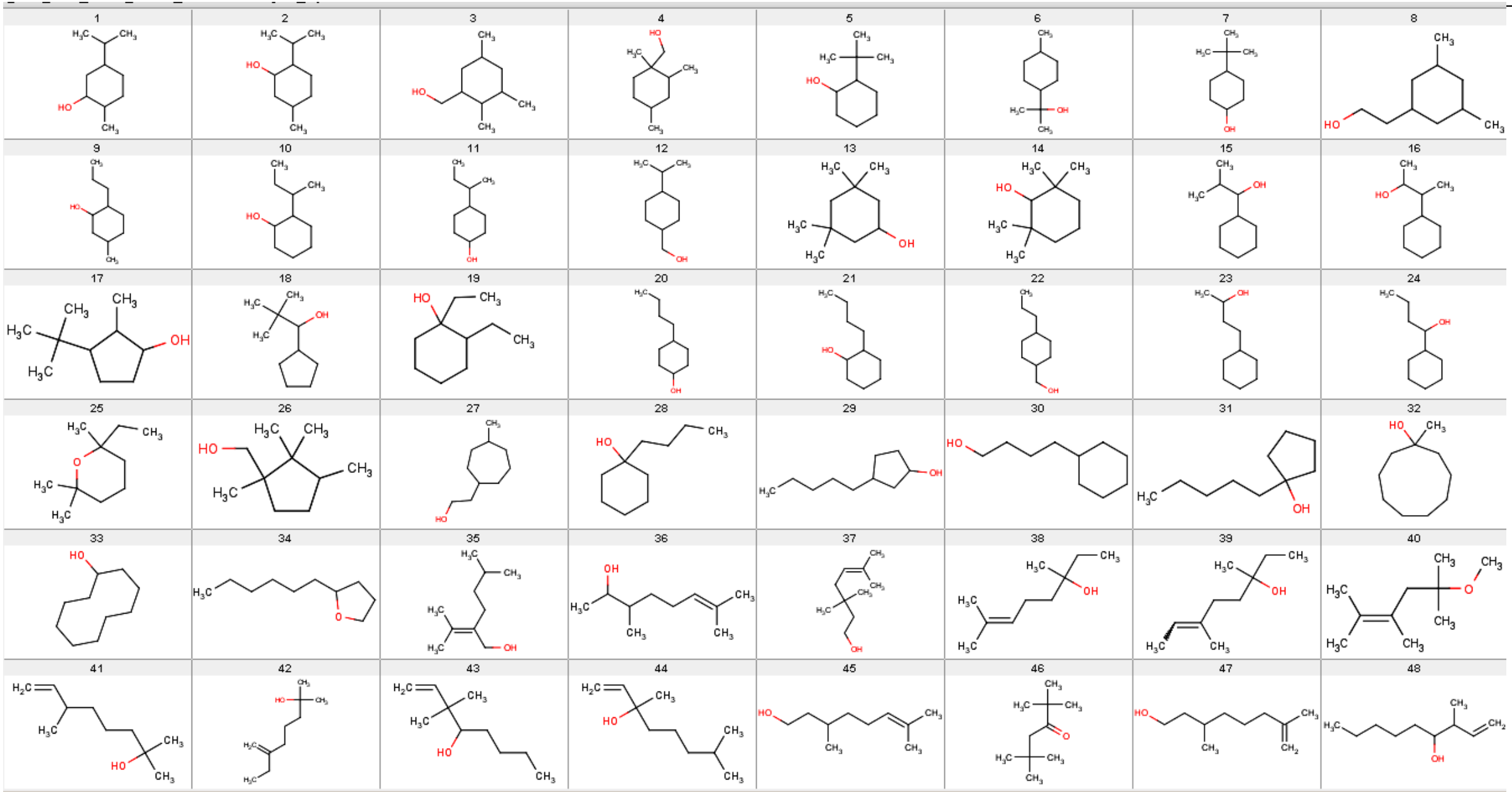


47
CBD_{MQN} = 7
antagonist
IC_{50} = 6.2 μM

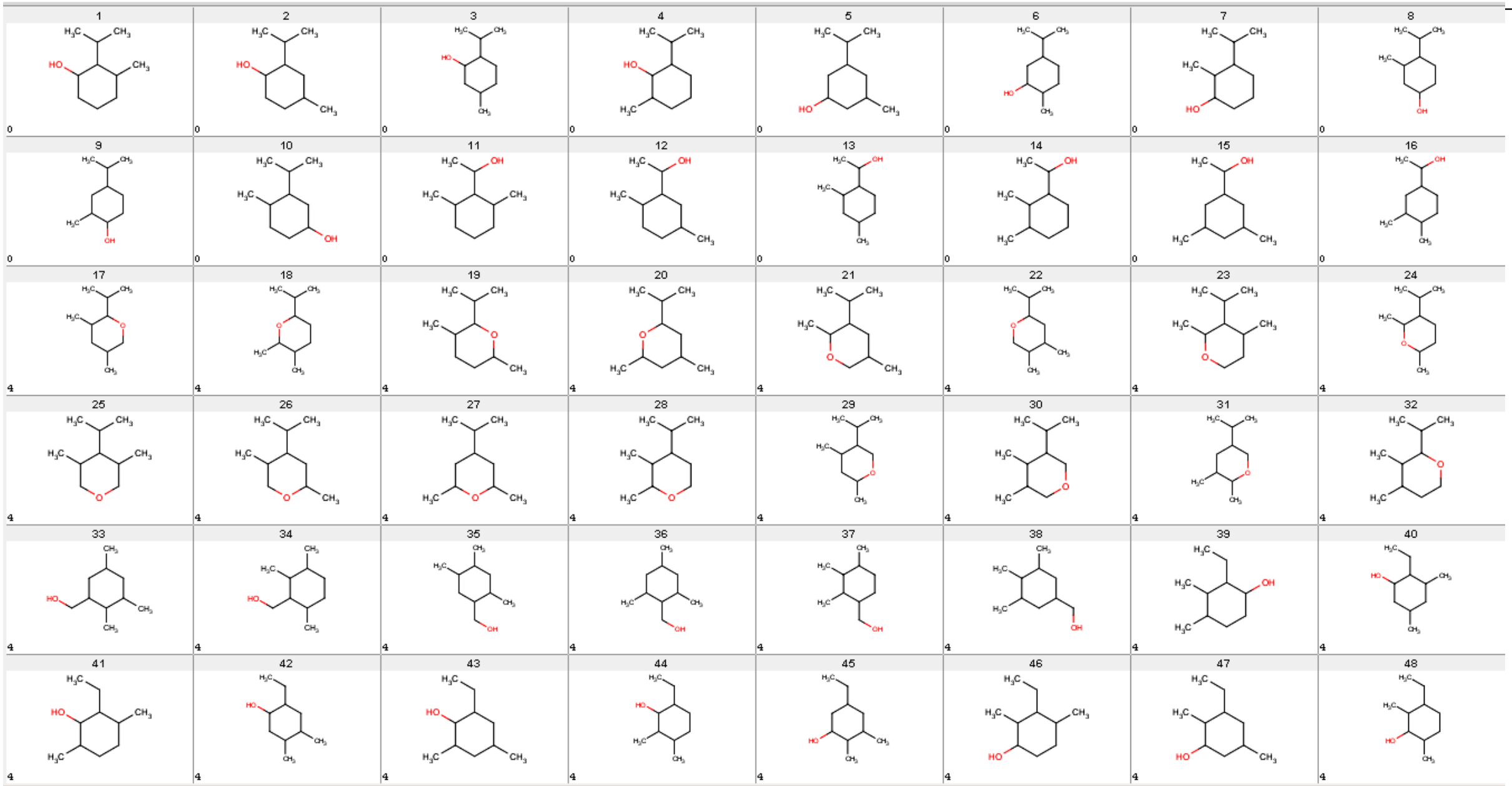


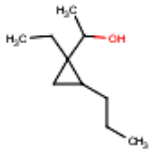
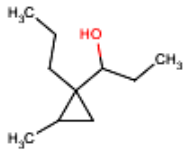
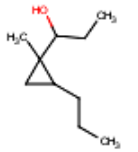
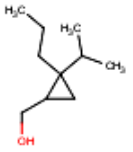
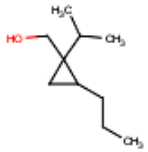
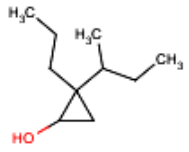
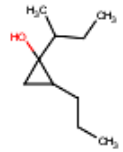
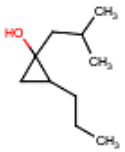
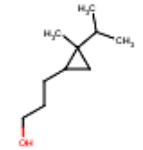
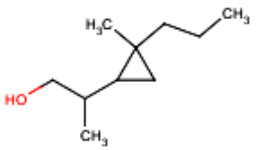
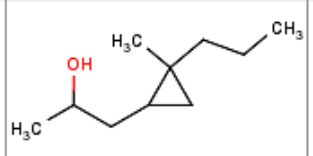
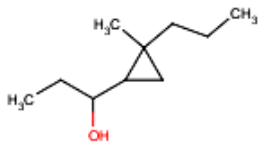
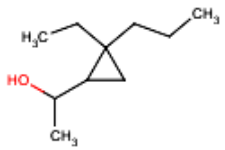
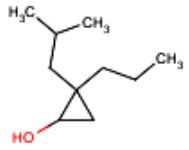
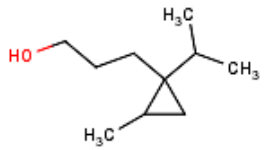
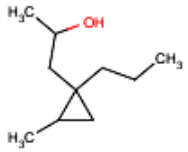
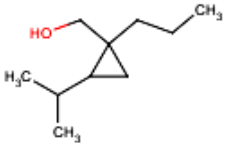
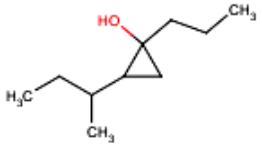
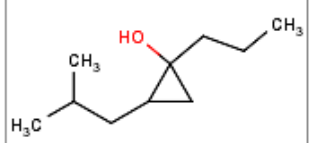
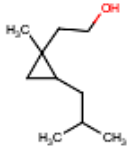
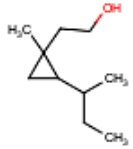
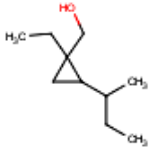
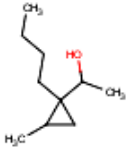
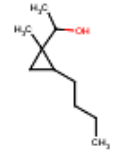
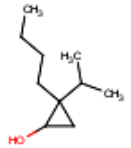
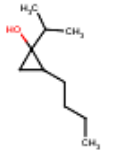
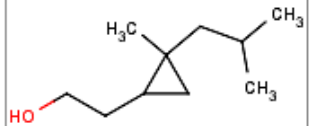
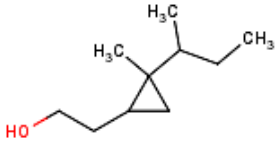
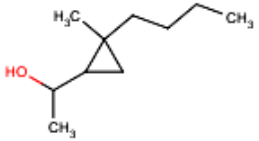
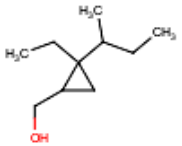
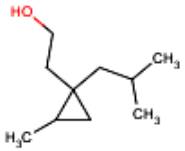
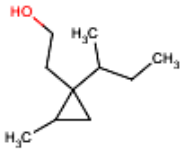
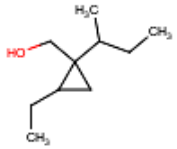
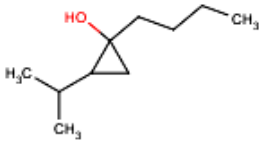
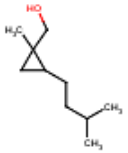
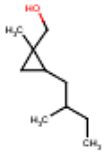
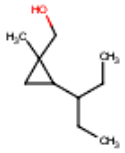
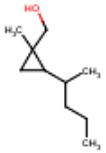
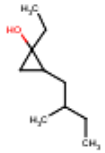
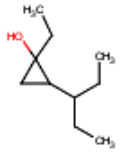
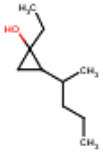
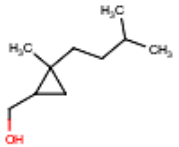
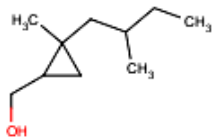
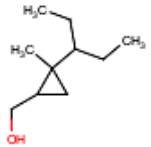
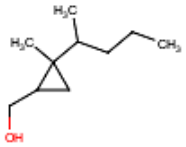
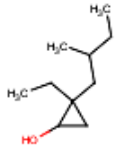
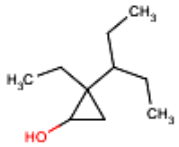
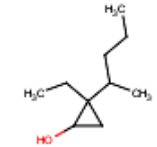
Lorenz C. Blum et al., *J. Chem. Inf. Model.* **2011,** *51*, 3105-3112

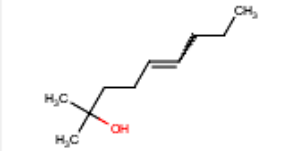
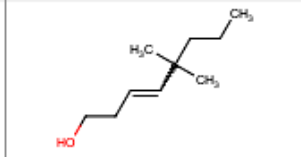
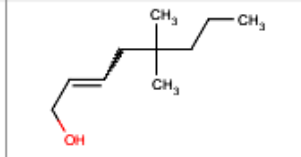
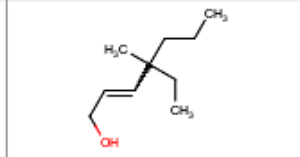
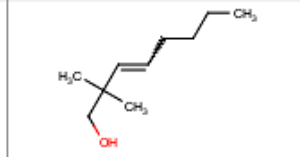
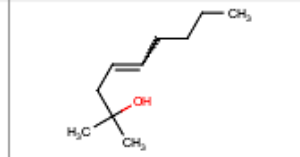
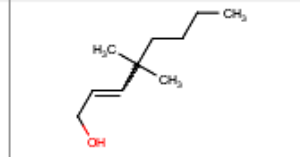
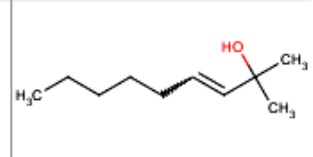
Menthol – ZINC isomers

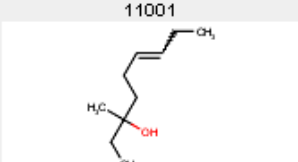
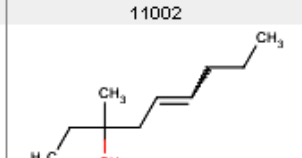
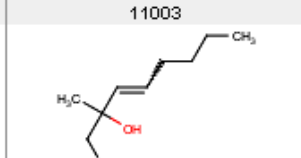
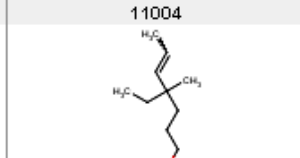
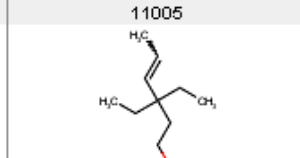
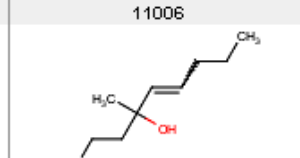
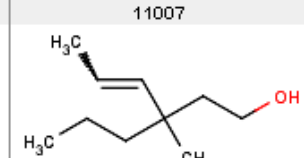
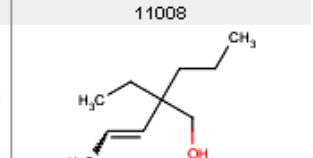


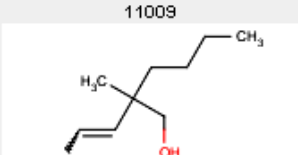
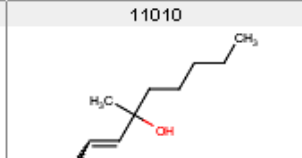
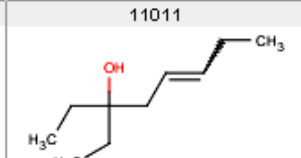
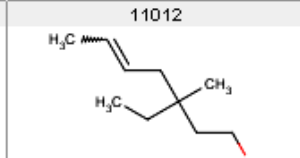
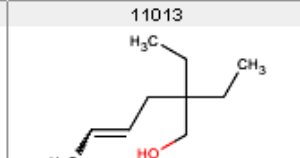
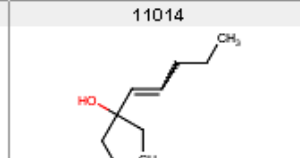
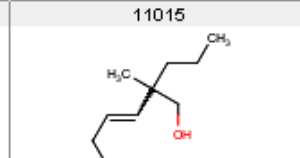
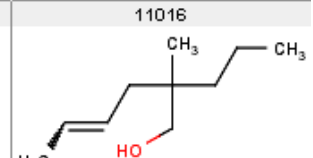
Menthol – GDB-13 isomers

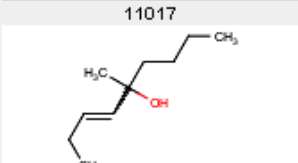
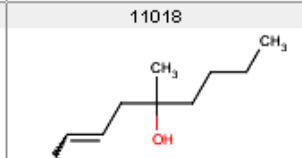
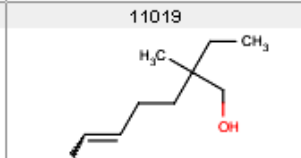
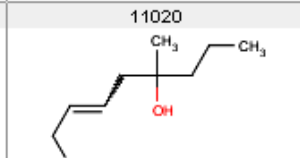
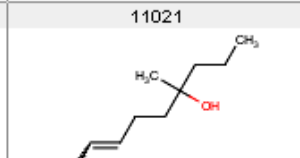
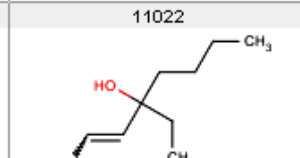
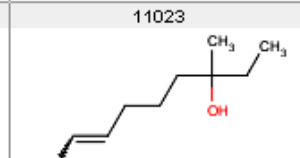
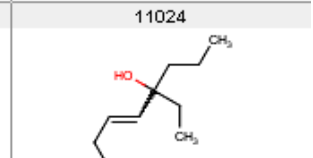


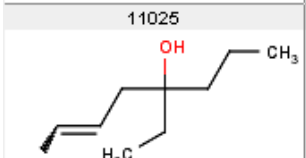
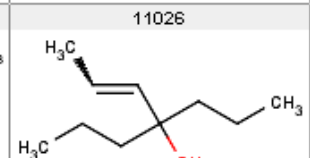
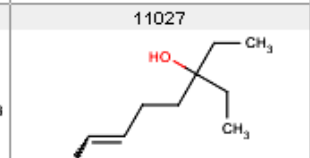
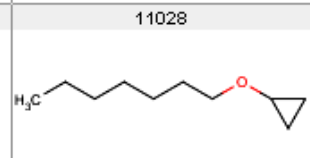
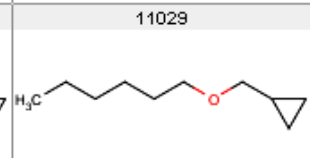
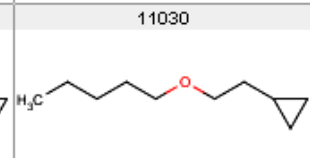
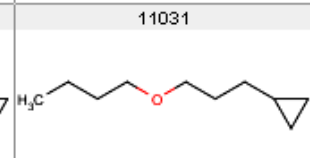
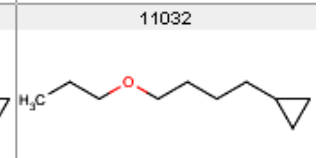
4993	4994	4995	4996	4997	4998	4999	5000
							
5001	5002	5003	5004	5005	5006	5007	5008
							
5009	5010	5011	5012	5013	5014	5015	5016
							
5017	5018	5019	5020	5021	5022	5023	5024
							
5025	5026	5027	5028	5029	5030	5031	5032
							
5033	5034	5035	5036	5037	5038	5039	5040
							

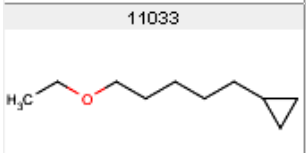
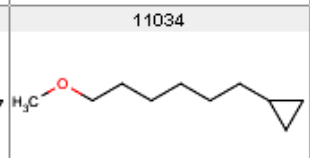
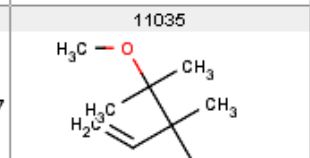
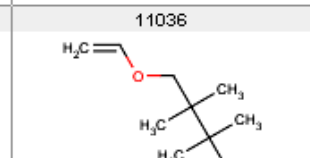
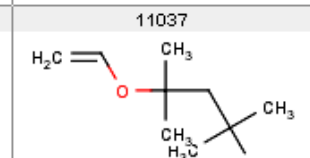
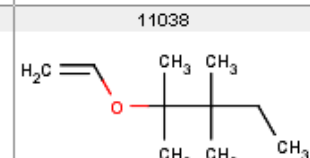
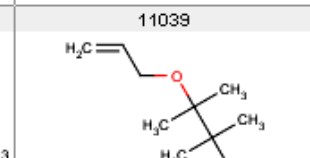
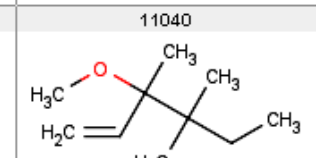
10993	10994	10995	10996	10997	10998	10999	11000
							
30	30	30	30	30	30	30	30

11001	11002	11003	11004	11005	11006	11007	11008
							
30	30	30	30	30	30	30	30

11009	11010	11011	11012	11013	11014	11015	11016
							
30	30	30	30	30	30	30	30

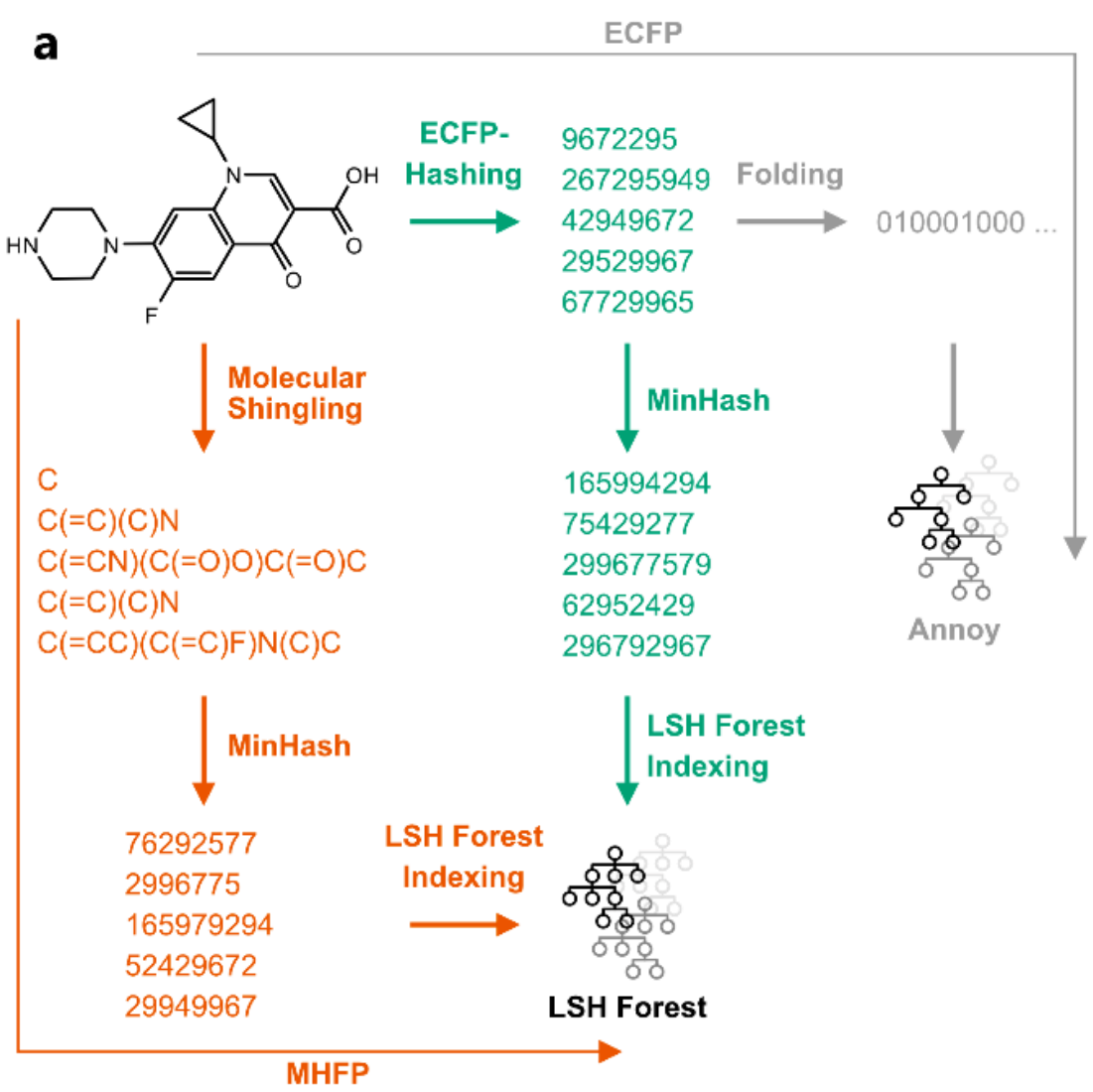
11017	11018	11019	11020	11021	11022	11023	11024
							
30	30	30	30	30	30	30	30

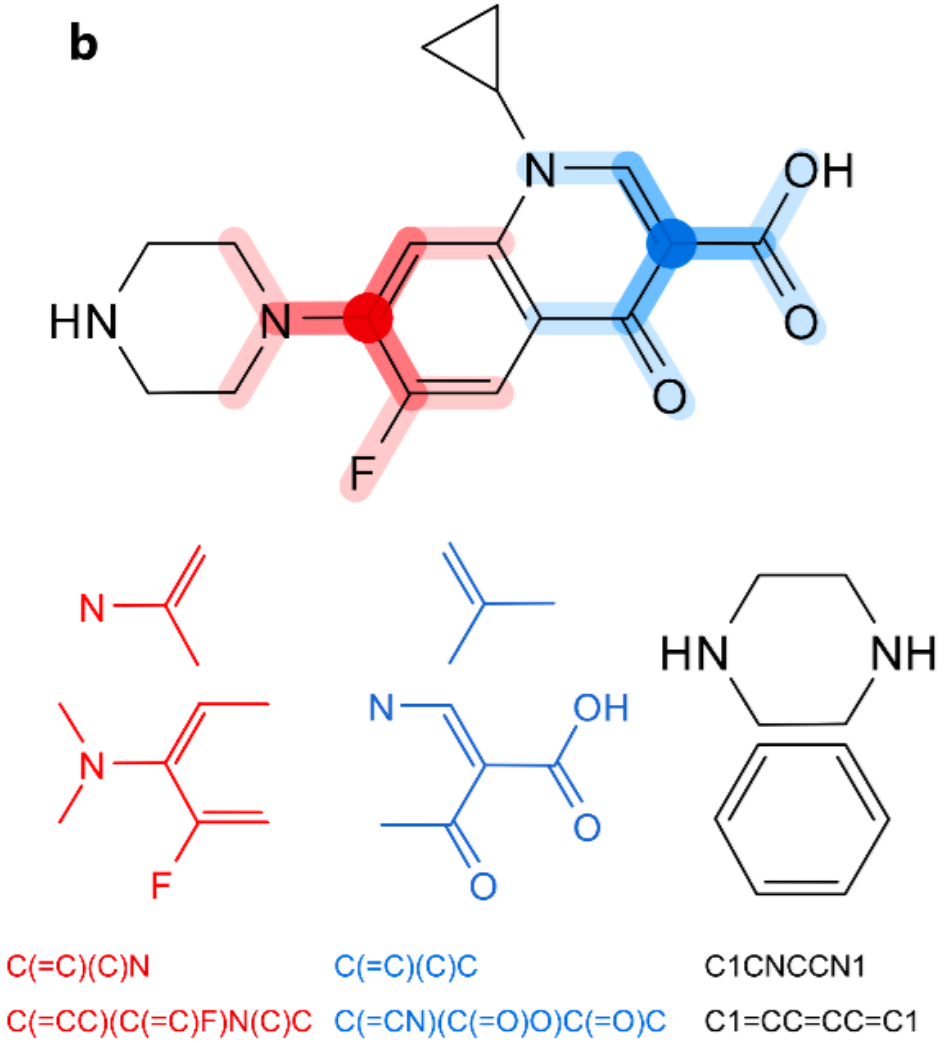
11025	11026	11027	11028	11029	11030	11031	11032
							
30	30	30	30	30	30	30	30

11033	11034	11035	11036	11037	11038	11039	11040
							
30	30	31	31	31	31	31	31

→ MHFP6

encoding precise atom environments





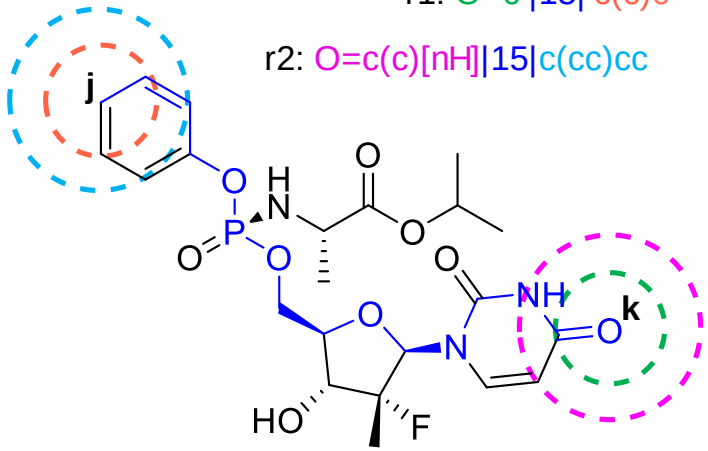
Daniel Probst et al., *J. Cheminf.* **2018,** *10*, 66

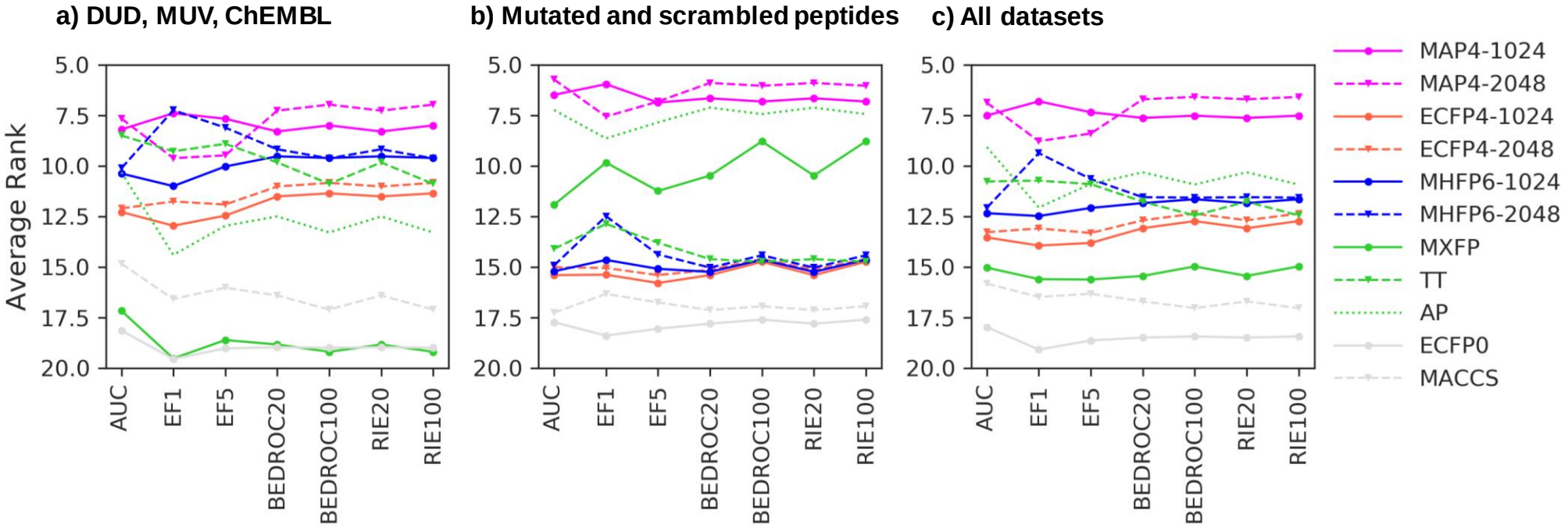
→ MAP4

MAP4 encoding of jk

r1: O=c |15| c(c)c

r2: O=c(c)[nH]|15|c(cc)cc





Alice Capecchi *et al*, *J. Cheminf.* **2020**, *12*, 43

MHFP → LSH → *k*-NN-graph → MST → 2D-layout

UNIVERSITÄT BERN

tmap

Environments
Installation
Laying out a Simple Graph
MinHash
LSH Forest

EXAMPLES

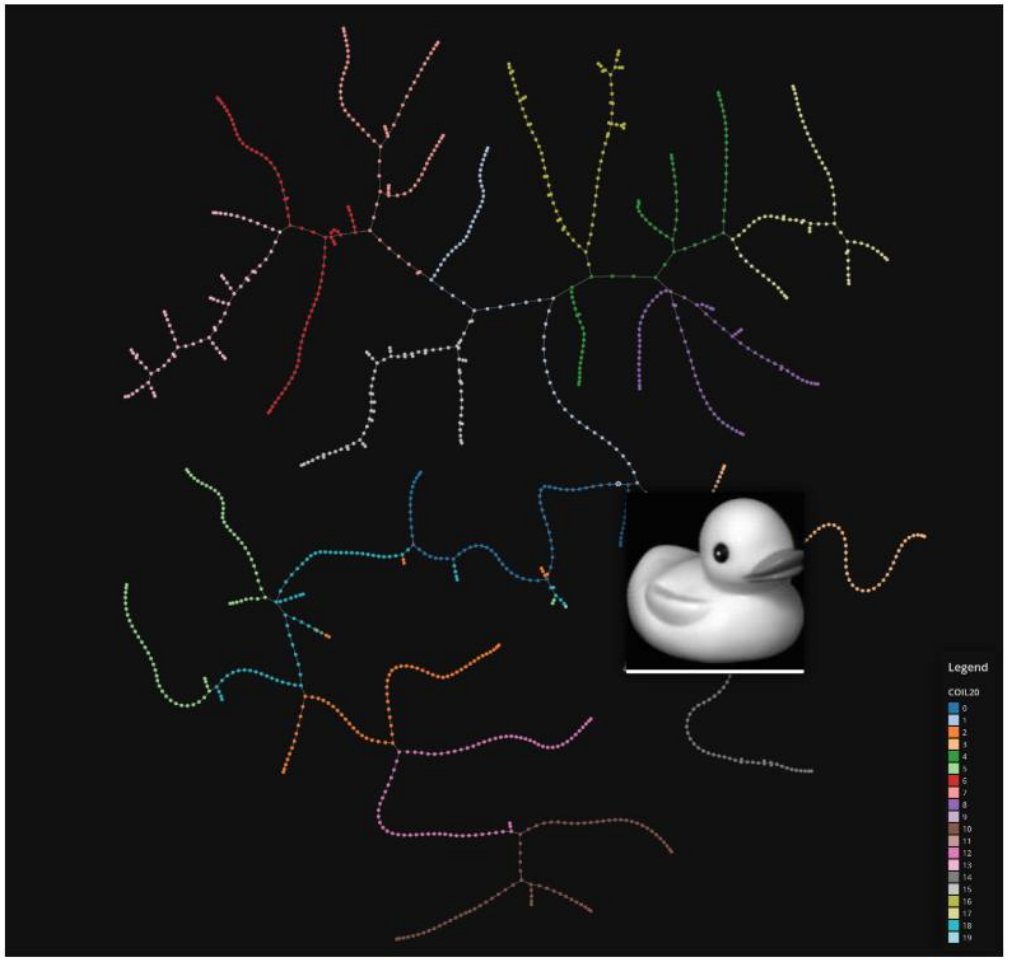
COIL20
MNIST
Fashion MNIST
ChEMBL
FDB17 and ChEMBL
Protein DataBank
Natural Product Atlas
DSSTox
PubMed Central
Gutenberg
RNA Sequencing
MiniBooNE
NIPS
Drugbank
Flowcytometry

DOCUMENTATION

Minhash
LSHForest
Layout

COIL 20

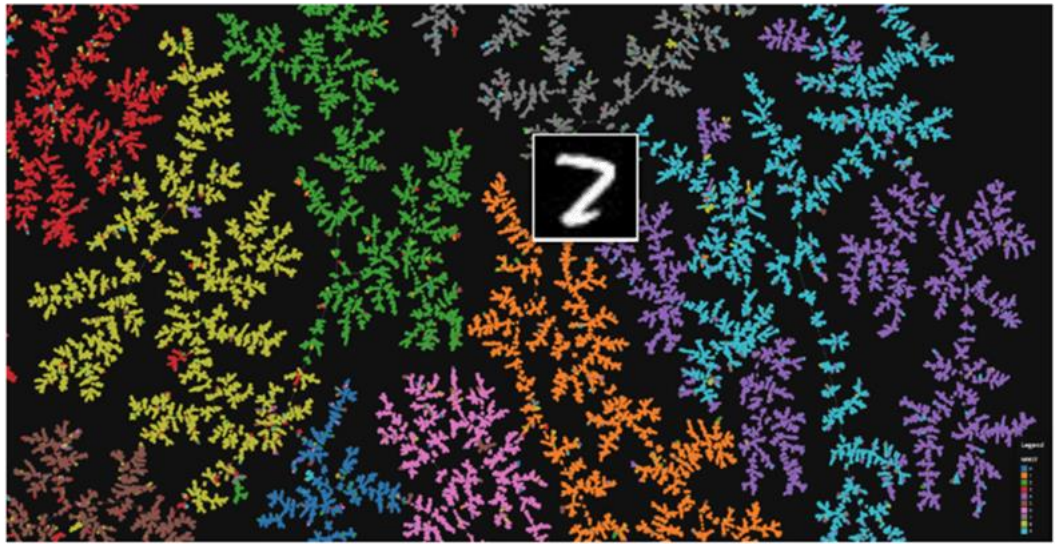
Data set dimensions: 1,440×16,384



View Download

MNIST

Data set dimensions: 70,000×764

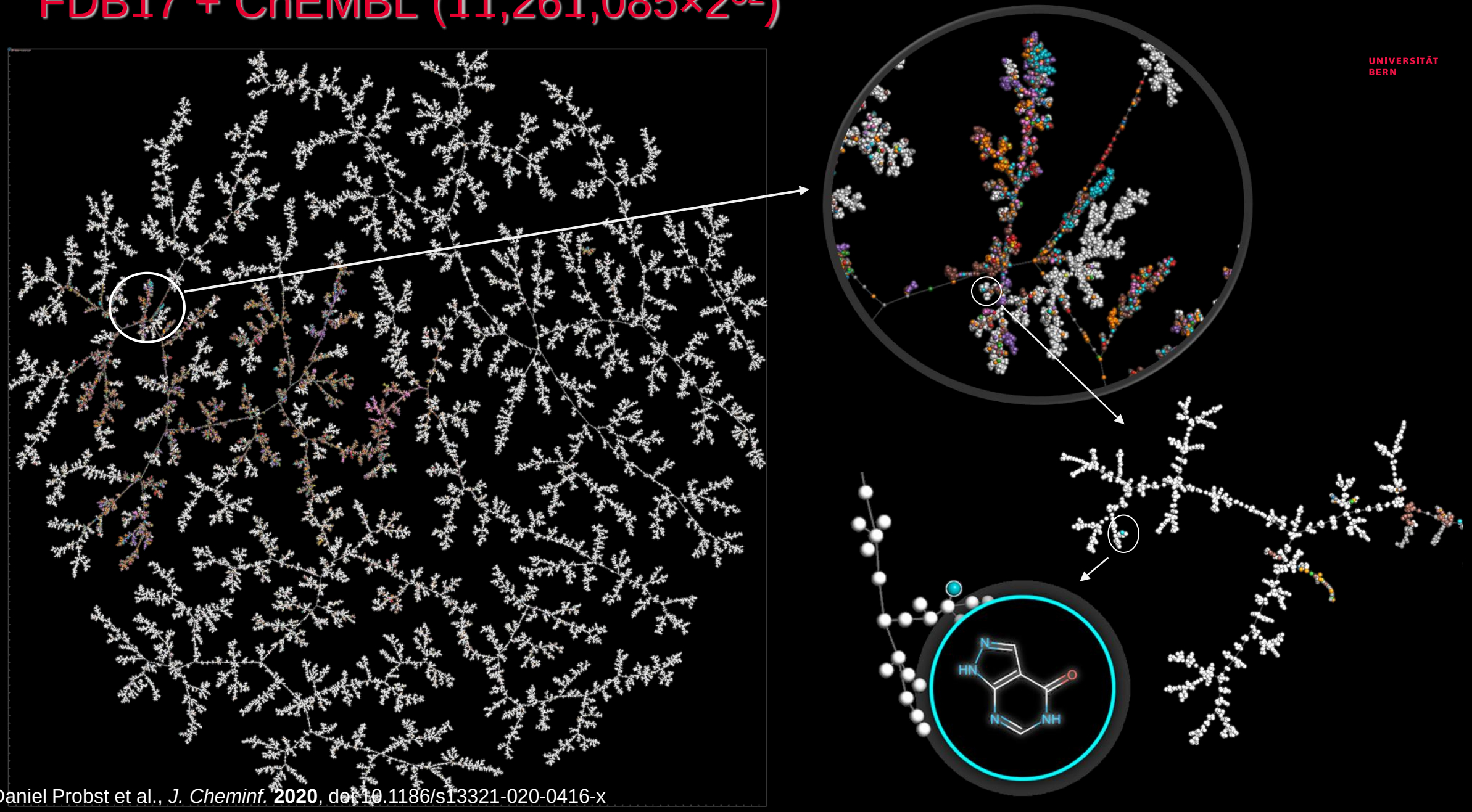


View Download

http://tmap.gdb.tools

Daniel Probst et al., *J. Cheminf.* **2020**, doi:10.1186/s13321-020-0416-x

FDB17 + ChEMBL ($11{,}261{,}085 \times 2^{32}$)



UNIVERSITÄT BERN

Daniel Probst et al., *J. Cheminf.* **2020**, doi:10.1186/s13321-020-0416-x

CHEMICAL & ENGINEERING NEWS
APRIL 6, 2020

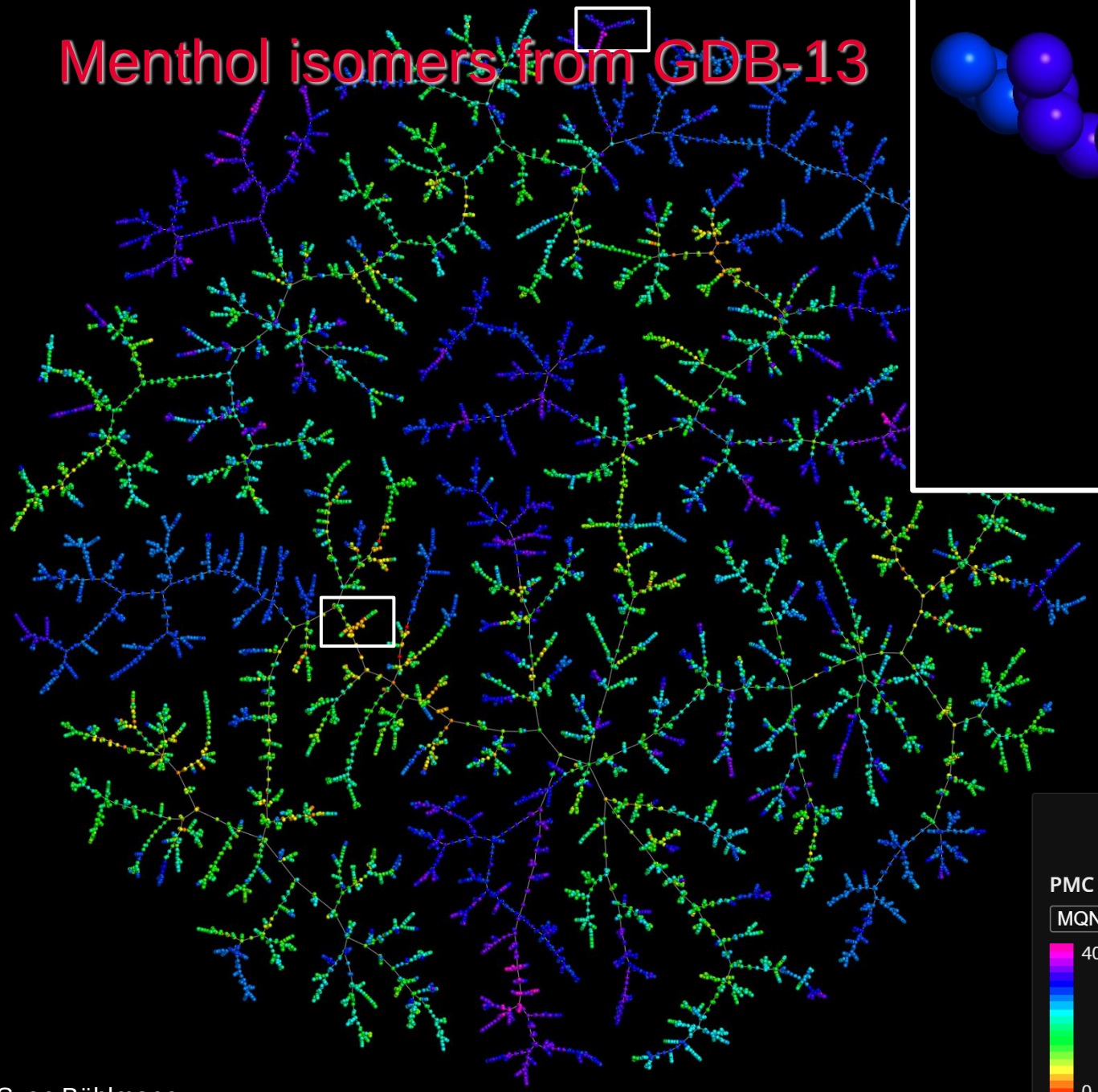
UNIVERSITÄT BERN

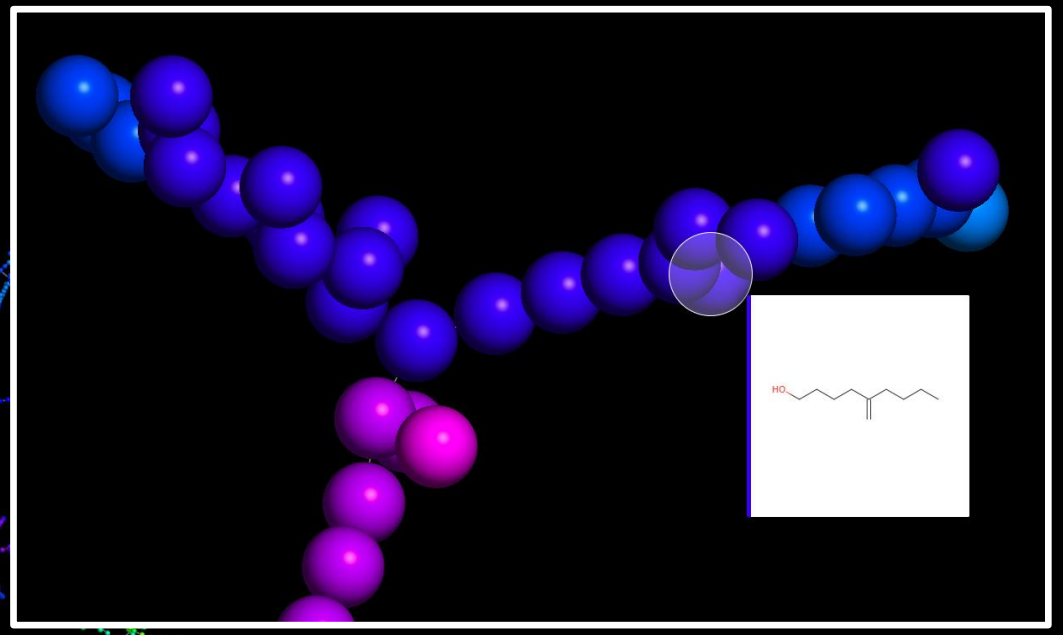
AI explores chemical space

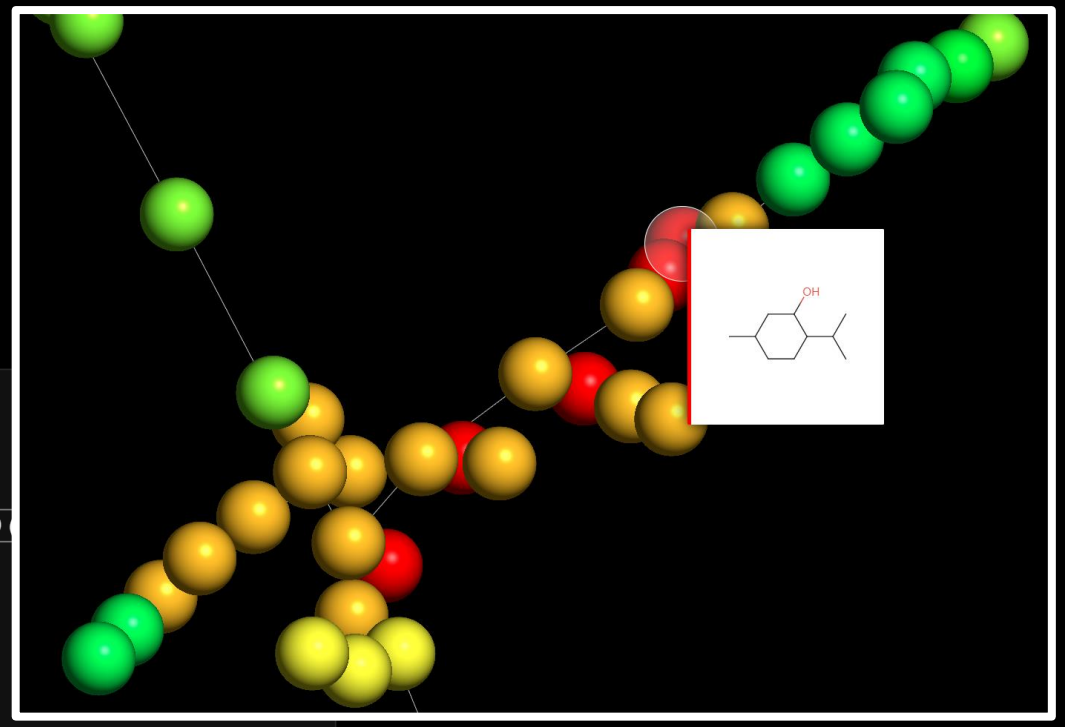
It could take 10^{172} C&EN covers to show every molecule possible in chemical space

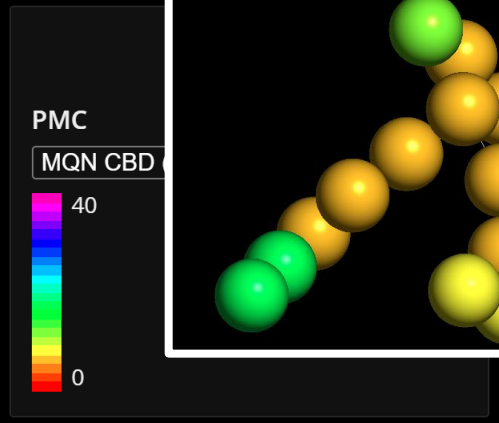
P.30

Menthol isomers from GDB-13









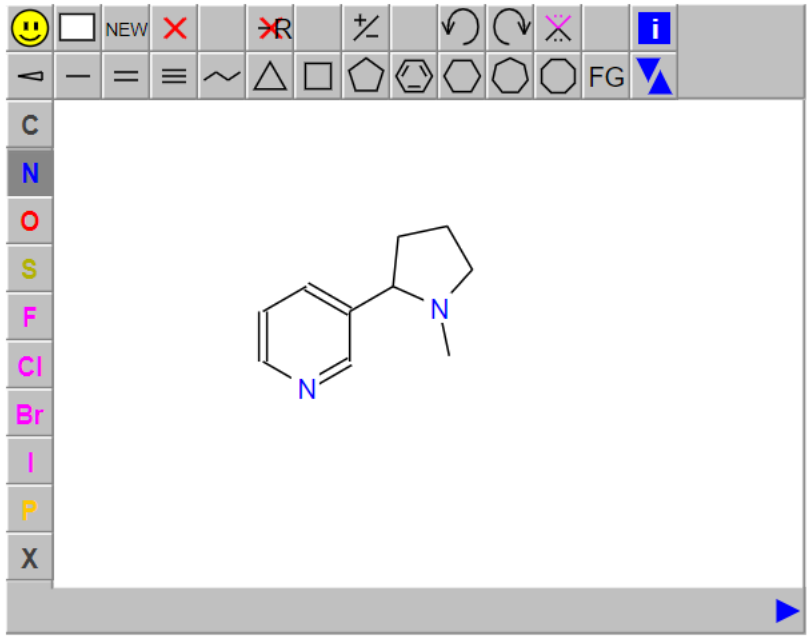
UNIVERSITÄT BERN

Sven Bühlmann

GDBChEMBL Similarity Search

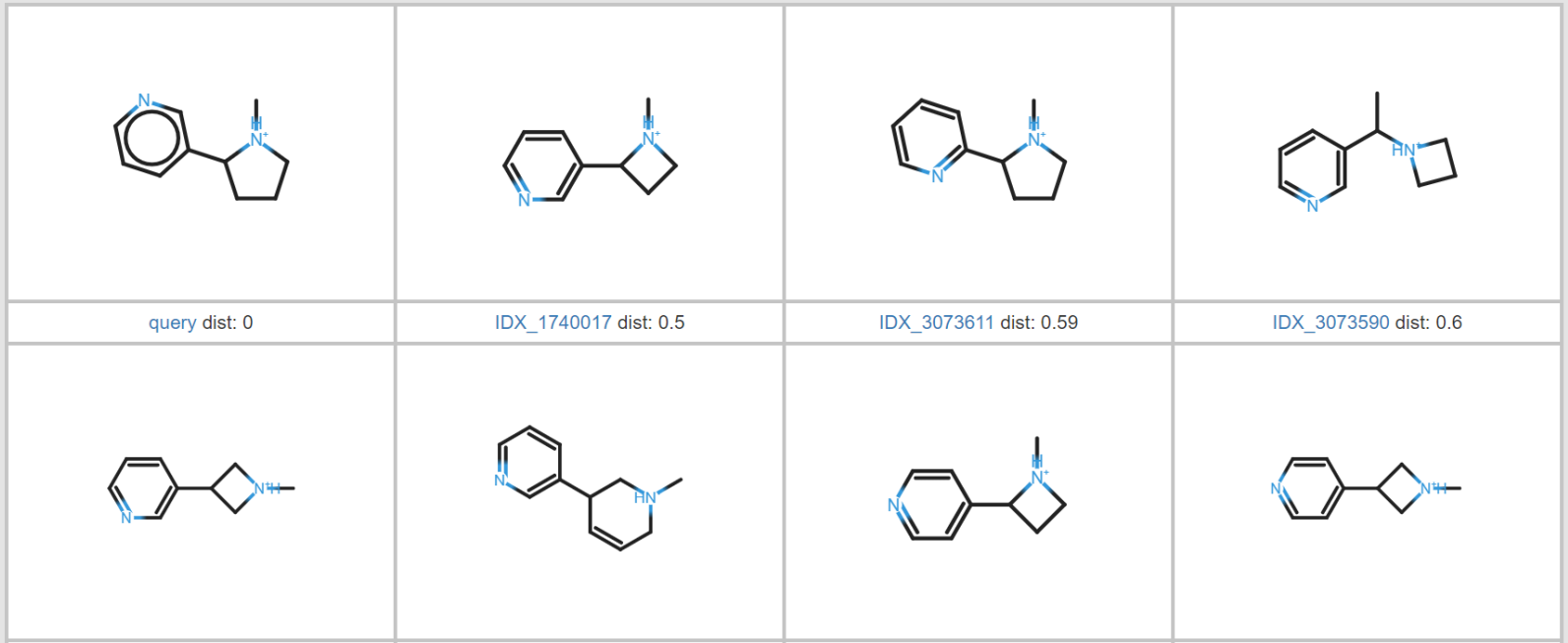
Home | Tutorial | FAQ

Draw or paste your query molecule here: (Click here to load test compound)



GDBChEMBL | MQN-MHFP6 | 500 | Submit

Retrieved 500 nearest neighbors from GDBChEMBL database using MQN-MHFP6. Save Table



query dist: 0	IDX_1740017 dist: 0.5	IDX_3073611 dist: 0.59	IDX_3073590 dist: 0.6
IDX_1494042 dist: 0.66	IDX_4886374 dist: 0.72	IDX_1740016 dist: 0.74	IDX_1494041 dist: 0.76

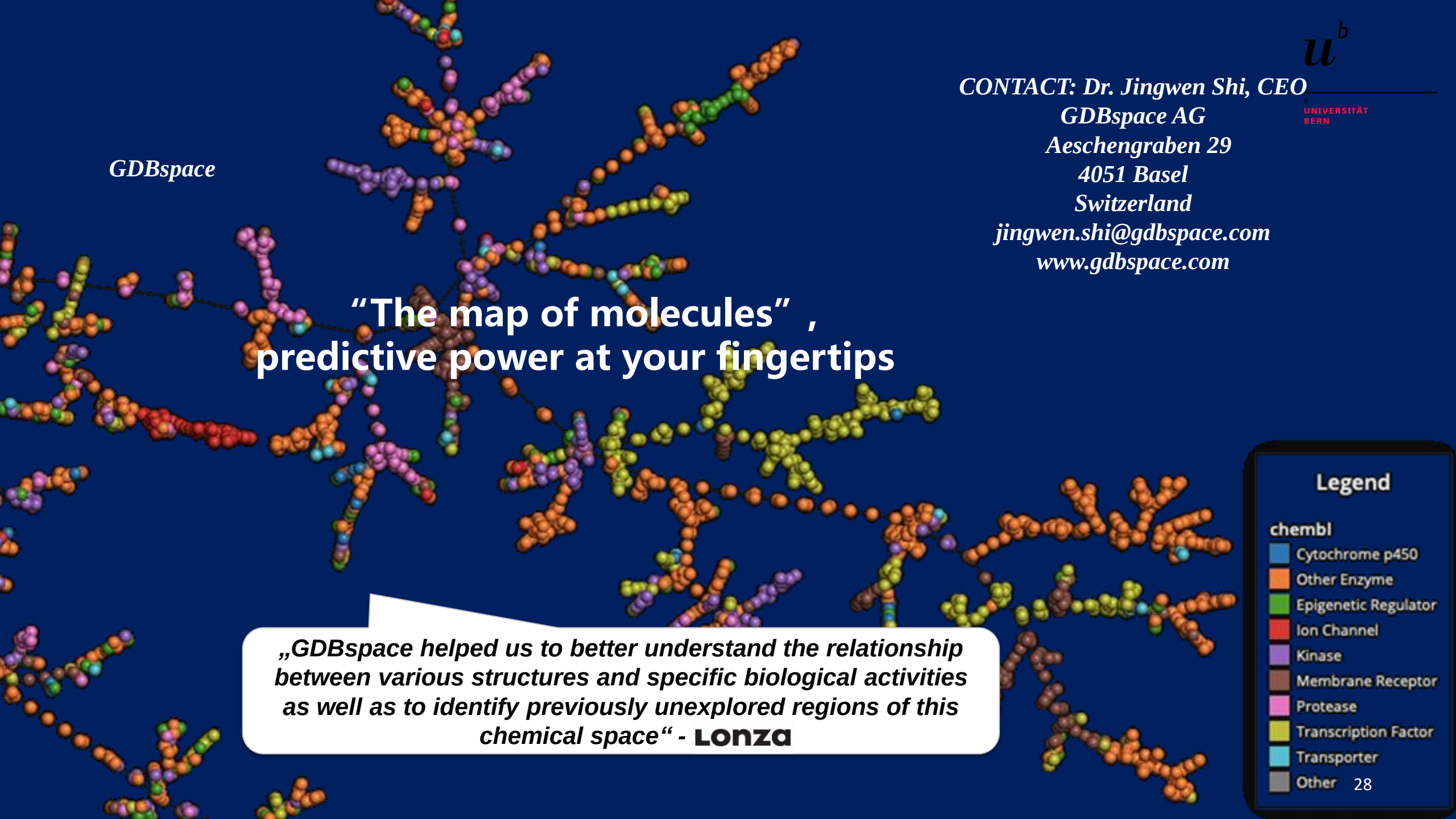
Mahendra Awale et al., *Mol. Inf.*. **2019**, *38*, 1900031

GDBspace

CONTACT: Dr. Jingwen Shi, CEO
GDBspace AG
Aeschengraben 29
4051 Basel
Switzerland
jingwen.shi@gdbspace.com
www.gdbspace.com

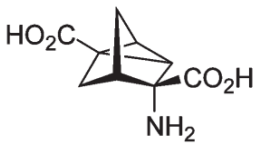
UNIVERSITÄT BERN

"The map of molecules" , predictive power at your fingertips

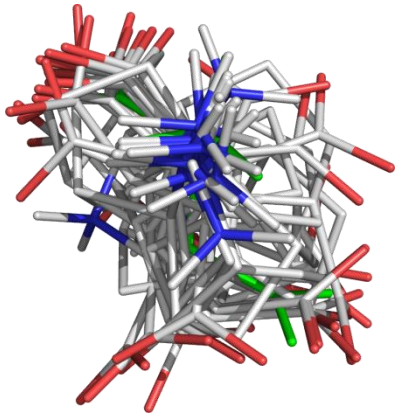


„GDBspace helped us to better understand the relationship between various structures and specific biological activities as well as to identify previously unexplored regions of this chemical space" - Lonza

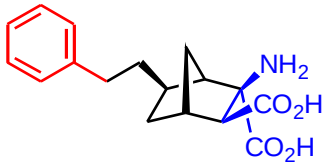
3. Synthesizing GDB



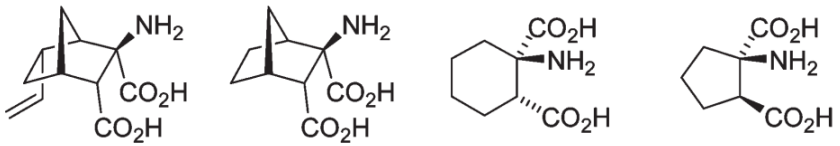
7 (WAY-855)
IC_{50} = 1.3 μM



1. Diversify: 101026 Asp and 151282 Glu from GDB-11
2. Virtual screening: HT **Docking** to GLT-1 (1XFH)
3. Synthesis, testing, optimization



rac-**25a**
IC_{50} = 1.4 μM

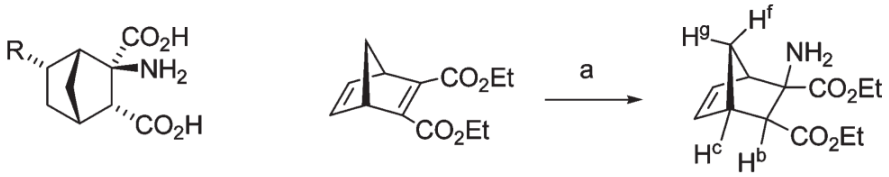


8 (rank 1)
-13.34 kcal/mol

9d (rank 19)
-12.52 kcal/mol

10 (rank 4888)
-10.46 kcal/mol

11 (rank 5176)
-10.42 kcal/mol



8 (R = CH=CH$_2$)
9d (R = H)

12

13

Table 1. IC_{50} Values for Inhibition of [^{3}H]-Glutamate Uptake by GLT-1 and EAAC1 Expressed in *Xenopus laevis* Oocytes

	IC_{50} (GLT-1) (μM)	IC_{50} (EAAC1) (μM)
rac-**23a** (vinyl)	130 ± 70	no inhibition
rac-**25a** (phenethyl)	1.4 ± 0.7	no inhibition
rac-**25d** (phenethyl)	19 ± 5	no inhibition
rac-**28a** (propyl)	25 ± 3	no inhibition
rac-**28b** (butyl)	14 ± 8	nd
rac-**28c** (*o*-HOPh)	21 ± 11	no inhibition
rac-**28d** (*p*-ClPh)	17 ± 11	no inhibition
7[22]	1.3	53
17a[41]	3.8	7.0
17b[43]	0.017	0.3

E. Luethi et al., *J. Med. Chem.* **2010**, *53*, 7236-7250

Check for updates

Cite this: *Chem. Commun.*, 2017, 53, 11399

Synthesis of trinorbornane†

Lorenzo Delarue Bizzini,[a] Thomas Müntener,[a] Daniel Häussinger,[a] Markus Neuburger[a] and Marcel Mayor*[abc]

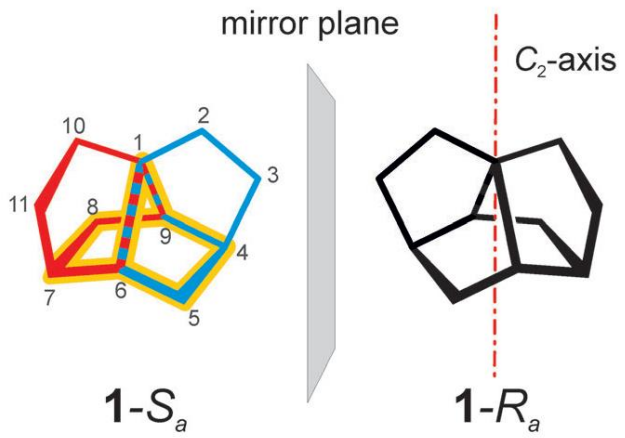
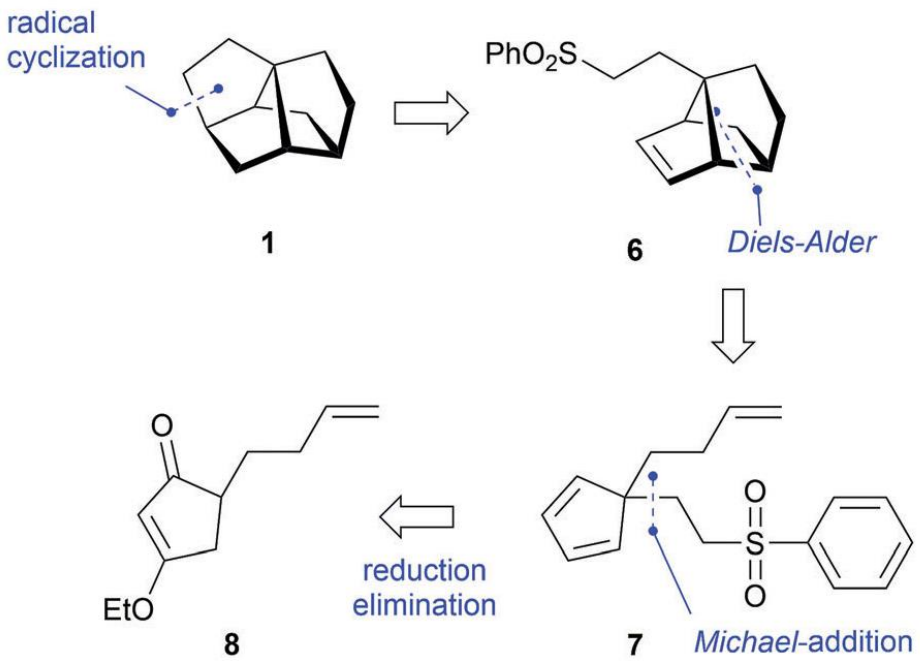
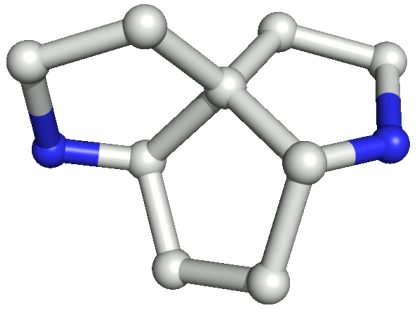
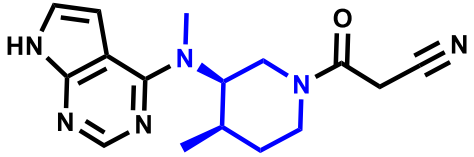


Fig. 1 The two enantiomers (1S_a,4S,6R,7S,9R)tetracyclo[5.2.2.0^{1,6}.0^{4,9}]-undecane (**1**-S_a) and (1R_a,4R,6S,7R,9S)-tetracyclo[5.2.2. 0^{1,6}.0^{4,9}]undecane (**1**-R_a) displaying the axial chirality. For **1**-S_a, on the left side the three norbornane subunits of the structure are displayed in red, blue, and yellow together with the numbering of the eleven carbon atoms corresponding to the IUPAC name. The enantiomer **1**-R_a is displayed together with the molecules on the C_2-axis.

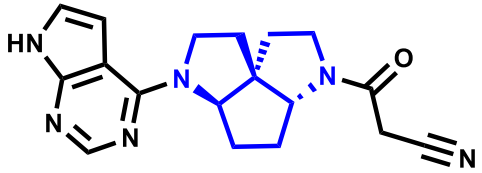


Scheme 1 Retrosynthetic analysis for racemic trinorbornane **1** (to improve the visibility, only the **1**-R_a enantiomer is displayed).

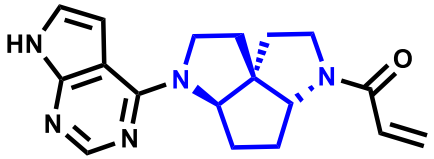




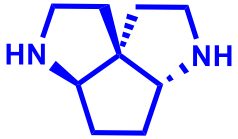
Tofacitinib



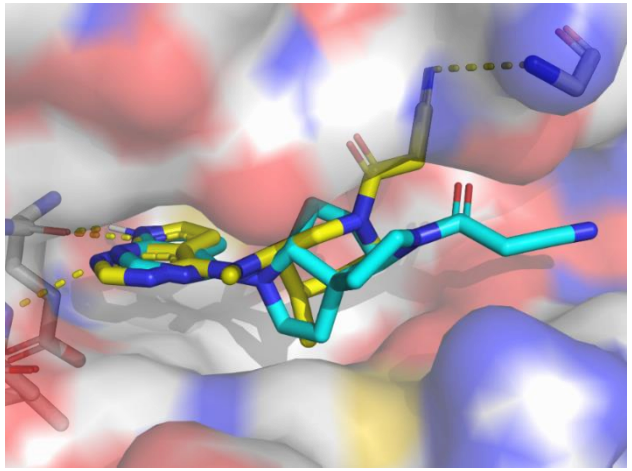
9 (KMC420)



10 (KMC423)



5-5-5



PDB: 3LXK (JAK3)

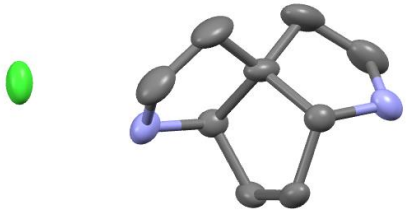
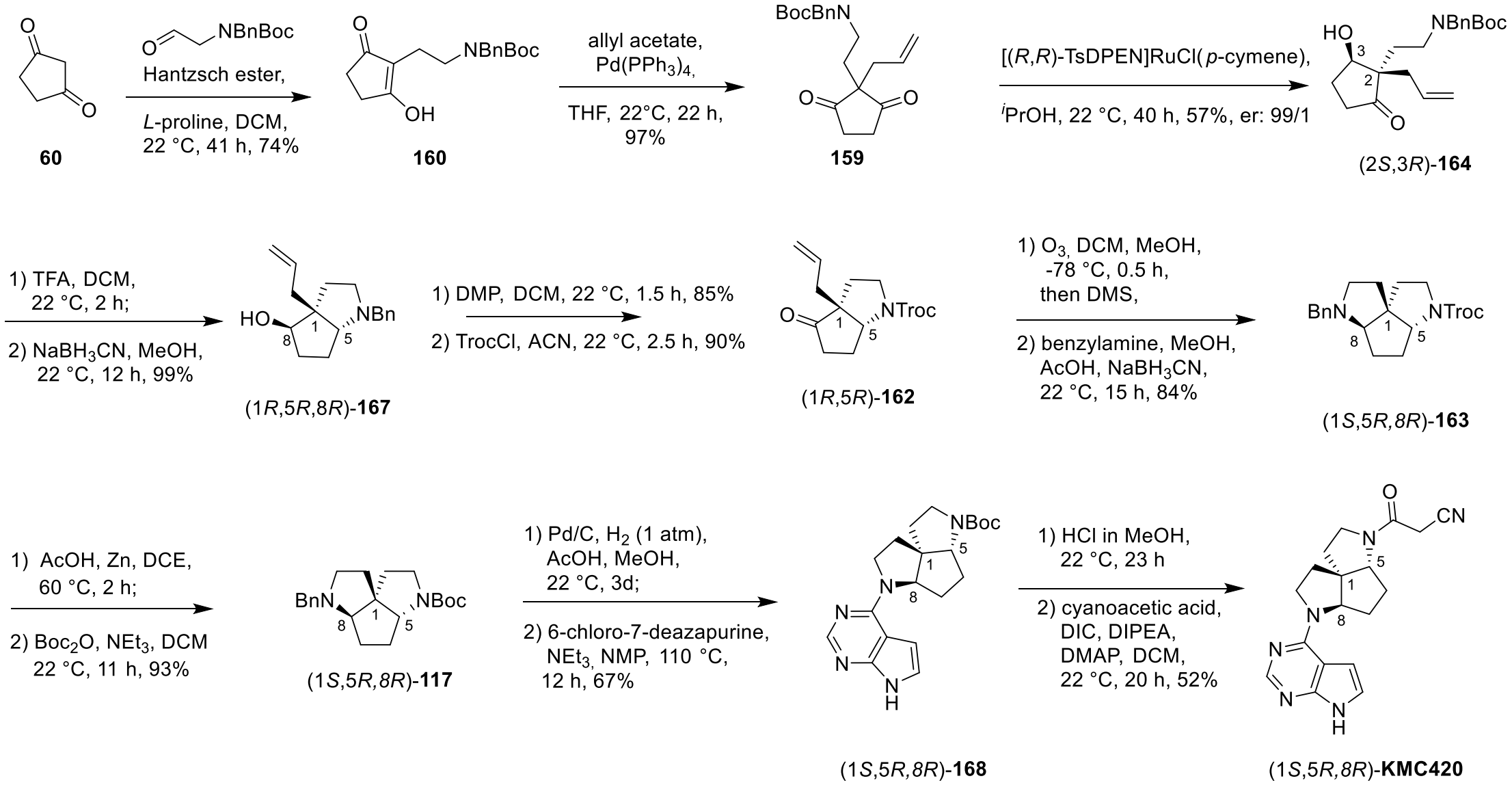


Table 1: JAK kinase inhibition profiles.

IC_{50} (nM)	JAK1	JAK2	JAK3	TYK2
Delgocitinib[45]	2.8	2.6	13	58
Tofacitinib[45]	2.9	1.2	1.1	42
PF-06651600[41]	>10,000	>10,000	33 nM	>10,000
rac-**9**[a)]	1.0	13	29	26
rac-**10**[a)]	30	53	9	183
(*R*)-**9**[a)]	0.8	5.4	14	52
(*S*)-**9**[b)]	>1,000	>1,000	>1,000	>1,000

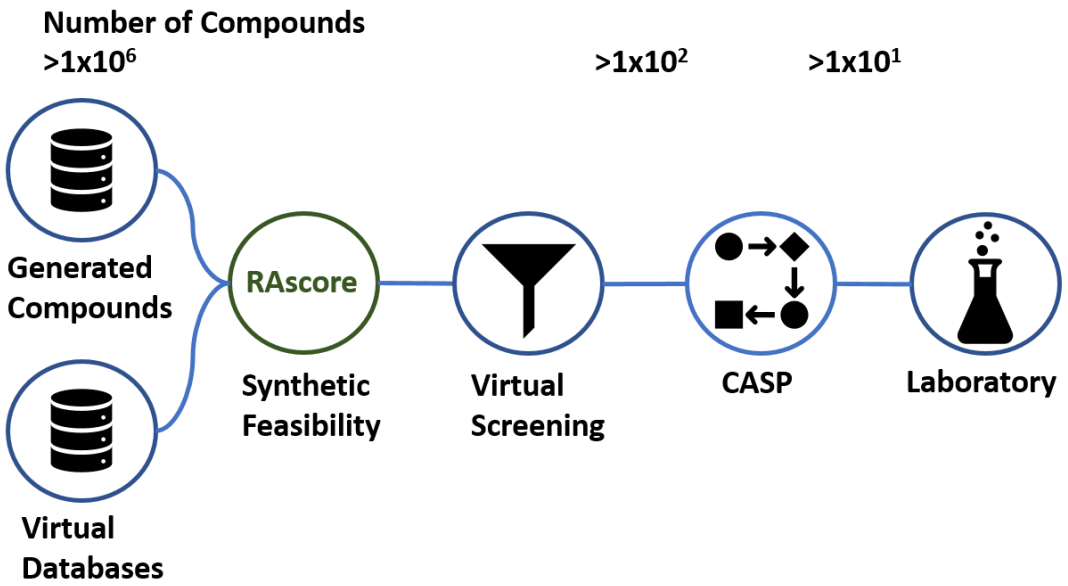
Kris Meier et al., *Angew. Chem., Int. Ed. Engl.* **2020**, doi:10.1002/anie.202012049

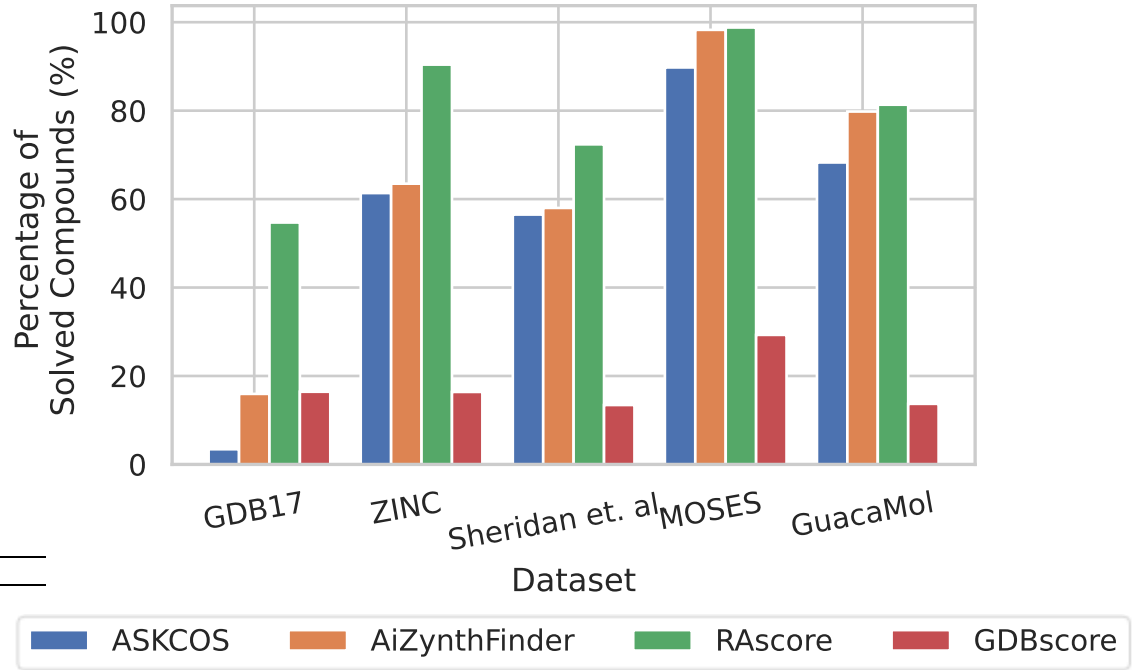
Enantioselective synthesis of KMC420



Kris Meier et al., *Angew. Chem., Int. Ed. Engl.* **2020**, doi:10.1002/anie.202012049

Predicting GDB synthesizability with RAscore





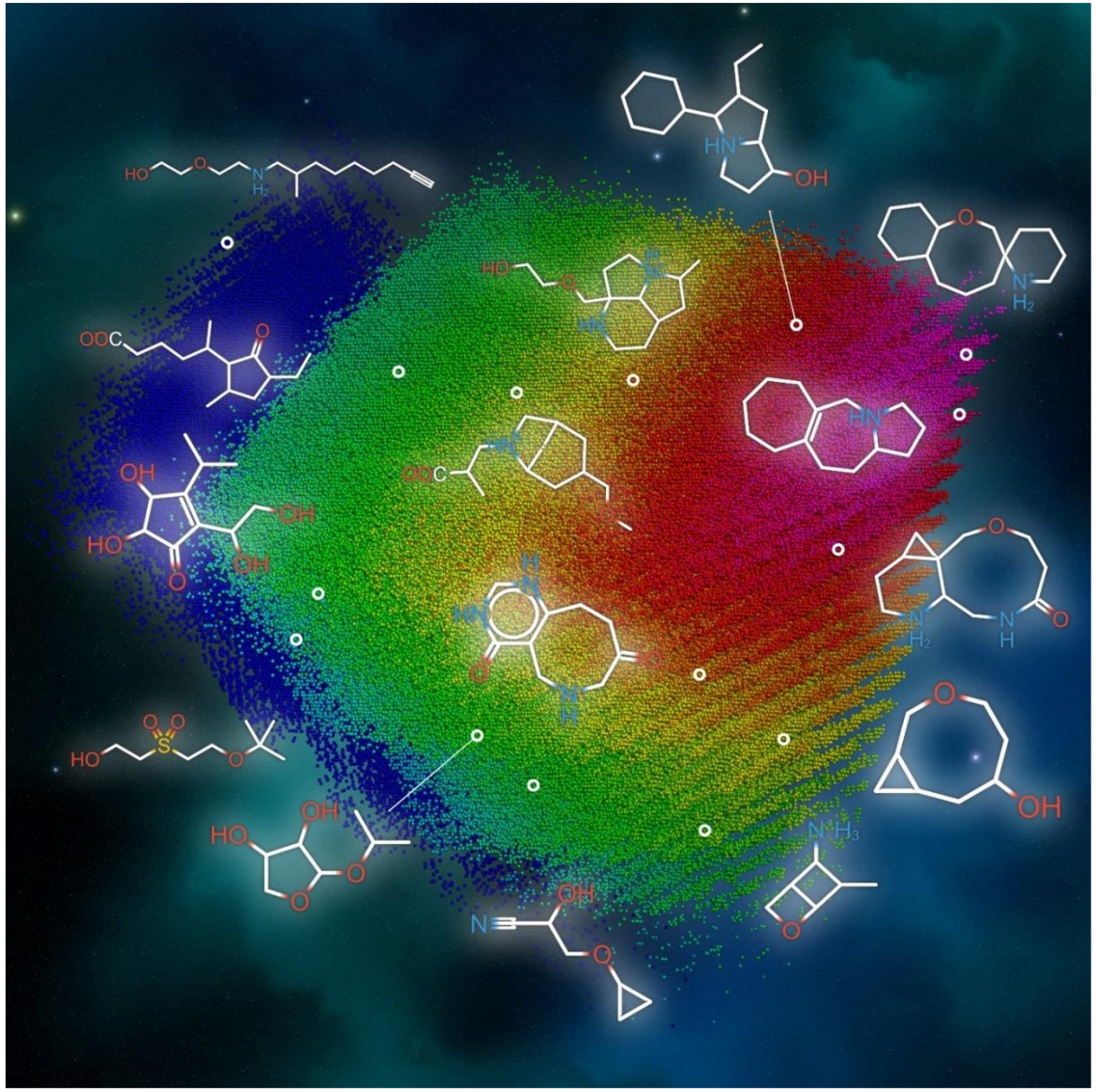
AiZynthfinder

	ChEMBL	GDBChEMBL	GDBMedChem
Percentage Solved	75.21	25.54	20.79
Size	200,000	100,000	100,000
AiZynthFinder Run Time (days)	239	149	151
Score Run Time (mins)	79 [b)]	30 [c)]	30 [c)]

[a)] Expressed in days taken on a single machine with 8 CPUs and 64 GB of RAM, rounded to the nearest day. The time taken in minutes for the neural network classifier with ECFP6 counted fingerprints is also given for comparative purposes. The neural network classifier, RAscore, is able to reproduce the results obtained from AiZynthFinder in a fraction of the time taken to predict full retrosynthetic routes.
[b)]RAscore
[c)]GDBscore

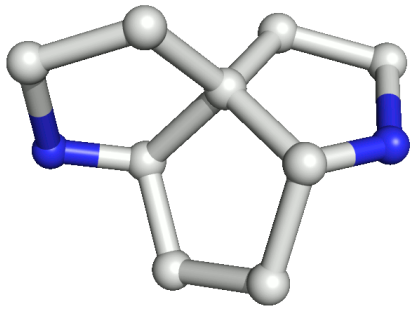
Amol Thakkar et al., *ChemRxiv.* **2020**, https://doi.org/10.26434/chemrxiv.13019993.v1

Summary and Outlook



> Constructing GDB
- From graphs to molecules
- Exhaustive enumeration vs. Sampling
- Novelty in ring systems and scaffolds

> Visualizing and searching GDB
- MQN
- Faerun
- MHFP6 and MAP4
- TMAP
- NN-searches with MQN



> Synthesizing GDB
- Virtual screening hits
- Novelty oriented synthesis
- CASP