3rd conference on machine learning
FALL into ML 2024

Exploring the Potential of AI in Ligand-Based Drug Design

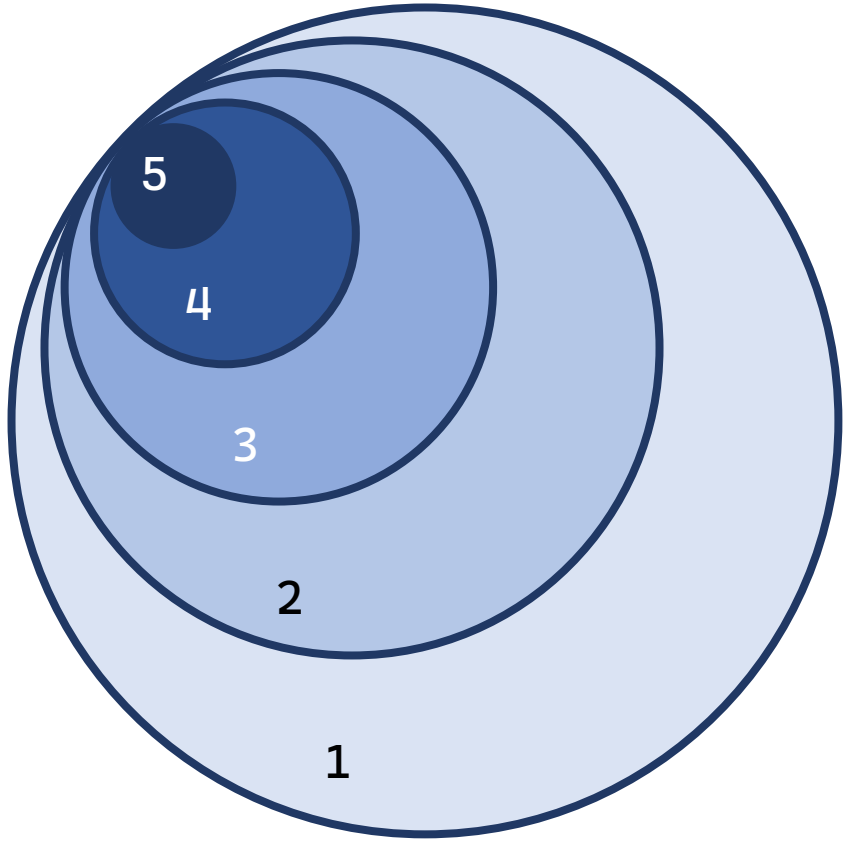
Medicinal Chemistry Toolbox

26.10.2024

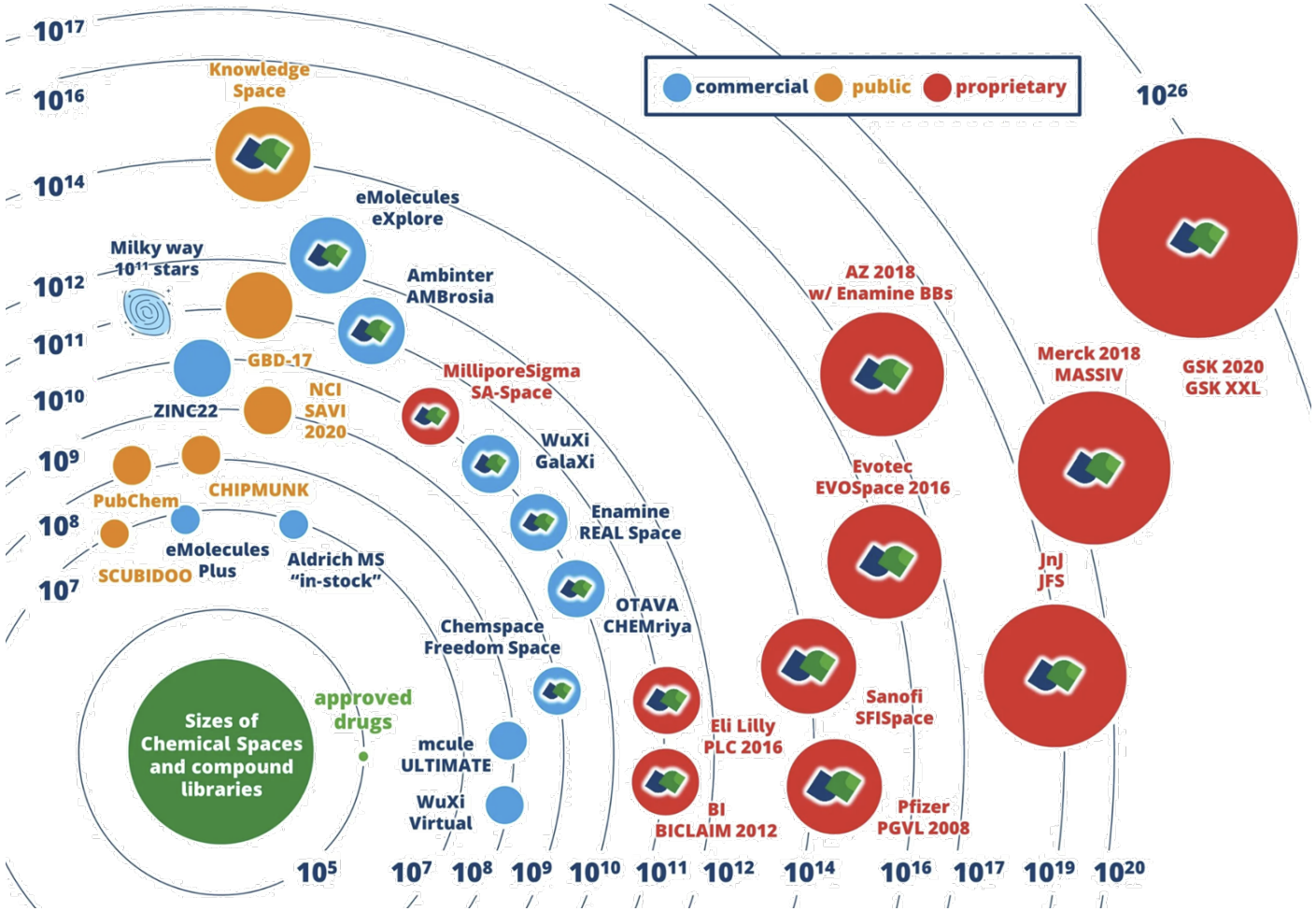
Alex Malyshev

Chemical space

1. **Theoretically possible molecules**
   ~10^{100}-10^{200}
2. **Drug-like molecules**
   ~10^{50}-10^{60}
3. **Synthetically accessible drug-like molecules**
   ~10^{10}-10^{11}
4. **Synthesized drug-like molecules**
   ~10^{7}-10^{8}
5. **Diverse hits**
   ~10^{4}-10^{5}



Drug-like libraries. Chemical databases



Cortellis Drug Discovery Intelligence

CAS SciFinder

Reaxys

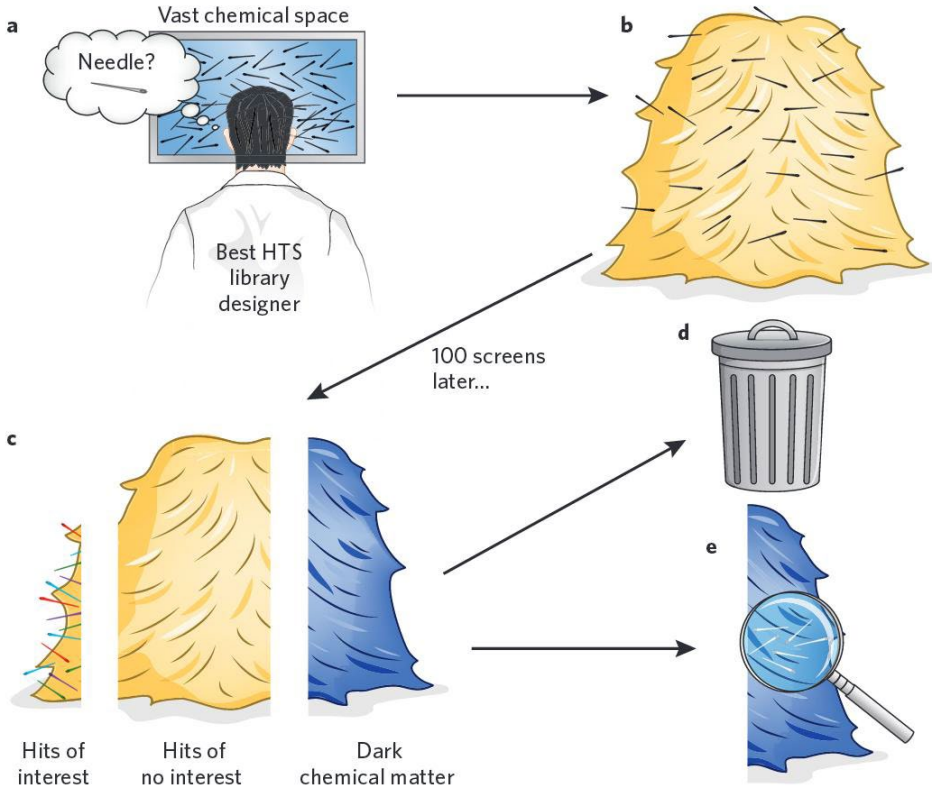
GOSTAR

ChEMBL

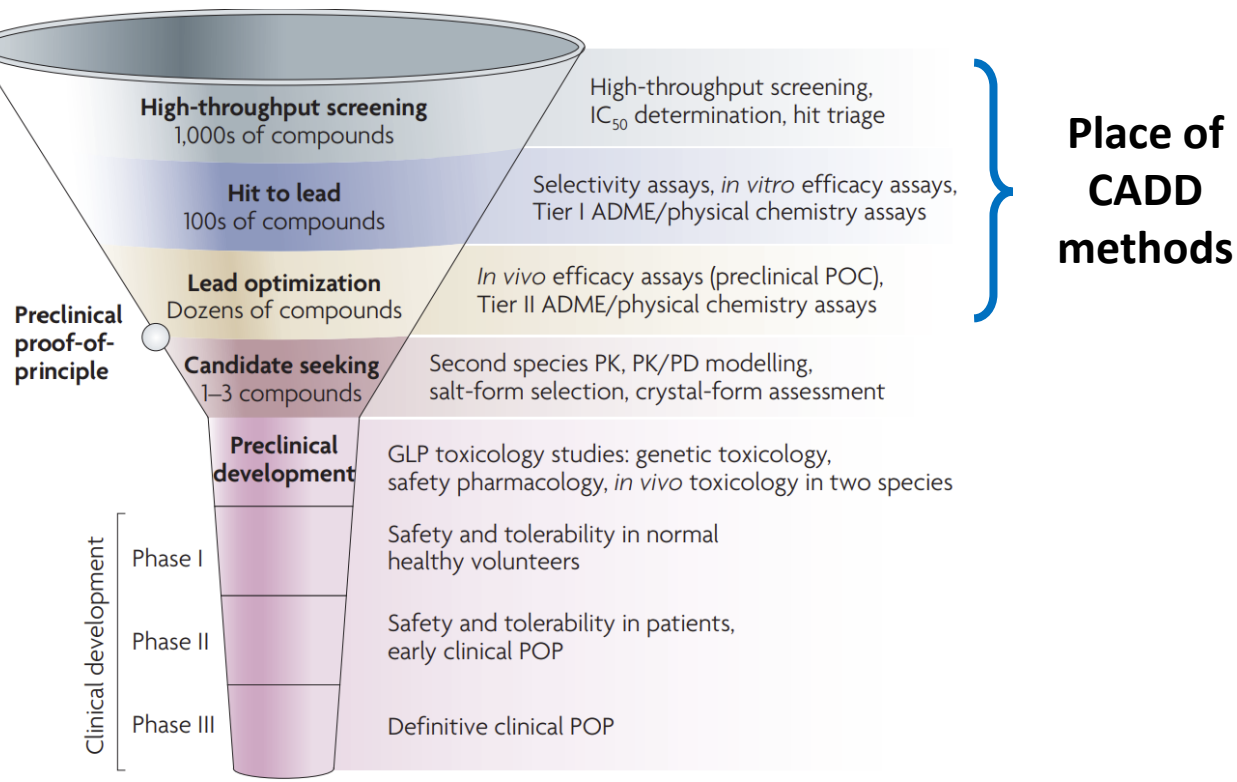
Cambridge Structural Database

https://www.biosolveit.de/chemical-spaces/

Drug discovery



10.1021/cb100420r

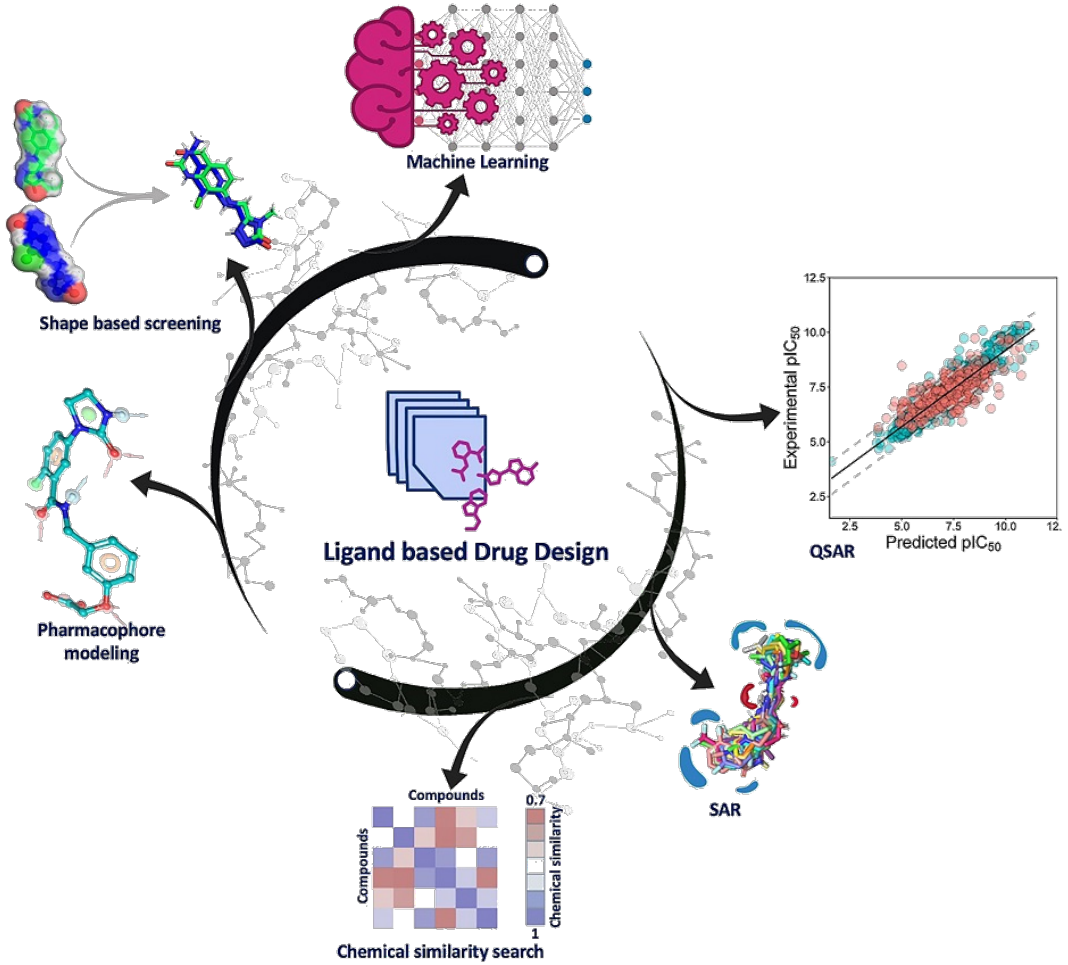


Place of CADD methods

10.1038/nrd2378

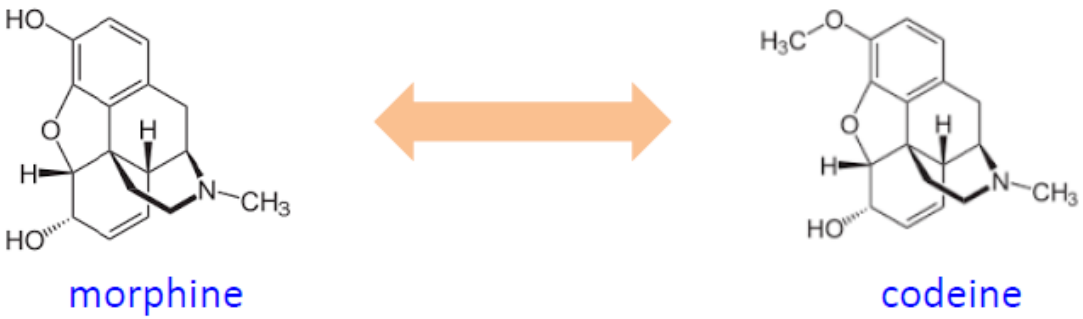
Rational drug design techniques

	Known ligand	Unknown ligand
Known target structure	**Structure-based drug design (SBDD)** Docking	***De novo* design**
Unknown target structure	**Ligand-based drug design (LBDD)** *1 or more ligands* • Similarity search *Several ligands* • Pharmacophore *Large number of ligands (20+)* • Quantitative Structure-Activity Relationships (QSAR)	**CADD not possible** some experimental data needed ADMET filtering



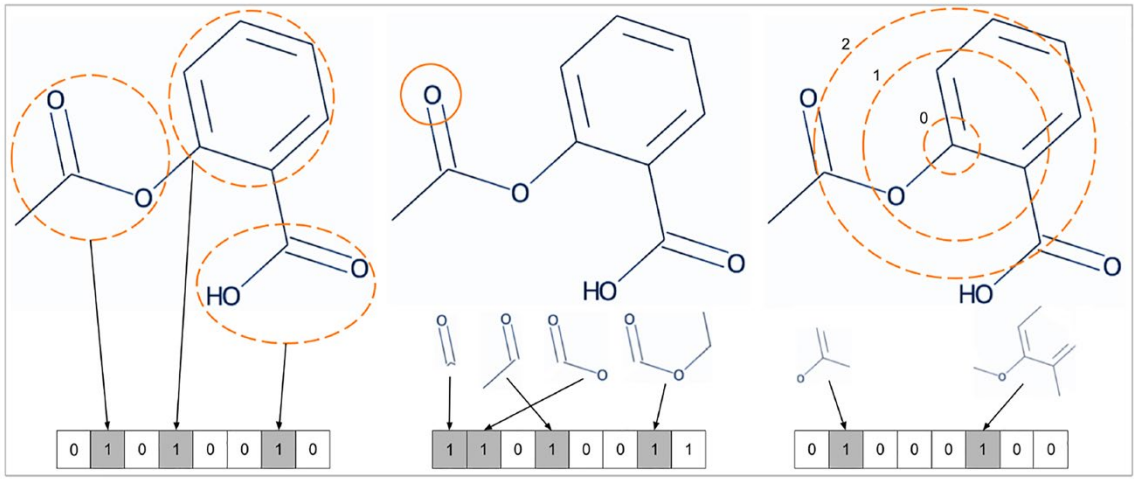
2D similarity

Similar compounds have similar properties



morphine codeine

Molecular Fingerprints

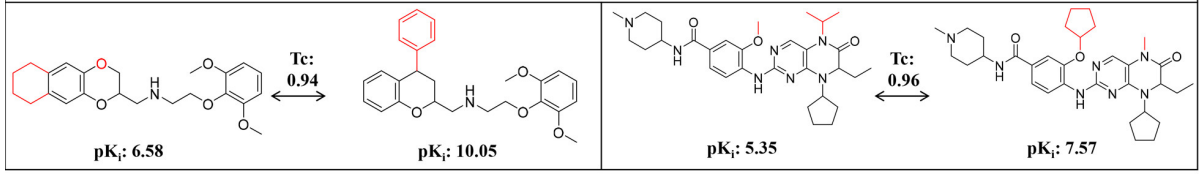


Structural Keys Path-based Circular

Name	Continuous	Dichotomous
Tanimoto coefficient	$T(x_a,x_b)=\frac{\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sum_{i=0}^{N}x_{ai}^{\ 2}+\sum_{i=0}^{N}x_{bi}^{\ 2}-\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}$	$T(x_a,x_b)=\frac{c}{a+b-c}$
Euclidean distance	$D(x_a,x_b)=\sqrt{\sum_{i=0}^{N}(x_{ai}\cdot x_{bi})^2}$	$D(x_a,x_b)=\sqrt{a+b-2c}$
Hamming distance	$D(x_a,x_b)=\sum_{i=0}^{N}\lvert x_{ai}-x_{bi}\rvert$	$D(x_a,x_b)=a+b-2c$
Cosine coefficient	$C(x_a,x_b)=\frac{\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N}x_{ai}^{\ 2}}\cdot\sqrt{\sum_{i=0}^{N}x_{bi}^{\ 2}}}$	$C(x_a,x_b)=\frac{c}{\sqrt{ab}}$
Dice coefficient	$D(x_a,x_b)=\frac{2\sum_{i=0}^{N}x_{ai}\cdot x_{bi}}{\sqrt{\sum_{i=0}^{N}x_{ai}^{\ 2}}\cdot\sqrt{\sum_{i=0}^{N}x_{bi}^{\ 2}}}$	$D(x_a,x_b)=\frac{2c}{a+b}$
Soergel	$S(x_a,x_b)=\frac{\sum_{i=0}^{N}\lvert x_{ai}-x_{bi}\rvert}{\sum_{i=0}^{N}\max(x_{ai},x_{bi})}$	$S(x_a,x_b)=\frac{a+b-2c}{a+b-c}$

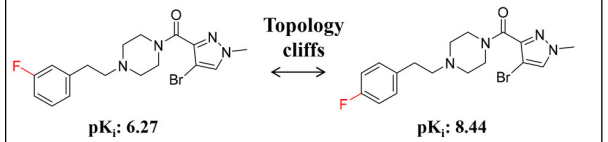
Activity and property cliffs

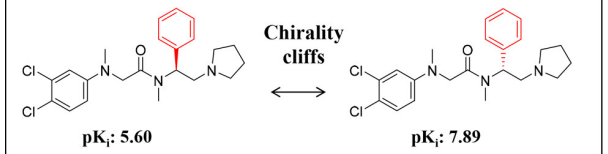
Fingerprint-based



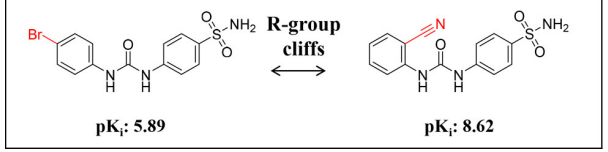
Substructure-based

Category 1: same scaffold, same R-groups

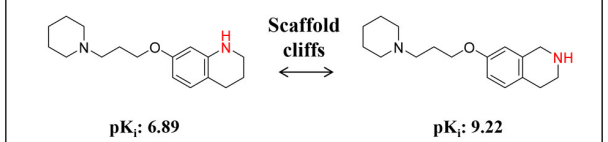


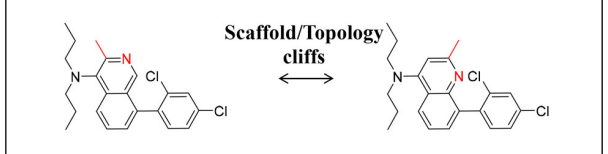


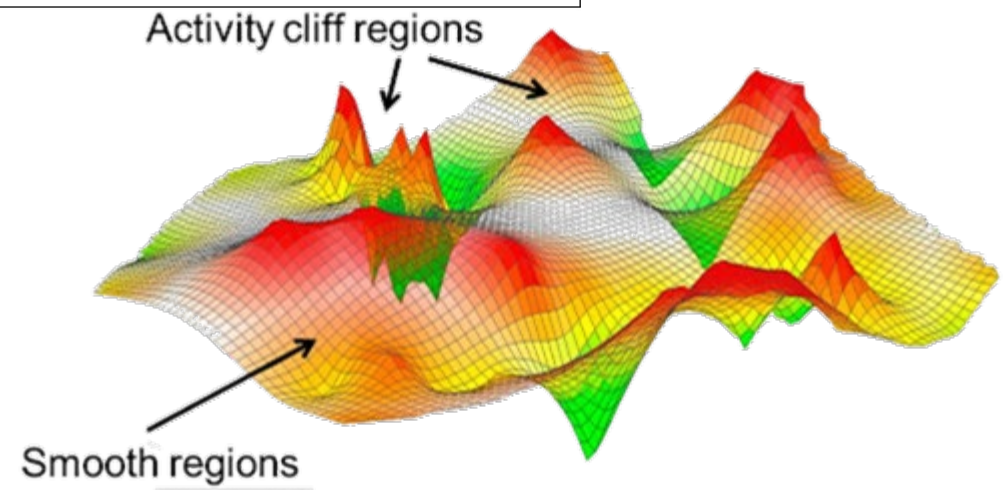
Category 2: same scaffold, different R-groups

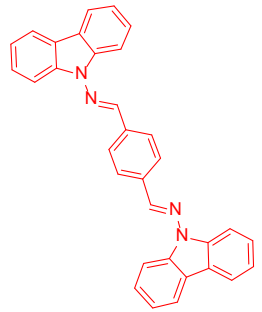
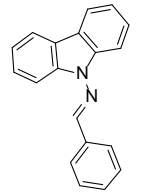
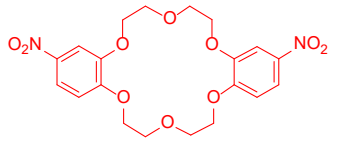
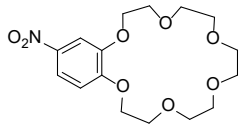
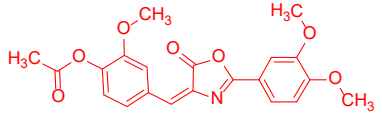
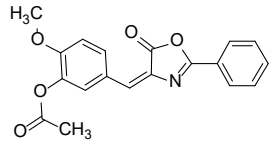
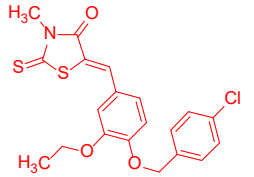
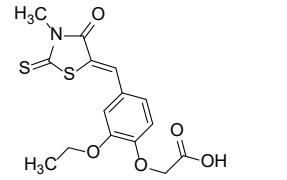
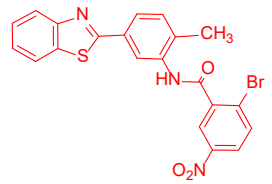
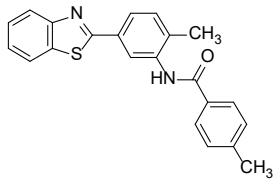


Category 3: different scaffolds, same R-groups





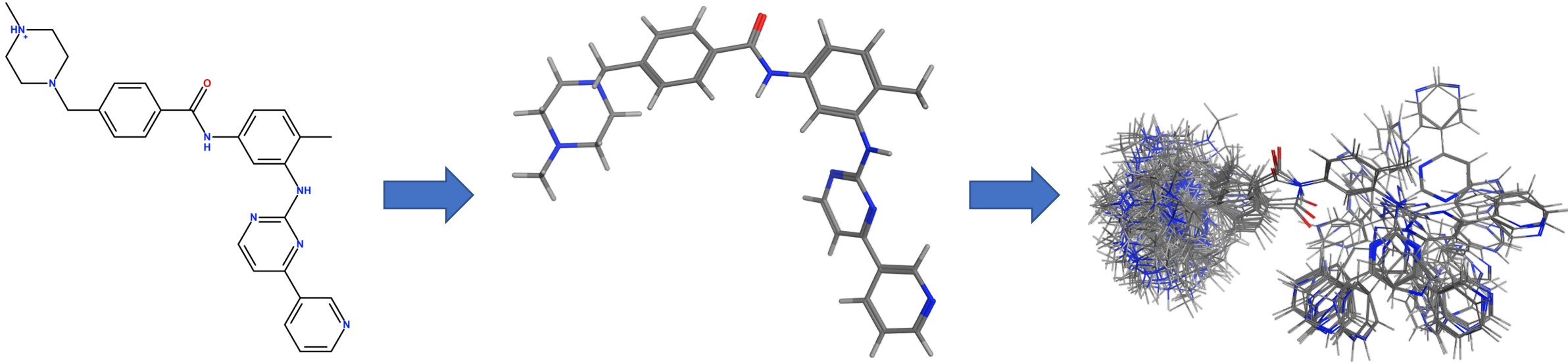


DMSO(-)	Similarity	DMSO(+)
	0.99	
	0.99	
	0.98	
	0.83	
	0.81	

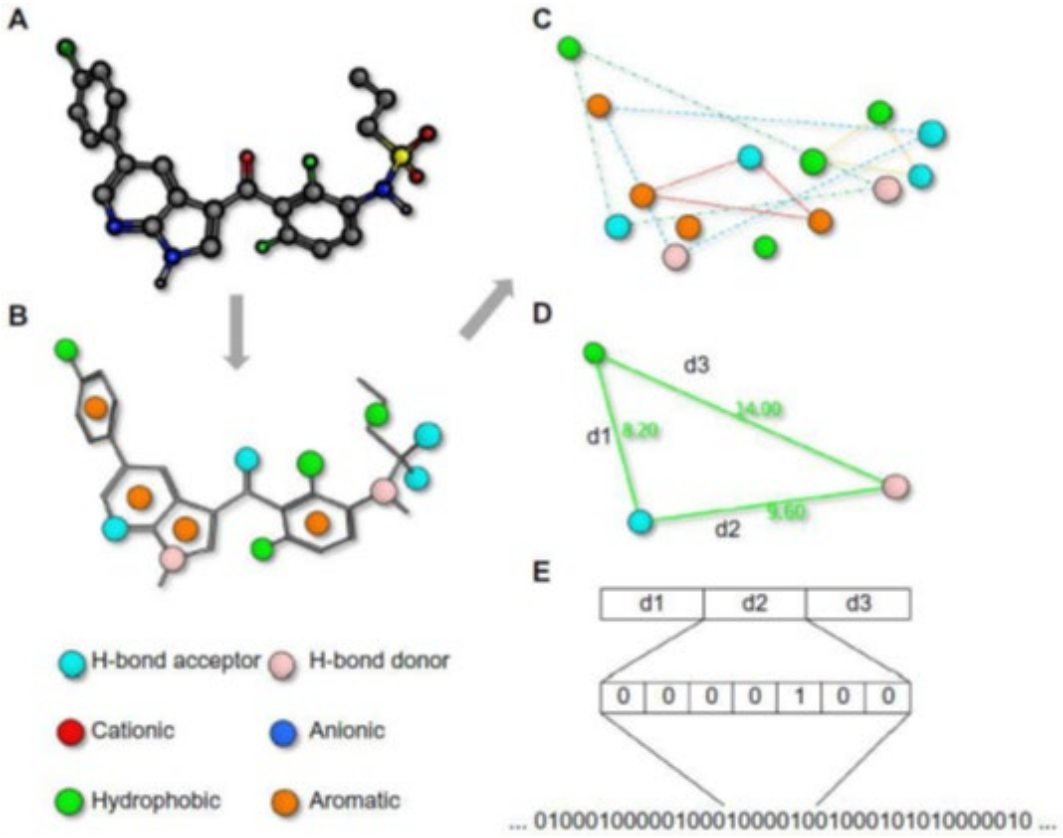
3D approaches. Conformational search

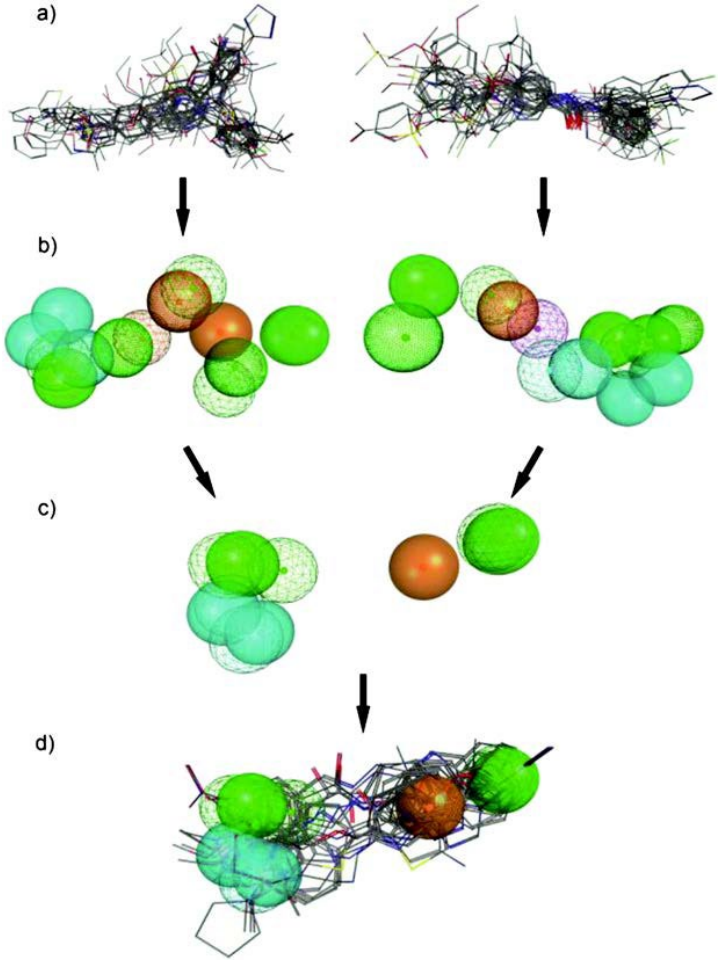
Methods

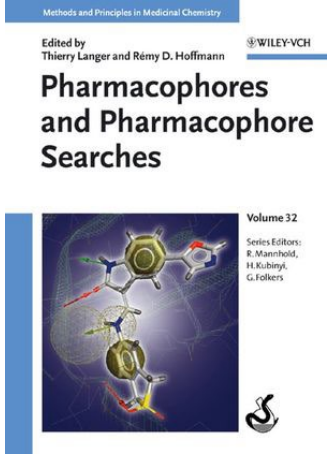
- Fragment-based (Corina)
- MM/MD-based
- QM-based



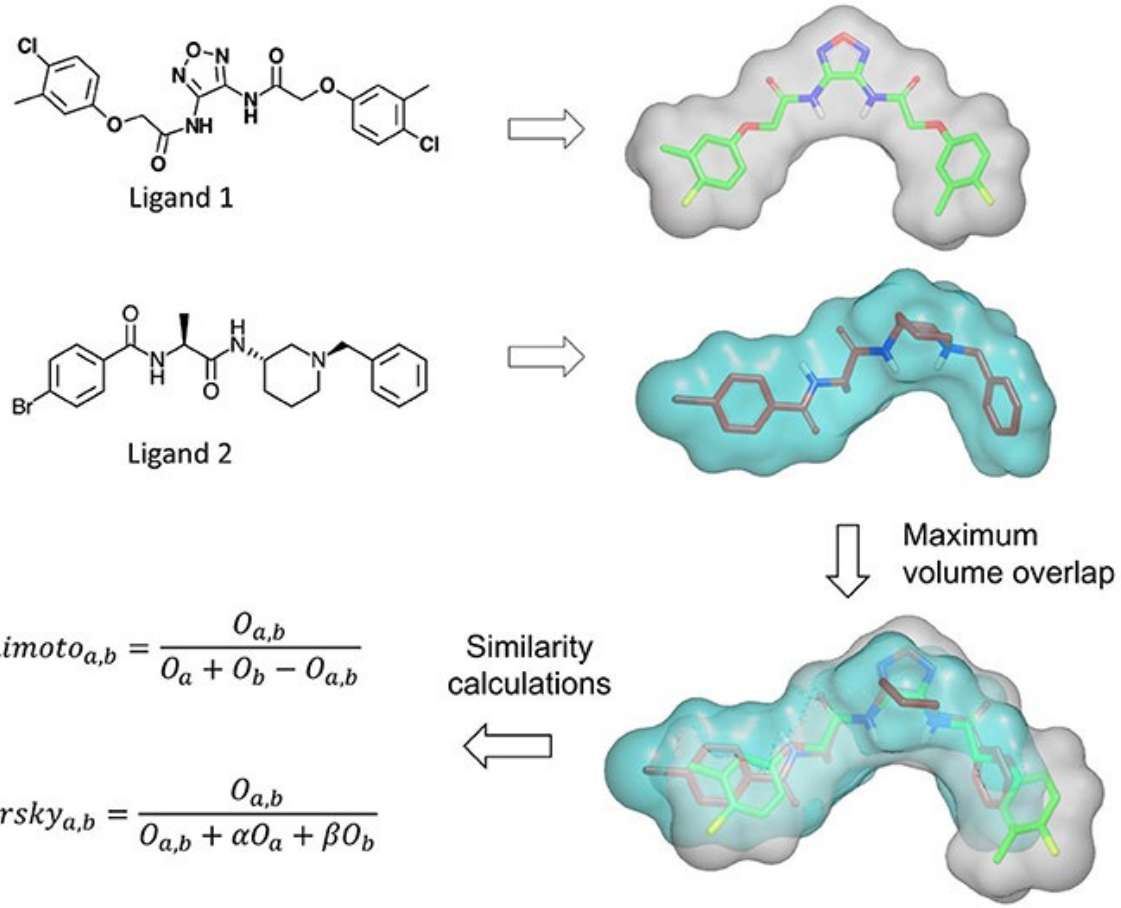
Pharmacophore modeling





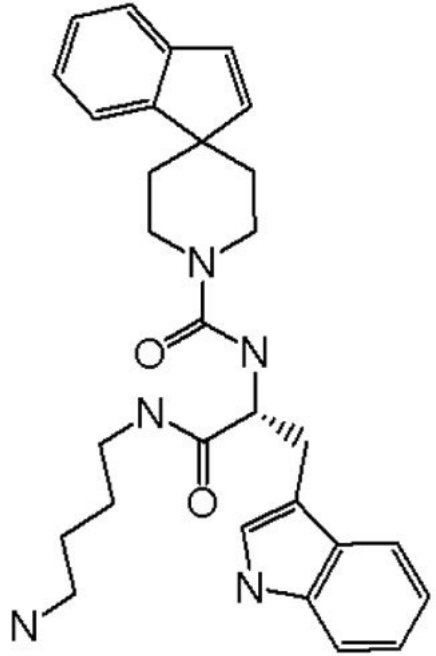


Shape similarity

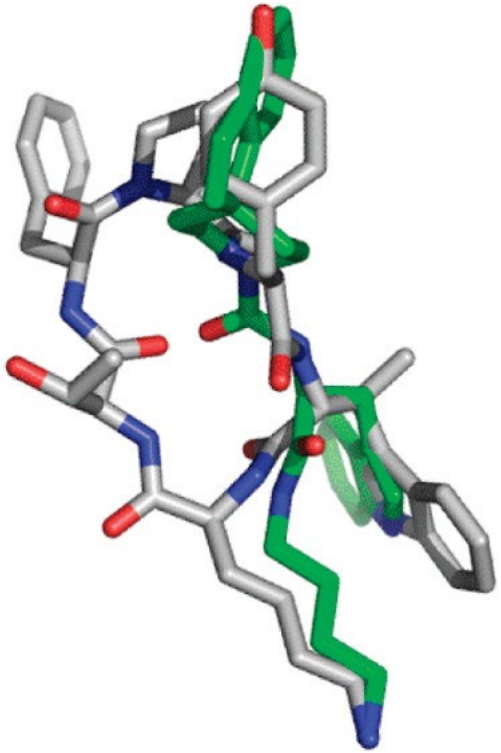


$$Tanimoto_{a,b} = \frac{O_{a,b}}{O_a + O_b - O_{a,b}}$$

$$Tversky_{a,b} = \frac{O_{a,b}}{O_{a,b} + \alpha O_a + \beta O_b}$$



K_i (SST_2) = 300 nM
Top-41 out of 1 M
(Merck database)



10.1021/jm900818s

ML and AI techniques

Input data

Bioassays

Databases

Preprocessing

Data normalization & curation

Feature extraction

Feature engineering

$$x_i' = \frac{x_i - \bar{x}}{\sum_j z_j}$$

Feature selection

Feature combination

Model training

Classification

Regression

Clustering

Model validation

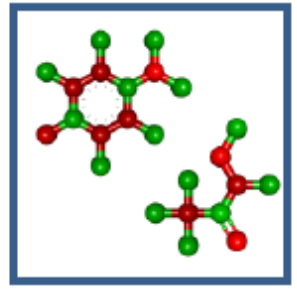
Cross-validation

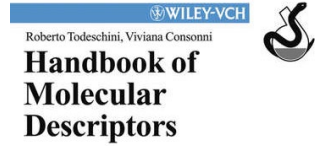
Bootstrap

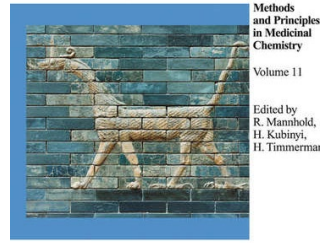
Test set

Applicability Domain

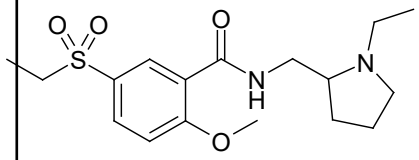
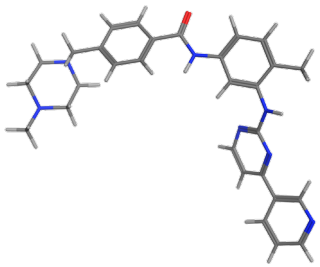
Interpretation







Descriptors as molecular features

Category	Input	Examples
1D	$C_{17}H_{26}N_2O_4S$	MW, N of atoms (by types), etc.
2D	2D structure 	Topological indices, logP, logS, structural fragments, topological pharmacophores, etc.
3D	3D structure 	VdW surface/volume, polar surface area (PSA), moment of inertia, CATS, MoRSE, etc.

Descriptors calculation

RDKit

Alvadesc

MOE

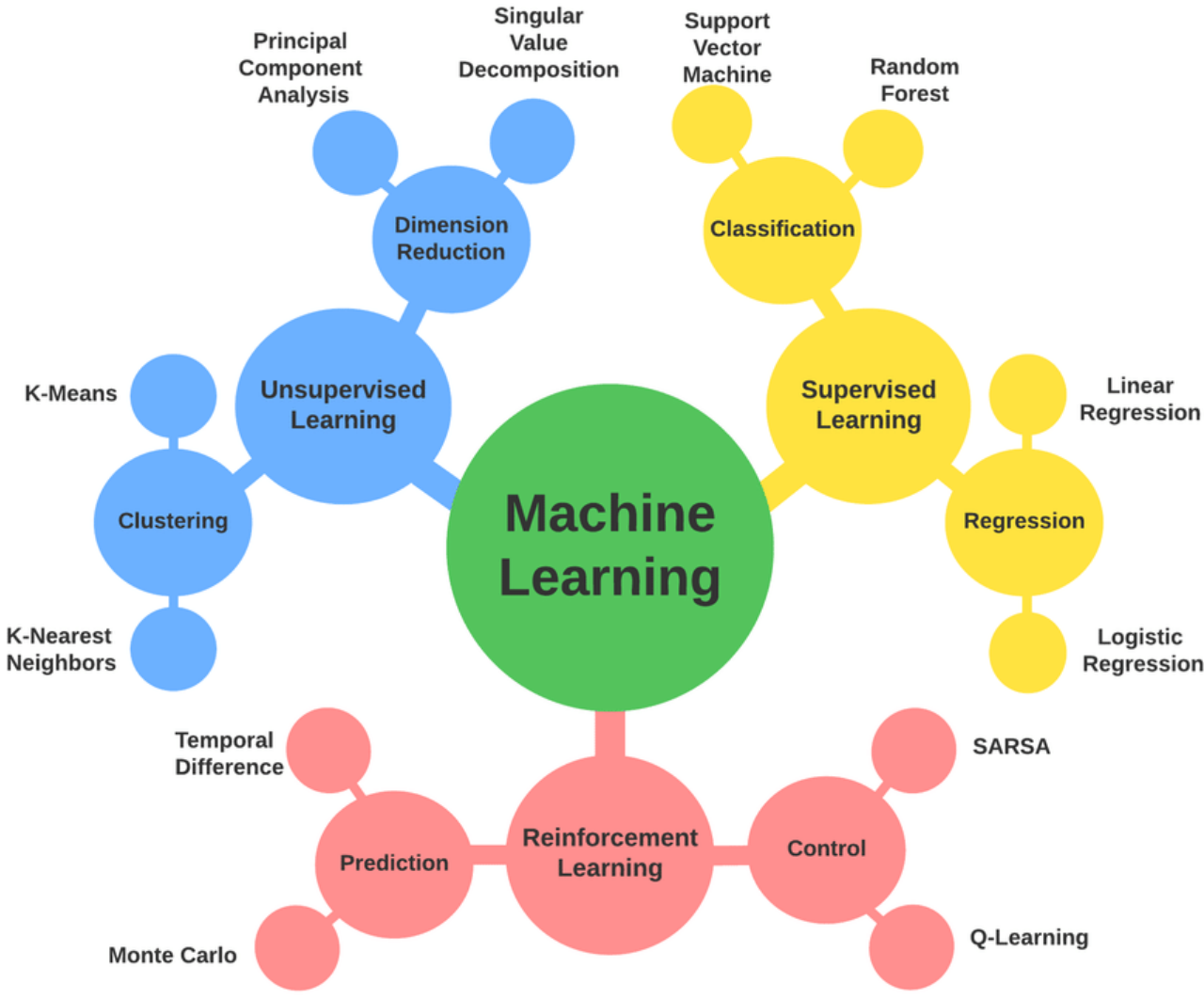
Dragon

SmartMining

AI impacting medicinal chemistry and drug design

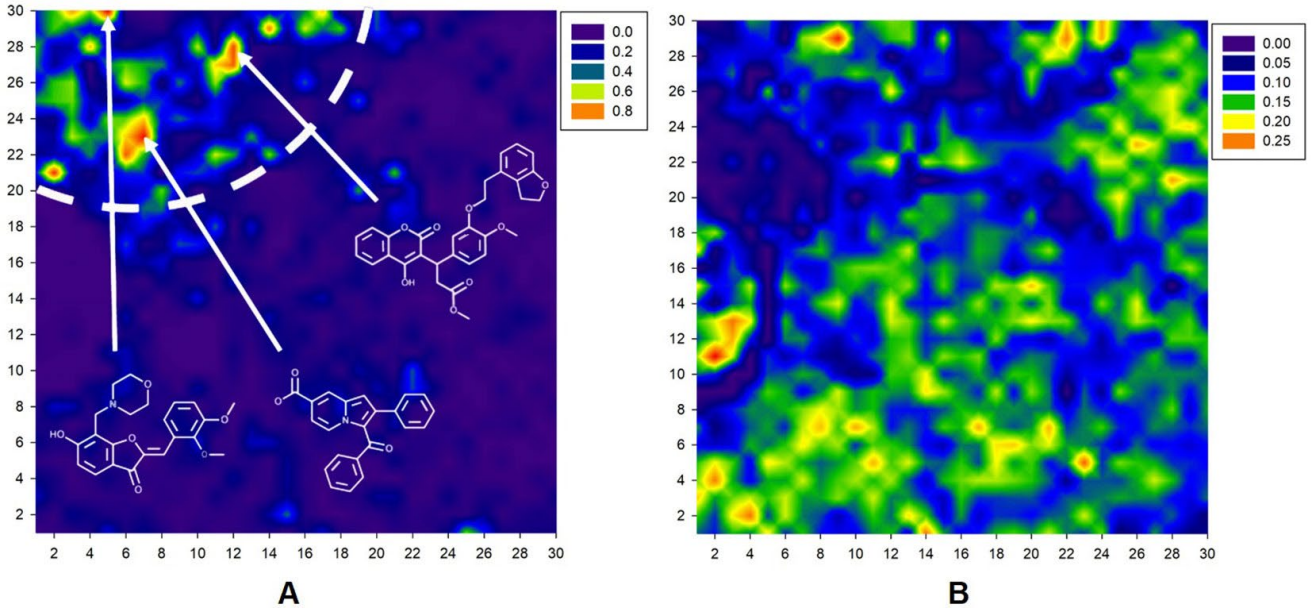
- ML/DL ADMET and other models
- Synthetic accessibility prediction
- *De novo* design:
  - Hit ID and hit-to-lead optimization (H2L)
  - Lead optimization (LeadOpt)

and many other implementations…

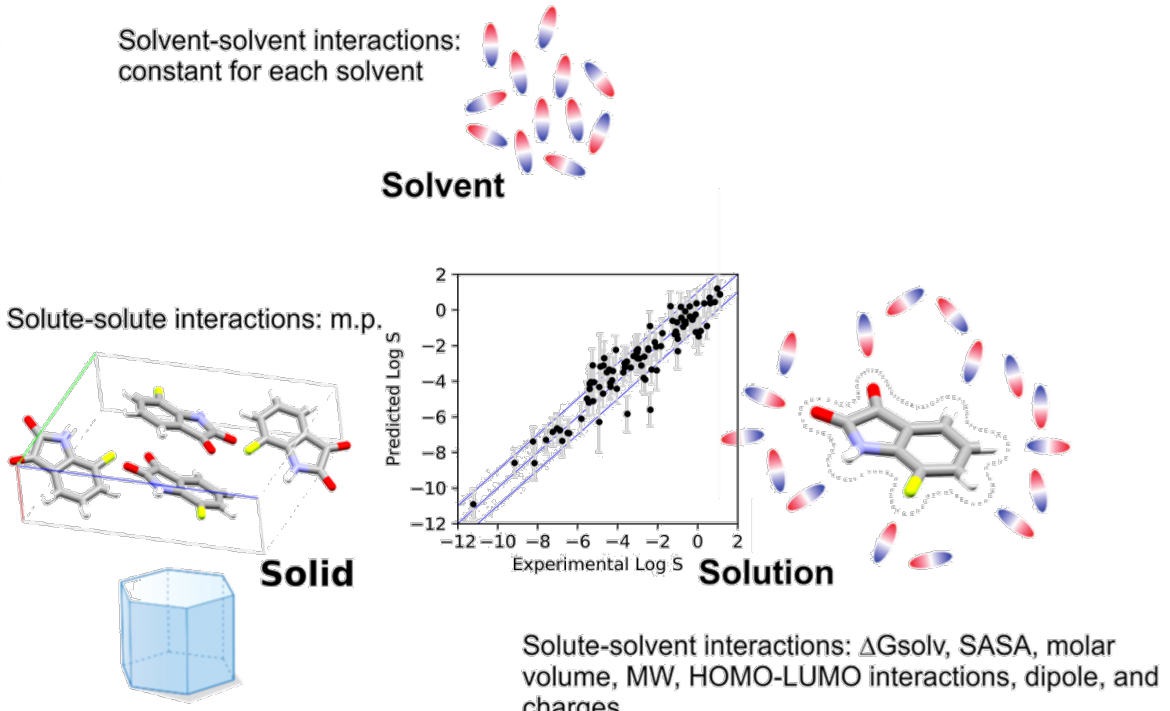


QSAR and QSPR

Antibacterial activity prediction

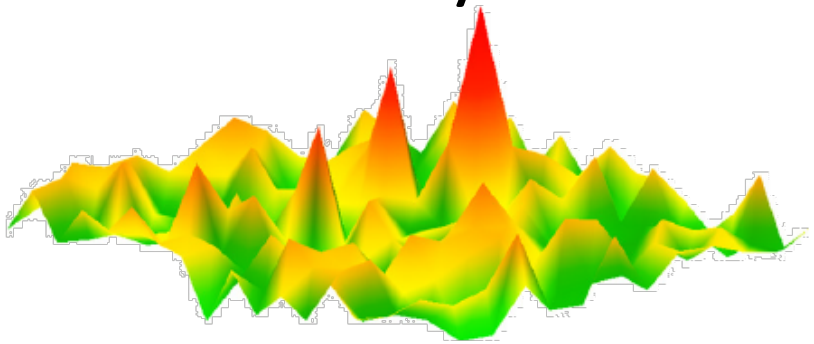


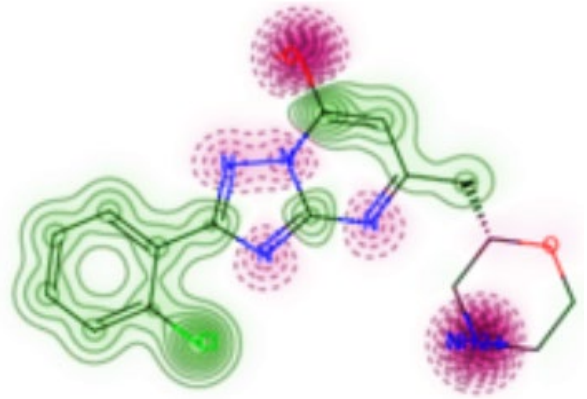
Solubility prediction



10.3389/fphar.2019.00913

Activity Cliffs





- **Molecule and fragment scoring**
- **Rapid prognosis**
- **Fully automatic platform**

✓ *In silico* PoC study (ABL1 ligands)

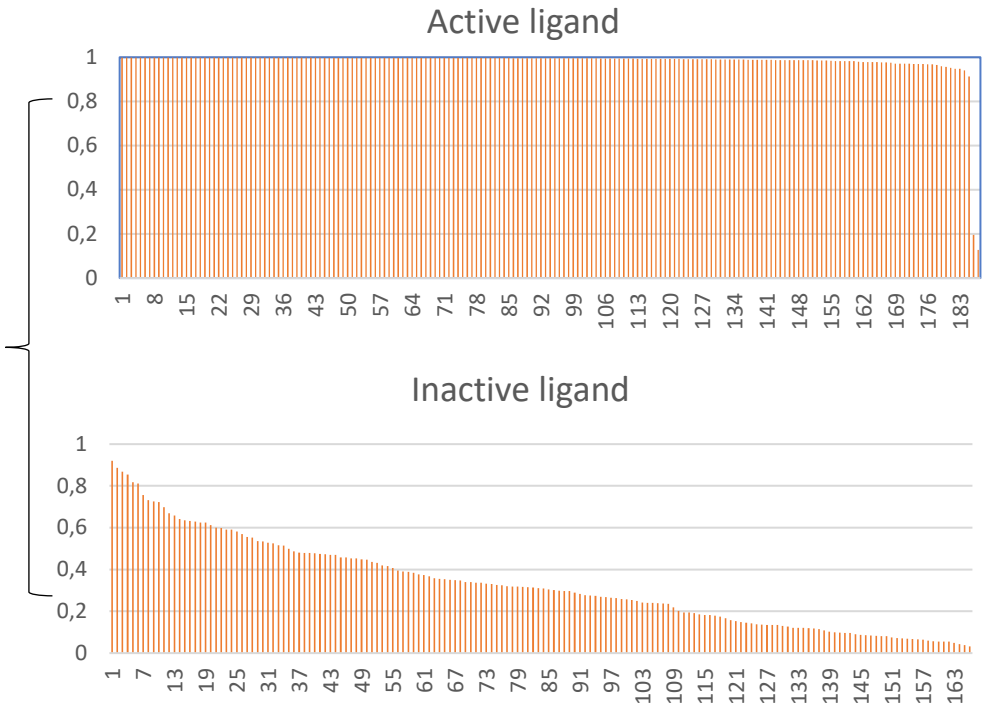
F1-metrics for both classes

Class	F1-score
Active	0.93
Inactive	0.82

Retrospective validation on active and inactive molecules

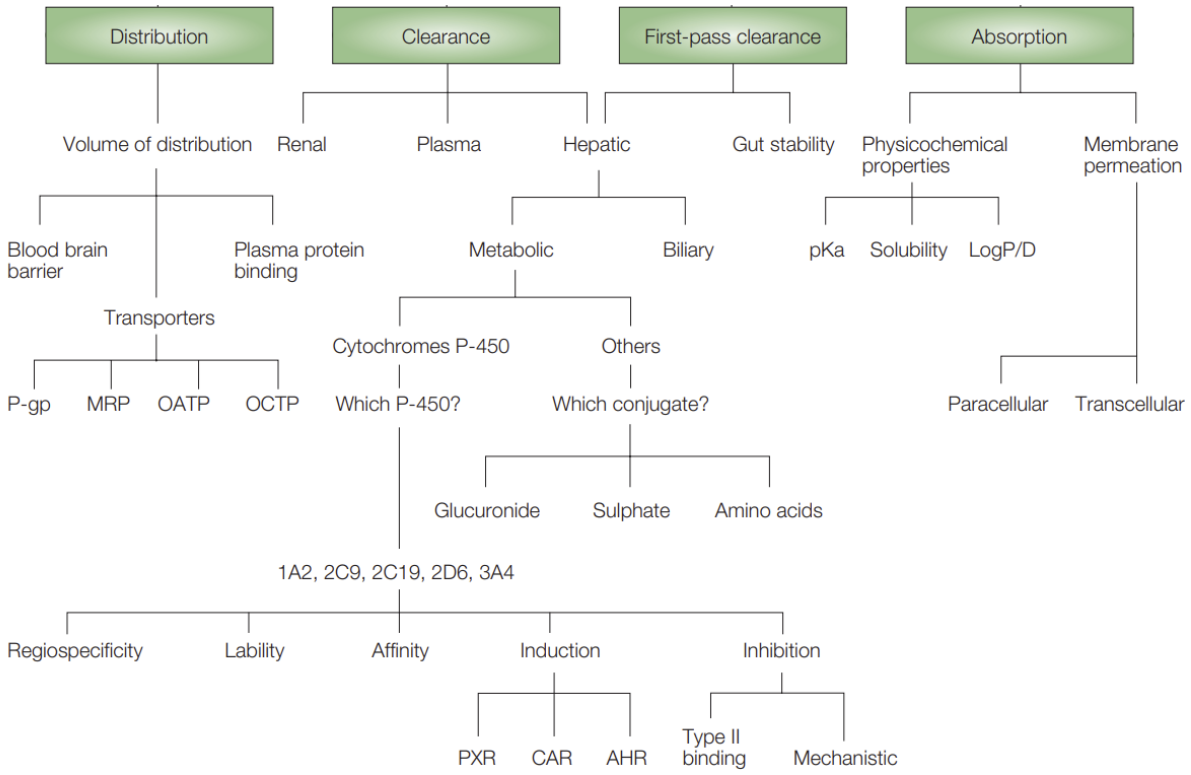
Class	Probability of being active
Active ligand	0.98
Inactive ligand	0.33

Fragment scores

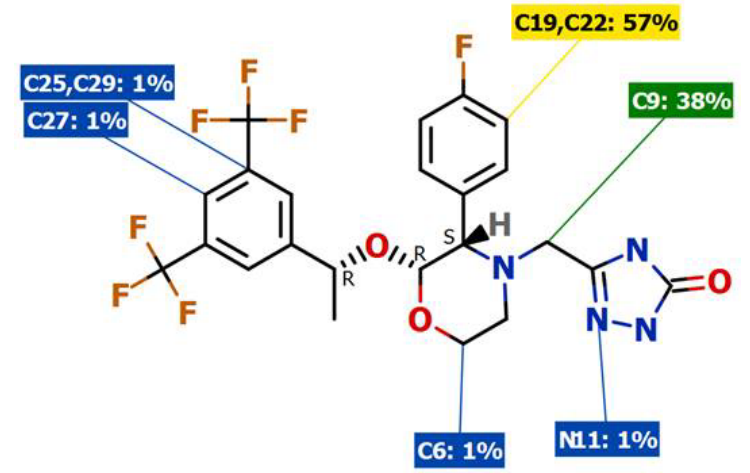


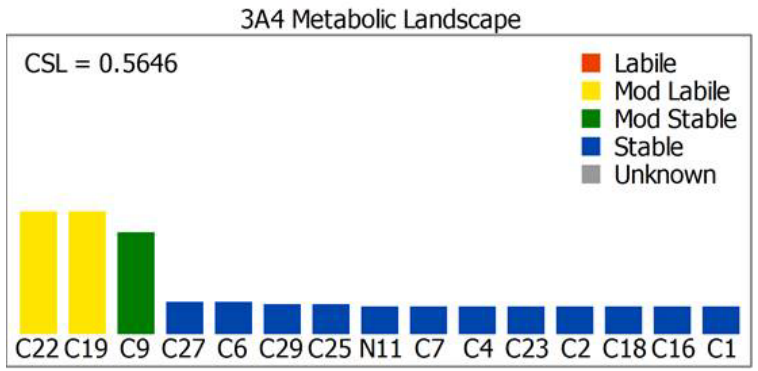
ADMET properties prediction

Classification of ADME properties

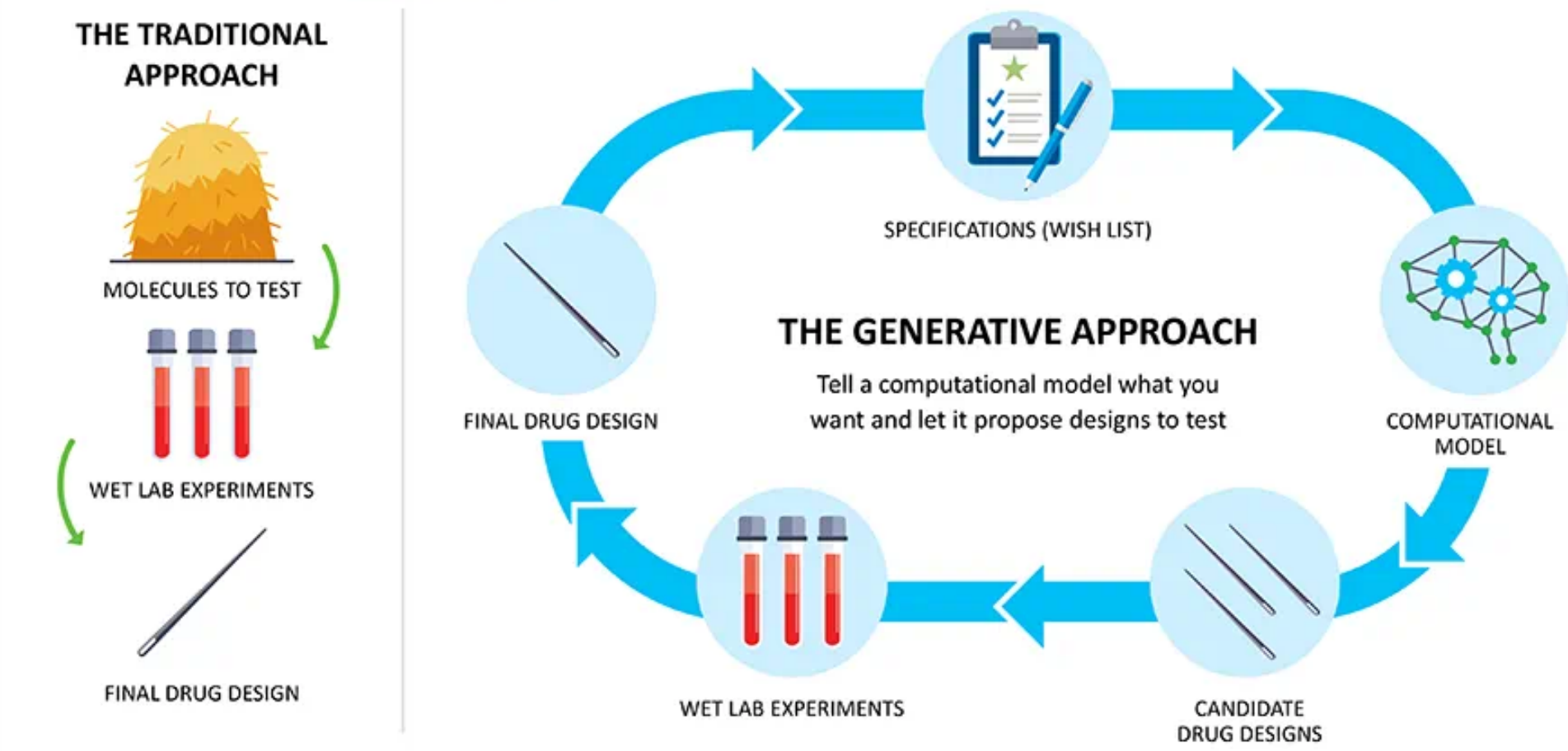


CYP3A4 metabolism prediction



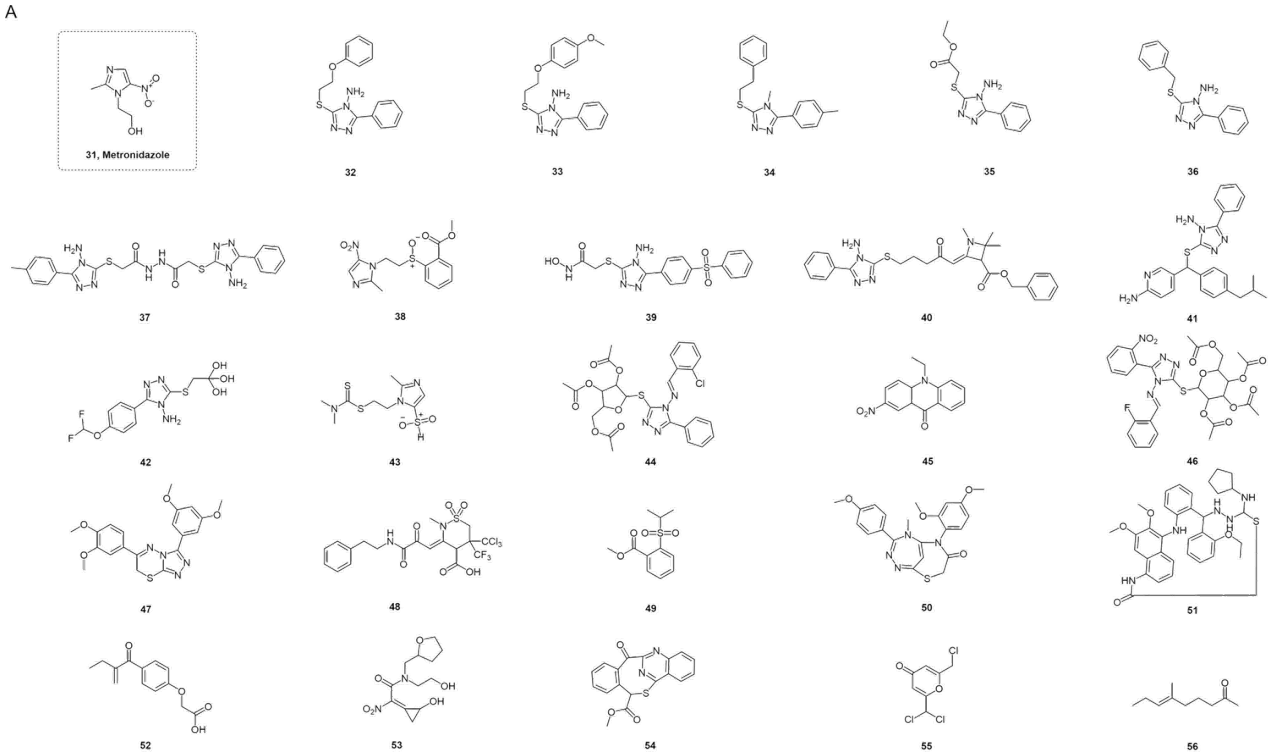


Generation of novel structures



The Hitchhiker's Guide to Deep Learning Driven Generative Chemistry

Output of RNN-based model



Cliff's Notes

IP position and Novelty

Filter out structural alerts

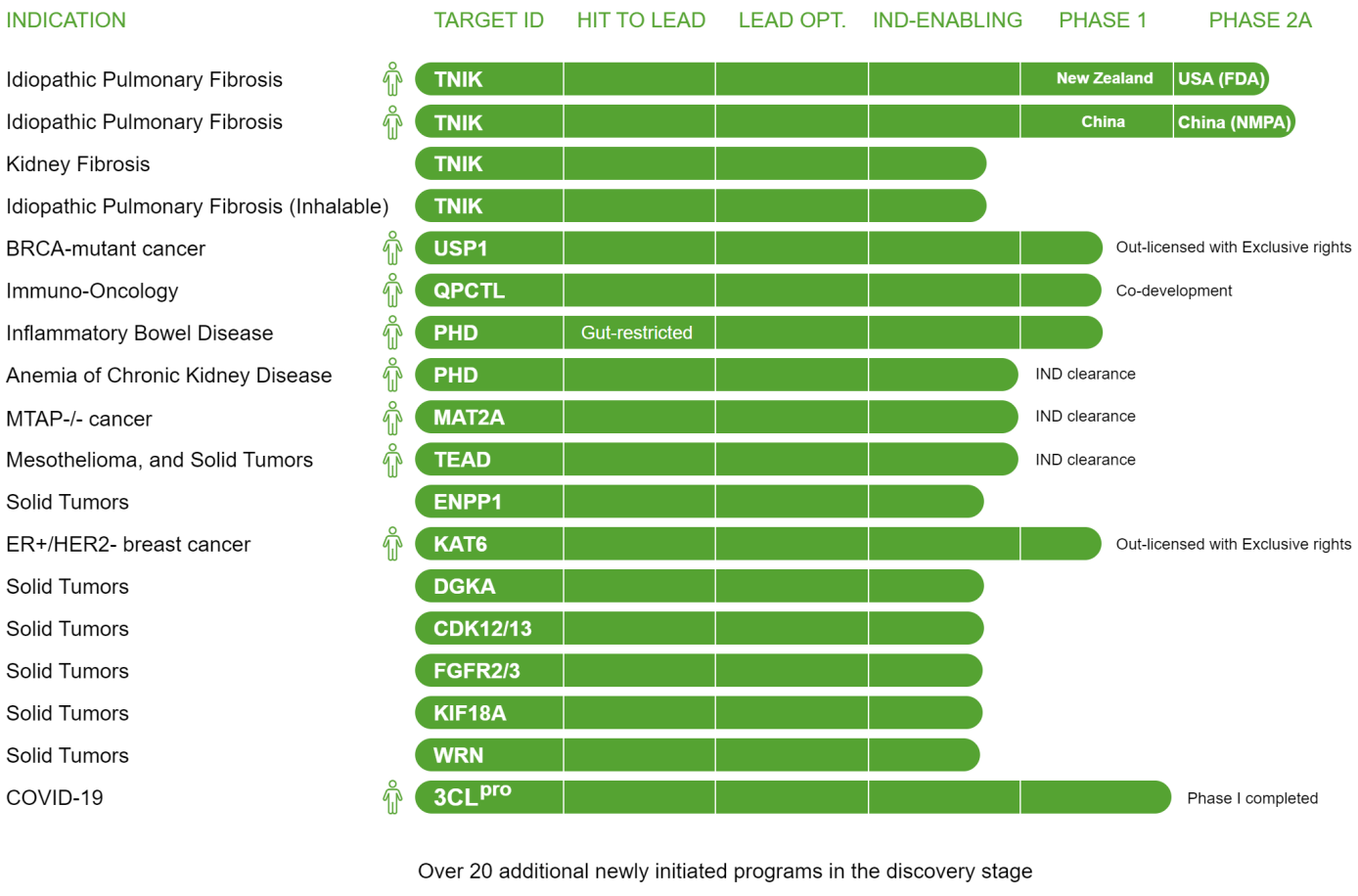
Generate molecules targeting other than kinases (e.g. PPI or GPCRs)

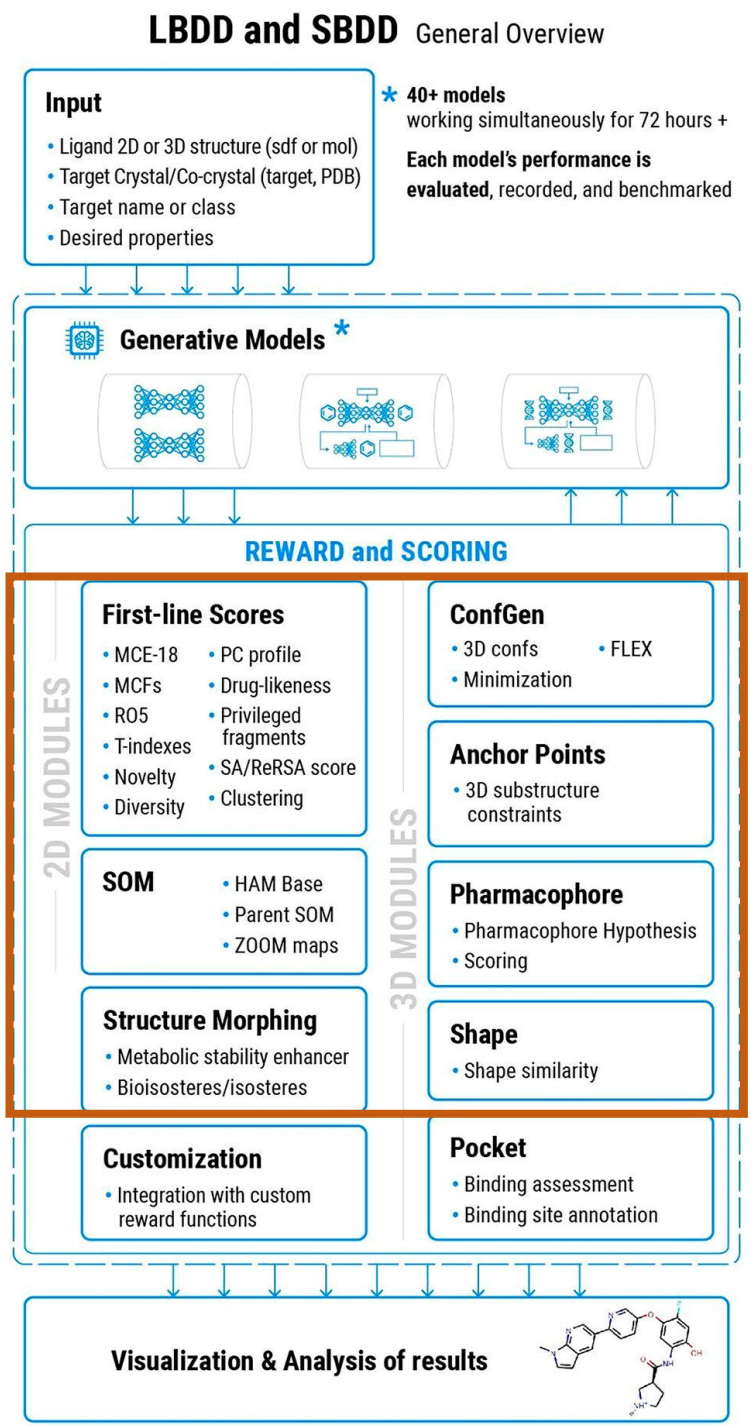
Target-specific profiling before testing

Synthetic accessibility

10.1021/acsmedchemlett.3c00041

Chemistry42 platform



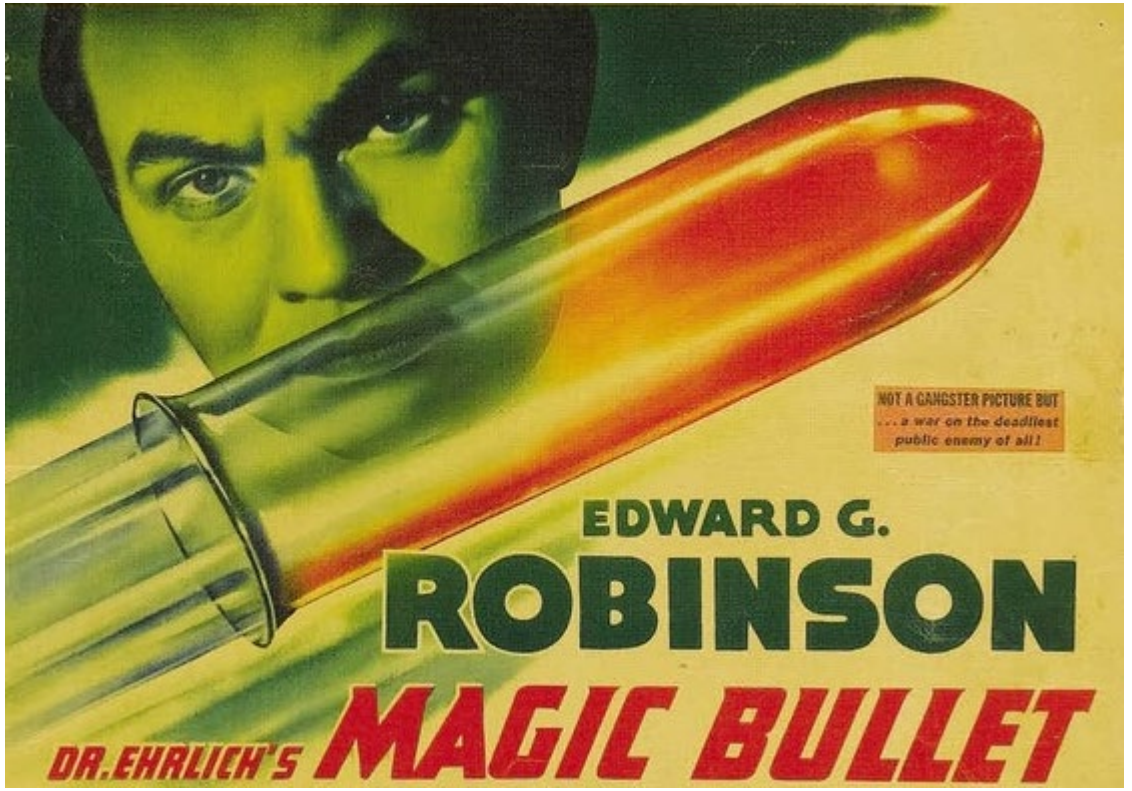


10.1021/acs.jcim.2c01191

Multimodality of drugs

PAST

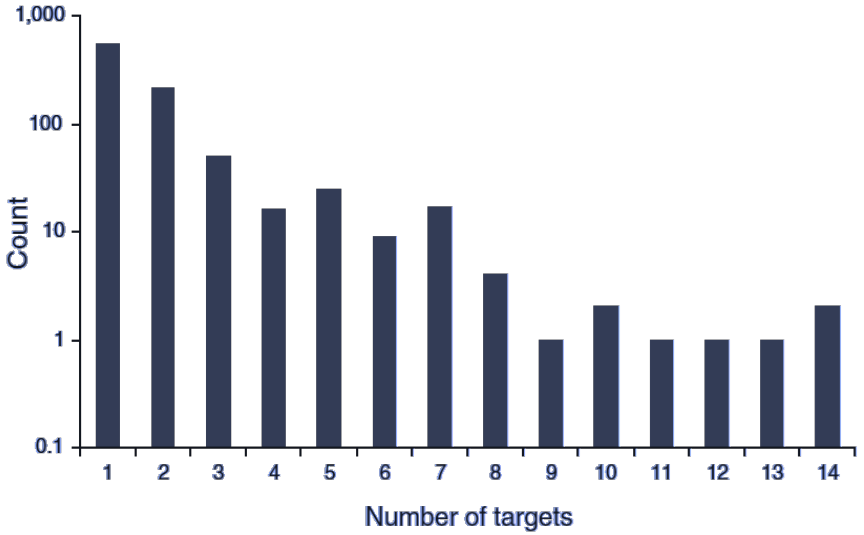
MONOpharmacology – drug is a «magic bullet»



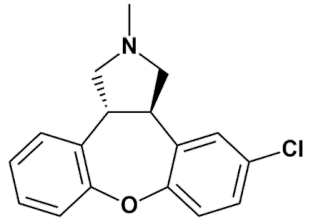
Multimodality of drugs

PAST — MONOpharmacology – drug is a «magic bullet»

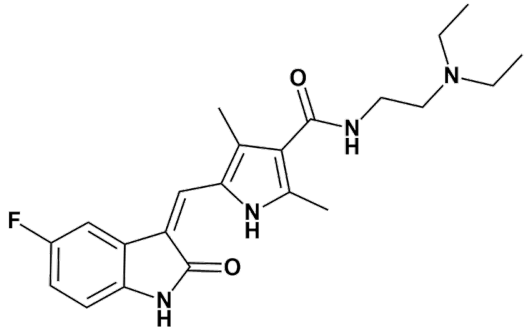
NOW — **POLYpharmacology – drug is a «magic shrapnel»**



Distribution of drugs & number of their targets



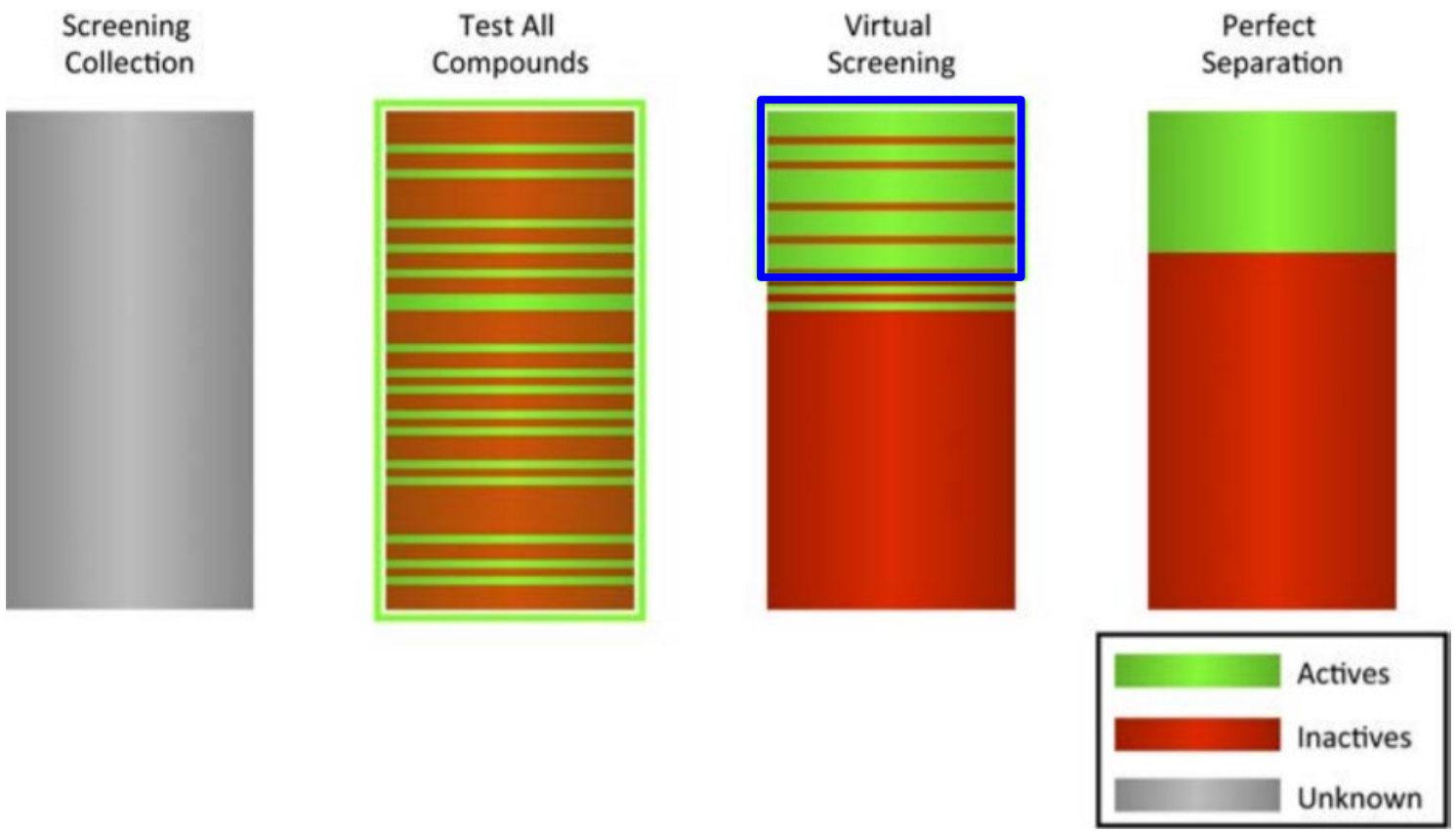
Asenapine
Schering-Plough (2009)
Schizophrenia
Low nM affinity for at least 18 GPCRs



Sunitinib
Pfizer (2006)
Cancer
Inhibition of 79 kinases (K_d < 10 µM)

Nat Biotechnol. 2007 25:1119–1126

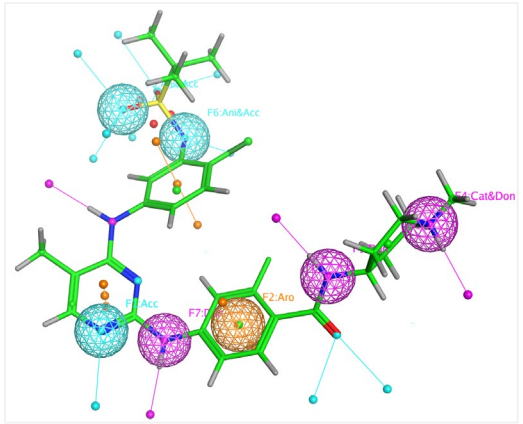
Strategies for ligand-target profiling

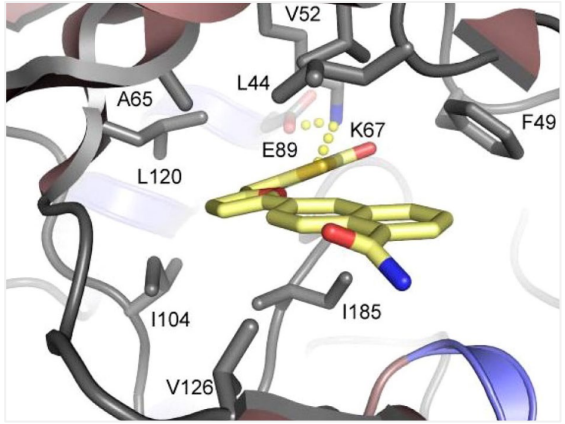


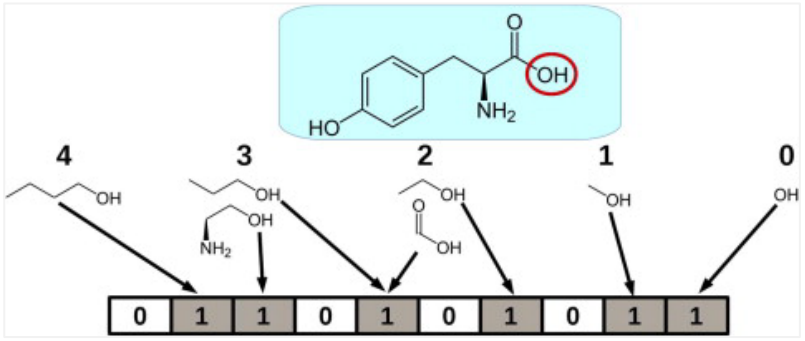
In Silico Medicinal Chemistry: Computational Methods to Support Drug Design, 2016, N. Brown

Strategies for ligand-target profiling

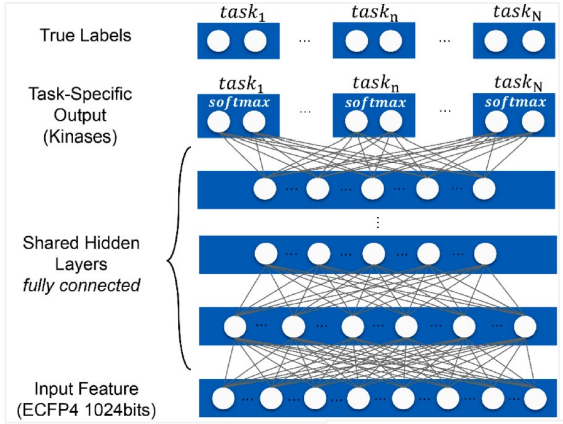
Classical approaches

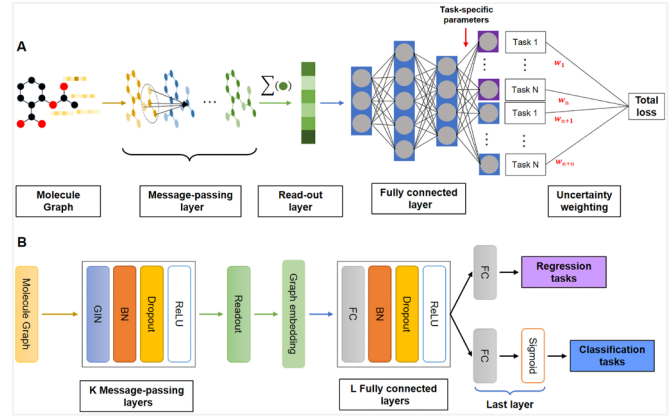


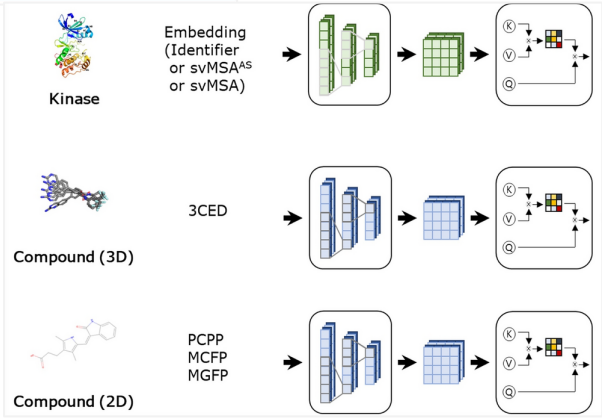




AI-based approaches







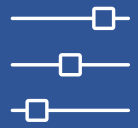
Insilico Medicine

GoldenCubes

Use cases

Predict Kinase Selectivity

Identify potential on- and off-targets for compounds during hit and lead optimization stages



Design Multimodal Compounds

Manage the kinome promiscuity of the molecules depending on your needs

Drug Repurposing

Discover novel targets for existing compounds and reduce costs for their development timeline

Core features

Big Data-Driven Approach

Large-scale and precisely annotated molecular datasets

>500K 2D structures

>2K 3D ligands

Descriptor-Based Engine

Evaluate molecules beyond the explored kinase chemical space

Interpretable descriptors

Diverse Kinase Target Set

Annotate structures across all existing kinase families

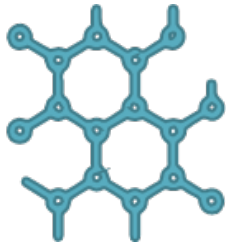
100 kinases

Golden Cubes scoring workflow

Input data

SMILES

2D/3D SDF

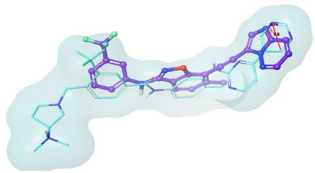


Preprocessing

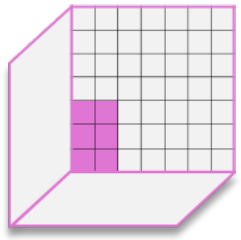
Scaffold selection

ConfGen

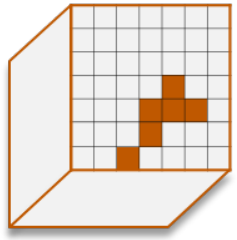
Descriptors calculation



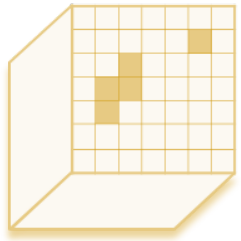
ANN-based HTVS engine



General 3D SOM

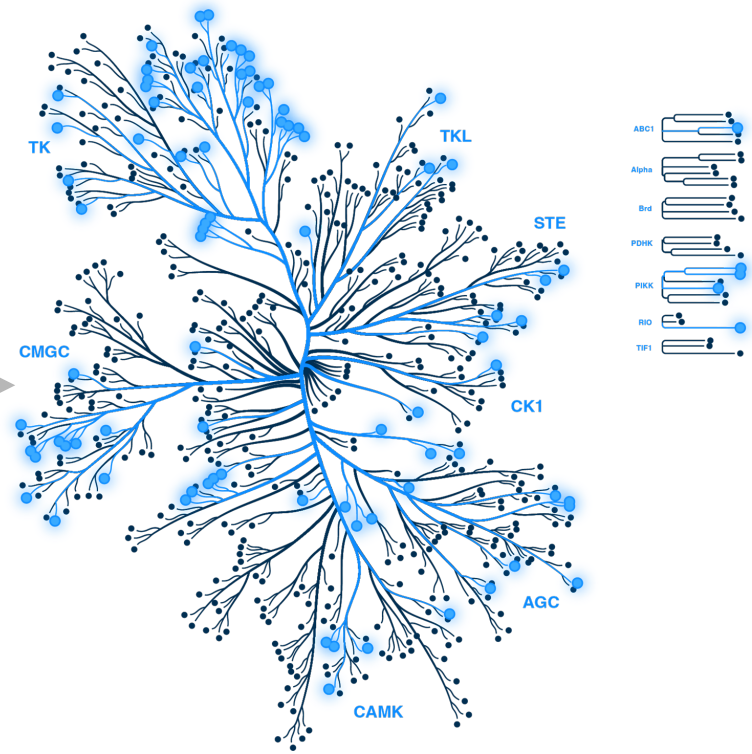


Group 3D SOMs



Atomic 3D SOMs

Scoring function



Retrospective validation

Test set

103 molecules with full-kinome profiles

LINCS KINOME scan

Full-kinome metrics

Metrics	Golden Cubes	Benchmark (QSAR)
BA	0.72	0.71
Precision	0.52	0.43
Specificity	0.9	0.85
ROC-AUC	0.72	0.75

Heatmap of predictions

Prospective validation

Test set

5 clinical candidates – kinase inhibitors with unreported profiles
TanSim **< 0.7**
Available for purchase

KINOME scan

Full-kinome metrics

Metrics	Golden Cubes
BA	0.57
Precision	0.33
Specificity	0.98
ROC-AUC	0.57

Heatmap of predictions

PATENT BUSTERS

Inside the patents tangle

US 20220119419A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2022/0119419 A1**

Zavoronkovs et al. (43) **Pub. Date: Apr. 21, 2022**

(54) **BYCYCLIC JAK INHIBITORS AND USES THEREOF**

(71) Applicant: **Insilico Medicine IP Limited**, Hong Kong (HK)

(72) Inventors: **Aleksandrs Zavoronkovs**, Pak Shek Kok (HK); **Yan Ivanenkov**, Moscow (RU); **Aleksandr Aliper**, Moscow (RU); **Anton S. Vantskul**, Moscow (RU)

(21) Appl. No.: **17/478,152**

(22) Filed: **Sep. 17, 2021**

Related U.S. Application Data

(63) Continuation of application No. PCT/US2020/025206, filed on Mar. 27, 2020.

(60) Provisional application No. 62/824,485, filed on Mar. 27, 2019.

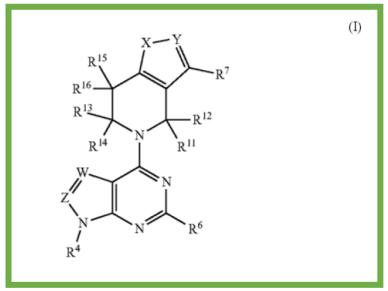
Publication Classification

(51) **Int. Cl.**
C07D 519/00 (2006.01)

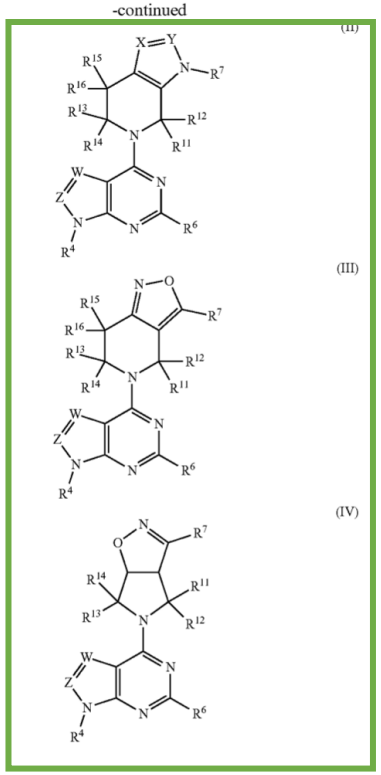
(52) **U.S. Cl.**
CPC ***C07D 519/00*** (2013.01)

(57) **ABSTRACT**

Provided herein are compounds of Formulas (I), (II), (III), and (IV)



-continued



and subformulas thereof, wherein the variables are defined herein. Also provided herein are pharmaceutical compositions comprising a compound of Formula (I), (II), (III), or (IV) and methods of using the compounds, e.g., in the treatment of immune disorders, inflammatory disorders, and cancer.

For formula I, formula II, formula III, and formula (IV):

W is N or CR^5;

Y is N or CR^2;

Z is N or CR^3;

wherein W and Z are not both N;

R^1 is selected from the group consisting of cyano, hydroxyl, NR^aR^b, C_{1-6}alkoxy, and -A-L^1-R^9;

R^2, R^3, R^4, and R^6 are each independently selected from the group consisting of hydrogen, deuterium, halogen, cyano, hydroxyl, —NR^aR^b, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, and C_{1-6}alkoxy;

R^5 is selected from the group consisting of hydrogen, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, aryl, heteroaryl, cycloalkyl, heterocyclyl, -aryl-C_{1-6}alkyl, -heteroaryl-C_{1-6}alkyl, -heterocyclyl-C_{1-6}alkyl, halogen, cyano, hydroxyl, C_{1-6}alkoxy, C_{1-6}haloalkoxy, amino, carboxy, aminocarbonyl, —C_{1-6}alkyl-aminocarbonylamino, C_{1-6}alkyl-aminocarbonyl, —S(O)—R^8, —$S(O)_2$—R^8, —NR^8—$S(O)_2$—R^8, —$S(O)_2$—NR^aR^b, —NR^8—$S(O)_2$—NR^aR^b, —C_{1-6}alkyl-aryl, —C_{1-6}alkyl-heteroaryl, —C_{1-6}alkyl-heterocycle, and —C_{1-6}alkyl-cycloalkyl, wherein said alkyl, aryl, and heteroaryl is optionally substituted with one or substituents independently selected from the group consisting of halo, hydroxyl, methoxy, amino, cyano, alkylamino, dialkylamino, CF_3, aminocarbonyl, —C_{1-6}alkyl-aminocarbonylamino, and C_{3-6}cycloalkyl;

R^7 is B-L^2-R^{10}, or R^7 is aryl or heteroaryl, wherein the aryl or heteroaryl is optionally substituted with one to four R^{17};

each R^8 is independently selected from the group consisting of hydrogen, C_{1-6}alkyl, C_{1-6} haloalkyl, hydroxy, C_{1-6}alkoxy, and —O—C_{1-6}haloalkyl;

R^9 is selected from the group consisting of hydrogen, cycloalkyl, heterocycloalkyl, aryl, and heteroaryl, wherein any non-hydrogen R^9 is optionally substituted with one to four R^{17};

R^{10} is selected from the group consisting of hydrogen, substituted by 1-3 substituents independently selected from the group consisting of halogen, C_{1-6}alkyl, and C_{1-6}haloalkyl;

R^{17} is, independently for each occurrence, selected from the group consisting of halogen, cyano, hydroxyl, —NR^aR^b, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, C_{1-6}haloalkyl, C_{1-6}alkoxy, CF_3, —SH, —S—C_{1-6}alkyl, —COOH, —CO_2—C_{1-6}alkyl, —C_{1-6}alkyl-CN, —C(O)NR^aR^b, —C(O)—C_{1-6} alkyl-NR^aR^b, —C(O)—NR^a—$S(O)_2$—C_{1-6}alkyl, —$S(O)_2$—C_{1-6}alkyl, —$S(O)_2$—NR^aR^b, —$S(O)_2$—C_{1-6} alkyl-NR^aR^b;

A is selected from the group consisting of —C(O)—, —S(O)—, and —$S(O)_2$—, or A is absent;

B is selected from the group consisting of —C(O)—, —$S(O)_2$—NR^8—, —CH_2—NR—, and —C(O)NR^8—;

L^1 is selected from the group consisting of a bond, C_{1-6}alkylene, C_{1-6}heteroalkylene, C_{2-6}alkenylene, and C_{2-6}alkynylene, wherein L^1 is optionally substituted with one to four R^{17} groups;

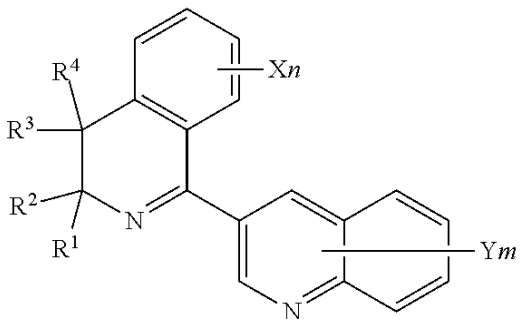
L^2 is selected from the group consisting of a bond, C_{1-6}alkylene, C_{2-6}alkenylene, and C_{2-6}alkynylene, wherein any CH_2 group of C_{1-6}alkylene can be replaced with a moeity selected from the group consisting of —O—, —NR^a—, and —$S(O)_2$—, and one CH_2 group of C_{1-6} alkylene can be replaced with a moiety selected from the group consisting of cycloalkylene, heterocycloalkylene, arylene, and heteroarylene, and wherein L^2 is optionally substituted with one to four R^{17} groups; or

when B is —$S(O)_2$—NR^8—, —CH_2—NR^8—, or —C(O)NR^8—, R^8 and L^2 can be taken together including the nitrogen atom to which they are attached to form a 3-7-membered heterocycloalkyl optionally substituted with one to four R^{17} groups; and

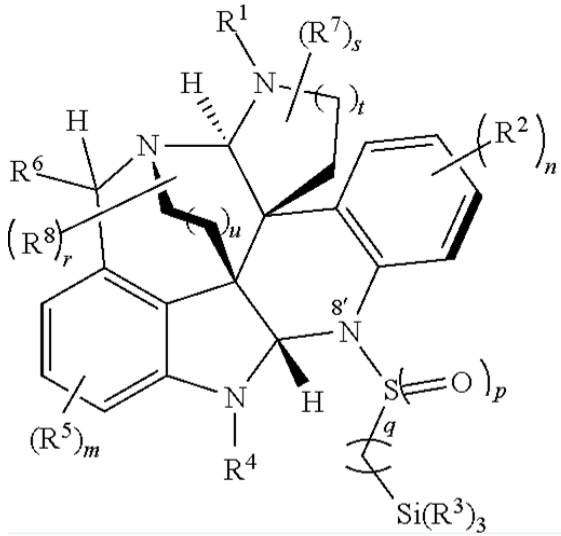
each of R^a and R^b are, independently for each occurrence, selected from the group consisting of hydrogen, C_{1-6}alkyl, and C_{1-6}haloalkyl, or R^a and R^b are taken together, including the nitrogen to which they are attached, to form a heterocycloalkyl ring.

Markush structures: the art of chemistry

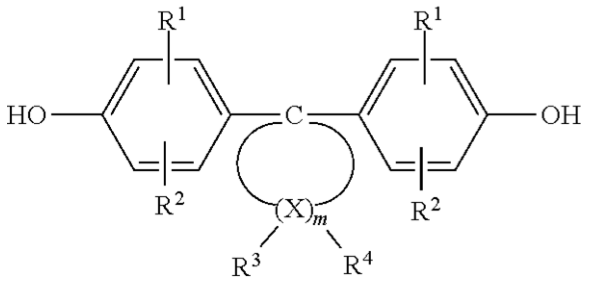
US20210000112A1



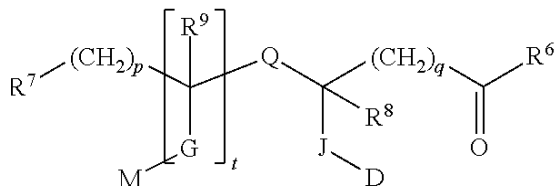
US20210000119A1



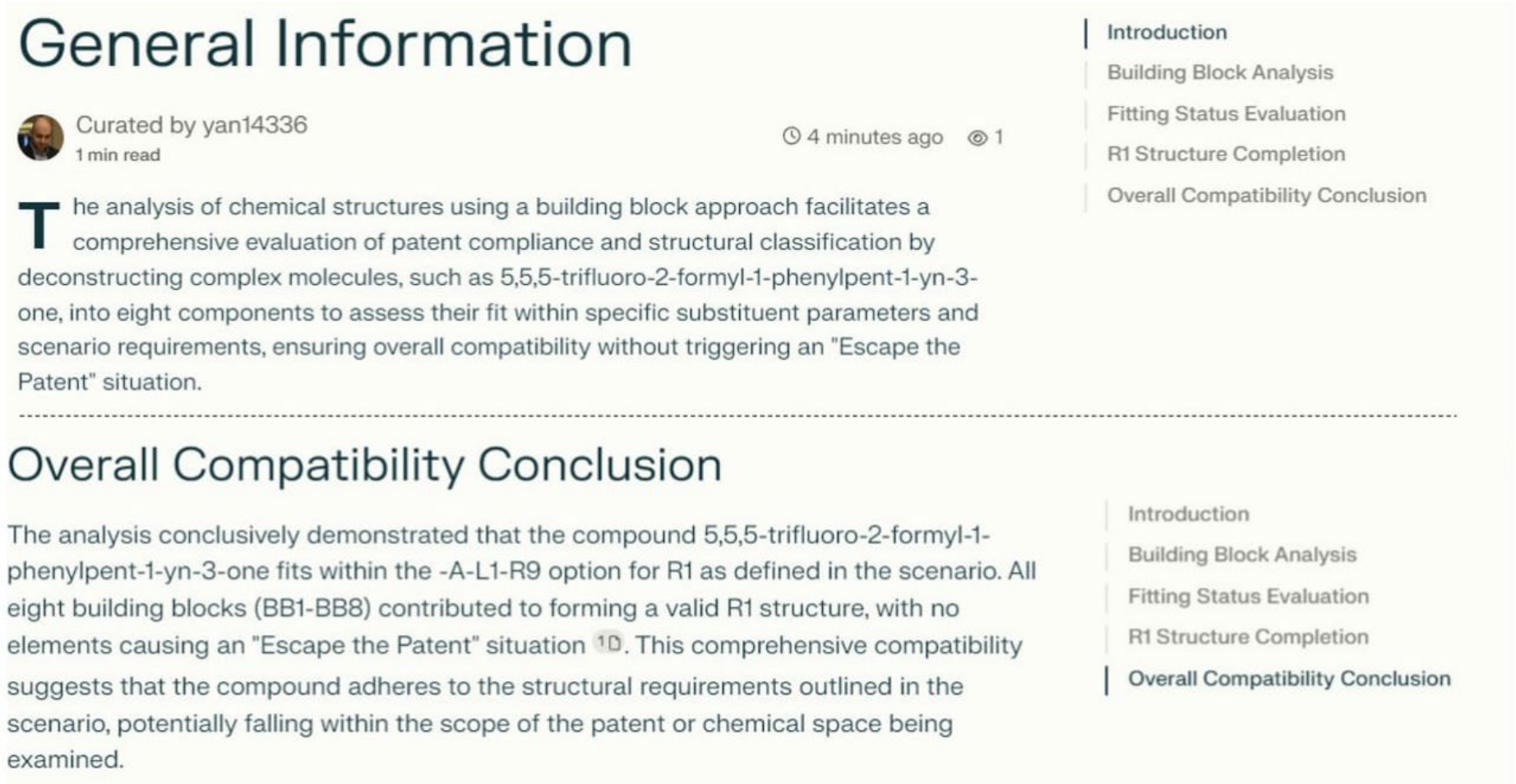
US20210001614A1



US20200360258A1



Output of the Patent Busters



General Information

Curated by yan14336
1 min read

4 minutes ago 1

The analysis of chemical structures using a building block approach facilitates a comprehensive evaluation of patent compliance and structural classification by deconstructing complex molecules, such as 5,5,5-trifluoro-2-formyl-1-phenylpent-1-yn-3-one, into eight components to assess their fit within specific substituent parameters and scenario requirements, ensuring overall compatibility without triggering an "Escape the Patent" situation.

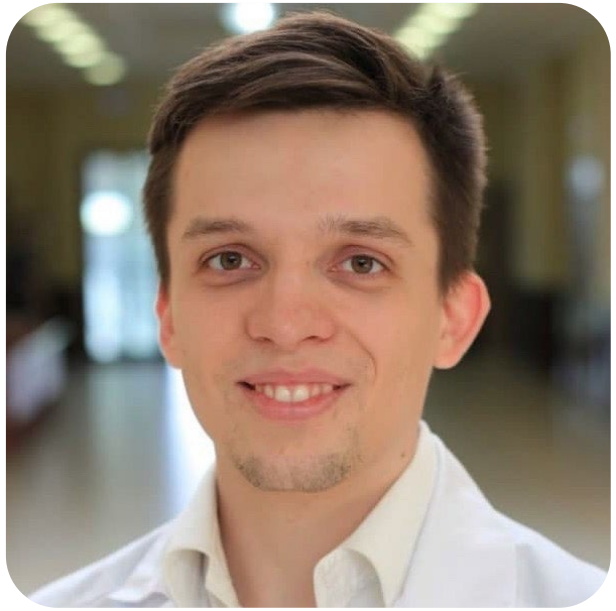
Introduction
Building Block Analysis
Fitting Status Evaluation
R1 Structure Completion
Overall Compatibility Conclusion

Overall Compatibility Conclusion

The analysis conclusively demonstrated that the compound 5,5,5-trifluoro-2-formyl-1-phenylpent-1-yn-3-one fits within the -A-L1-R9 option for R1 as defined in the scenario. All eight building blocks (BB1-BB8) contributed to forming a valid R1 structure, with no elements causing an "Escape the Patent" situation 1D. This comprehensive compatibility suggests that the compound adheres to the structural requirements outlined in the scenario, potentially falling within the scope of the patent or chemical space being examined.

Introduction
Building Block Analysis
Fitting Status Evaluation
R1 Structure Completion
Overall Compatibility Conclusion

Thank you!



Alex Malyshev

@ALEXMALYSHEV95

alexmalyshev95@gmail.com