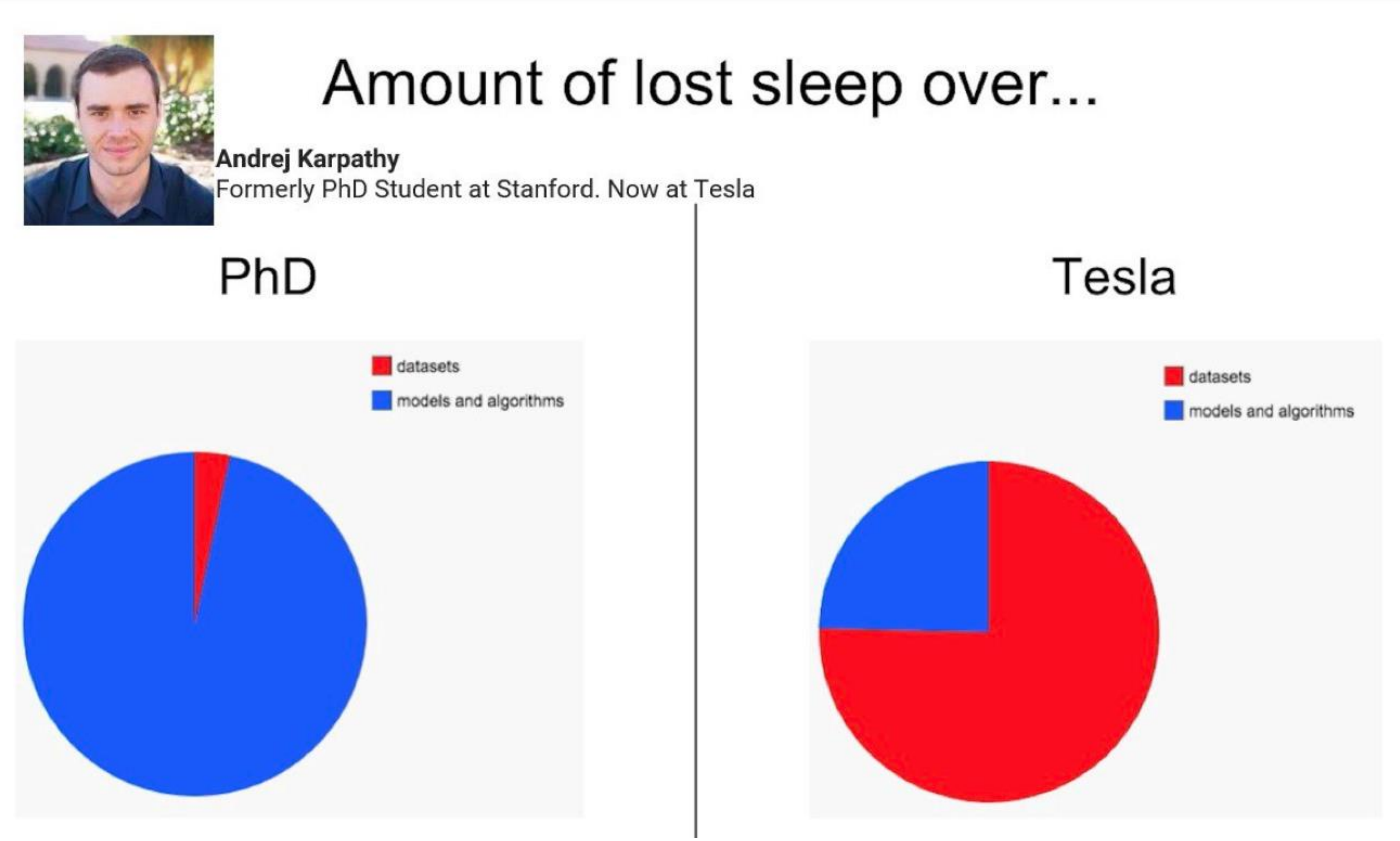
Review : Molecular de-novo design through deep reinforcement learning + α

RL study 8th

2022.6.27.(月)

P.S. Park

Data.... Data.... Data.... Data....



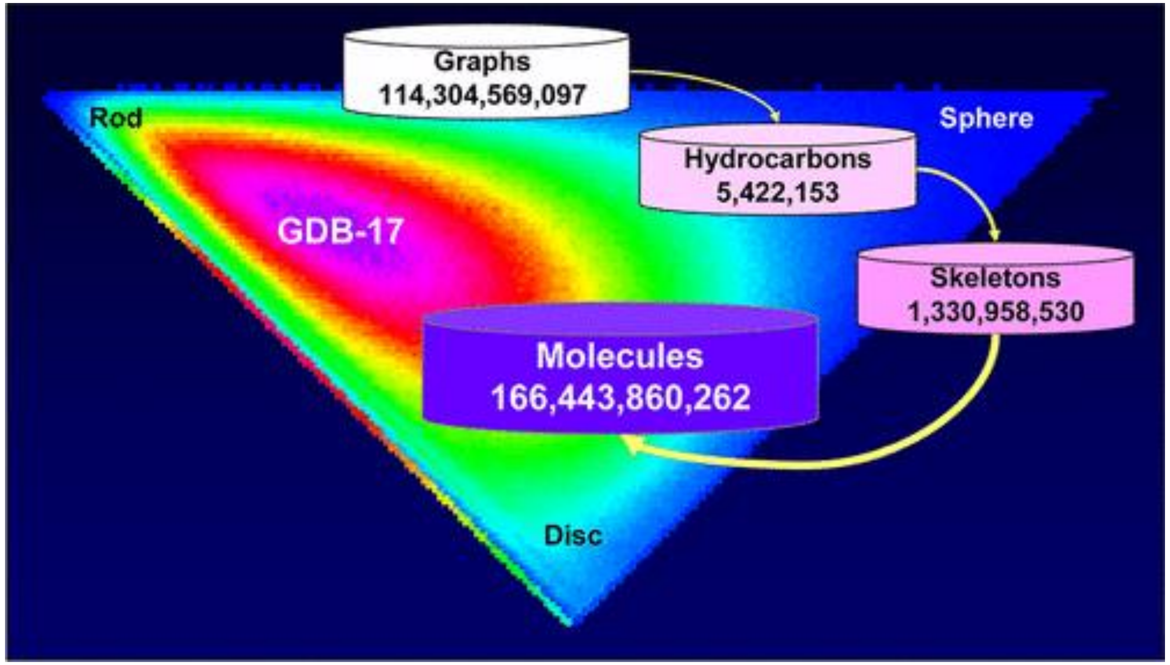
Amount of lost sleep over...

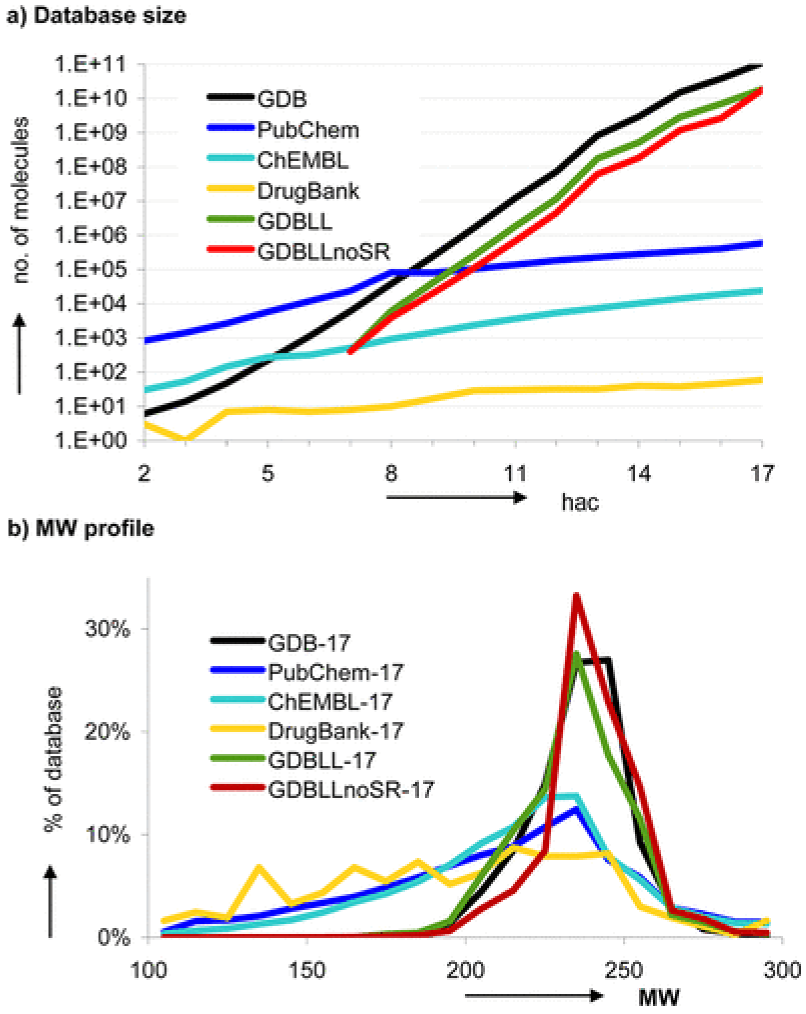
Andrej Karpathy
Formerly PhD Student at Stanford. Now at Tesla

PhD

Tesla

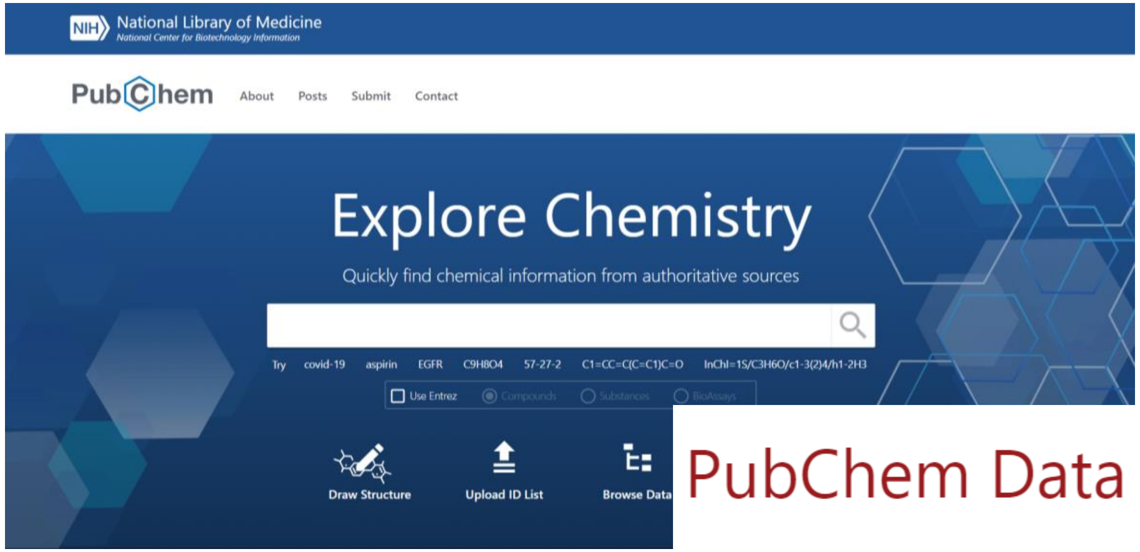
Chemical data

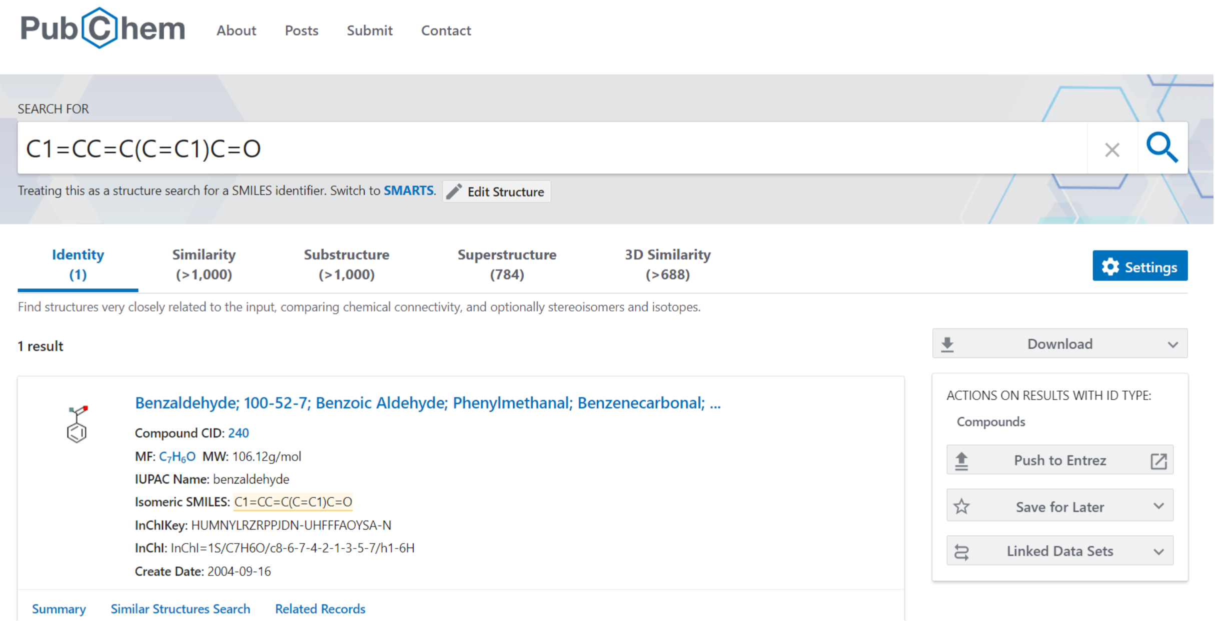




"Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17"
J. Chem. Inf. Model. 2012, 52, 11, 2864–2875

Chemical data


NIH National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Posts
Submit
Contact
Explore Chemistry
Quickly find chemical information from authoritative sources
Try covid-19 aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)4/h1-2H3
□ Use Entrez ○ Compounds ○ Substances ○ BioAssays
Draw Structure
Upload ID List
Browse Data



SEARCH FOR

C1=CC=C(C=C1)C=O

Treating this as a structure search for a SMILES identifier. Switch to **SMARTS**. Edit Structure

Identity (1) | Similarity (>1,000) | Substructure (>1,000) | Superstructure (784) | 3D Similarity (>688)

Find structures very closely related to the input, comparing chemical connectivity, and optionally stereoisomers and isotopes.

1 result

Benzaldehyde; 100-52-7; Benzoic Aldehyde; Phenylmethanal; Benzenecarbonal; ...

Compound CID: 240

MF: C_7H_6O MW: 106.12g/mol

IUPAC Name: benzaldehyde

Isomeric SMILES: C1=CC=C(C=C1)C=O

InChIKey: HUMNYLRZRPPJDN-UHFFFAOYSA-N

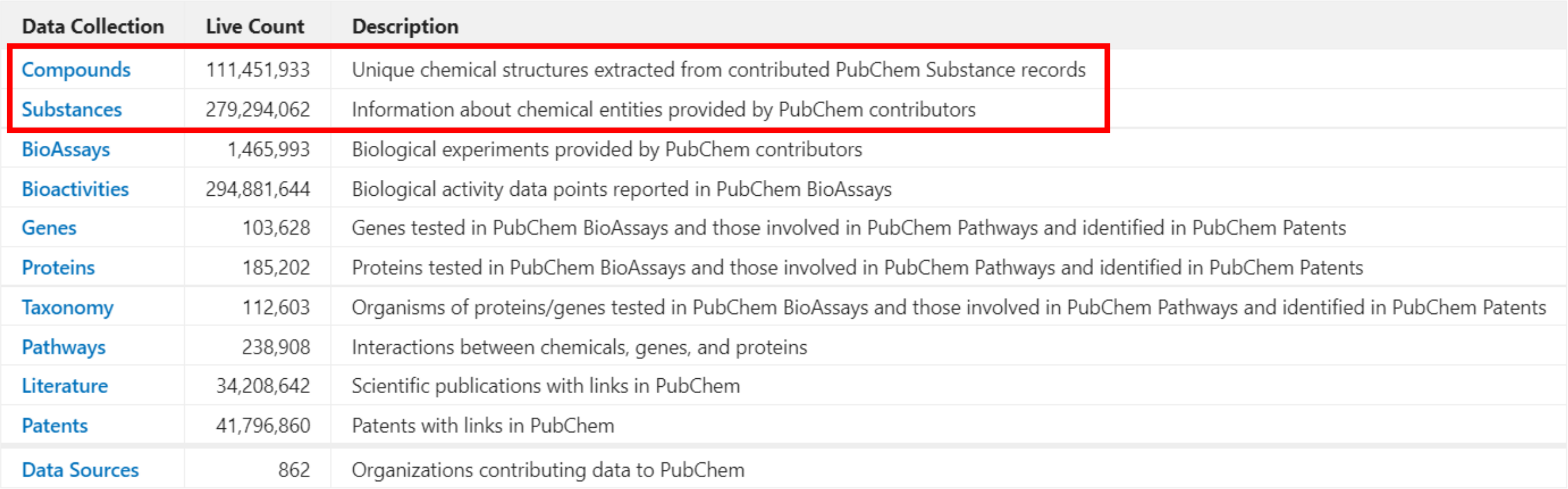
InChI: InChI=1S/C7H6O/c8-6-7-4-2-1-3-5-7/h1-6H

Create Date: 2004-09-16

ACTIONS ON RESULTS WITH ID TYPE:

Compounds

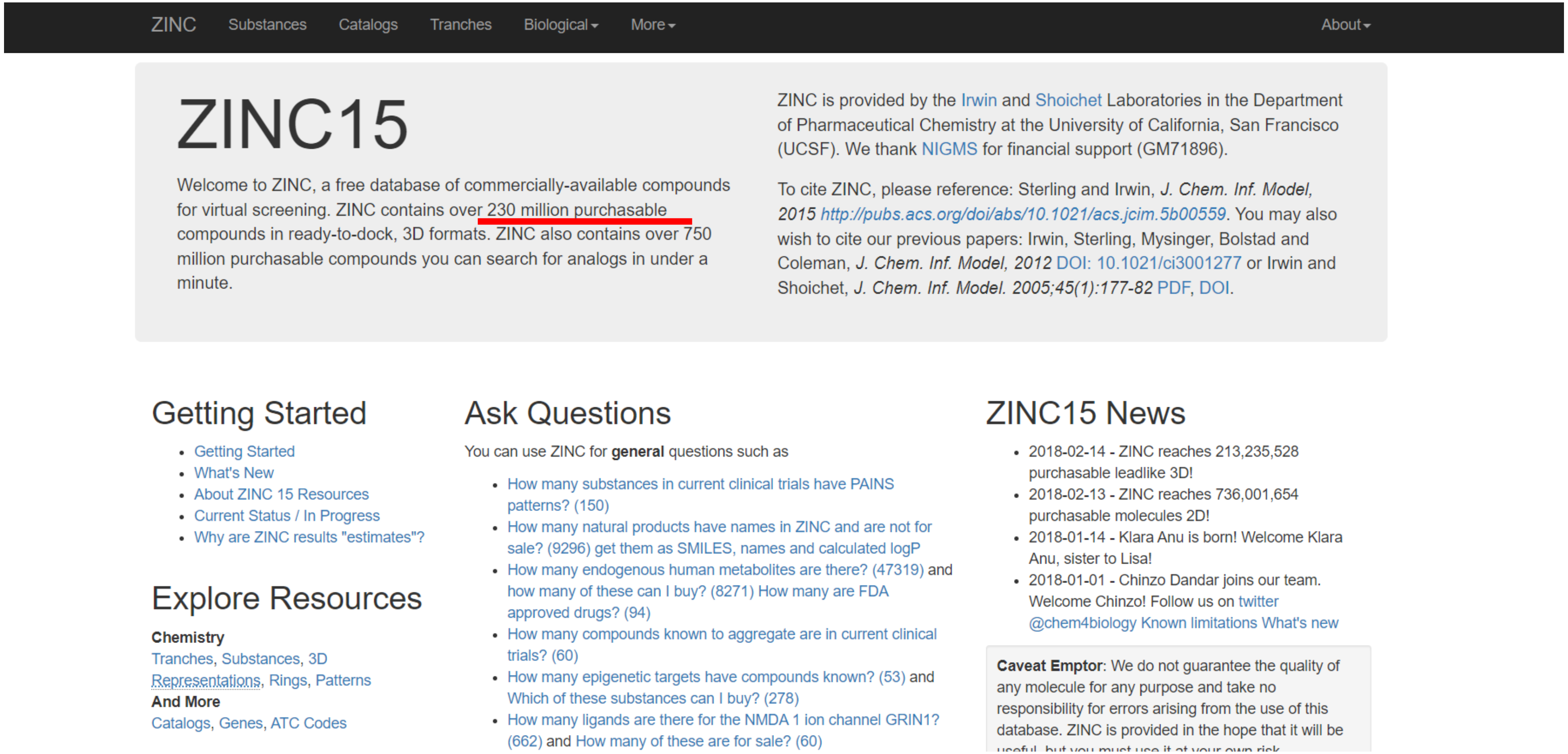
PubChem Data Counts



Data Collection	Live Count	Description
Compounds	111,451,933	Unique chemical structures extracted from contributed PubChem Substance records
Substances	279,294,062	Information about chemical entities provided by PubChem contributors
BioAssays	1,465,993	Biological experiments provided by PubChem contributors
Bioactivities	294,881,644	Biological activity data points reported in PubChem BioAssays
Genes	103,628	Genes tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents
Proteins	185,202	Proteins tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents
Taxonomy	112,603	Organisms of proteins/genes tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents
Pathways	238,908	Interactions between chemicals, genes, and proteins
Literature	34,208,642	Scientific publications with links in PubChem
Patents	41,796,860	Patents with links in PubChem
Data Sources	862	Organizations contributing data to PubChem

https://pubchem.ncbi.nlm.nih.gov/

Chemical data



ZINC15

Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.

ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). We thank NIGMS for financial support (GM71896).

To cite ZINC, please reference: Sterling and Irwin, *J. Chem. Inf. Model, 2015 http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559*. You may also wish to cite our previous papers: Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model, 2012* DOI: 10.1021/ci3001277 or Irwin and Shoichet, *J. Chem. Inf. Model. 2005;45(1):177-82* PDF, DOI.

Getting Started

- Getting Started
- What's New
- About ZINC 15 Resources
- Current Status / In Progress
- Why are ZINC results "estimates"?

Explore Resources

Chemistry
Tranches, Substances, 3D Representations, Rings, Patterns
And More
Catalogs, Genes, ATC Codes

Ask Questions

You can use ZINC for **general** questions such as

- How many substances in current clinical trials have PAINS patterns? (150)
- How many natural products have names in ZINC and are not for sale? (9296) get them as SMILES, names and calculated logP
- How many endogenous human metabolites are there? (47319) and how many of these can I buy? (8271) How many are FDA approved drugs? (94)
- How many compounds known to aggregate are in current clinical trials? (60)
- How many epigenetic targets have compounds known? (53) and Which of these substances can I buy? (278)
- How many ligands are there for the NMDA 1 ion channel GRIN1? (662) and How many of these are for sale? (60)

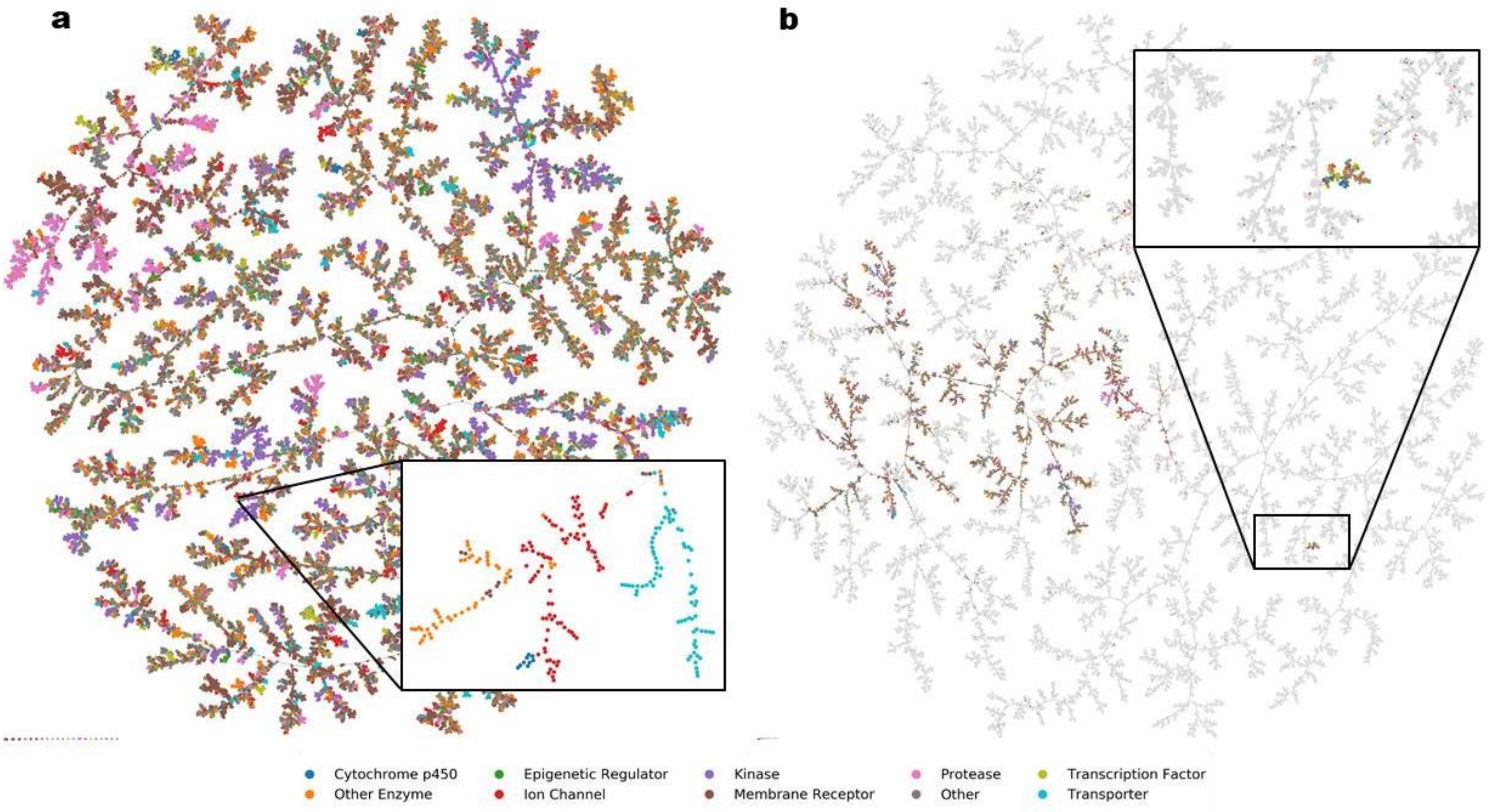
ZINC15 News

- 2018-02-14 - ZINC reaches 213,235,528 purchasable leadlike 3D!
- 2018-02-13 - ZINC reaches 736,001,654 purchasable molecules 2D!
- 2018-01-14 - Klara Anu is born! Welcome Klara Anu, sister to Lisa!
- 2018-01-01 - Chinzo Dandar joins our team. Welcome Chinzo! Follow us on twitter @chem4biology Known limitations What's new

Caveat Emptor: We do not guarantee the quality of any molecule for any purpose and take no responsibility for errors arising from the use of this database. ZINC is provided in the hope that it will be useful, but you must use it at your own risk.

https://zinc.docking.org/

“TMAP” visualization of ChEMBL, FDB17, DSSTox, and the Natural Products Atlas in the MHFP6 chemical space



Probst, D., Reymond, JL. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* **12,** 12 (2020).

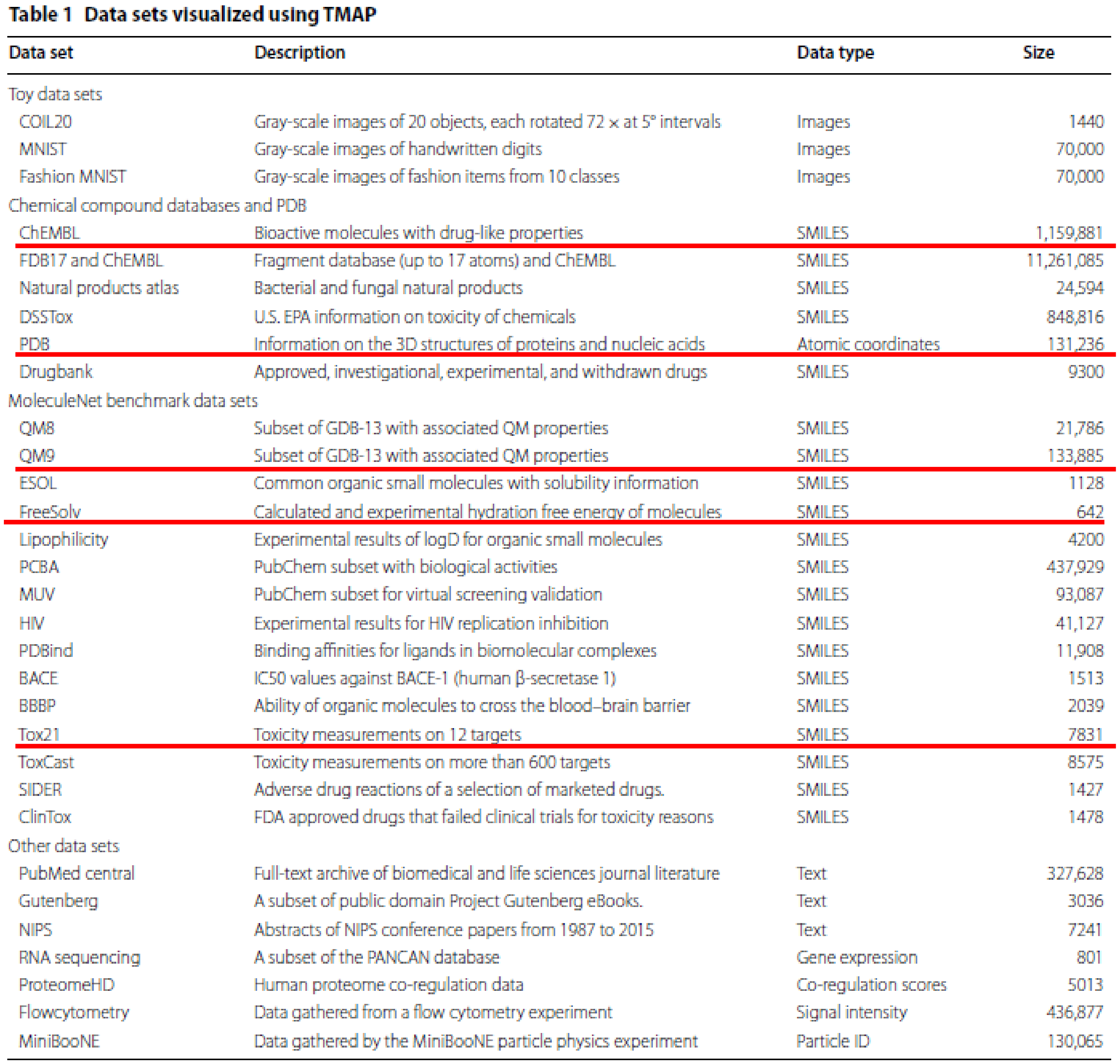
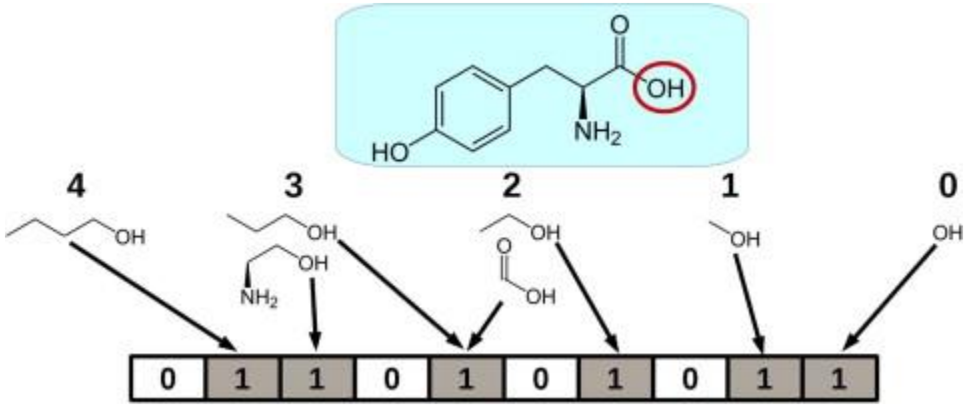


Table 1 Data sets visualized using TMAP

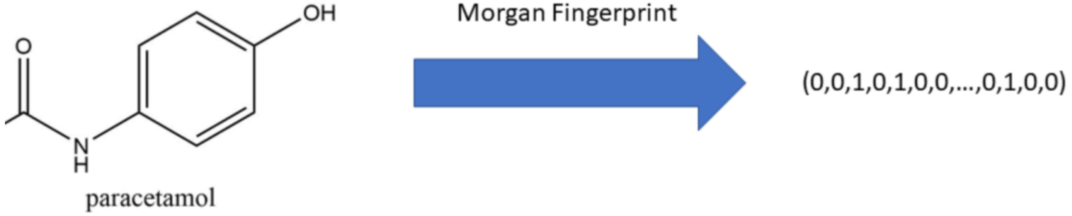
Data set	Description	Data type	Size
Toy data sets			
COIL20	Gray-scale images of 20 objects, each rotated 72 × at 5° intervals	Images	1440
MNIST	Gray-scale images of handwritten digits	Images	70,000
Fashion MNIST	Gray-scale images of fashion items from 10 classes	Images	70,000
Chemical compound databases and PDB			
ChEMBL	Bioactive molecules with drug-like properties	SMILES	1,159,881
FDB17 and ChEMBL	Fragment database (up to 17 atoms) and ChEMBL	SMILES	11,261,085
Natural products atlas	Bacterial and fungal natural products	SMILES	24,594
DSSTox	U.S. EPA information on toxicity of chemicals	SMILES	848,816
PDB	Information on the 3D structures of proteins and nucleic acids	Atomic coordinates	131,236
Drugbank	Approved, investigational, experimental, and withdrawn drugs	SMILES	9300
MoleculeNet benchmark data sets			
QM8	Subset of GDB-13 with associated QM properties	SMILES	21,786
QM9	Subset of GDB-13 with associated QM properties	SMILES	133,885
ESOL	Common organic small molecules with solubility information	SMILES	1128
FreeSolv	Calculated and experimental hydration free energy of molecules	SMILES	642
Lipophilicity	Experimental results of logD for organic small molecules	SMILES	4200
PCBA	PubChem subset with biological activities	SMILES	437,929
MUV	PubChem subset for virtual screening validation	SMILES	93,087
HIV	Experimental results for HIV replication inhibition	SMILES	41,127
PDBind	Binding affinities for ligands in biomolecular complexes	SMILES	11,908
BACE	IC50 values against BACE-1 (human β-secretase 1)	SMILES	1513
BBBP	Ability of organic molecules to cross the blood–brain barrier	SMILES	2039
Tox21	Toxicity measurements on 12 targets	SMILES	7831
ToxCast	Toxicity measurements on more than 600 targets	SMILES	8575
SIDER	Adverse drug reactions of a selection of marketed drugs.	SMILES	1427
ClinTox	FDA approved drugs that failed clinical trials for toxicity reasons	SMILES	1478
Other data sets			
PubMed central	Full-text archive of biomedical and life sciences journal literature	Text	327,628
Gutenberg	A subset of public domain Project Gutenberg eBooks.	Text	3036
NIPS	Abstracts of NIPS conference papers from 1987 to 2015	Text	7241
RNA sequencing	A subset of the PANCAN database	Gene expression	801
ProteomeHD	Human proteome co-regulation data	Co-regulation scores	5013
Flowcytometry	Data gathered from a flow cytometry experiment	Signal intensity	436,877
MiniBooNE	Data gathered by the MiniBooNE particle physics experiment	Particle ID	130,065

Probst, D., Reymond, JL. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* **12,** 12 (2020).

Molecular descriptor : Fingerprint



0	1	1	0	1	0	1	0	1	1



paracetamol

(0,0,1,0,1,0,0,...,0,1,0,0)

Typical fingerprint sizes: 1K-4K bits.

Molecular descriptor : Fingerprint

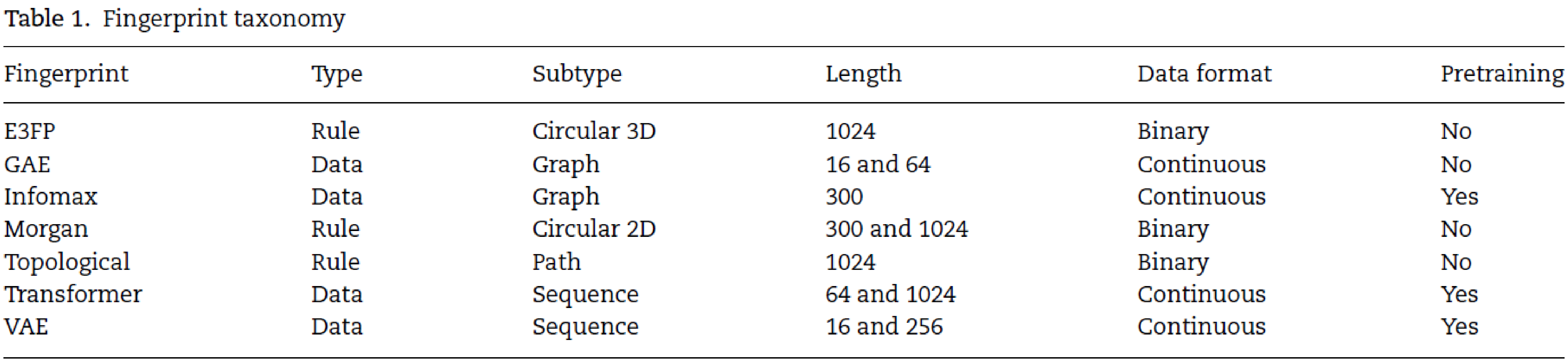
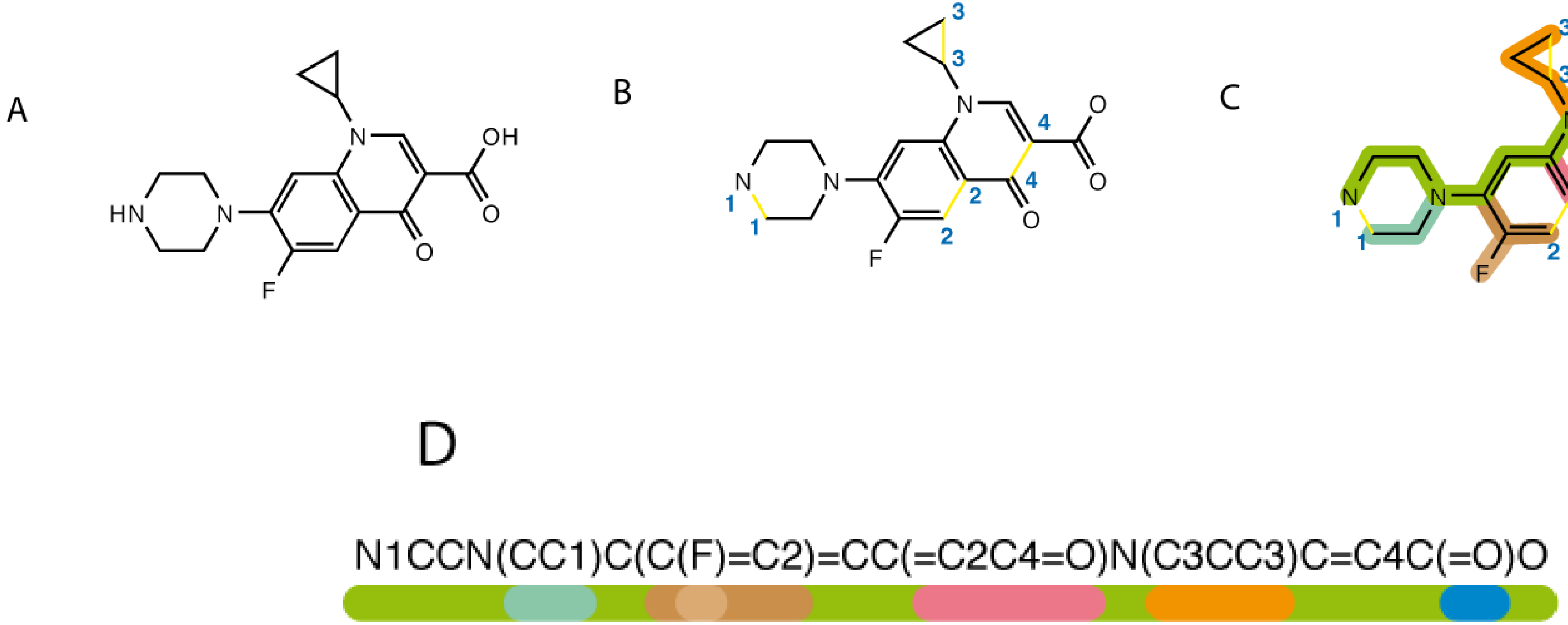


Table 1. Fingerprint taxonomy

Fingerprint	Type	Subtype	Length	Data format	Pretraining
E3FP	Rule	Circular 3D	1024	Binary	No
GAE	Data	Graph	16 and 64	Continuous	No
Infomax	Data	Graph	300	Continuous	Yes
Morgan	Rule	Circular 2D	300 and 1024	Binary	No
Topological	Rule	Path	1024	Binary	No
Transformer	Data	Sequence	64 and 1024	Continuous	Yes
VAE	Data	Sequence	16 and 256	Continuous	Yes

Ref: “Comparative analysis of molecular fingerprints in prediction of drug combination effects”, *Briefings in Bioinformatics*, Volume 22, Issue 6, November 2021

Molecular descriptor : SMILES(Simplified molecular-input line-entry system)



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Molecular descriptor : SMILES(**S**implified **m**olecular-**i**nput **l**ine-**e**ntry **s**ystem)

1. 원자는 C, N, O, Cl처럼 원자 기호를 직접 넣어도 되고, [#7]처럼 대괄호 안에 원자 기호를 넣어도 됩니다.

2. Bond는 -(single), =(double), #(triple), $(quadruple), : (aromatic) 이 있습니다. 그리고 공유결합이 없는 (이온 결합 같은) 경우는 . 으로 표기됩니다. 예시 ([Na+].[Cl-])

3. 수소원자와 단일 결합은 보통 생략됩니다.

4. CCC 라고 한다면, C-C-C, [CH3]-[CH2]-[CH3] 을 의미합니다. 괄호를 쓰면, 원자를 좀 더 명확히 규정해서 쓸 수 있습니다.

5. Formal charge가 없을 때 C, N, O, 에서 원자의 결합수가 4,3,2가 아니라면, 부족한 만큼이 수소로 채워져 있다고 생각할 수 있습니다. Ex) O → H_2O

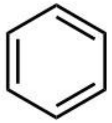
Molecular descriptor : SMILES(Simplified **m**olecular-**i**nput **l**ine-**e**ntry **s**ystem**)**

1. Main chain 을 정합니다. (일반적으로 제일 긴 체인을 잡습니다.)

2. 나머지는 side chain 취급을 합니다. Side-chain은 main chain에 ()를 삽입해서 만듭니다.

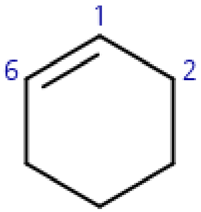
3. Ring은 side-chain만으로는 표기가 안되고, 원래 자리로 돌아와야 하기 때문에 결합이 필요한 위치에 숫자를 붙입니다.

   Ex) 벤젠 → c1ccccc1

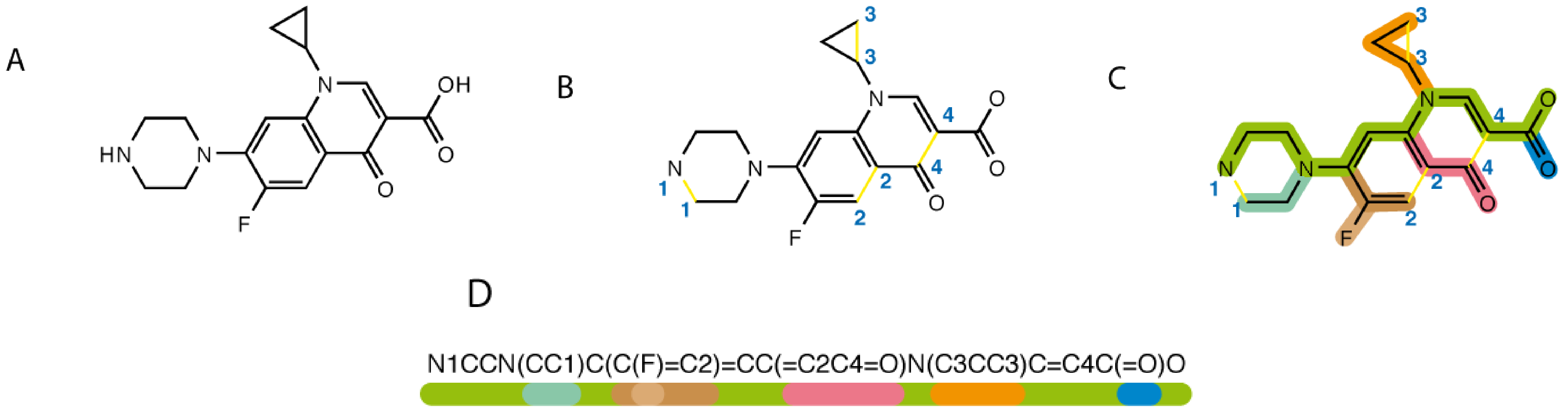


benzene

4. single, double 본드로 ring 표기하는 경우 인덱스를 표기하기 위해서 원자에 괄호 안에 :0, :1, :6 의 숫자를 붙였습니다. 1과 6 사이만 =이고, 나머지는 - 입니다. 원자1과 원자6 사이의 bond type을 적을 때는 ring index 숫자 바로 앞에 적습니다. 숫자 뒤에 적으면 숫자 다음 원자와의 bond type이 됩니다.



[CH1:1]=1[CH2:2]-C-C-C-[CH1:6]1



C의 main chain만 적어보겠습니다. → NCCNC=CCNC=CCO

여기에 side chain 괄호를 삽입해봅시다. 가지의 깊이는 1입니다. →
NCCN(CC)C(C=C)=CC(=CC=O)N(CCC)C=CC(=O)O

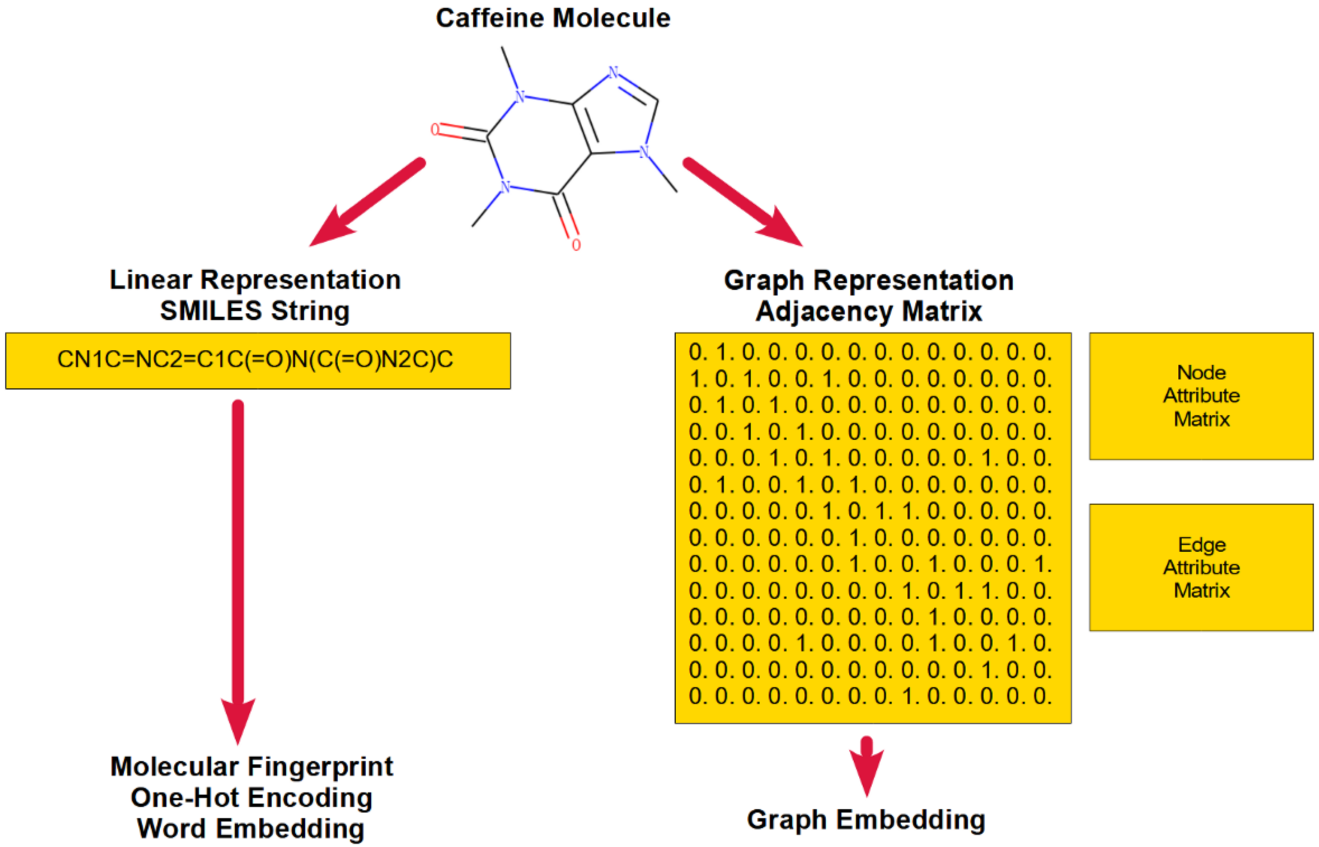
가지의 깊이를 2인 경우에는 괄호안에 괄호가 들어옵니다. 여기서는 C(F)=C 밖에 없습니다. →
NCCN(CC)C(C(F)=C)=CC(=CC=O)N(CCC)C=CC(=O)O

Ring 연결을 위해서 ring 인덱스를 삽입해줍니다. →
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

이러면 위와 같은 SMILES을 얻을 수 있습니다.

https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Molecular descriptor : Graph



Caffeine Molecule

Linear Representation
SMILES String

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Molecular Fingerprint
One-Hot Encoding
Word Embedding

Graph Representation
Adjacency Matrix

```
0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
1. 0. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.
0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0.
0. 1. 0. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.
0. 0. 0. 0. 0. 0. 1. 0. 0. 1. 0. 0. 0. 1.
0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0.
0. 0. 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 1. 0.
0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.
0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.
```

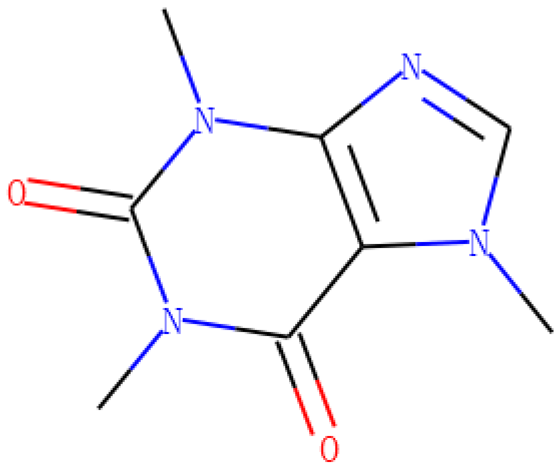
Node
Attribute
Matrix

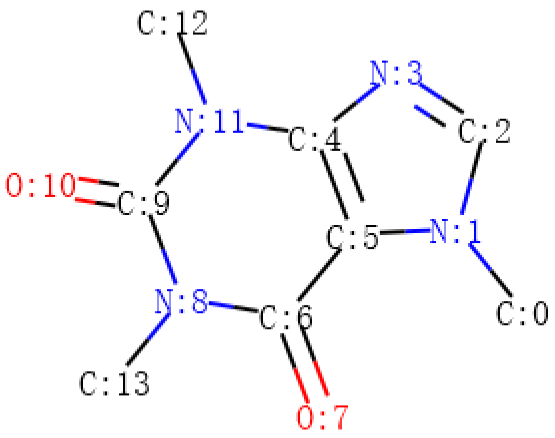
Edge
Attribute
Matrix

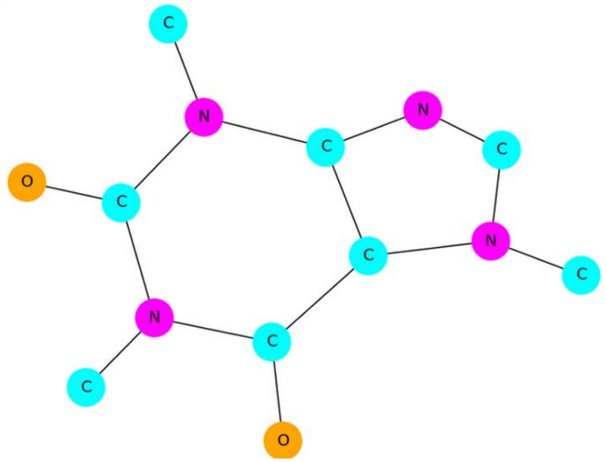
Graph Embedding

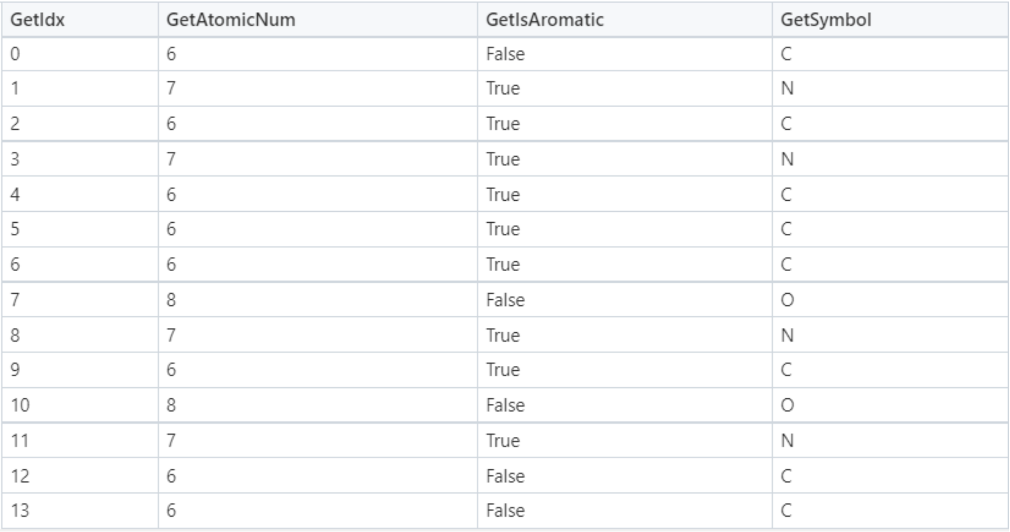
https://pasaentuciudad.com.mx/basic-molecular-representation-for-machine-learning/

Molecular descriptor : Graph

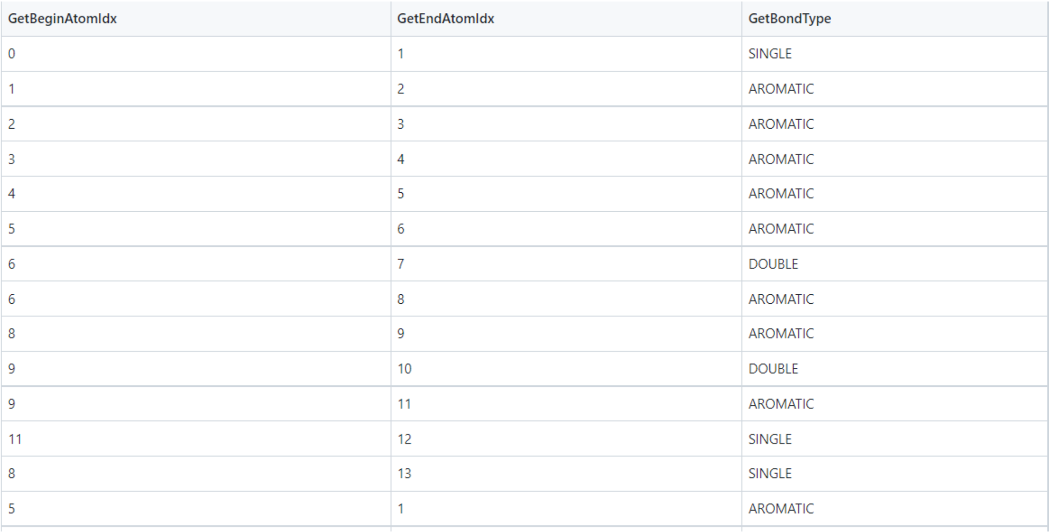






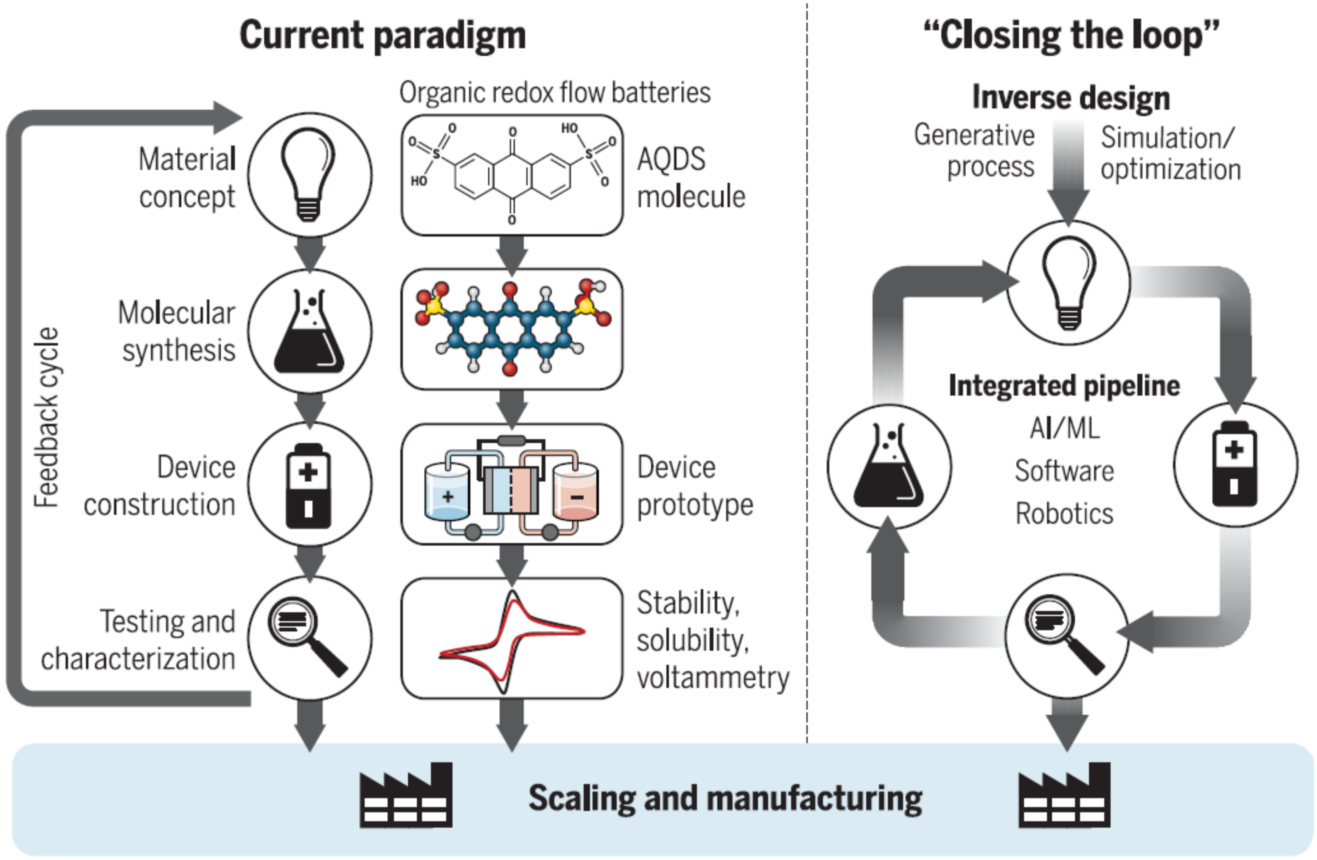


GetIdx	GetAtomicNum	GetIsAromatic	GetSymbol
0	6	False	C
1	7	True	N
2	6	True	C
3	7	True	N
4	6	True	C
5	6	True	C
6	6	True	C
7	8	False	O
8	7	True	N
9	6	True	C
10	8	False	O
11	7	True	N
12	6	False	C
13	6	False	C



GetBeginAtomIdx	GetEndAtomIdx	GetBondType
0	1	SINGLE
1	2	AROMATIC
2	3	AROMATIC
3	4	AROMATIC
4	5	AROMATIC
5	6	AROMATIC
6	7	DOUBLE
6	8	AROMATIC
8	9	AROMATIC
9	10	DOUBLE
9	11	AROMATIC
11	12	SINGLE
8	13	SINGLE
5	1	AROMATIC

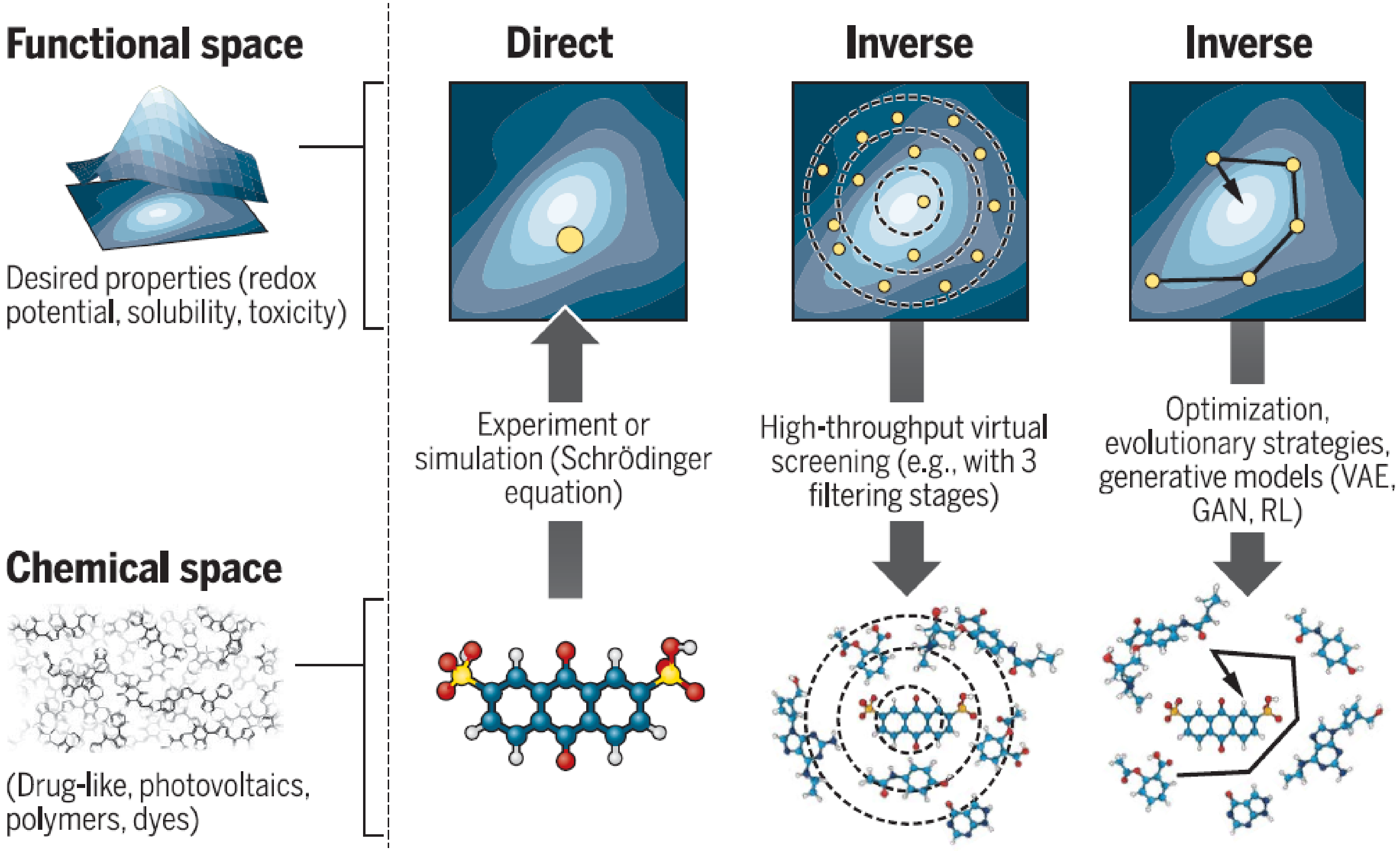
Inverse molecular design by machine learning



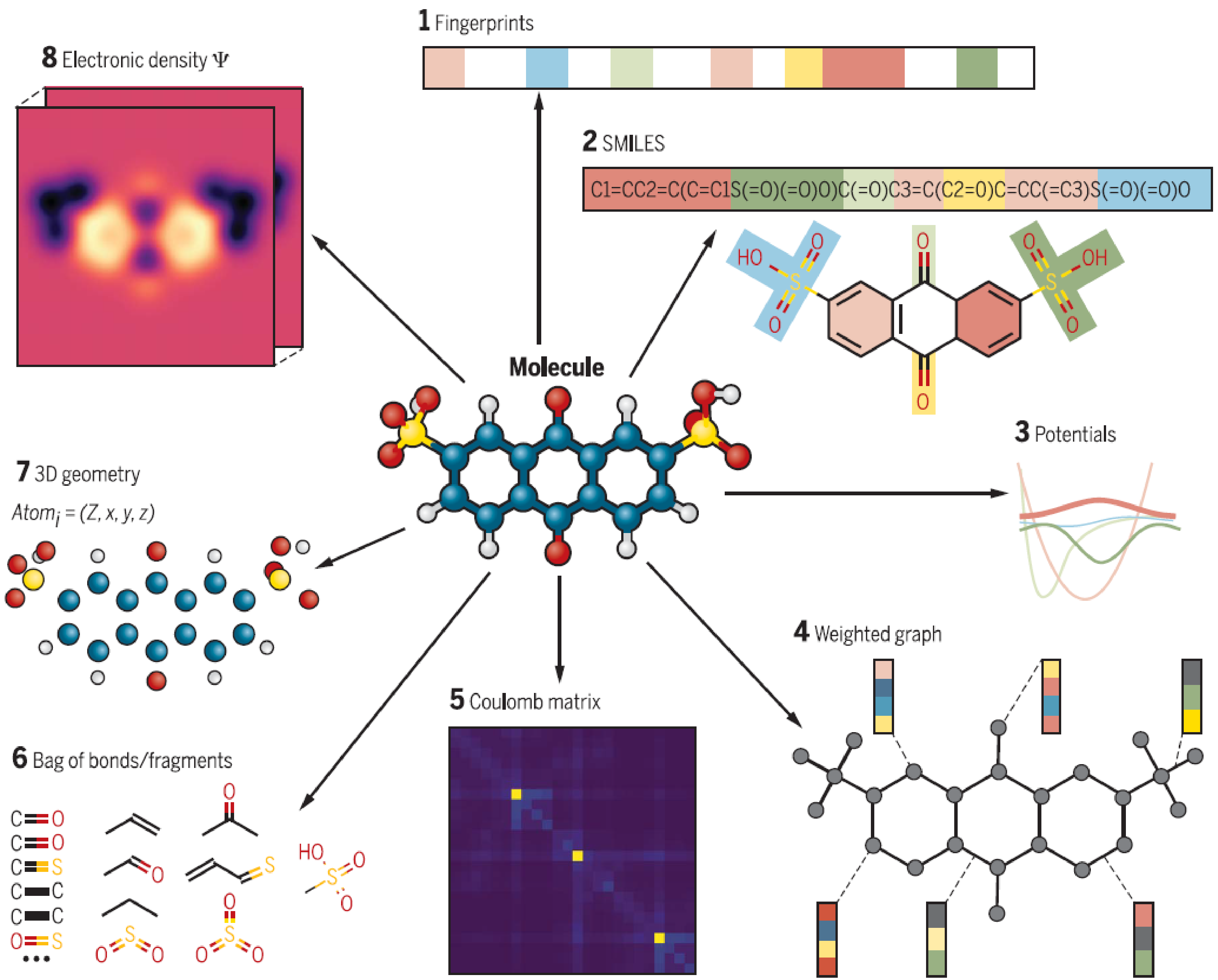
Schematic comparison of material discovery paradigms.

Sanchez-Lengeling et al., Science 361, 360–365 (2018)

Inverse molecular design by machine learning

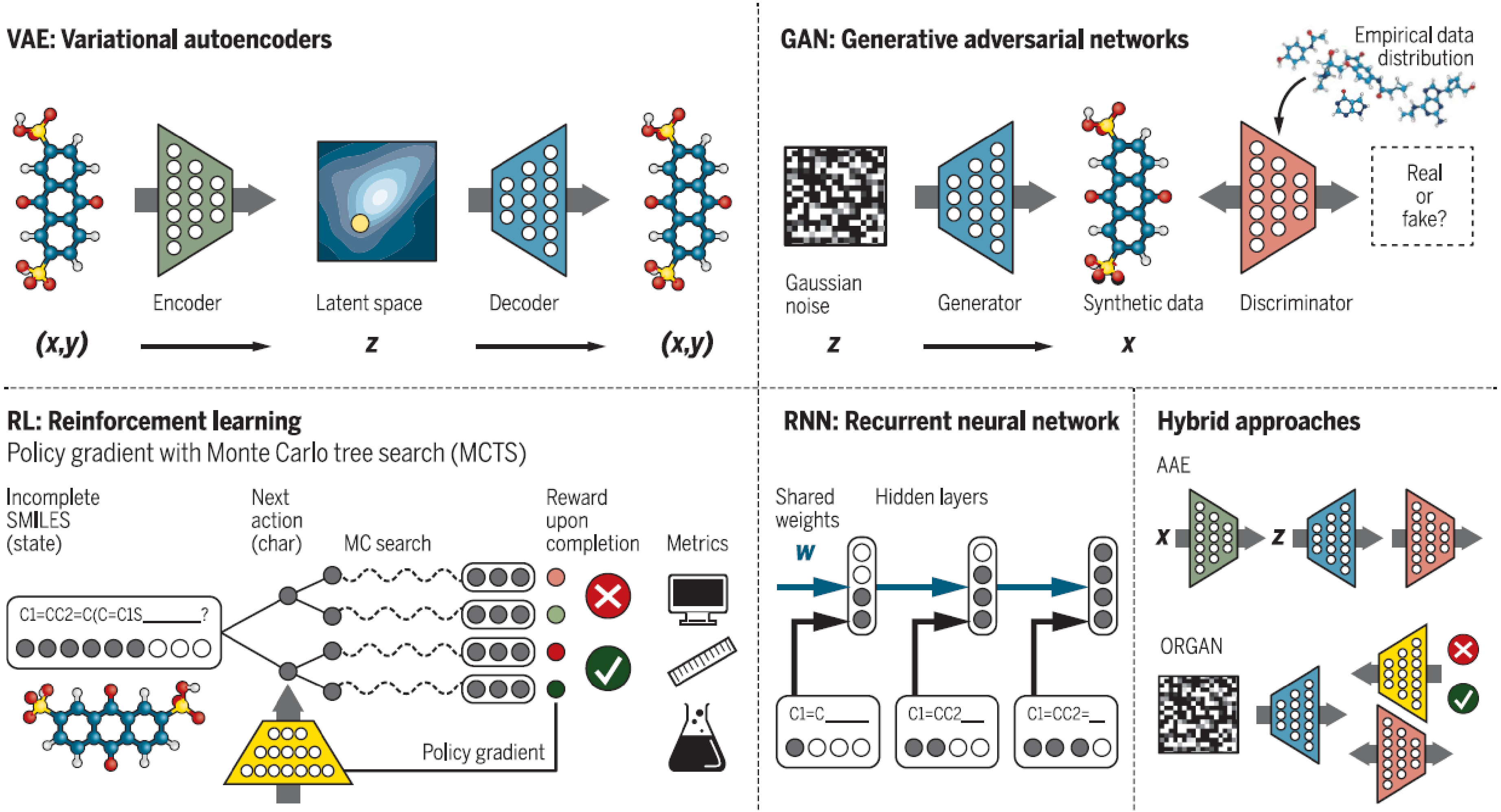

Functional space
Desired properties (redox potential, solubility, toxicity)
Direct
Inverse
Inverse
Experiment or simulation (Schrödinger equation)
High-throughput virtual screening (e.g., with 3 filtering stages)
Optimization, evolutionary strategies, generative models (VAE, GAN, RL)
Chemical space
(Drug-like, photovoltaics, polymers, dyes)

Sanchez-Lengeling et al., Science 361, 360–365 (2018)



Sanchez-Lengeling et al., Science 361, 360–365 (2018)

Inverse molecular design by machine learning



Sanchez-Lengeling et al., Science 361, 360–365 (2018)

Molecular de-novo design through deep reinforcement learning

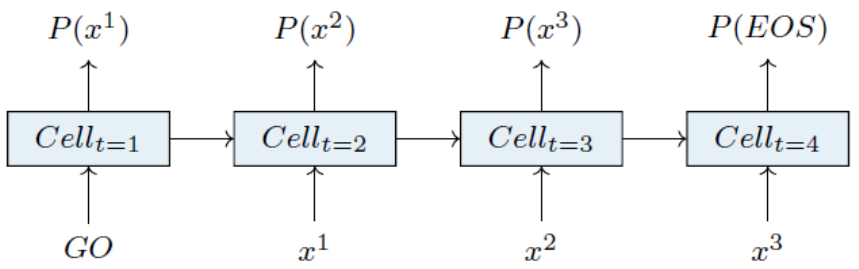


Fig. 1 Learning the data. Depiction of maximum likelihood training of an RNN. x^t are the target sequence tokens we are trying to learn by maximizing $P(x^t)$ for each step

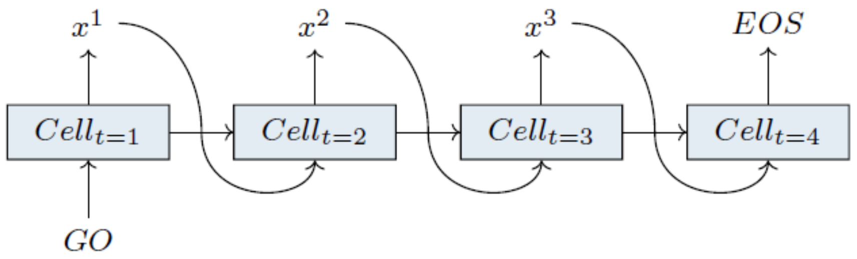
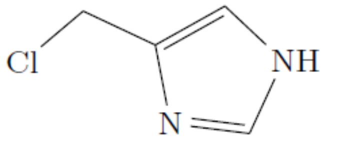


Fig. 2 Generating sequences. Sequence generation by a trained RNN. Every timestep t we sample the next token of the sequence x^t from the probability distribution given by the RNN, which is then fed in as the next input

Graph:



SMILES: ClCc1c[nH]cn1

One-hot encoding:

	Cl	C	c	1	c	nH	c	n	1
C	0	1	0	0	0	0	0	0	0
c	0	0	1	0	1	0	1	0	0
n	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
nH	0	0	0	0	0	1	0	0	0
Cl	1	0	0	0	0	0	0	0	0

Fig. 3 Three representations of 4-(chloromethyl)-1H-imidazole. Depiction of a one-hot representation derived from the SMILES of a molecule. Here a reduced vocabulary is shown, while in practice a much larger vocabulary that covers all tokens present in the training data is used

Olivecrona *et al. J Cheminform (2017) 9:48*

Molecular de-novo design through deep reinforcement learning

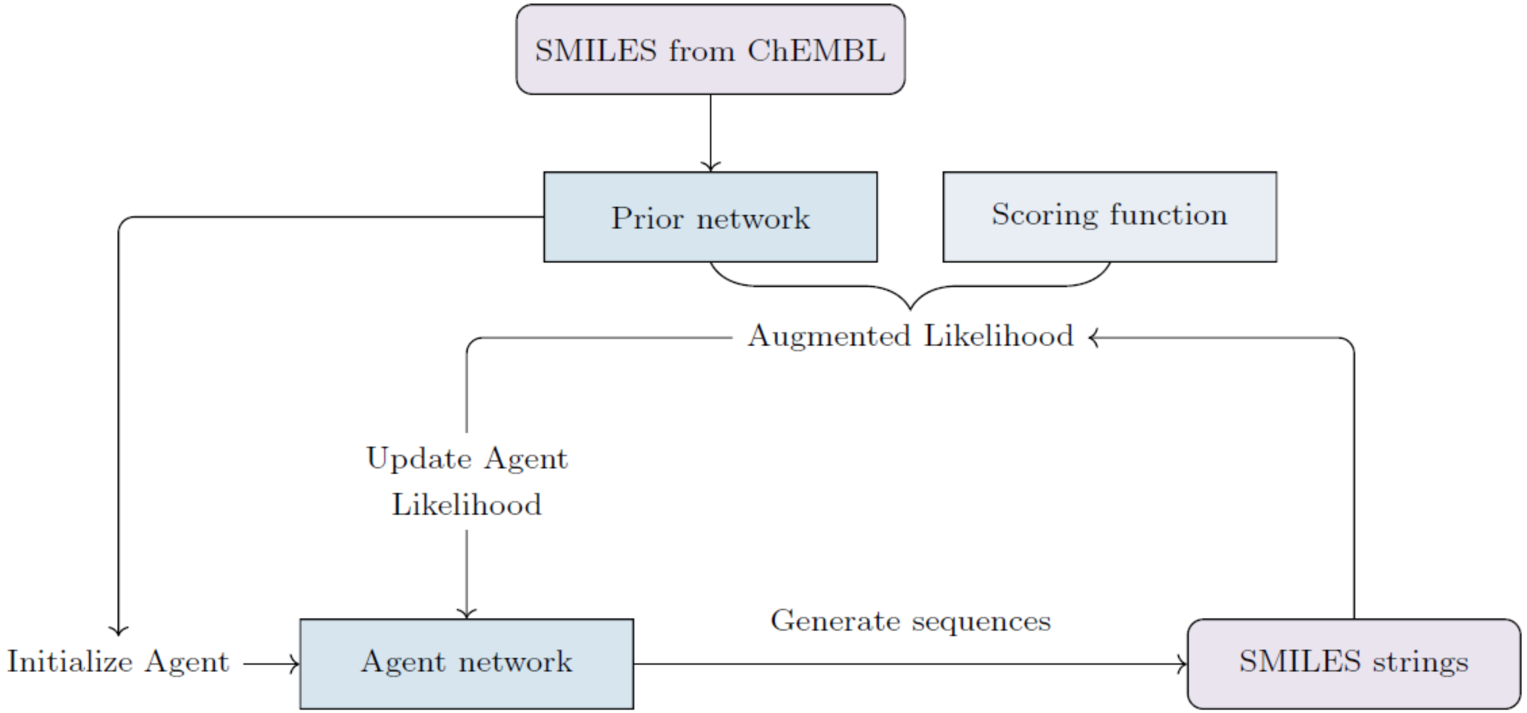


Fig. 4 The Agent. Illustration of how the model is constructed. Starting from a Prior network trained on ChEMBL, the Agent is trained using the augmented likelihood of the SMILES generated

Olivecrona *et al. J Cheminform (2017) 9:48*

Molecular de-novo design through deep reinforcement learning

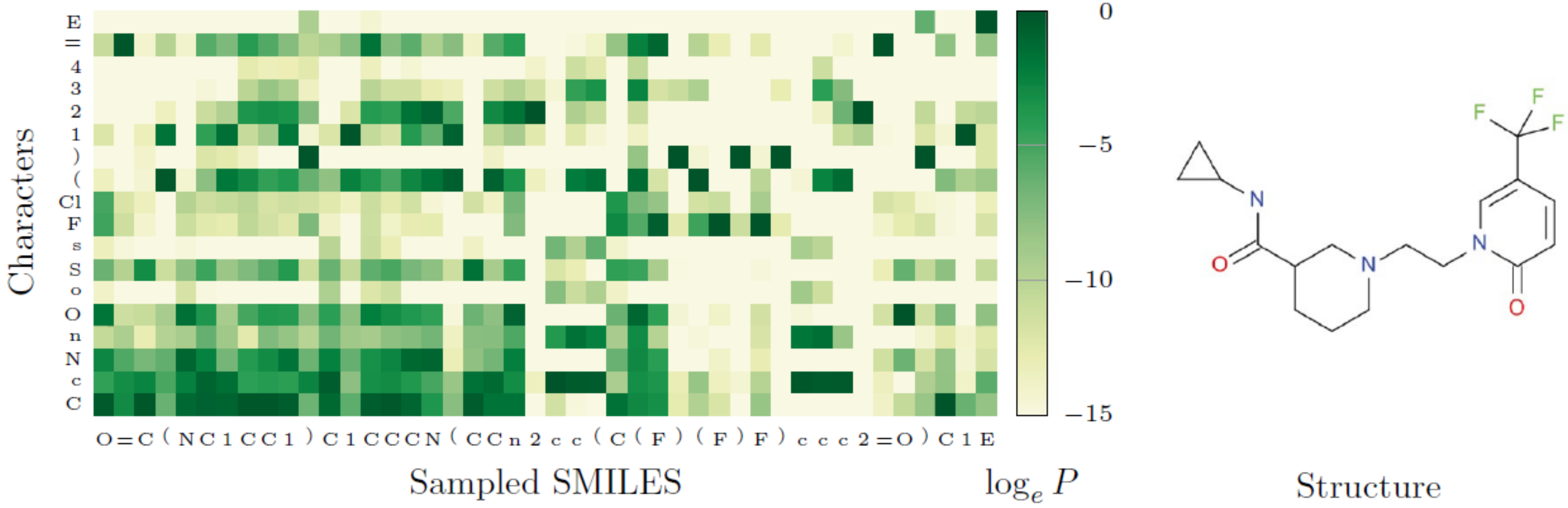


Fig. 5 How the model thinks while generating the molecule on the right. Conditional probability over the next token as a function of previously chosen ones according to the model. On the y-axis is shown the probability distribution for the character to be choosen at the current step, and on the x-axis is shown the character that in this instance was sampled. E = EOS

Olivecrona *et al. J Cheminform (2017) 9:48*

Molecular de-novo design through deep reinforcement learning

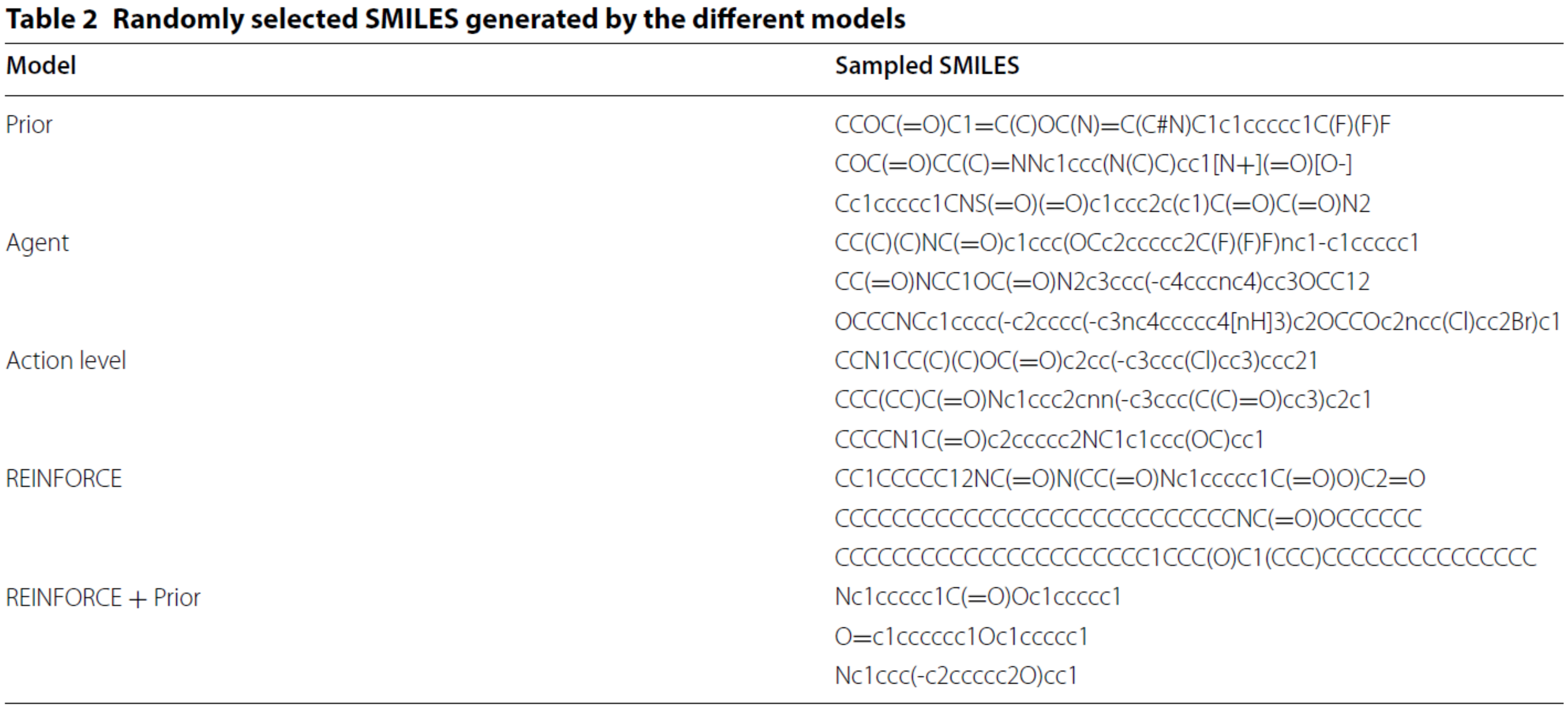


Table 2 Randomly selected SMILES generated by the different models

Model	Sampled SMILES
Prior	CCOC(=O)C1=C(C)OC(N)=C(C#N)C1c1ccccc1C(F)(F)F
	COC(=O)CC(C)=NNc1ccc(N(C)C)cc1[N+](=O)[O-]
	Cc1ccccc1CNS(=O)(=O)c1ccc2c(c1)C(=O)C(=O)N2
Agent	CC(C)(C)NC(=O)c1ccc(OCc2ccccc2C(F)(F)F)nc1-c1ccccc1
	CC(=O)NCC1OC(=O)N2c3ccc(-c4cccnc4)cc3OCC12
	OCCCNCc1cccc(-c2cccc(-c3nc4ccccc4[nH]3)c2OCCOc2ncc(Cl)cc2Br)c1
Action level	CCN1CC(C)(C)OC(=O)c2cc(-c3ccc(Cl)cc3)ccc21
	CCC(CC)C(=O)Nc1ccc2cnn(-c3ccc(C(C)=O)cc3)c2c1
	CCCCN1C(=O)c2ccccc2NC1c1ccc(OC)cc1
REINFORCE	CC1CCCCC12NC(=O)N(CC(=O)Nc1ccccc1C(=O)O)C2=O
	CCCCCCCCCCCCCCCCCCCCCCCCCCCCNC(=O)OCCCCCC
	CCCCCCCCCCCCCCCCCCCCCC1CCC(O)C1(CCC)CCCCCCCCCCCCCCC
REINFORCE + Prior	Nc1ccccc1C(=O)Oc1ccccc1
	O=c1cccccc1Oc1ccccc1
	Nc1ccc(-c2ccccc2O)cc1

Olivecrona *et al. J Cheminform (2017) 9:48*

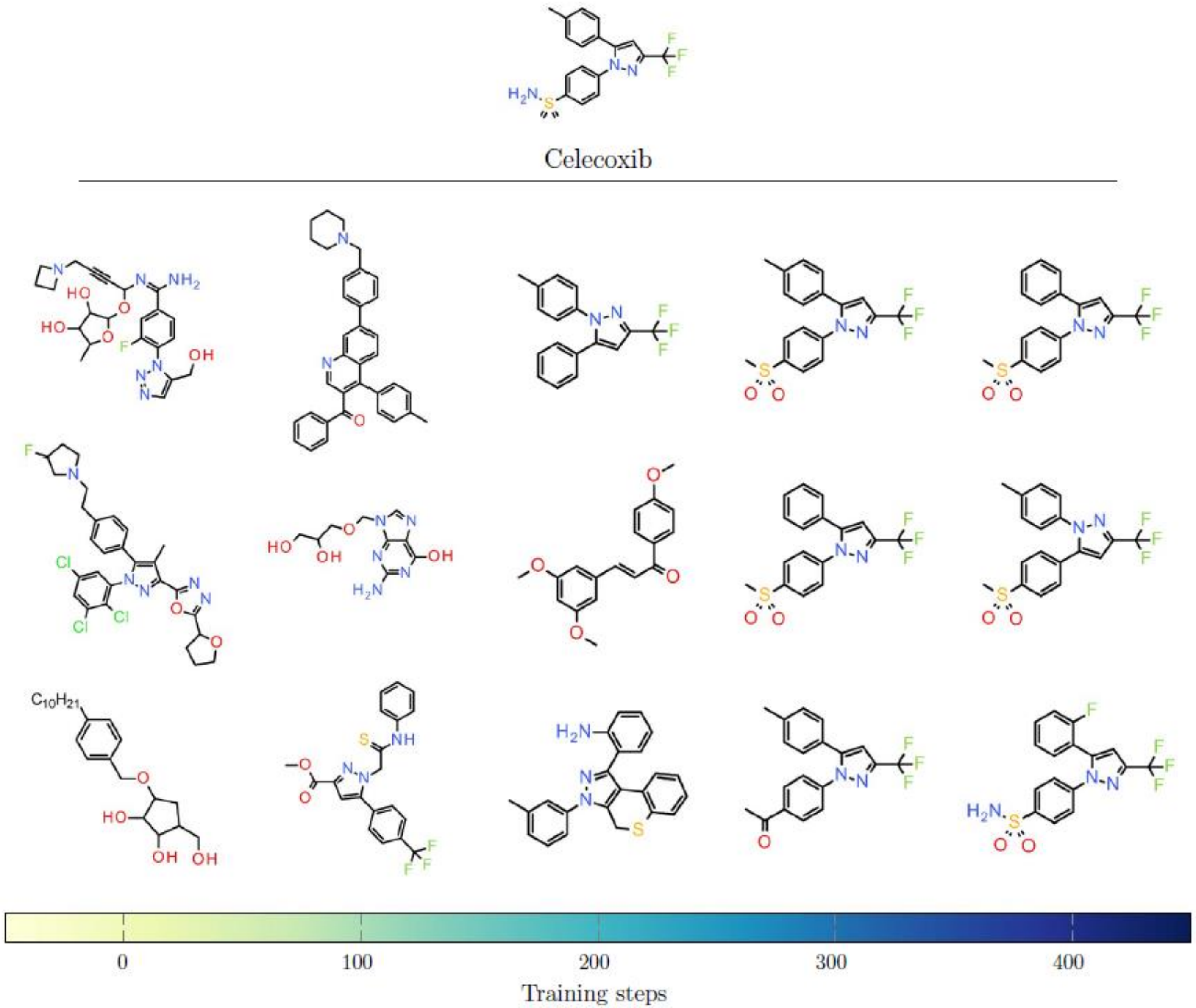
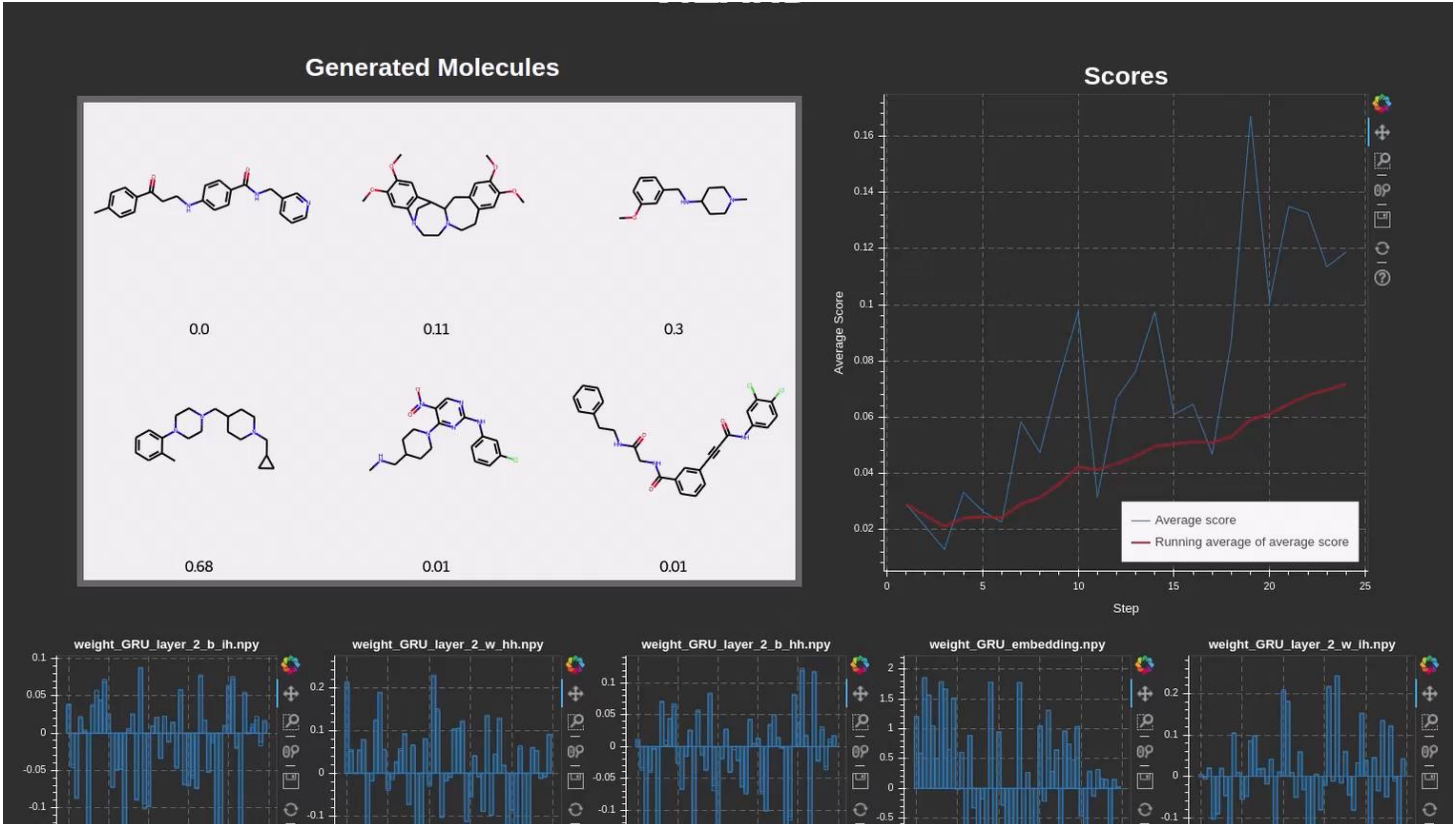


Fig. 7 Evolution of generated structures during training Structures sampled every 100 training steps during the training of the Agent towards similarity to Celecoxib with $k = 0.7$ and $\sigma = 15$

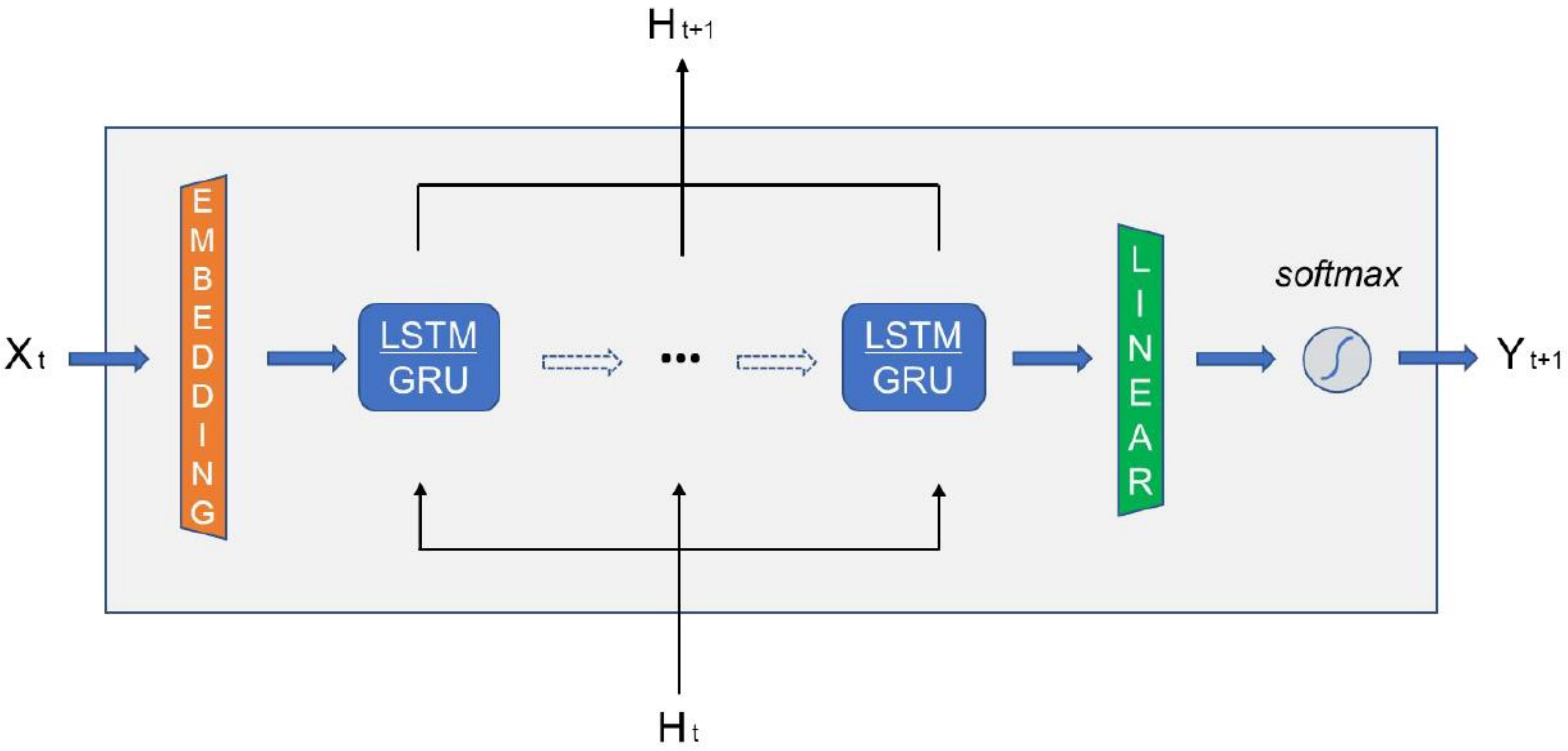
Olivecrona *et al. J Cheminform (2017) 9:48*



Generated Molecules

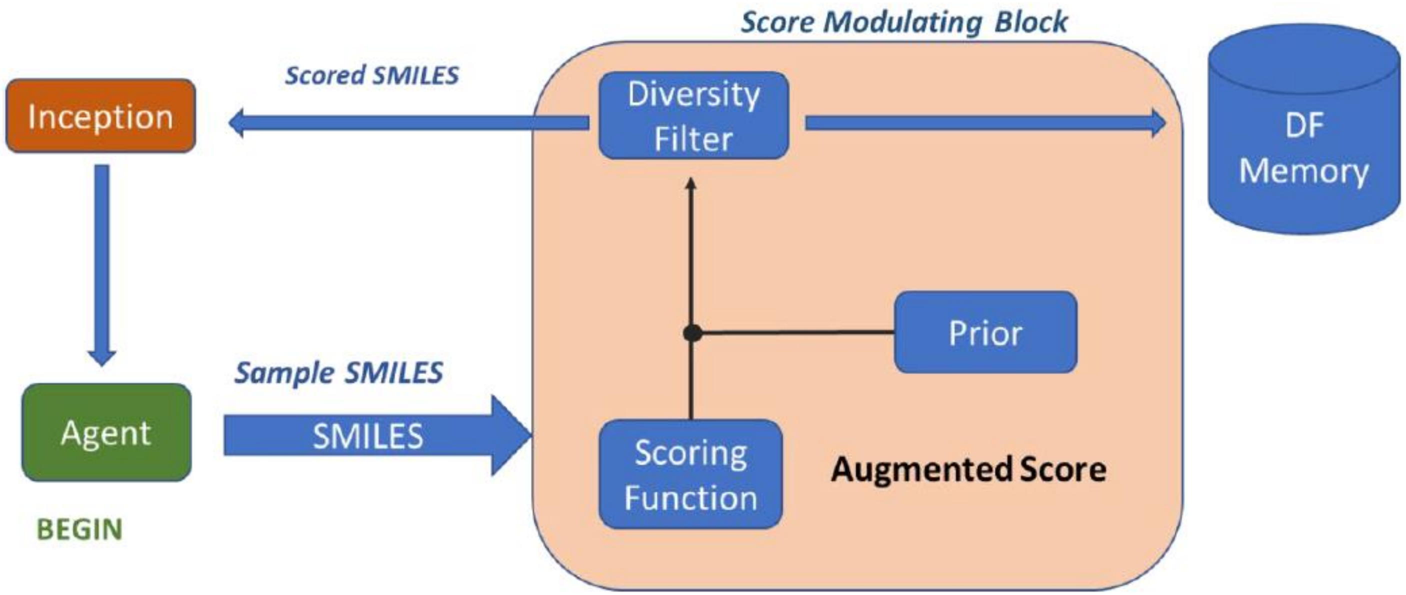
Scores

REINVENT 2.0 – an AI tool for de novo drug design



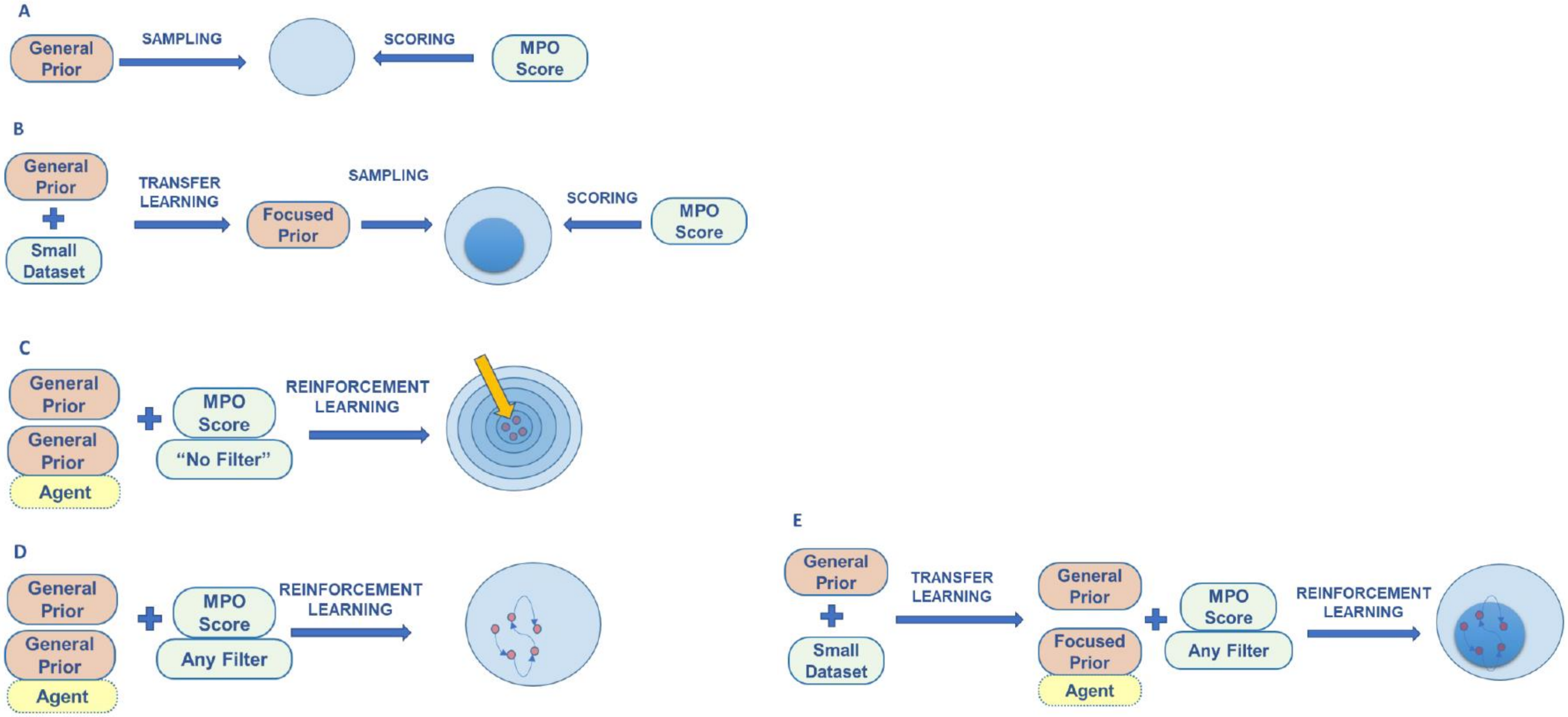
J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

REINVENT 2.0 – an AI tool for de novo drug design



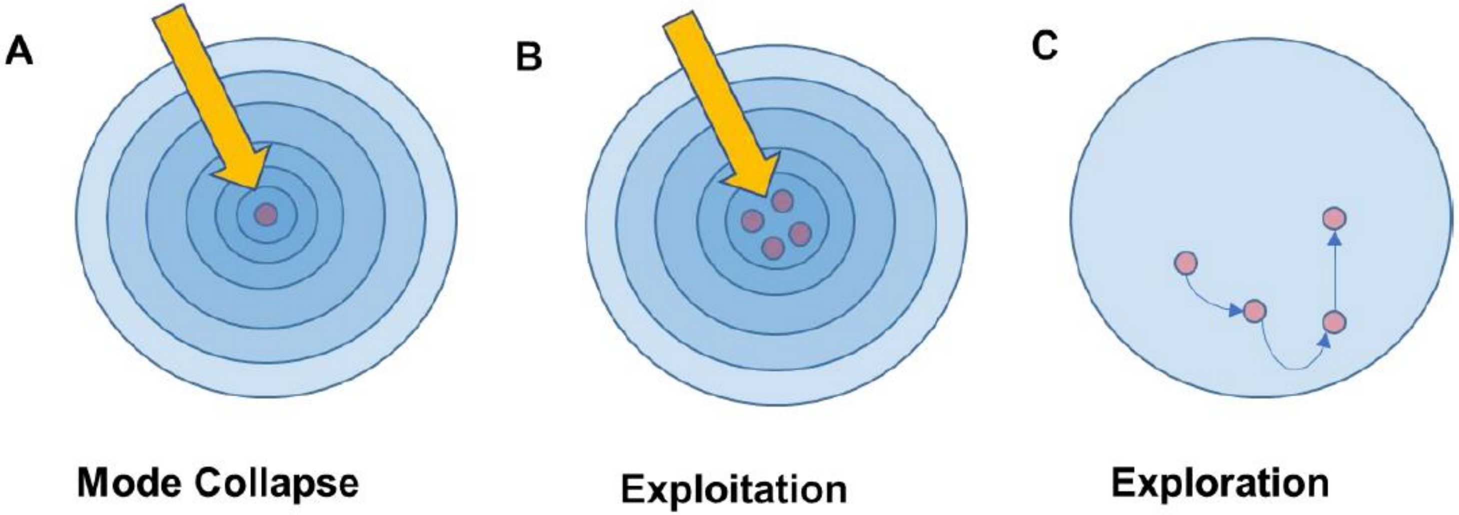
J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

REINVENT 2.0 – an AI tool for de novo drug design



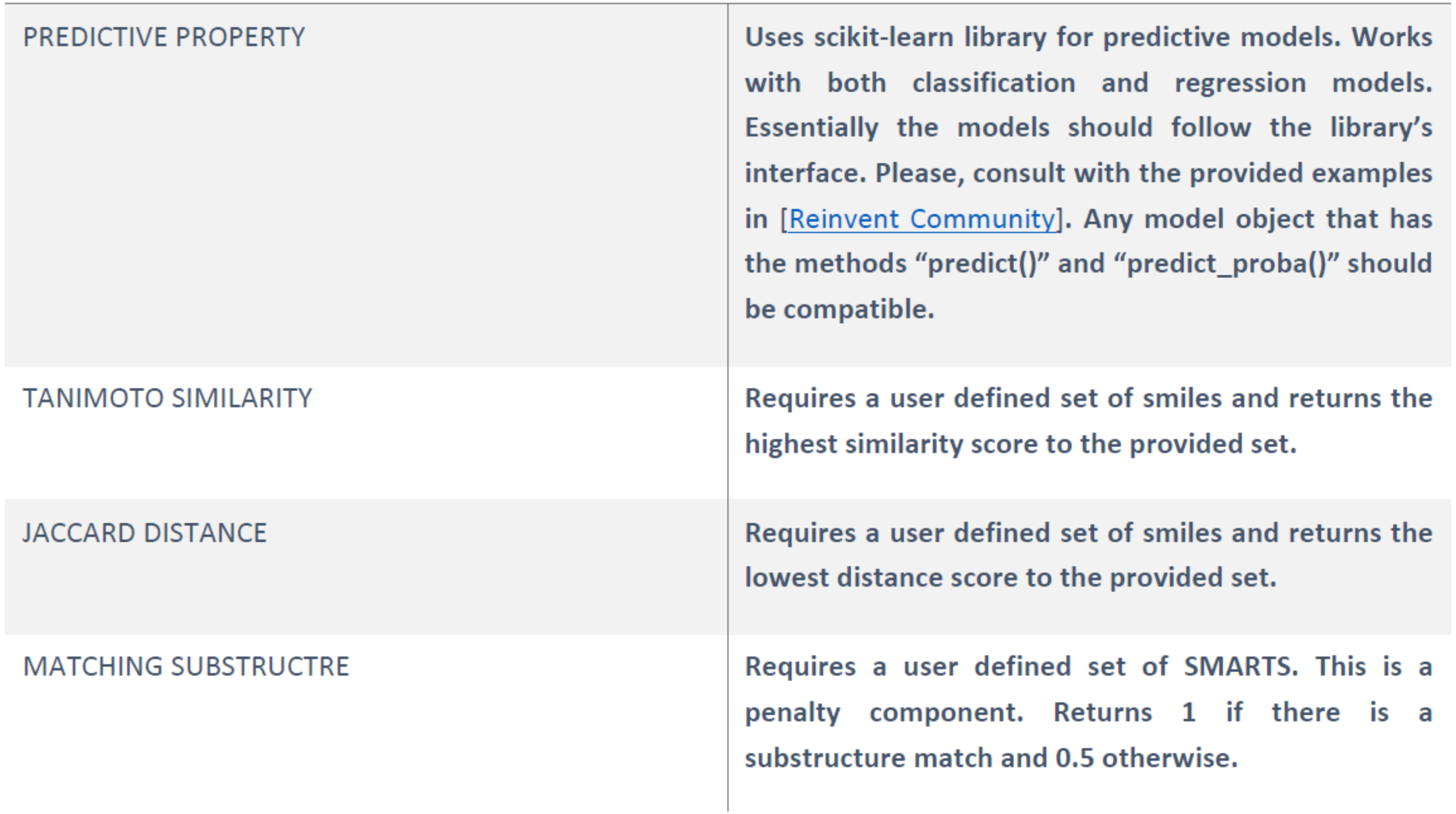
J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

REINVENT 2.0 – an AI tool for de novo drug design



J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

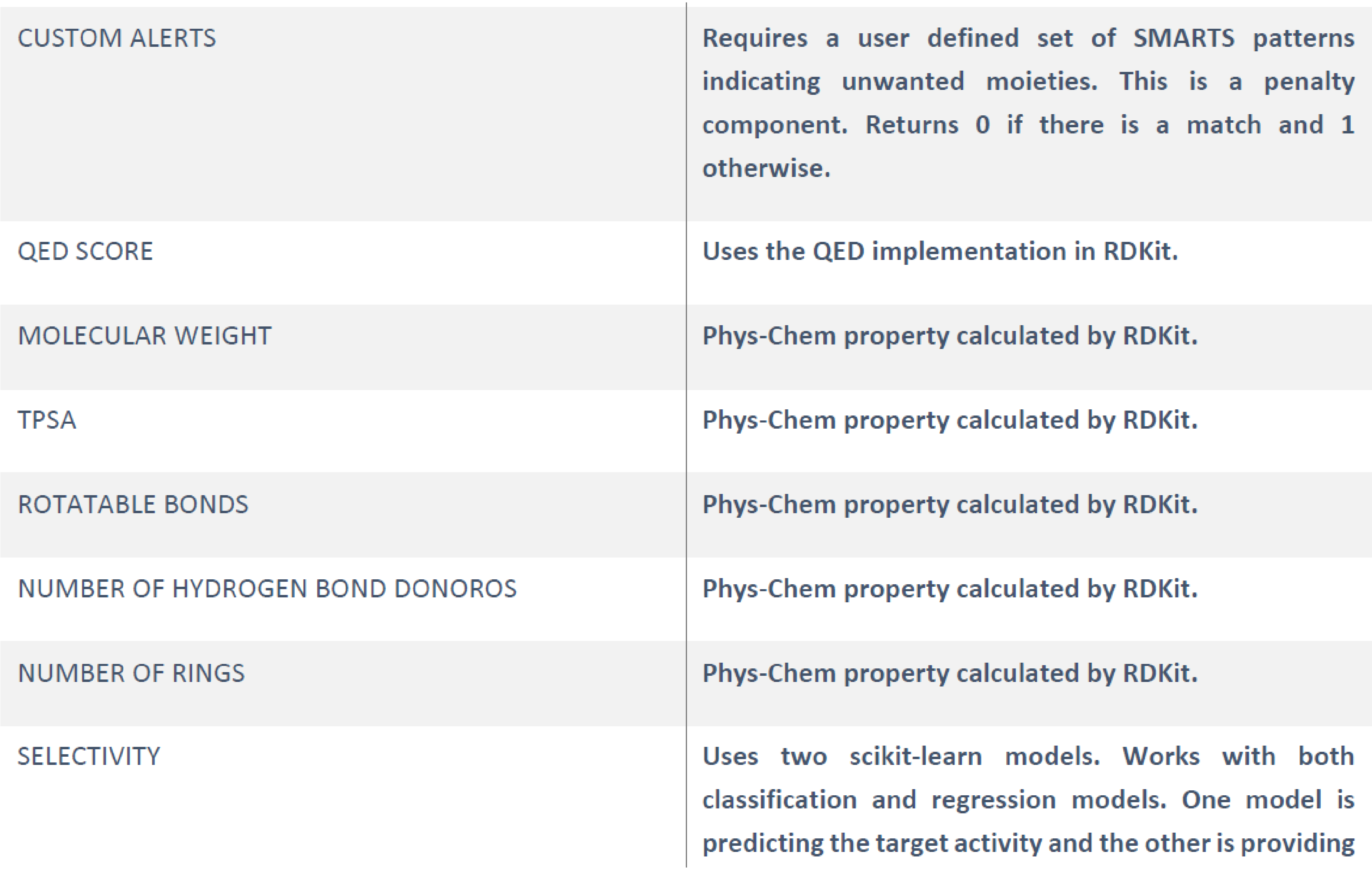
REINVENT 2.0 – an AI tool for de novo drug design



PREDICTIVE PROPERTY	Uses scikit-learn library for predictive models. Works with both classification and regression models. Essentially the models should follow the library's interface. Please, consult with the provided examples in [Reinvent Community]. Any model object that has the methods "predict()" and "predict_proba()" should be compatible.
TANIMOTO SIMILARITY	Requires a user defined set of smiles and returns the highest similarity score to the provided set.
JACCARD DISTANCE	Requires a user defined set of smiles and returns the lowest distance score to the provided set.
MATCHING SUBSTRUCTRE	Requires a user defined set of SMARTS. This is a penalty component. Returns 1 if there is a substructure match and 0.5 otherwise.

J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

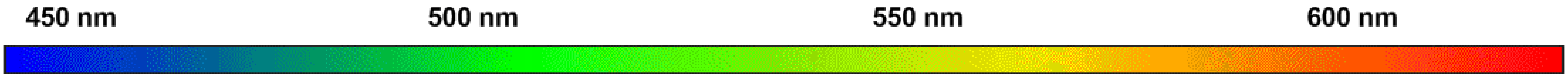
REINVENT 2.0 – an AI tool for de novo drug design

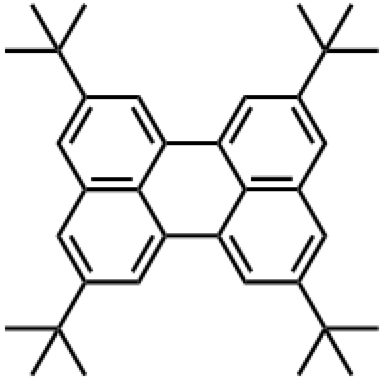


CUSTOM ALERTS	Requires a user defined set of SMARTS patterns indicating unwanted moieties. This is a penalty component. Returns 0 if there is a match and 1 otherwise.
QED SCORE	Uses the QED implementation in RDKit.
MOLECULAR WEIGHT	Phys-Chem property calculated by RDKit.
TPSA	Phys-Chem property calculated by RDKit.
ROTATABLE BONDS	Phys-Chem property calculated by RDKit.
NUMBER OF HYDROGEN BOND DONOROS	Phys-Chem property calculated by RDKit.
NUMBER OF RINGS	Phys-Chem property calculated by RDKit.
SELECTIVITY	Uses two scikit-learn models. Works with both classification and regression models. One model is predicting the target activity and the other is providing

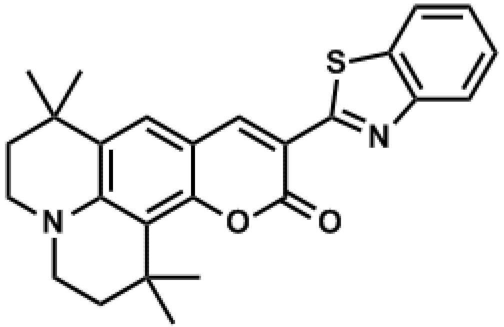
J. Chem. Inf. Model. 2020, 60, 12, 5918−5922

~~Conclusion~~ → Issue?

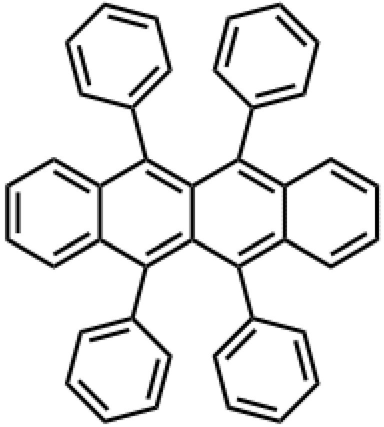




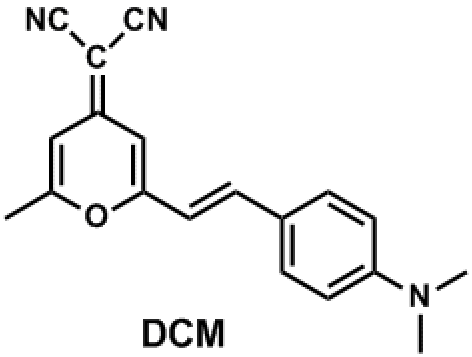
TBP
[T3053]



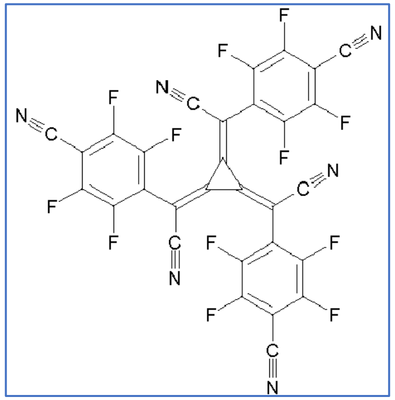
Coumarin 545T
[B4257]

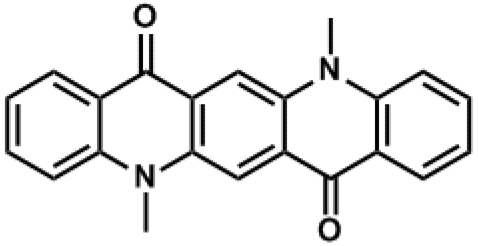


Rubrene
[T0561, T2233]

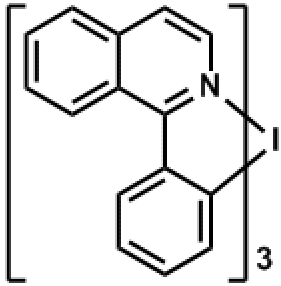


DCM
[D2849]





DMQA
[D2687, D3227]



Ir(piq)$_3$
[T2685]