CDD WEBINAR

Generative Chemistry, Deep Learning and Traditional Models

Moderated by Dr. Eric Gifford

LIVE

Thursday, September 26th 2024
8:00 AM (PT) | 11:00 AM (ET) | 16:00 (BST)

Eric Gifford, PhD.
Business Development Consultant
Collaborative Drug Discovery

Pat Walters, PhD.
Chief Data Officer
Relay Therapeutics

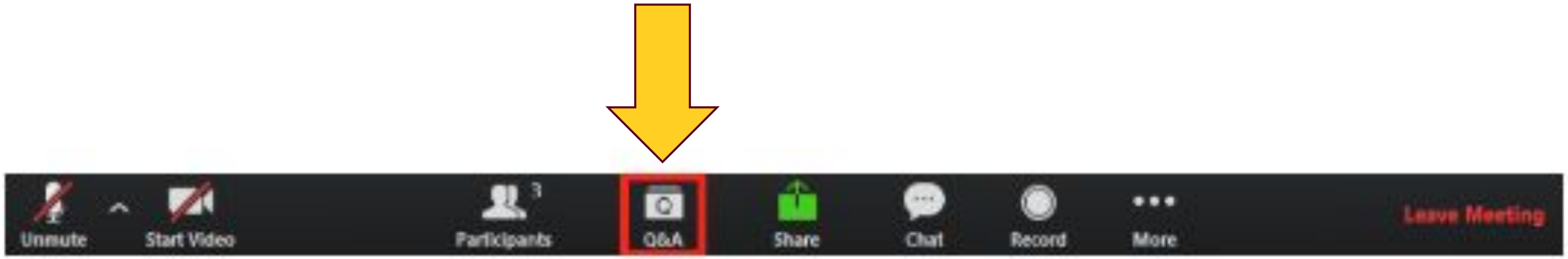
Greg Landrum, PhD.
Senior Scientist
ETH Zürich

Peter Gedeck, PhD.
Senior Informatician
Collaborative Drug Discovery

© All Rights Reserved Collaborative Drug Discovery

Do you have a question to ask our panel?

Open the ZOOM Q&A
type in your question during the webinar



We will reserve time and answer as many questions as we can at the end

© All Rights Reserved Collaborative Drug Discovery

CDD WEBINAR

Generative Chemistry, Deep Learning and Traditional Models

Moderated by Dr. Eric Gifford

Eric Gifford, PhD.
Business Development Consultant
Collaborative Drug Discovery

LIVE

Thursday, September 26th 2024
8:00 AM (PT) | 11:00 AM (ET) | 16:00 (BST)

Pat Walters, PhD.
Chief Data Officer
Relay Therapeutics

Greg Landrum, PhD.
Senior Scientist
ETH Zürich

Peter Gedeck, PhD.
Senior Informatician
Collaborative Drug Discovery

© All Rights Reserved Collaborative Drug Discovery

What does our field need to move forward?

(Some topics for today's discussion)

◆ Comparable Data ◆

◆ Open-source tool kits and commercial software ◆

◆ Robust Models & AI ◆

© All Rights Reserved Collaborative Drug Discovery

Pop quiz

Suppose you’re doing machine learning and you need some bioactivity (or ADME) data sets to work with.

You want/need to use public data, so you start with ChEMBL

You notice that you can increase the size of the data sets available to you *significantly* if you combine results taken from different assays but run on the same target

Is it safe to do this?

Unfortunately common in our field

“I really need more data, so I’m just going to combine the assays. How bad can it possibly be?”

Aside: the Cheng-Prusoff equation

Relates IC50 to Ki for competitive inhibition assays:

$$Ki = IC50/(1+[S]/Km)$$

https://pharmacologycanada.org/Cheng-Prusoff-equation

where:

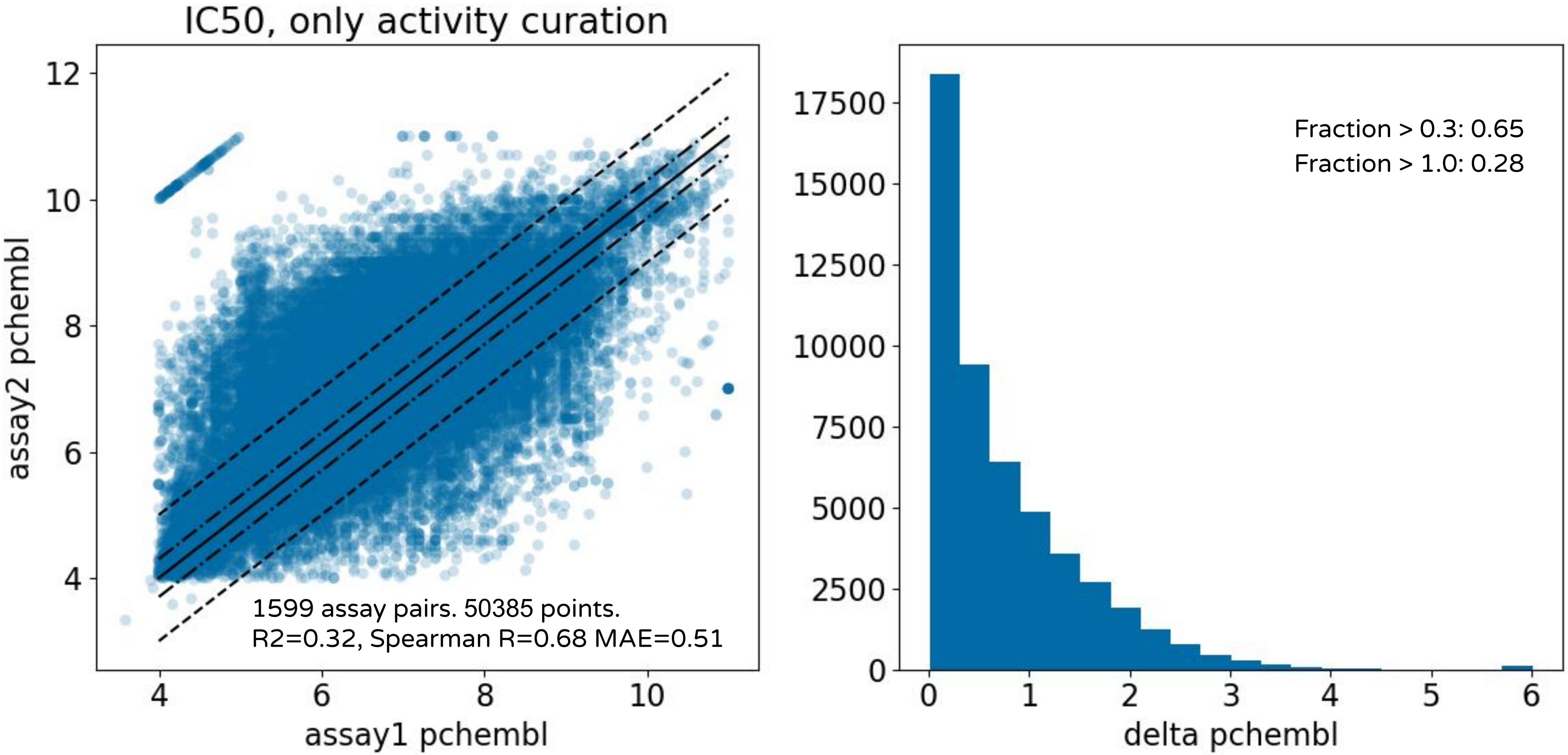
[S] = concentration of the substrate

Km = the affinity constant of the substrate

This does *not* magically let you convert IC50 (generally unsafe to compare across assays) to Ki (safe to compare across assays): you need to know both [S] and Km in order to use it, and if you know [S] and Km then you *already* know whether or not it's safe to compare

A fun take on the relationship between Ki and IC50:
https://krhornberger.substack.com/p/tweetorial-ic50-vs-ki

It’s pretty bad.

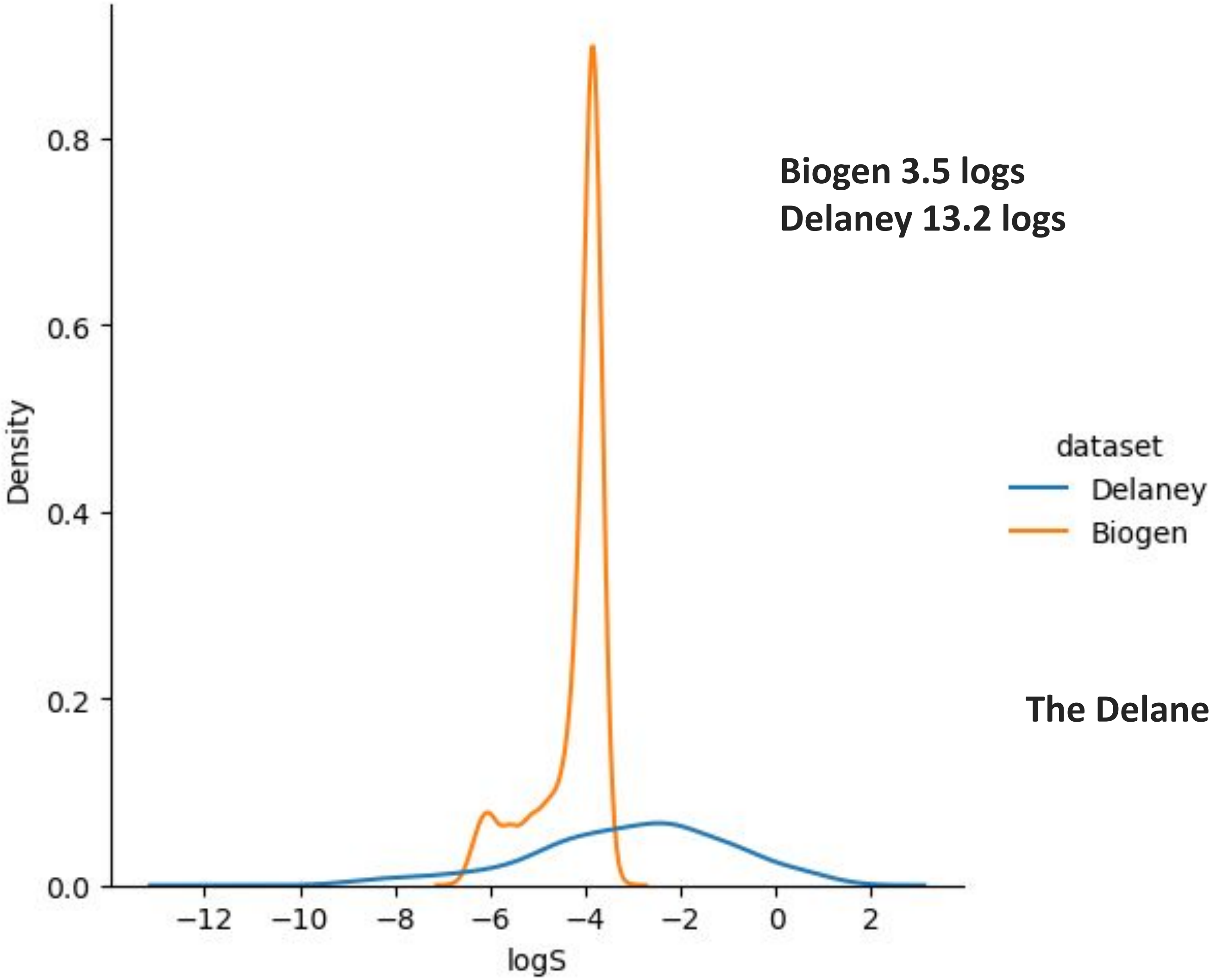


IC50, only activity curation

1599 assay pairs. 50385 points.
R2=0.32, Spearman R=0.68 MAE=0.51

Fraction > 0.3: 0.65
Fraction > 1.0: 0.28

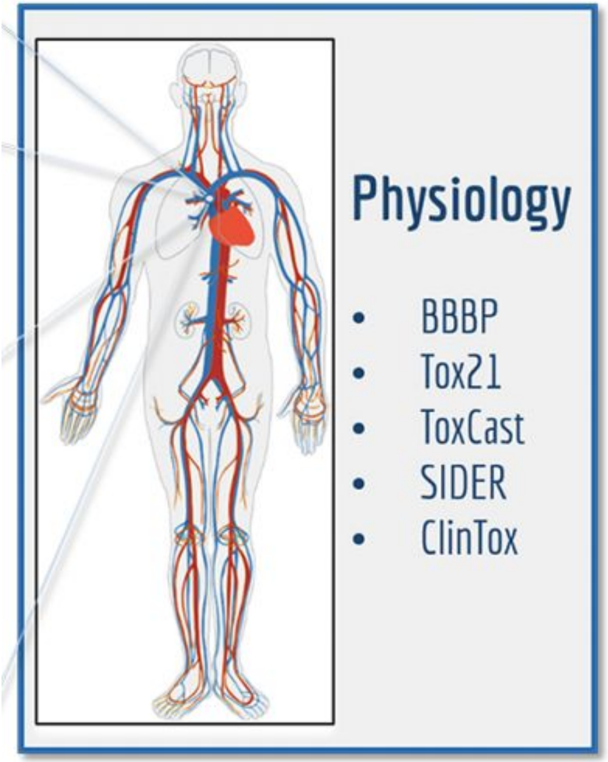
Having a Realistic Dynamic Range is Important



Biogen 3.5 logs
Delaney 13.2 logs

The Delaney solubility dataset is a component of MoleculeNet

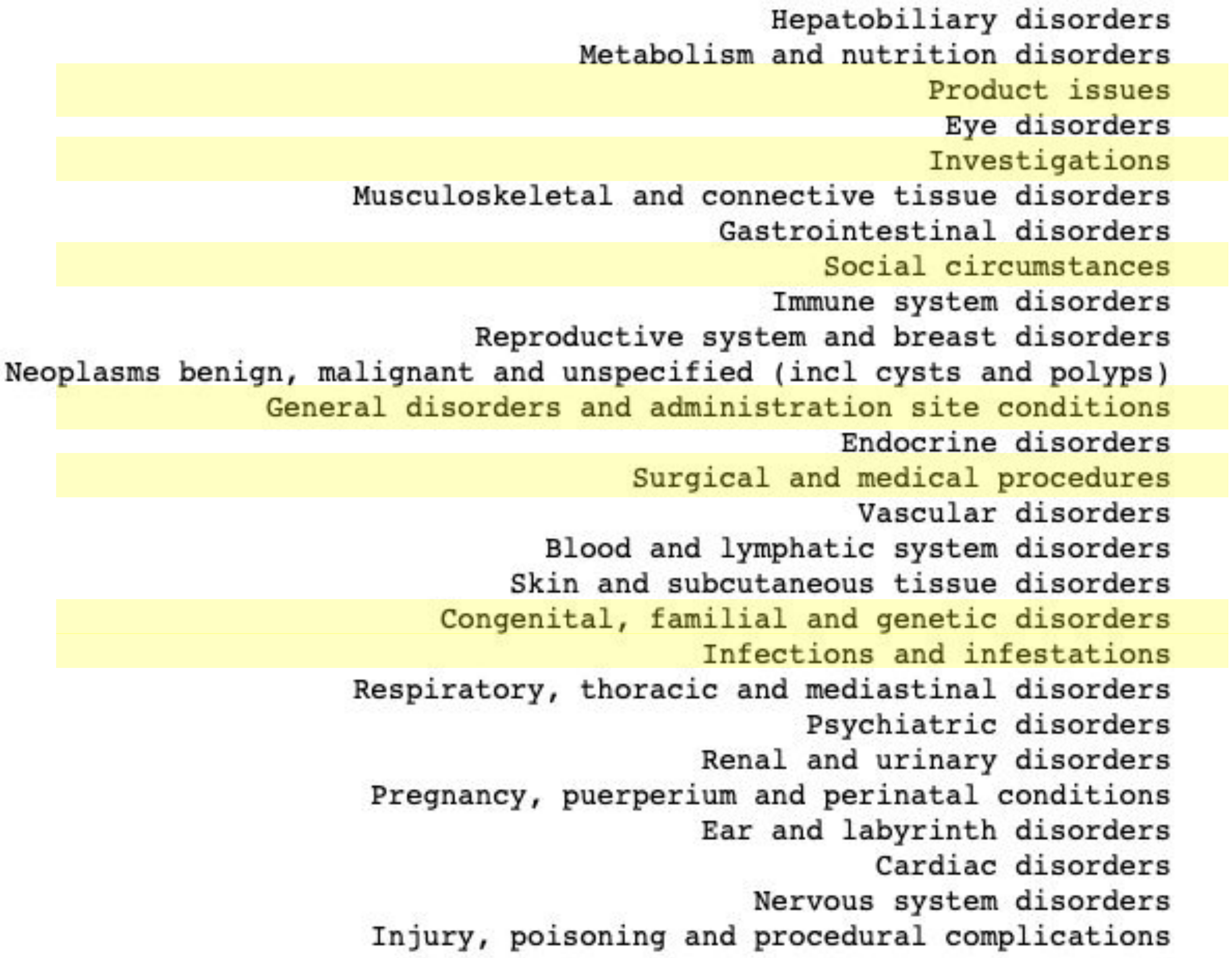
Focus on Simple, Consistent, Well-Defined Endpoints



Physiology

- BBBP
- Tox21
- ToxCast
- SIDER
- ClinTox

SIDER Categories



```
                                           Hepatobiliary disorders
                                  Metabolism and nutrition disorders
                                                     Product issues
                                                      Eye disorders
                                                     Investigations
                     Musculoskeletal and connective tissue disorders
                                         Gastrointestinal disorders
                                               Social circumstances
                                            Immune system disorders
                         Reproductive system and breast disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
             General disorders and administration site conditions
                                                Endocrine disorders
                                       Surgical and medical procedures
                                                 Vascular disorders
                                 Blood and lymphatic system disorders
                               Skin and subcutaneous tissue disorders
                              Congenital, familial and genetic disorders
                                        Infections and infestations
                    Respiratory, thoracic and mediastinal disorders
                                             Psychiatric disorders
                                        Renal and urinary disorders
                     Pregnancy, puerperium and perinatal conditions
                                        Ear and labyrinth disorders
                                                  Cardiac disorders
                                           Nervous system disorders
                     Injury, poisoning and procedural complications
```

CDD and Open Source

CDD Chemistry Engine

- Backend uses RDKit
- Frontend uses RDKit and WebMolKit (Alex Clark)

CDD contributions to RDKit:

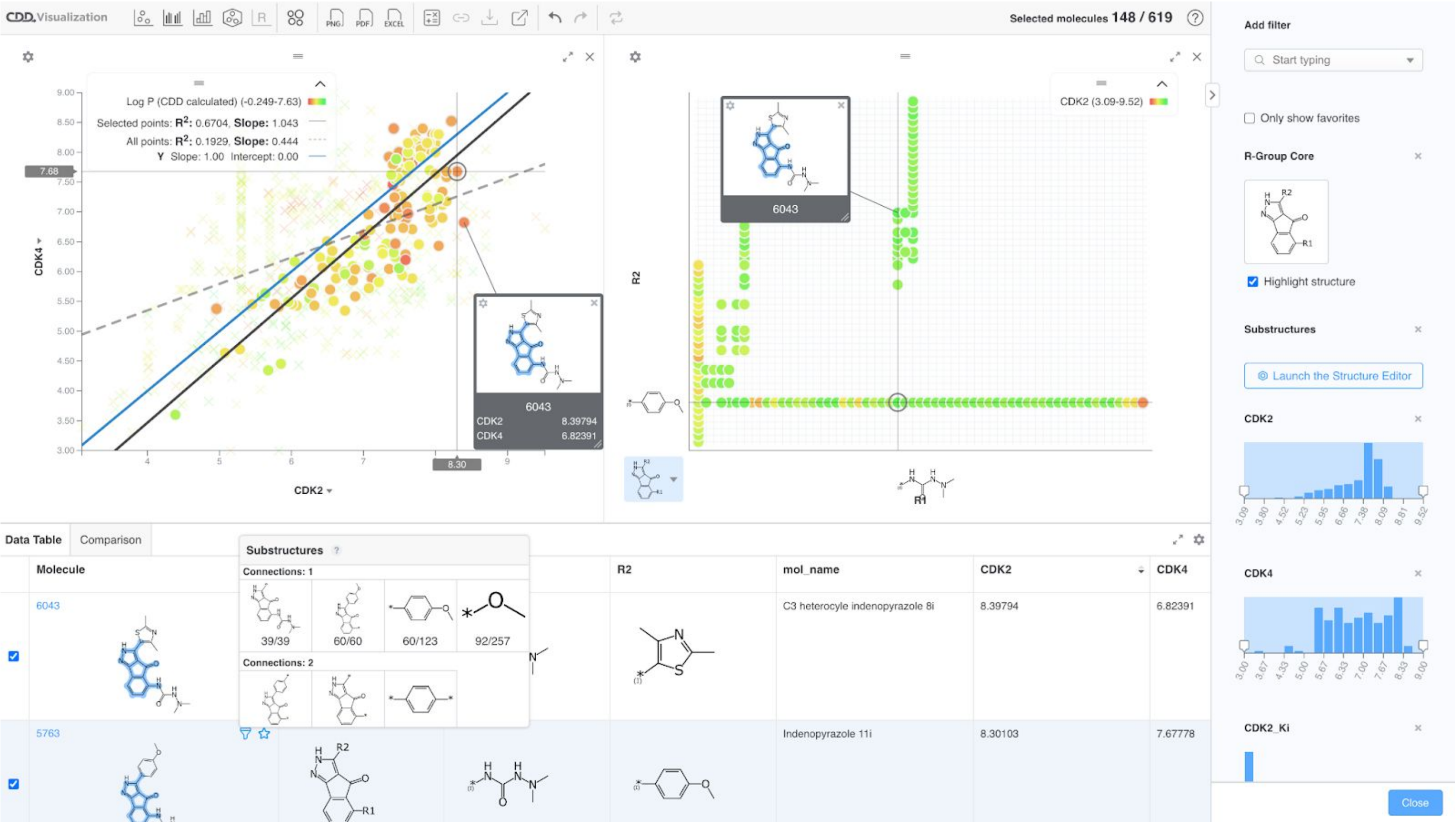
- Representation of atropisomers

Ketcher (open source structure editor):

- Work with *epam* to implement our requirements
- Focus on adding macromolecule functionality (non-natural peptides, ADC, …)

CDD Visualization

- Not open source, but scientists can use CDD Visualization for free



	Molecule	R2	mol_name	CDK2	CDK4
☑	6043		C3 heterocyle indenopyrazole 8i	8.39794	6.82391
☑	5763		Indenopyrazole 11i	8.30103	7.67778

https://www.collaborativedrug.com/scientific-data-visualization-software

RDKit: State of the Toolkit

Greg Landrum
@dr_greg_landrum@sciencemastodon.com
@greg_landrum.bsky.social

Usage in Commercial Tools

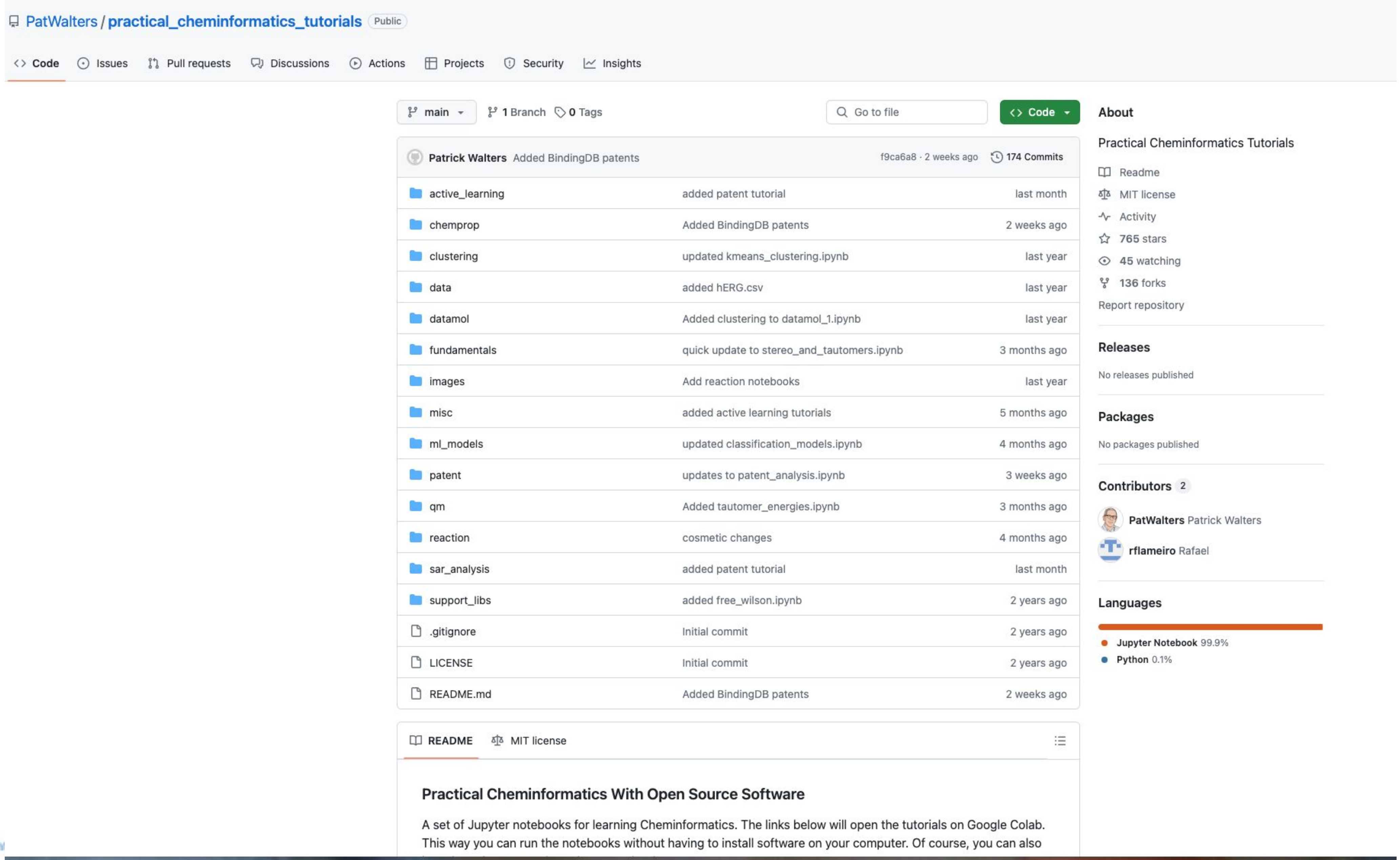
- Amazon Web Services
- Collaborative Drug Discovery
- Cresset Software
- Dalke Scientific Software
- Datagrok
- Glysade
- MedChemica
- NextMove Software
- Schrödinger
- SCM
- Wolfram Research

Disclaimer: this info is from public statements made by people from those companies. I almost certainly have forgotten someone

Adoption Measures

- Mailing lists: ~250 messages to rdkit-discuss from 2022.09 - 2023.08
- Google scholar: >2300 hits for "rdkit" in 2022, >2000 so far in 2023
- Searching github for ”from rdkit import Chem” returns >27000 code results
- Each of the last nine in-person UGMs at capacity with 40-150 attendees

Models, Tools, Technologies for Comparisons



PatWalters / practical_cheminformatics_tutorials Public

Patrick Walters Added BindingDB patents f9ca6a8 · 2 weeks ago 174 Commits

active_learning	added patent tutorial	last month
chemprop	Added BindingDB patents	2 weeks ago
clustering	updated kmeans_clustering.ipynb	last year
data	added hERG.csv	last year
datamol	Added clustering to datamol_1.ipynb	last year
fundamentals	quick update to stereo_and_tautomers.ipynb	3 months ago
images	Add reaction notebooks	last year
misc	added active learning tutorials	5 months ago
ml_models	updated classification_models.ipynb	4 months ago
patent	updates to patent_analysis.ipynb	3 weeks ago
qm	Added tautomer_energies.ipynb	3 months ago
reaction	cosmetic changes	4 months ago
sar_analysis	added patent tutorial	last month
support_libs	added free_wilson.ipynb	2 years ago
.gitignore	Initial commit	2 years ago
LICENSE	Initial commit	2 years ago
README.md	Added BindingDB patents	2 weeks ago

About

Practical Cheminformatics Tutorials

- Readme
- MIT license
- Activity
- 765 stars
- 45 watching
- 136 forks

Report repository

Releases

No releases published

Packages

No packages published

Contributors 2

- PatWalters Patrick Walters
- rflameiro Rafael

Languages

- Jupyter Notebook 99.9%
- Python 0.1%

Practical Cheminformatics With Open Source Software

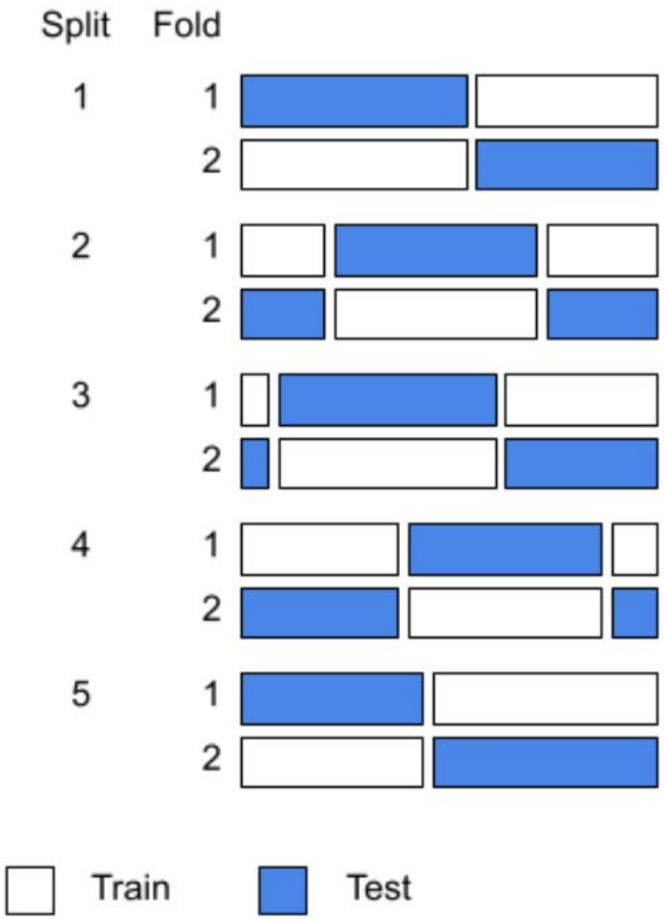
A set of Jupyter notebooks for learning Cheminformatics. The links below will open the tutorials on Google Colab. This way you can run the notebooks without having to install software on your computer. Of course, you can also

© All Righ

How can we compare different modeling methods?

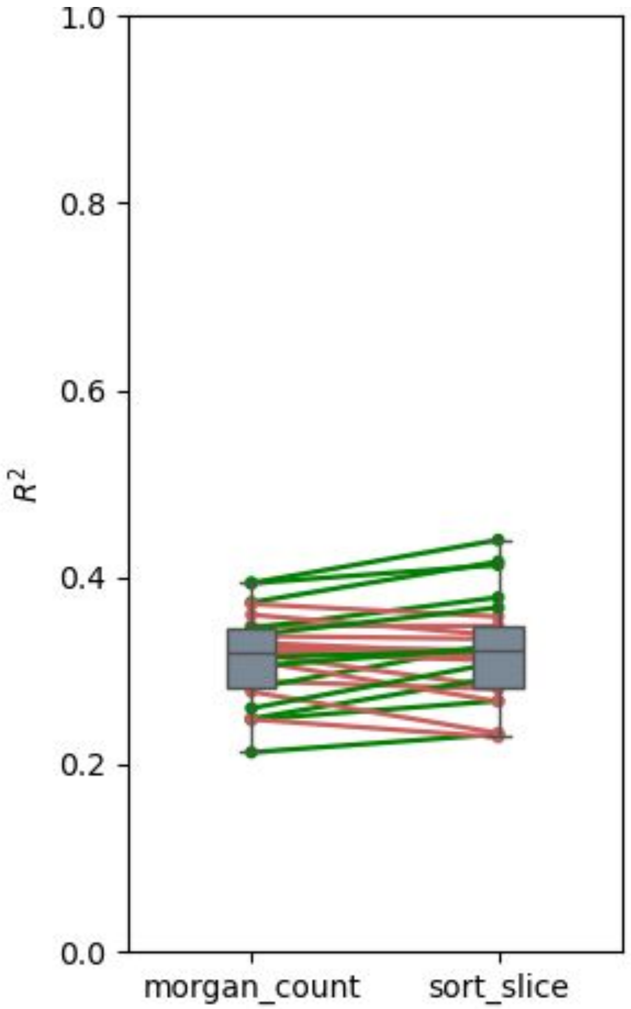
© All Rights Reserved Collaborative Drug Discovery

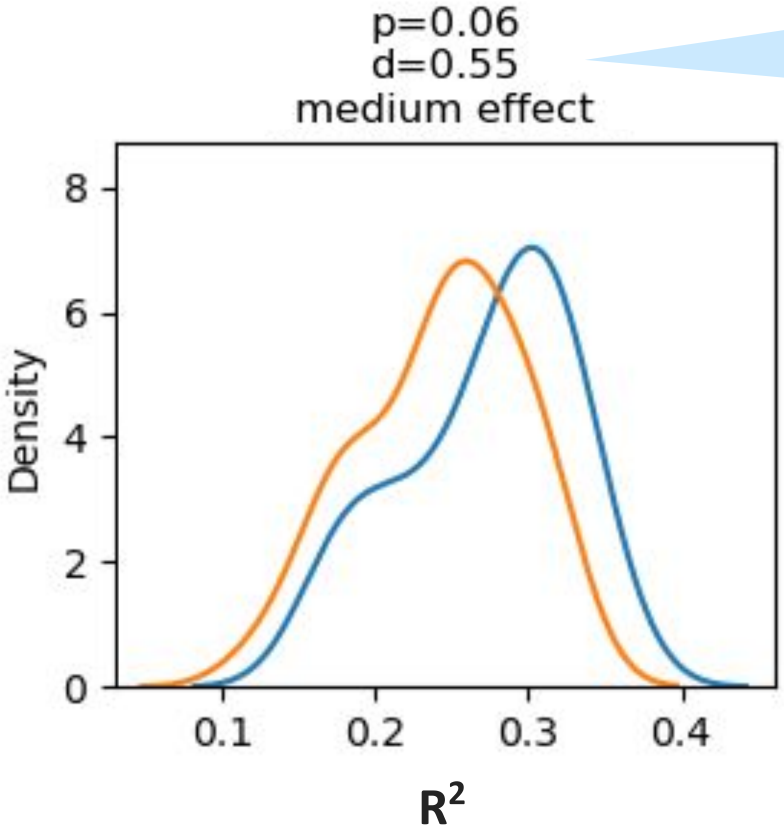
Don’t Compare Mean Performance Across Cross-Validation Folds



Method	Mean R^2
1	0.46
2	**0.48**

We Have Distributions Across Folds





There are well established statistical tests for comparing distributions

Thoughts on when it is best to use different methods: AI, Deep Learning, Generative QSAR and/or Structure-Based regression vs classification models?

What are the key differentiators?

© All Rights Reserved Collaborative Drug Discovery

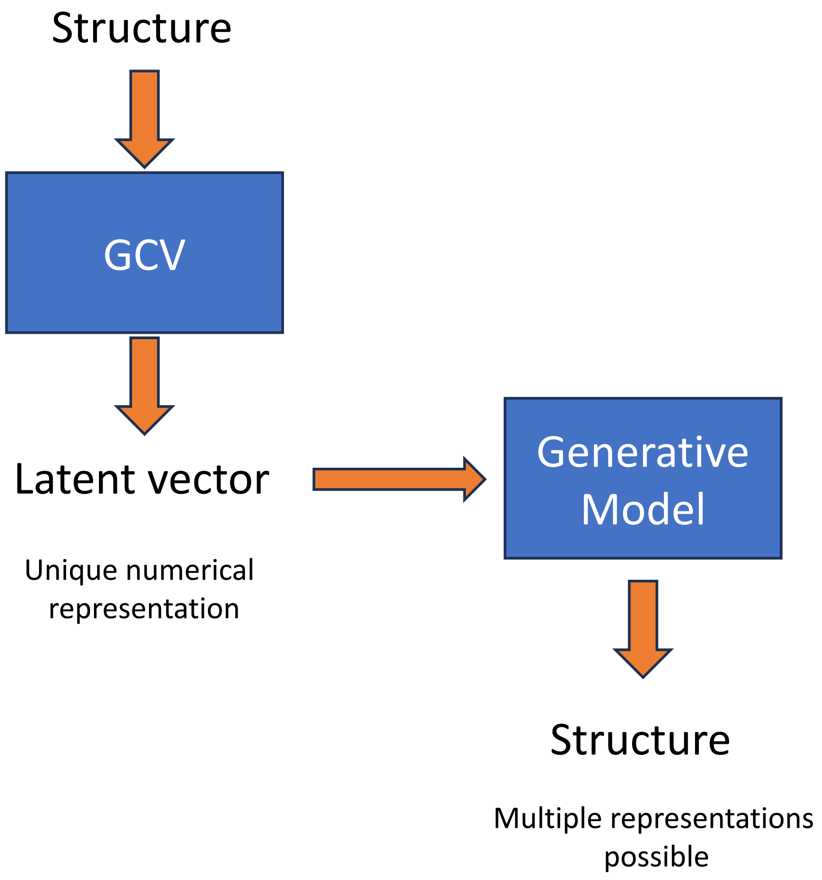
Deep Learning Technology

CDD Deep Learning model

- Create *unique* numerical representation of chemical structures (latent vector)
- Generate structures for a given latent vector

Validation study showed

- Latent vector encodes chemical structure
- Latent vector encodes structural relationships

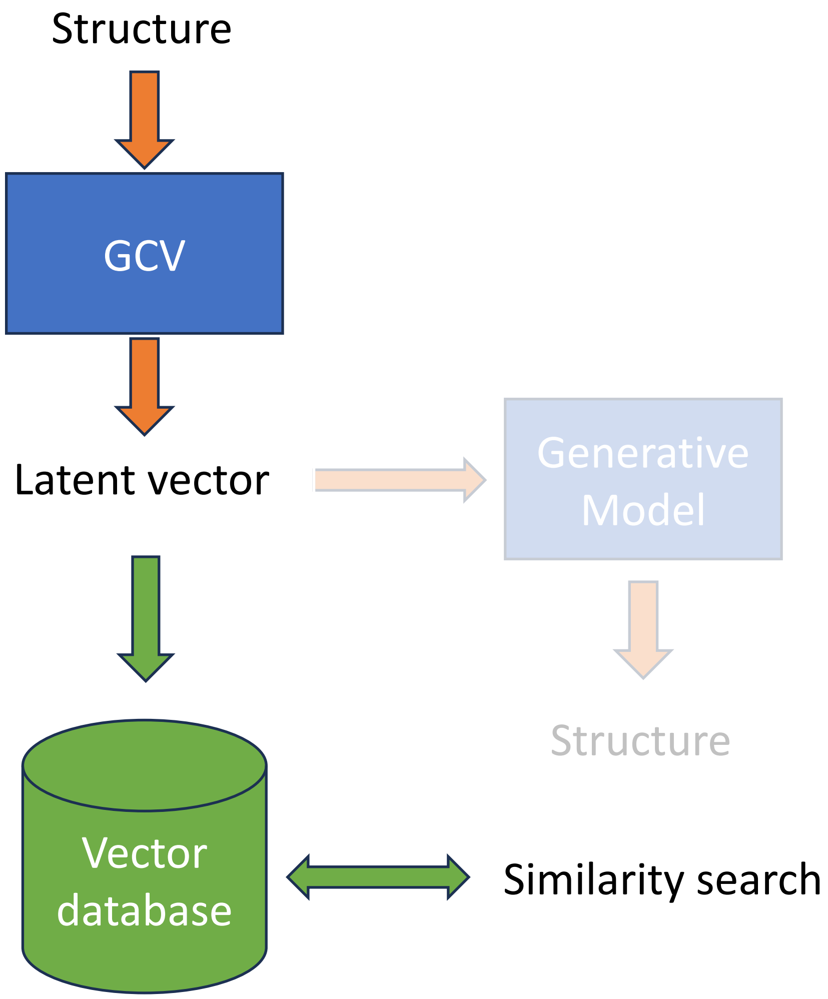


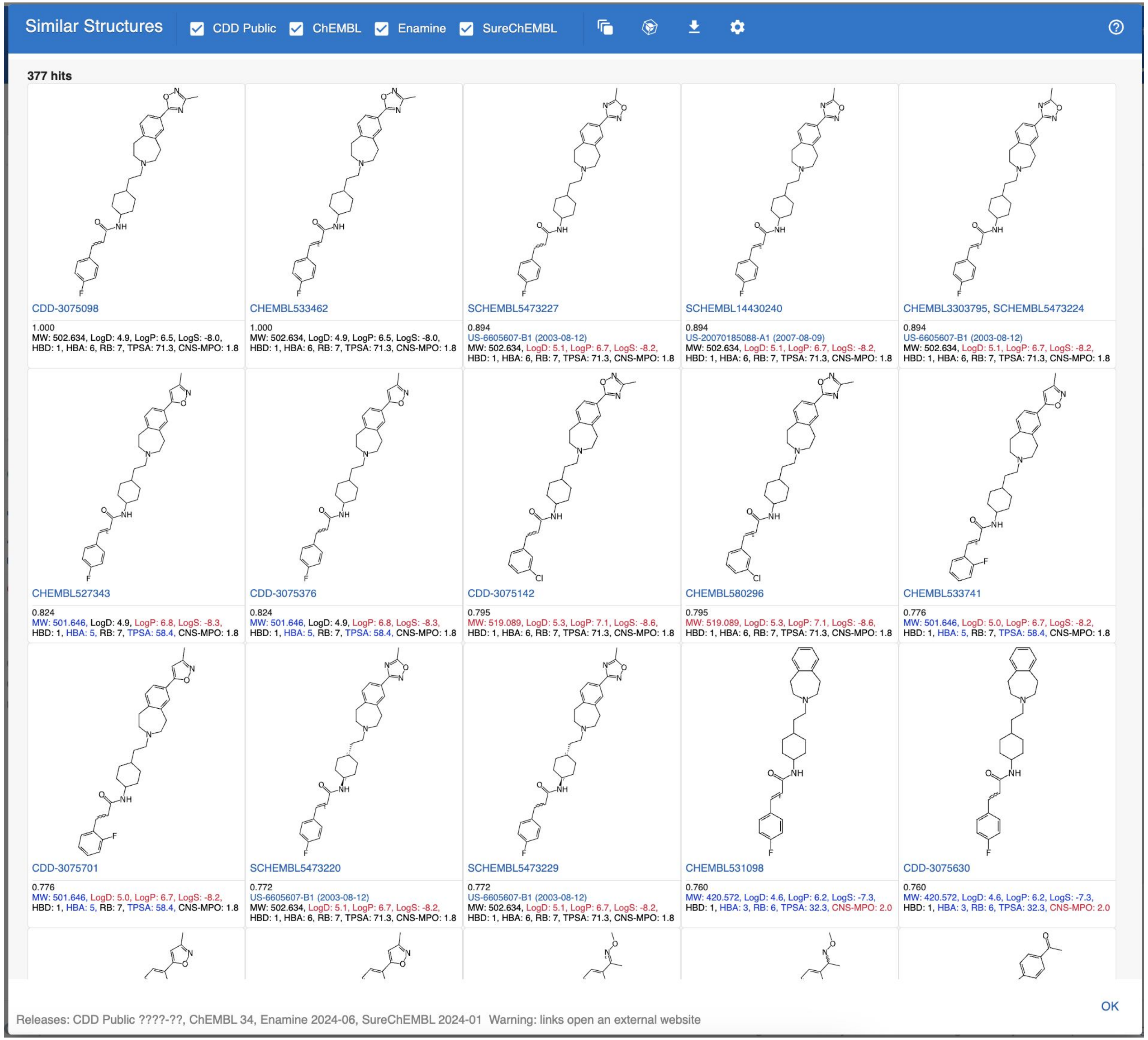
Unique numerical representation

Multiple representations possible

© All Rights Reserved Collaborative Drug Discovery

Deep Learning Similarity





377 hits

CDD-3075098
1.000
MW: 502.634, LogD: 4.9, LogP: 6.5, LogS: -8.0,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

CHEMBL533462
1.000
MW: 502.634, LogD: 4.9, LogP: 6.5, LogS: -8.0,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

SCHEMBL5473227
0.894
US-6605607-B1 (2003-08-12)
MW: 502.634, LogD: 5.1, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

SCHEMBL14430240
0.894
US-20070185088-A1 (2007-08-09)
MW: 502.634, LogD: 5.1, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

CHEMBL3303795, SCHEMBL5473224
0.894
US-6605607-B1 (2003-08-12)
MW: 502.634, LogD: 5.1, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

CHEMBL527343
0.824
MW: 501.646, LogD: 4.9, LogP: 6.8, LogS: -8.3,
HBD: 1, HBA: 5, RB: 7, TPSA: 58.4, CNS-MPO: 1.8

CDD-3075376
0.824
MW: 501.646, LogD: 4.9, LogP: 6.8, LogS: -8.3,
HBD: 1, HBA: 5, RB: 7, TPSA: 58.4, CNS-MPO: 1.8

CDD-3075142
0.795
MW: 519.089, LogD: 5.3, LogP: 7.1, LogS: -8.6,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

CHEMBL580296
0.795
MW: 519.089, LogD: 5.3, LogP: 7.1, LogS: -8.6,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

CHEMBL533741
0.776
MW: 501.646, LogD: 5.0, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 5, RB: 7, TPSA: 58.4, CNS-MPO: 1.8

CDD-3075701
0.776
MW: 501.646, LogD: 5.0, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 5, RB: 7, TPSA: 58.4, CNS-MPO: 1.8

SCHEMBL5473220
0.772
US-6605607-B1 (2003-08-12)
MW: 502.634, LogD: 5.1, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

SCHEMBL5473229
0.772
US-6605607-B1 (2003-08-12)
MW: 502.634, LogD: 5.1, LogP: 6.7, LogS: -8.2,
HBD: 1, HBA: 6, RB: 7, TPSA: 71.3, CNS-MPO: 1.8

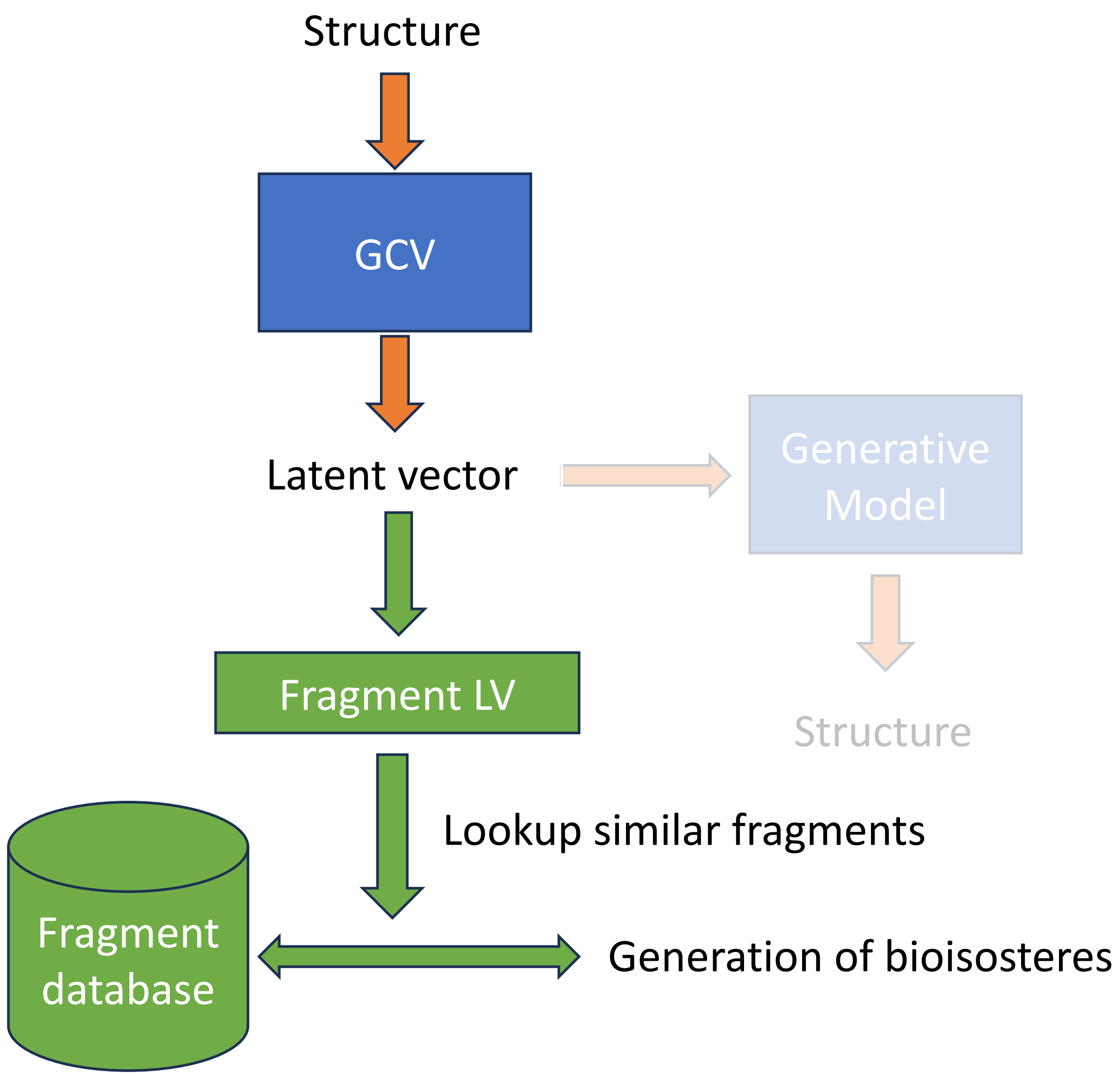
CHEMBL531098
0.760
MW: 420.572, LogD: 4.6, LogP: 6.2, LogS: -7.3,
HBD: 1, HBA: 3, RB: 6, TPSA: 32.3, CNS-MPO: 2.0

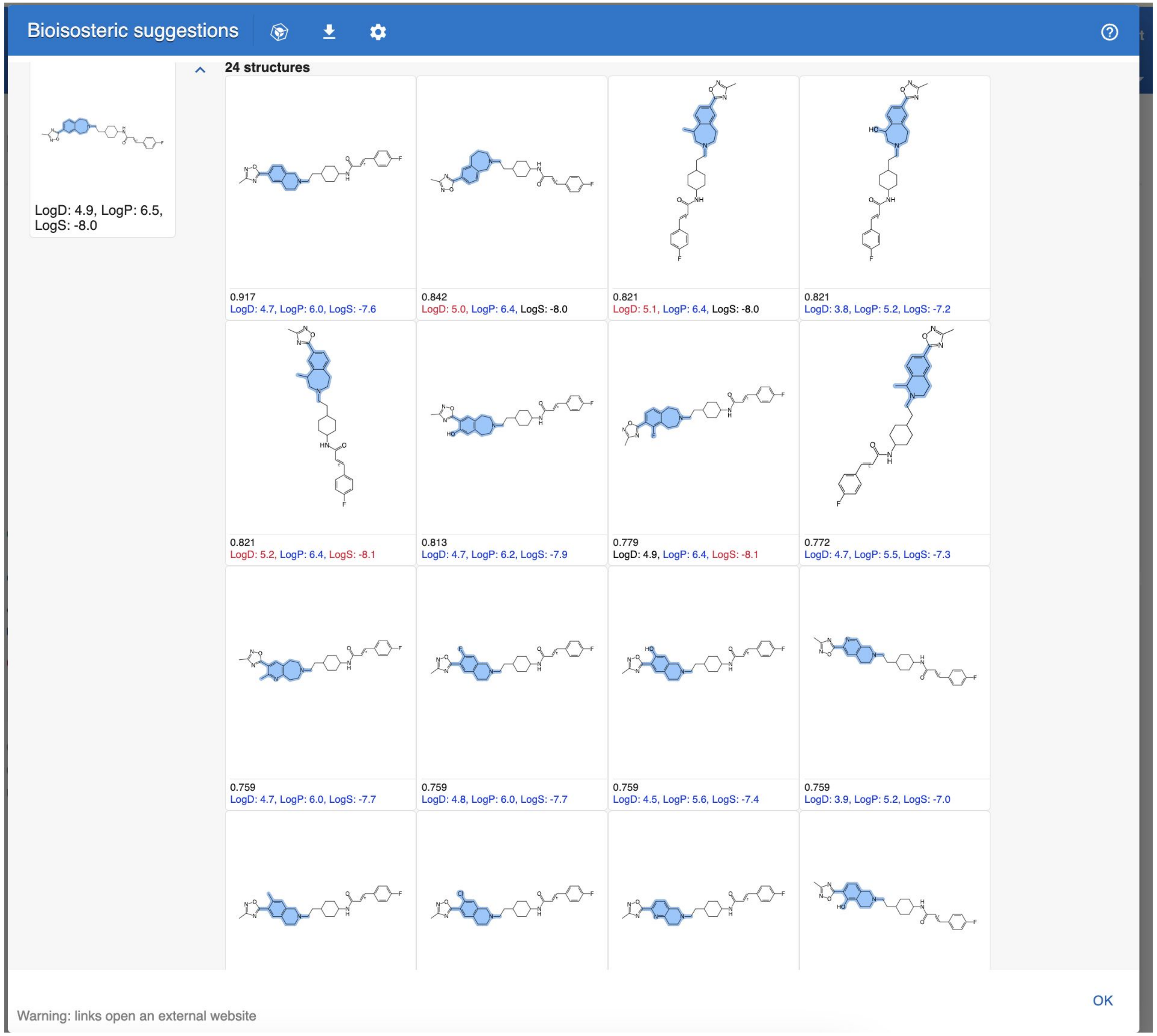
CDD-3075630
0.760
MW: 420.572, LogD: 4.6, LogP: 6.2, LogS: -7.3,
HBD: 1, HBA: 3, RB: 6, TPSA: 32.3, CNS-MPO: 2.0

Releases: CDD Public ????-??, ChEMBL 34, Enamine 2024-06, SureChEMBL 2024-01 Warning: links open an external website

© All Rights Reserved Collaborative Drug Discovery

Generative Molecular Bioisosteres





24 structures

LogD: 4.9, LogP: 6.5, LogS: -8.0

0.917
LogD: 4.7, LogP: 6.0, LogS: -7.6

0.842
LogD: 5.0, LogP: 6.4, LogS: -8.0

0.821
LogD: 5.1, LogP: 6.4, LogS: -8.0

0.821
LogD: 3.8, LogP: 5.2, LogS: -7.2

0.821
LogD: 5.2, LogP: 6.4, LogS: -8.1

0.813
LogD: 4.7, LogP: 6.2, LogS: -7.9

0.779
LogD: 4.9, LogP: 6.4, LogS: -8.1

0.772
LogD: 4.7, LogP: 5.5, LogS: -7.3

0.759
LogD: 4.7, LogP: 6.0, LogS: -7.7

0.759
LogD: 4.8, LogP: 6.0, LogS: -7.7

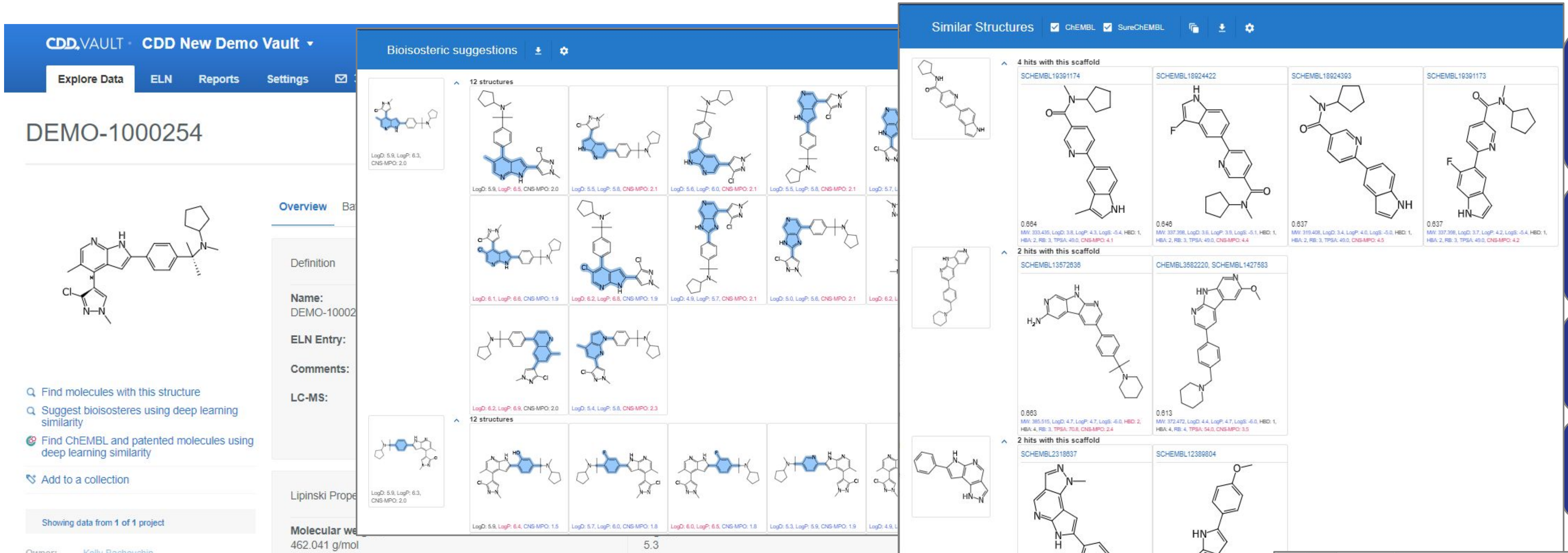
0.759
LogD: 4.5, LogP: 5.6, LogS: -7.4

0.759
LogD: 3.9, LogP: 5.2, LogS: -7.0

Warning: links open an external website

© All Rights Reserved Collaborative Drug Discovery

Deep Learning: ChEMBL, SureChEMBL, New IP



DEMO-1000254

- Find molecules with this structure
- Suggest bioisosteres using deep learning similarity
- Find ChEMBL and patented molecules using deep learning similarity
- Add to a collection

Showing data from 1 of 1 project

Owner: Kelly Bachovchin
Created: February 14, 2024
Updated: Februar

Definition

Name: DEMO-10002

ELN Entry:

Comments:

LC-MS:

Lipinski Prope

Molecular we 462.041 g/mol

5.3

log P:

log S:

LogD: 5.9, LogP: 6.3, CNS-MPO: 2.0

0.664
MW: 333.435, LogD: 3.8, LogP: 4.3, LogS: -5.4, HBD: 1, HBA: 2, RB: 3, TPSA: 49.0, CNS-MPO: 4.1

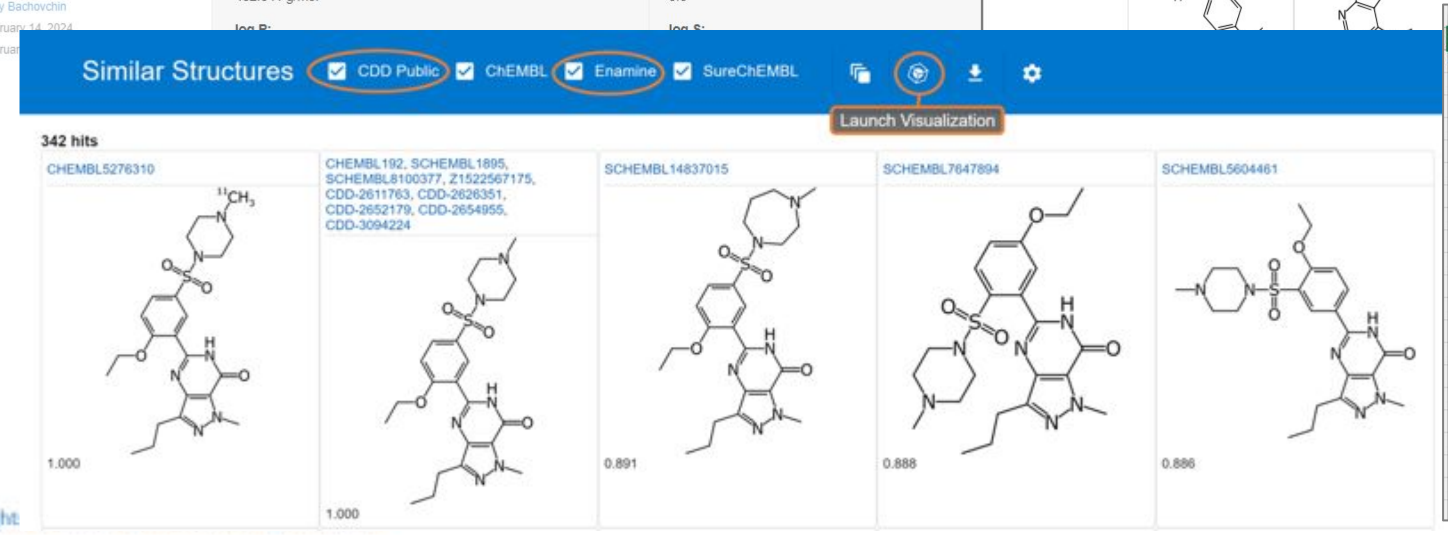
0.646
MW: 337.398, LogD: 3.6, LogP: 3.9, LogS: -5.1, HBD: 1, HBA: 2, RB: 3, TPSA: 49.0, CNS-MPO: 4.4

0.637
MW: 319.408, LogD: 3.4, LogP: 4.0, LogS: -5.0, HBD: 1, HBA: 2, RB: 3, TPSA: 49.0, CNS-MPO: 4.5

0.637
MW: 337.398, LogD: 3.7, LogP: 4.2, LogS: -5.4, HBD: 1, HBA: 2, RB: 3, TPSA: 49.0, CNS-MPO: 4.2

0.663
MW: 385.515, LogD: 4.7, LogP: 4.7, LogS: -6.0, HBD: 2, HBA: 4, RB: 3, TPSA: 70.8, CNS-MPO: 2.4

0.613
MW: 372.472, LogD: 4.4, LogP: 4.7, LogS: -6.0, HBD: 1, HBA: 4, RB: 4, TPSA: 54.0, CNS-MPO: 3.5



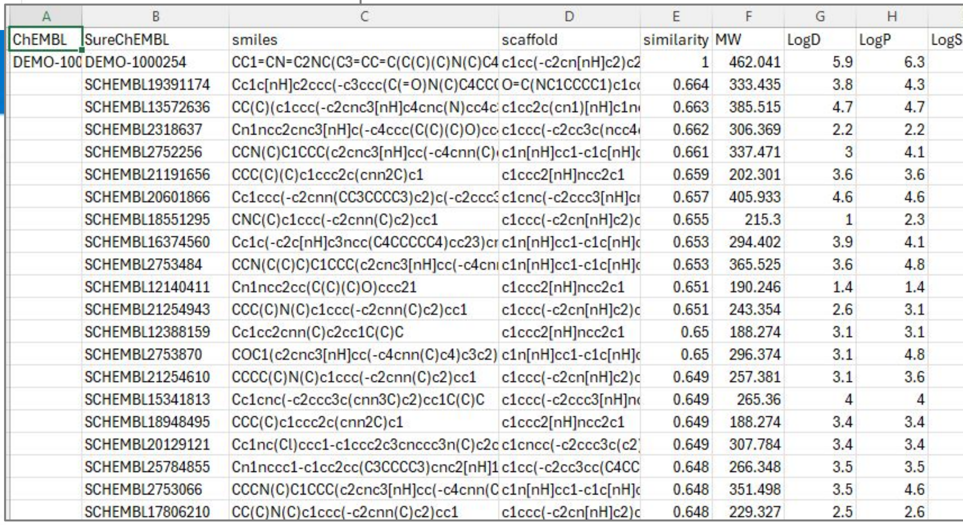
342 hits

1) Generate Bioisosteres
2) Ultra-Fast Deep Learning Similarity searching

2. Similar compounds
- Sort by scaffold
- Calculate properties

3. Export compounds to a spreadsheet

4. New hits can be used for further analysis or purchasing



ChEMBL	SureChEMBL	smiles	scaffold	similarity	MW	LogD	LogP	LogS
DEMO-100	DEMO-1000254	CC1=CN=C2NC(C3=CC=C(C(C)(C)N(C)C4	c1cc(-c2cn[nH]c2)c2	1	462.041	5.9	6.3	
	SCHEMBL19391174	Cc1c[nH]c2ccc(-c3ccc(C(=O)N(C)C4CC(	O=C(NC1CCCC1)c1cc	0.664	333.435	3.8	4.3	
	SCHEMBL13572636	CC(C)(c1ccc(-c2cnc3[nH]c4cnc(N)cc4c	c1cc2c(cn1)[nH]c1n	0.663	385.515	4.7	4.7	
	SCHEMBL2318637	Cn1ncc2cnc3[nH]c(-c4ccc(C(C)(C)O)cc	c1ccc(-c2cc3c(ncc4	0.662	306.369	2.2	2.2	
	SCHEMBL2752256	CCN(C)C1CCC(c2cnc3[nH]cc(-c4cnn(C)	c1n[nH]cc1-c1c[nH]c	0.661	337.471	3	4.1	
	SCHEMBL21191656	CCC(C)(C)c1ccc2c(cnn2C)c1	c1ccc2[nH]ncc2c1	0.659	202.301	3.6	3.6	
	SCHEMBL20601866	Cc1ccc(-c2cnn(CC3CCCC3)c2)c(-c2ccc3	c1cnc(-c2ccc3[nH]cr	0.657	405.933	4.6	4.6	
	SCHEMBL18551295	CNC(C)c1ccc(-c2cnn(C)c2)cc1	c1ccc(-c2cn[nH]c2)c	0.655	215.3	1	2.3	
	SCHEMBL16374560	Cc1c(-c2c[nH]c3ncc(C4CCCCC4)cc23)cr	c1n[nH]cc1-c1c[nH]c	0.653	294.402	3.9	4.1	
	SCHEMBL2753484	CCN(C(C)C)C1CCC(c2cnc3[nH]cc(-c4cn	c1n[nH]cc1-c1c[nH]c	0.653	365.525	3.6	4.8	
	SCHEMBL12140411	Cn1ncc2cc(C(C)(C)O)ccc21	c1ccc2[nH]ncc2c1	0.651	190.246	1.4	1.4	
	SCHEMBL21254943	CCC(C)N(C)c1ccc(-c2cnn(C)c2)cc1	c1ccc(-c2cn[nH]c2)c	0.651	243.354	2.6	3.1	
	SCHEMBL12388159	Cc1cc2cnn(C)c2cc1C(C)C	c1ccc2[nH]ncc2c1	0.65	188.274	3.1	3.1	
	SCHEMBL2753870	COC1(c2cnc3[nH]cc(-c4cnn(C)c4)c3c2)	c1n[nH]cc1-c1c[nH]c	0.65	296.374	3.1	4.8	
	SCHEMBL21254610	CCCC(C)N(C)c1ccc(-c2cnn(C)c2)cc1	c1ccc(-c2cn[nH]c2)c	0.649	257.381	3.1	3.6	
	SCHEMBL15341813	Cc1cnc(-c2ccc3c(cnn3C)c2)cc1C(C)C	c1ccc(-c2ccc3[nH]nc	0.649	265.36	4	4	
	SCHEMBL18948495	CCC(C)c1ccc2c(cnn2C)c1	c1ccc2[nH]ncc2c1	0.649	188.274	3.4	3.4	
	SCHEMBL20129121	Cc1nc(Cl)ccc1-c1ccc2c3cnccc3n(C)c2c	c1cncc(-c2ccc3c(c2)	0.649	307.784	3.4	3.4	
	SCHEMBL25784855	Cn1nccc1-c1cc2cc(C3CCCC3)cnc2[nH]1	c1cc(-c2cc3cc(C4CC	0.648	266.348	3.5	3.5	
	SCHEMBL2753066	CCCN(C)C1CCC(c2cnc3[nH]cc(-c4cnn(C	c1n[nH]cc1-c1c[nH]c	0.648	351.498	3.5	4.6	
	SCHEMBL17806210	CC(C)N(C)c1ccc(-c2cnn(C)c2)cc1	c1ccc(-c2cn[nH]c2)c	0.648	229.327	2.5	2.6	

© All Right

What are the roles for different ecosystem members (Academic, Commercial, Other)?

How has the field been impacted by advances in ML & LLMs?

Where do you see the field in a year and in a decade…

© All Rights Reserved Collaborative Drug Discovery

Questions?

CDD VAULT®

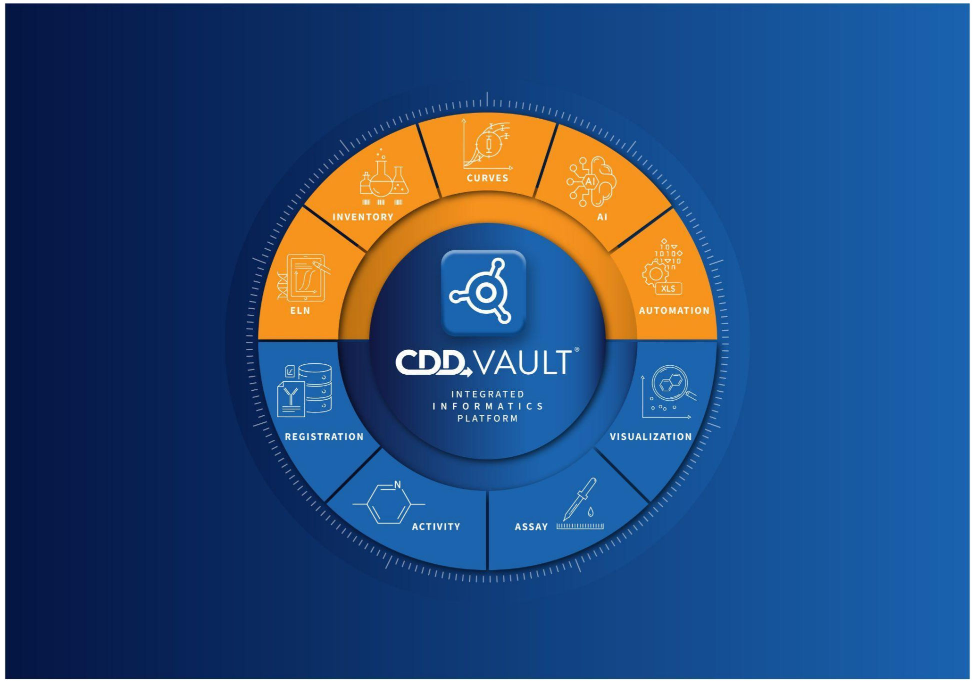
Complexity Simplified

© All Rights Reserved Collaborative Drug Discovery

Single Source of Truth

Inventory
Keep track of samples, biologicals and compounds

Activity
Manage and analyze experimental data



Automation
Connect data with our robust API, Parser, and Mapping Tools

Assays
Comparison of assays using standardized protocols

Registration
Store and organize your research data

Visualization
Plot datasets and mine them

Curves
Generate, QC, and analyze results

AI
Computer aided design

ELN
Document all your research.

www.collaborativedrug.com info@collaborativedrug.com

© All Rights Reserved Collaborative Drug Discovery