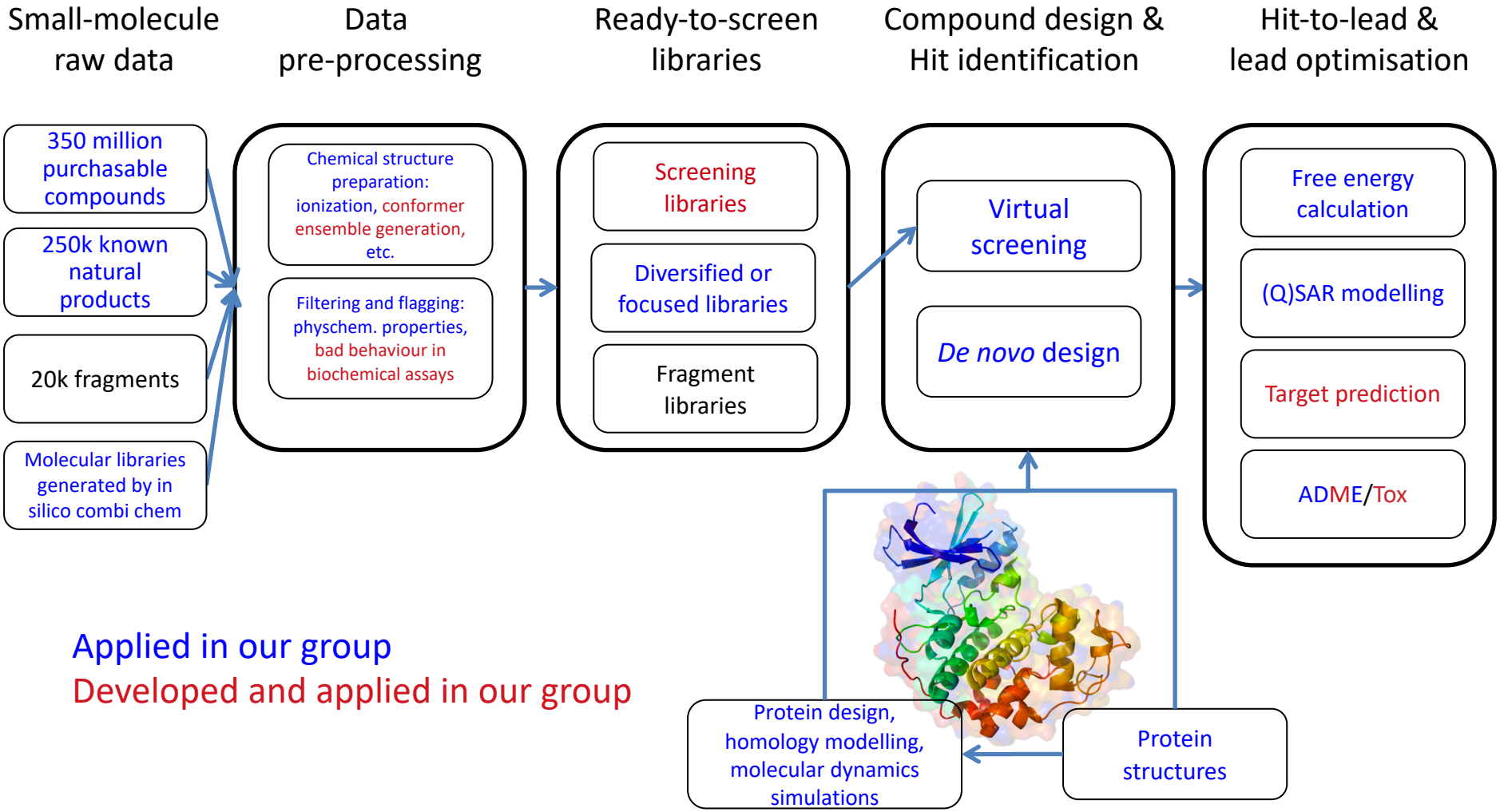
Applications of RDKit in Machine Learning

Ya Chen
Applied Cheminformatics and Molecular Design (ACM) group
Centre for Bioinformatics
Universität Hamburg
Bundesstraße 43
20146 Hamburg, Germany

www.zbh.uni-hamburg.de
chen@zbh.uni-hamburg.de

We develop and apply a wide range of computational methods that can **provide guidance to early drug discovery**



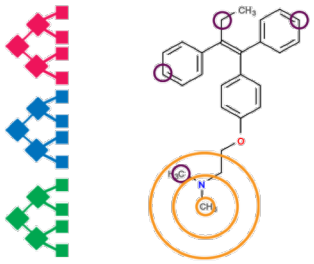
Applied in our group

Developed and applied in our group

NERDD

New E-Resource for Drug Discovery

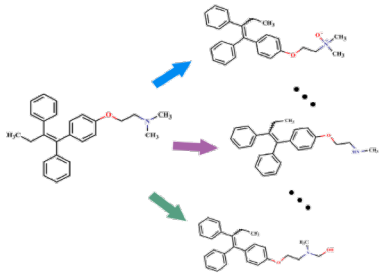
Sites of Metabolism



FAME 3

Regioselectivity prediction for phase 1 and phase 2 metabolism

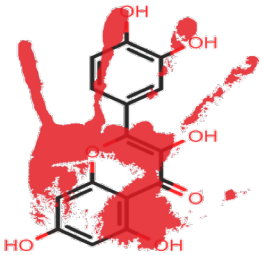
Metabolite Structures



GLORY

Metabolite structure prediction for cytochrome P450 metabolism

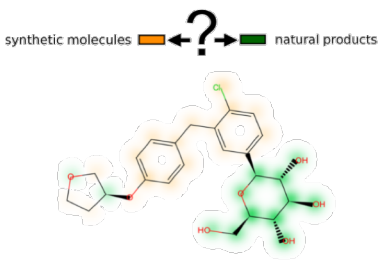
Frequent Hitters



Hit Dexter 2.0

Prediction of frequent hitters

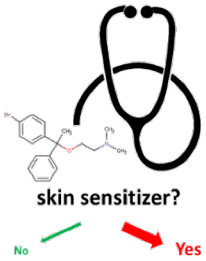
Natural Product-Likeness



NP-Scout

Identification and visualization of natural product-likeness

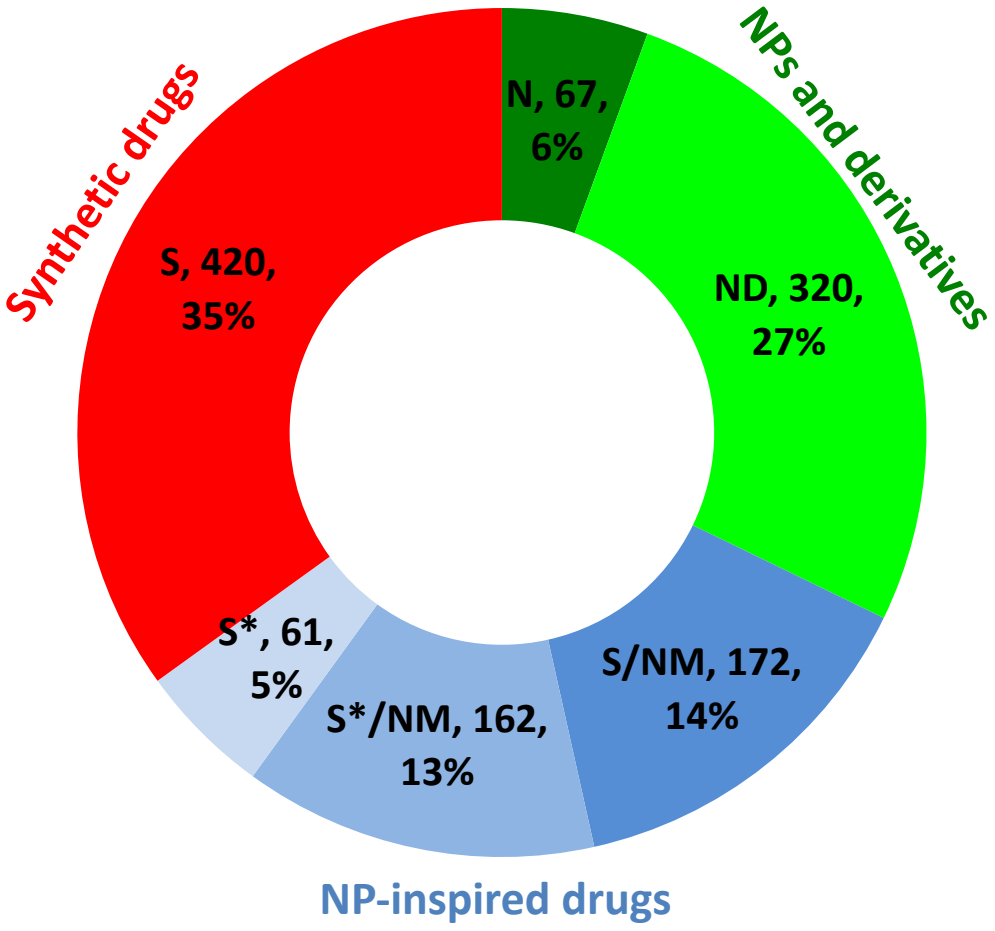
Skin Sensitization



Skin Doctor

Prediction of skin sensitization potential

Sources for small-molecule drugs (1981-2014)



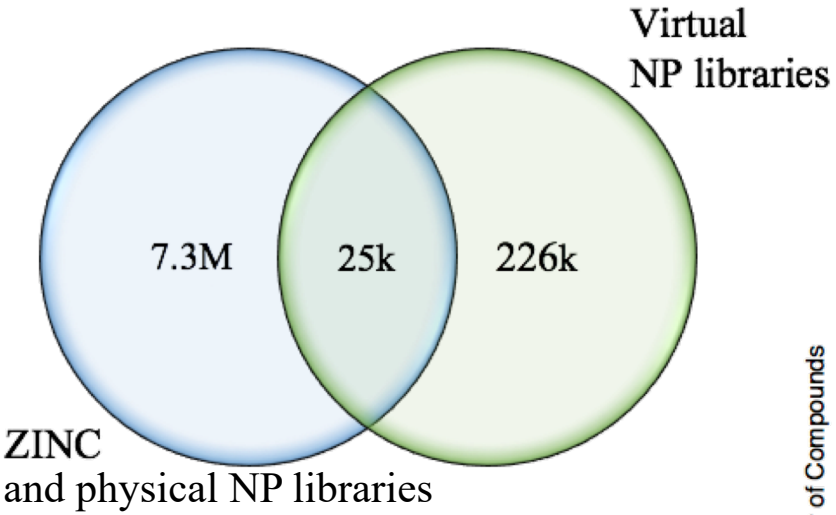
More than half of all modern small-molecule drugs are linked to natural products.

N	natural product
ND	natural product derivative (usually semi-synthetic)
S	synthetic drug
S*	synthetic drug with NP pharmacophore
/NM	mimic of natural product

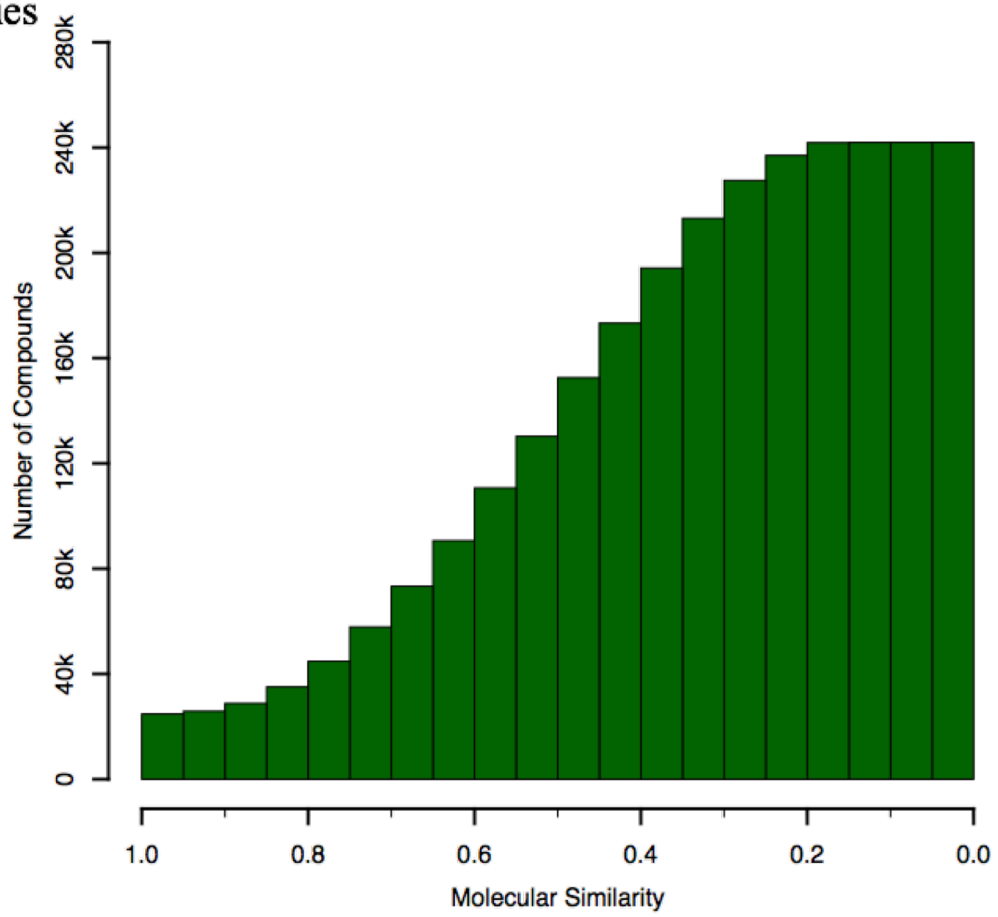
Newman D and Cragg G, *J Nat Prod*, 2016, 79, 629-661.

Readily obtainable NPs and derivatives

Structures of more than 250k NPs have been deposited to date



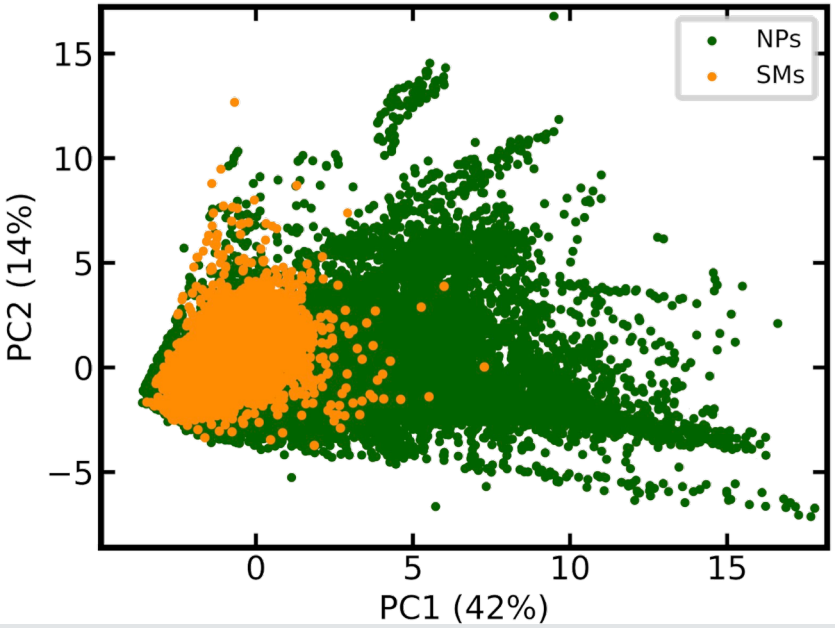
- Only approx. 10% of all known NPs are readily obtainable
- At a Morgan2 fingerprint-based TC of 0.7, about 25% of all known NPs are covered by obtainable NPs and their analogs



Chen Y. et al., *J Chem Inf Model* 2017, 57, 2099−2111.

NP-Scout: development of a method for the assessment of natural product-likeness

- Use cases:
  - Profiling of databases (% NPs; NP-likeness)
  - Identification of genuine NPs in commercial compound libraries, which often contain also synthetic molecules
  - Prioritization of compounds for experimental testing
  - Library design
- Data set:
  - 201,761 unique NPs (multiple free sources)
  - 201,761 unique synthetic compounds (ZINC)



Data preparation with RDKit

Open-Source Cheminformatics and Machine Learning

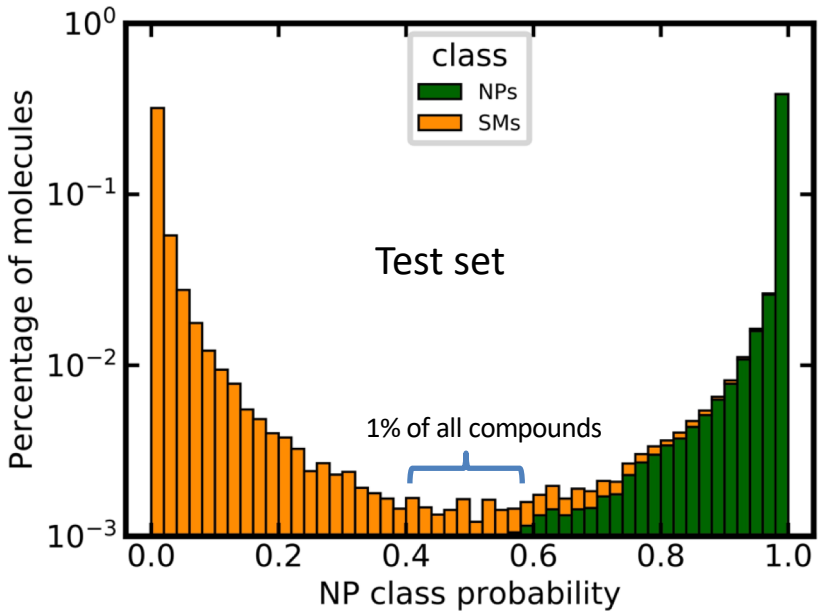
- Parsing molecules from different formats (SMILES, sdf, mol)
- ACM_wash (also used in Hit Dexter and Skin Doctor)
  - Salt filter
  - Molecular weight filter (NP-Scout: 150-1500 Da)
  - Element filter
  - Canonicalize tautomers (MolVS)
- Deduplicate compounds by canonical SMILES
- For different model, deduplicate compounds by calculated fingerprint (Morgan2 fingerprint, MACCS keys)

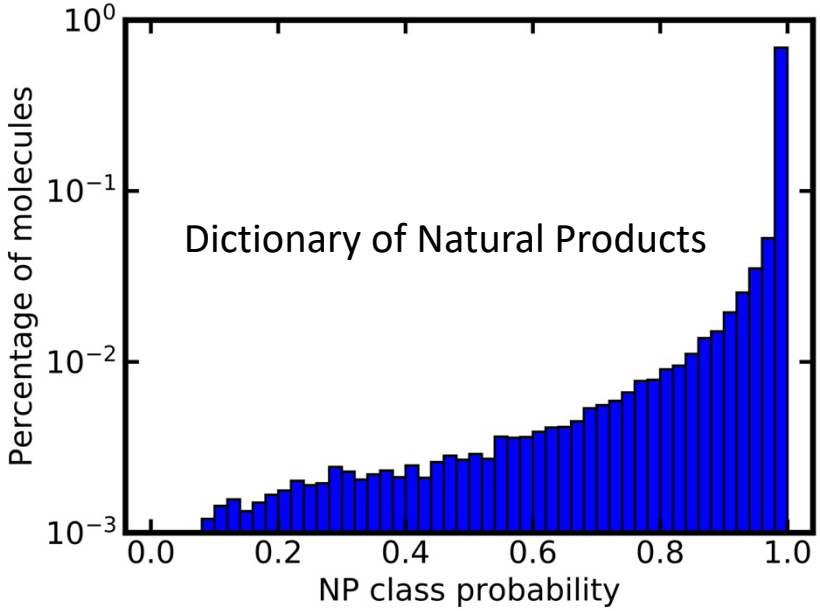
NP-Scout: Modelling approach

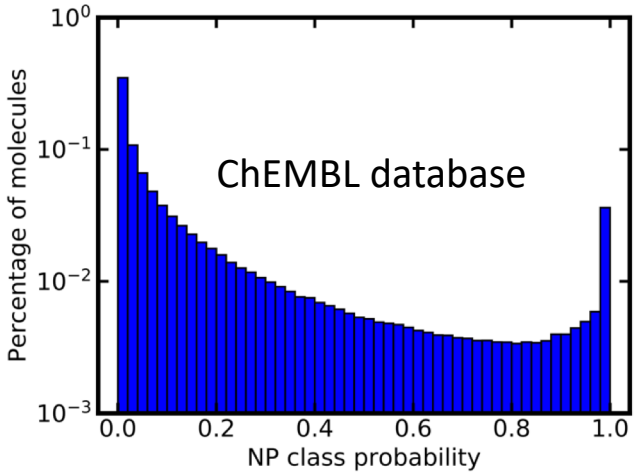
- Machine learning approach
  - Random forest classifiers
- Descriptors:
  - MOE 2D descriptors
  - Morgan2 fingerprint
  - MACCS keys

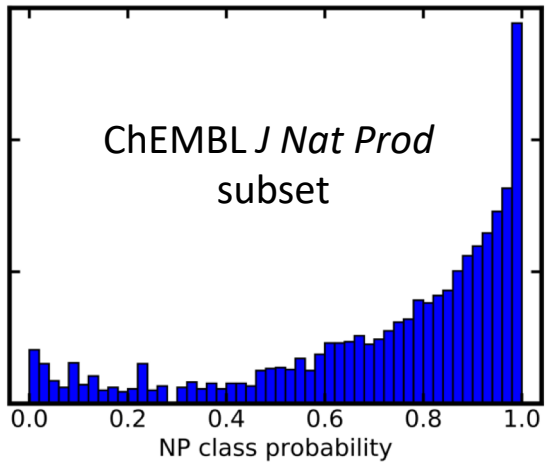
Test method	Metric	MOE 2D descriptors	Morgan2 fingerprint (1024 bits)	MACCS keys	NP-Likeness calculator
10-fold cross-validation	AUC	0.997	0.997	0.996	/
	MCC	0.953	0.958	0.950	/
External test set	AUC	0.997	0.997	0.997	0.997
	MCC	0.954	0.960	0.960	0.959

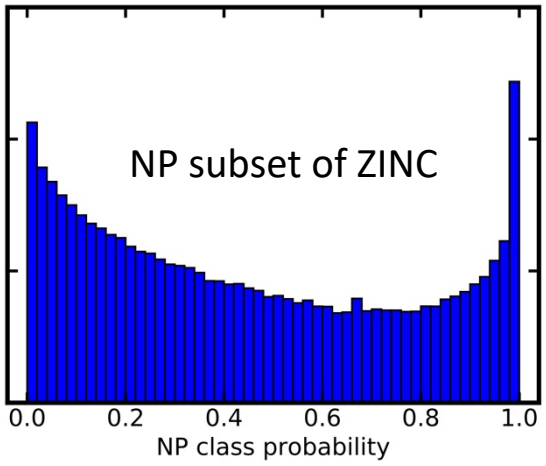
NP-Scout: Model validation





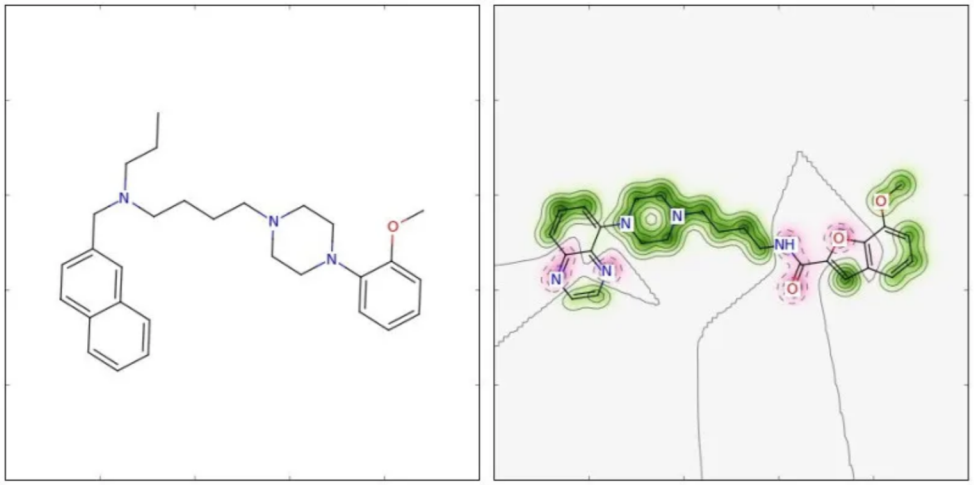






Similarity maps

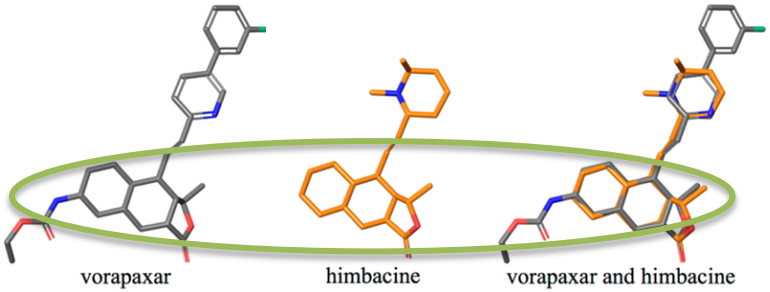
- Straightforward and general strategy to visualize the atomic contributions to
  - the similarity between two molecules or
  - the predicted probability of a machine learning model

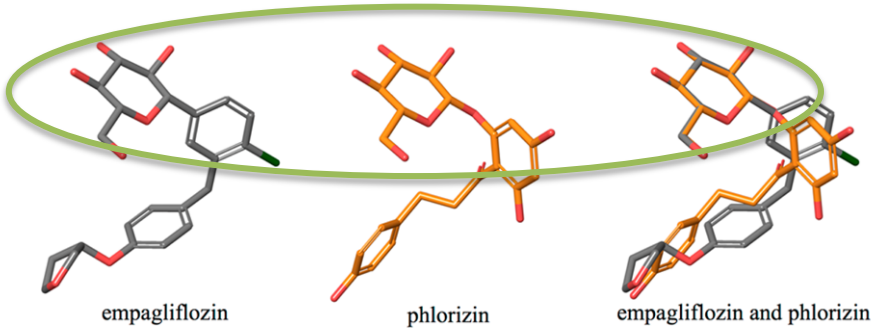


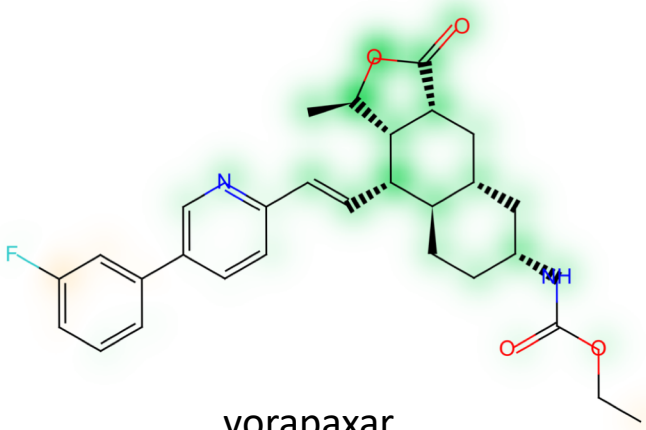
atom-pairs fingerprint

Riniker S. et al., *J. Cheminformatics.* 2013, 5, 43.

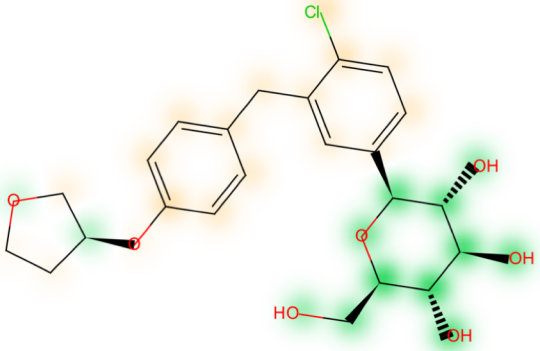
NP-Scout: Similarity maps from RDKit







vorapaxar



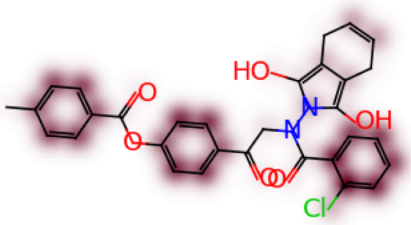
empagliflozin

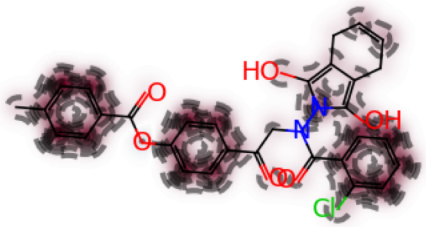
Chen Y. et al., *J Chem Inf Model* 2017, 57, 2099−2111.
Chen Y. et al., *Biomolecules* 2019, 9, 43.

Similarity maps usage

```
from rdkit.Chem.Draw import SimilarityMaps

fig, _ = SimilarityMaps.GetSimilarityMapForModel(
    molecule,
    SimilarityMaps.GetMorganFingerprint,
    predict_proba,
    size =(150,150),
    colorMap,
    contourLines=0,
    sigma=None,
    step=0.01)
```



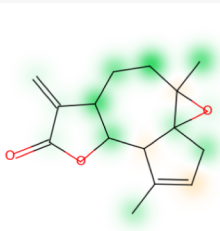
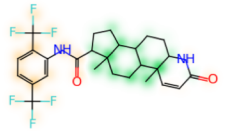


NP-Scout: web service

http://nerdd.zbh.uni-hamburg.de/npscout/

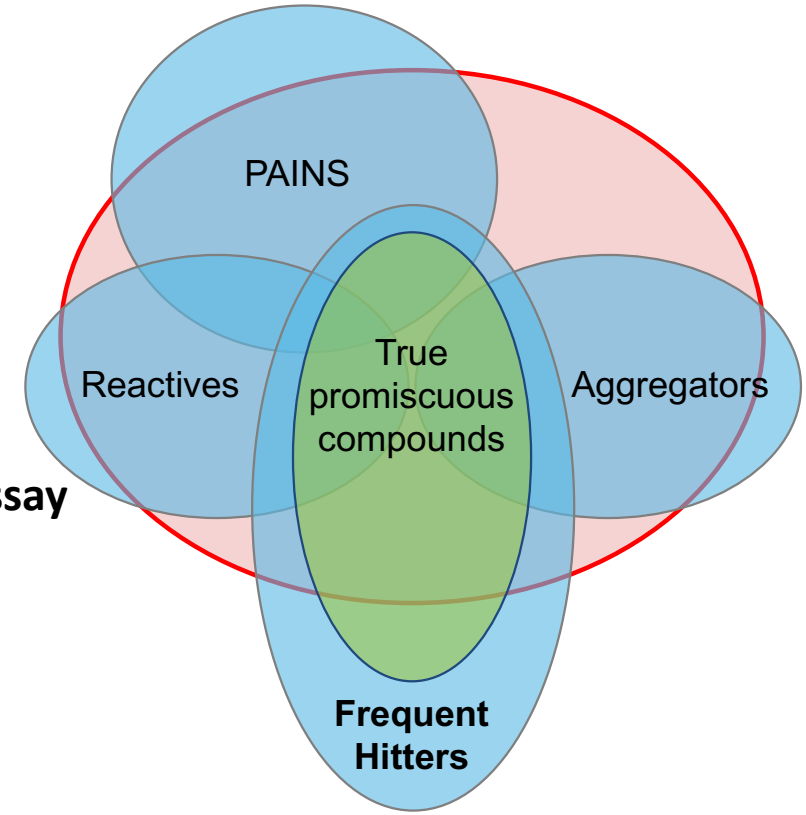
Show 10 entries

Search:

SMILES	Molecule name	Error/Warning	NP class probability	Similarity maps
C=C1C(=O)OC2C1CCC1(C)OC13CC=C(C)C23	arglabin	-	1.0	
CC12C=CC(=O)NC1CCC1C2CCC2(C)C(C(=O)Nc3cc(C(F)(F)F)ccc3C(F	dutaseride	-	0.18	

Hit Dexter 2.0: Machine-learning models for the prediction of frequent hitters

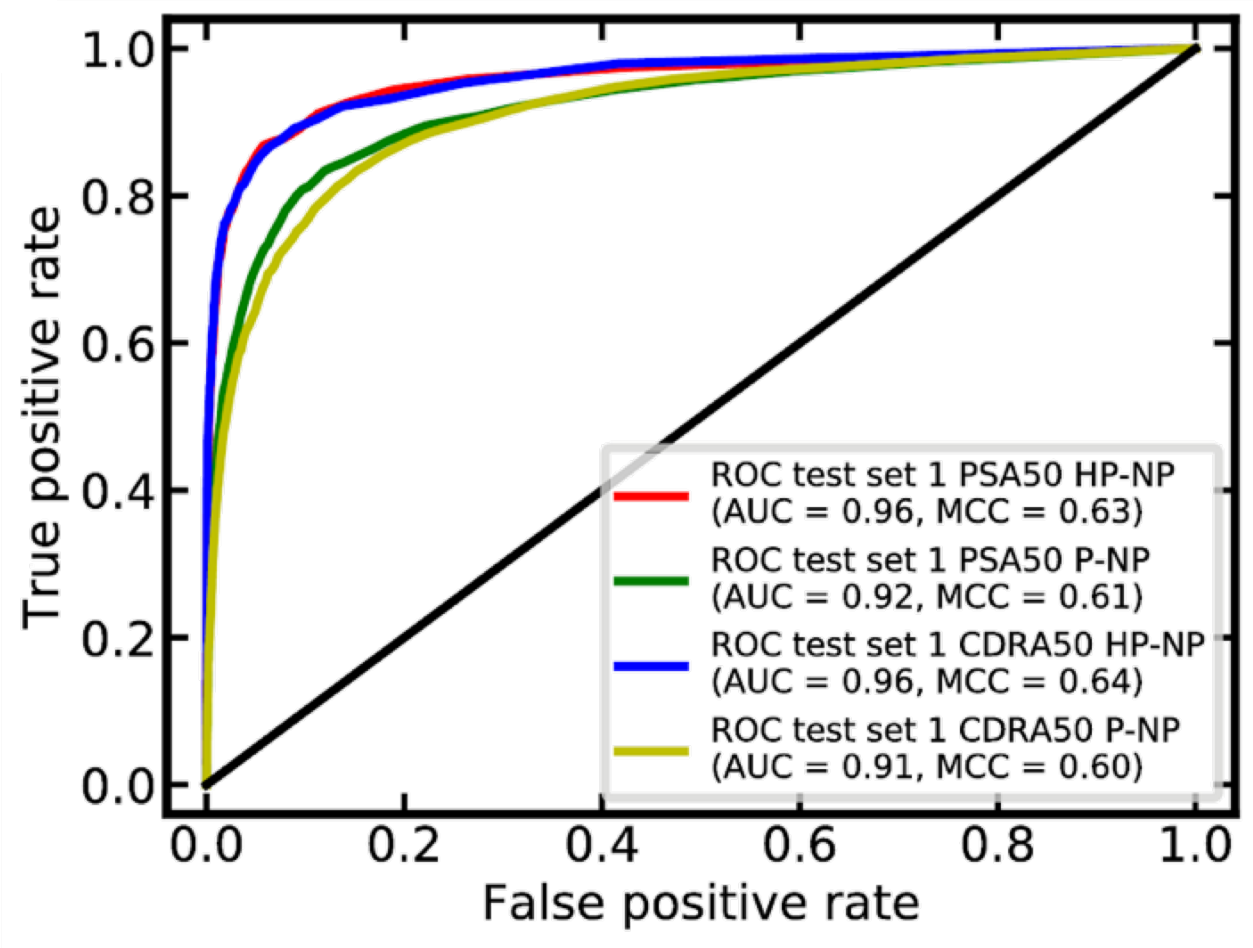
- **Frequent hitters:** compounds with a higher-than-expected activity rate recorded in historical screening data
  - Some compounds based on **PAINS** scaffolds
  - Some **aggregators**
  - **Reactive compounds** and others
  - **True promiscuous compounds** (sometimes related to privileged scaffolds)
- Aim:
  - Develop a **simple and robust tool for the identification of compounds for which extra caution should be exercised with positive assay readouts**
- Key components:
  - PubChem Bioassay data
  - Morgan2 fingerprints
  - Extremely randomized trees classifiers



Stork C. et al., *J Chem Inf Model* 2019, 59, 1030−1043.

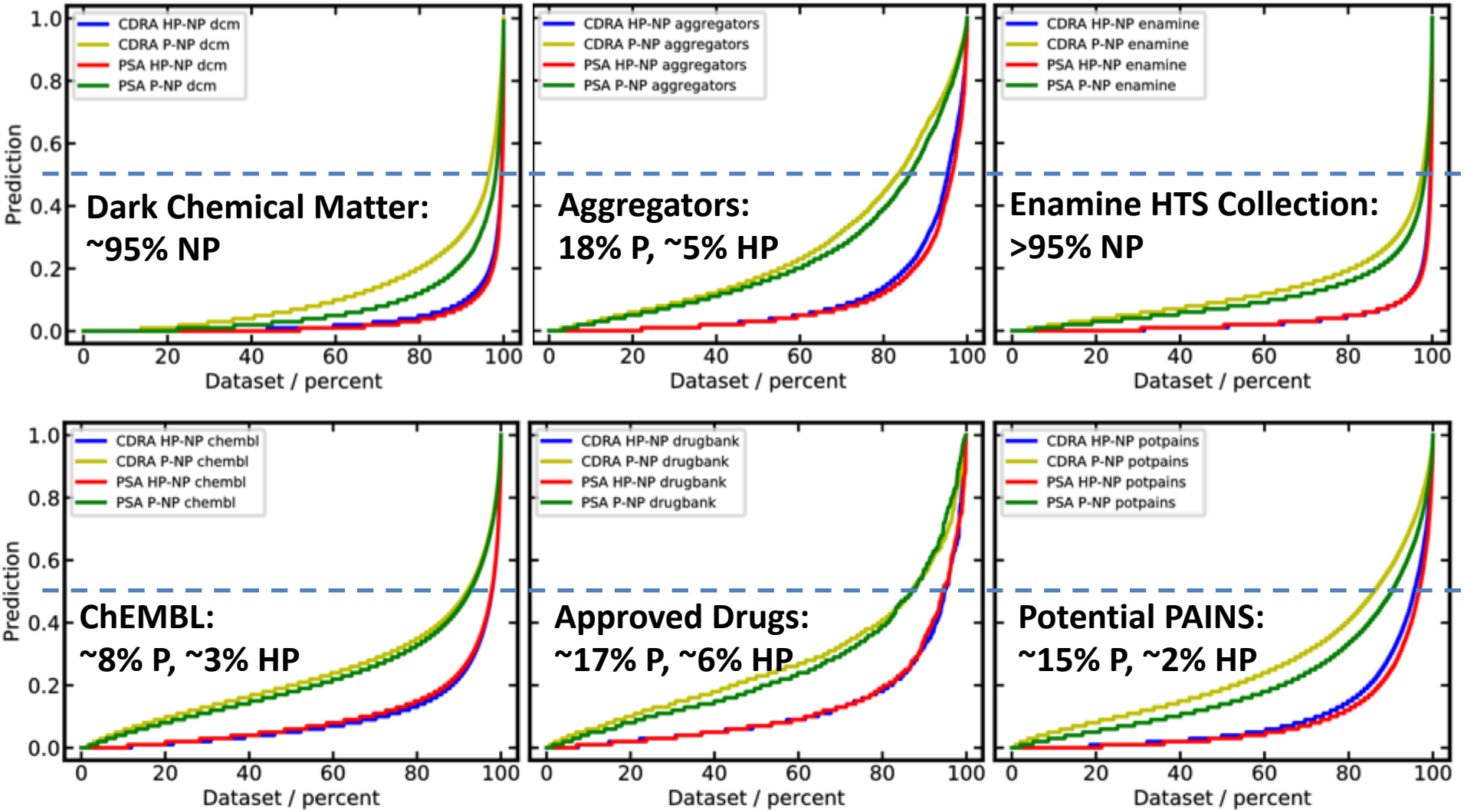
Hit Dexter 2.0: Model performance on external data

(10% of dataset selected by random split prior to any modelling)



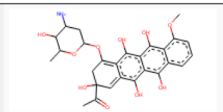
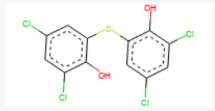
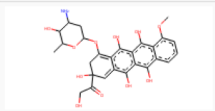
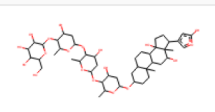
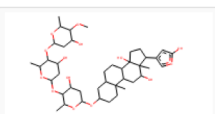
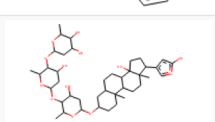
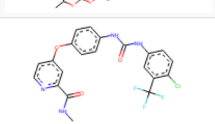
- PSA: primary screening assays
- CDRA: confirmatory dose-response assays
- HP: highly promiscuous
- P: promiscuous
- NP: non-promiscuous

Characterization of datasets with Hit Dexter 2.0



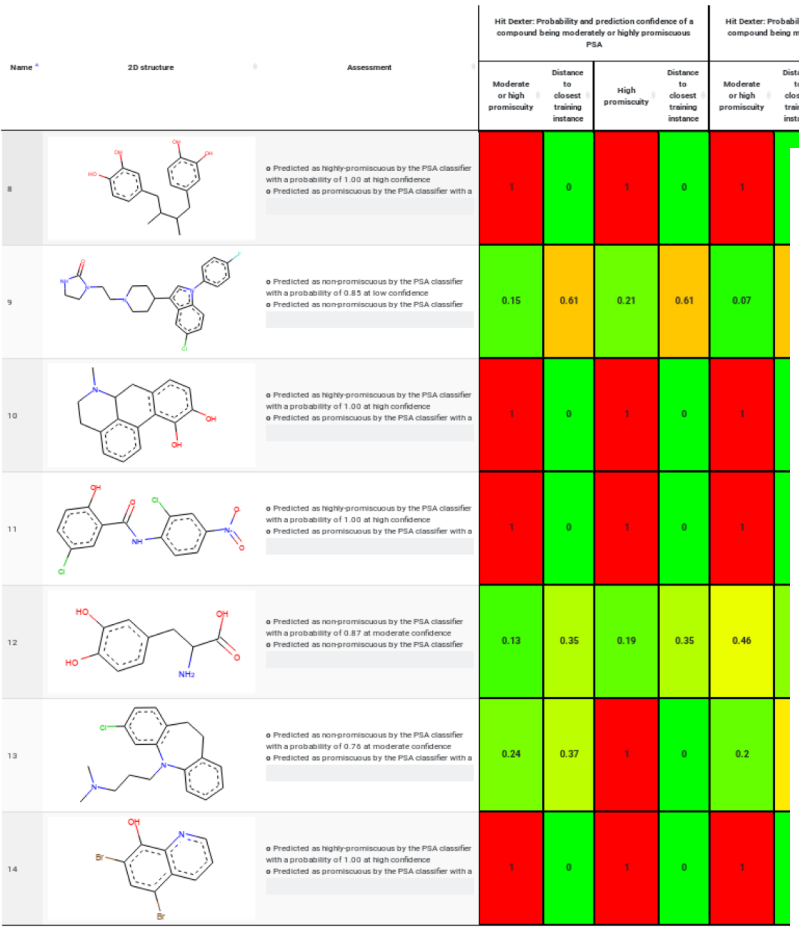
 Stork C. et al., *J Chem Inf Model* 2019, 59, 1030−1043

Case study on the 15 noisiest, approved drugs identified by GSK[1]

Name *	2D structure	Assessment	Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous PSA				Hit Dexter: Probability and prediction confidence of a compound being moderately or highly promiscuous CDRA				Similarity of a compound to known aggregators and dark chemical matter (DCM)		Number of undesired functional groups present in a compound							
			Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance	Moderate or high promiscuity	Distance to closest training instance	High promiscuity	Distance to closest training instance	Distance to closest aggregator	Distance to closest DCM	PAINS SMARTS (480 patterns)	BMS (180 patterns)	Dundee (105 patterns)	Glaxo (55 patterns)	Pfizer (57 patterns)	MLSMR (116 patterns)	SureChEMBL (166 patterns)	Chakravorty
0		o Predicted as highly-promiscuous by the PSA classifier with a probability of 1.00 at high confidence o Predicted as promiscuous by the PSA classifier with a	1	0	1	0	1	0	1	0	1	0.73								
1		o Predicted as non-promiscuous by the PSA classifier with a probability of 0.69 at low confidence o Predicted as non-promiscuous by the PSA classifier	0.31	0.55	0.48	0.55	1	0	1	0	1	0.66								
2		o Predicted as highly-promiscuous by the PSA classifier with a probability of 1.00 at high confidence o Predicted as promiscuous by the PSA classifier with a	1	0	1	0	1	0	1	0	1	0.71								
3		o Predicted as non-promiscuous by the PSA classifier with a probability of 0.64 at moderate confidence o Predicted as non-promiscuous by the PSA classifier	0.36	0.21	0.48	0.21	0.54	0.21	0.53	0.21	1	0.67								
4		o Predicted as highly-promiscuous by the PSA classifier with a probability of 0.93 at high confidence o Predicted as promiscuous by the PSA classifier with a	0.93	0.1	0.93	0.1	0.96	0.1	0.97	0.1	1	0.7								
5		o Predicted as highly-promiscuous by the PSA classifier with a probability of 0.63 at moderate confidence o Predicted as promiscuous by the PSA classifier with a	0.63	0.24	1	0	0.16	0.64	0.46	0.64	1	0.66								
6		o Predicted as highly-promiscuous by the PSA classifier with a probability of 1.00 at high confidence o Predicted as promiscuous by the PSA classifier with a	1	0	1	0	1	0	1	0	1	0.62								
7		o Predicted as highly-promiscuous by the PSA classifier with a probability of 1.00 at high confidence o Predicted as promiscuous by the PSA classifier with a	1	0	1	0	1	0	1	0	0	0.52								

[1]Chakravorty et al., *SLAS Discovery* 2018, 23, 532–544

Case study on the 15 noisiest, approved drugs identified by GSK[1]



Assessment (examples):

- Predicted as highly-promiscuous by the PSA classifier with a probability of 1.00 at high confidence
- Predicted as non-promiscuous by the PSA classifier with a probability of 0.33 at low confidence
- Predicted as highly-promiscuous by the CDRA classifier with a probability of 1.00 at high confidence
- Predicted as non-promiscuous by the CDRA classifier with a probability of 0.83 at moderate confidence
- Possibly an aggregator

[1]Chakravorty et al., *SLAS Discovery* 2018, 23, 532–544

Why do we love RDKit and for what tasks are we using it?

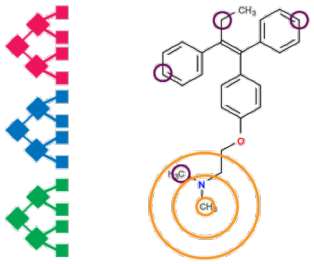
- Open source
- Robust
- Python
- Large, active developer and user community
- KNIME nodes

- Parsing of molecules from different formats
- Molecular structure processing
- Molecular descriptors, fingerprints
- Molecular similarity (fingerprints, Tanimoto coefficient)
- Visualization: similarity maps

NERDD

New E-Resource for Drug Discovery

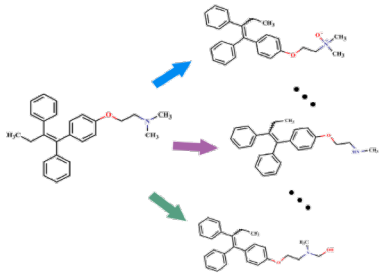
Sites of Metabolism



FAME 3

Regioselectivity prediction for phase 1 and phase 2 metabolism

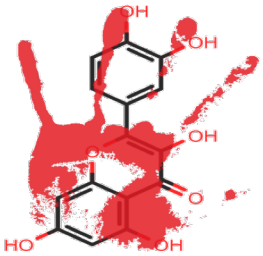
Metabolite Structures



GLORY

Metabolite structure prediction for cytochrome P450 metabolism

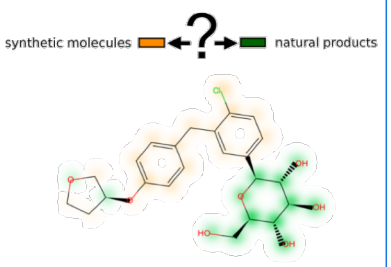
Frequent Hitters



Hit Dexter 2.0

Prediction of frequent hitters

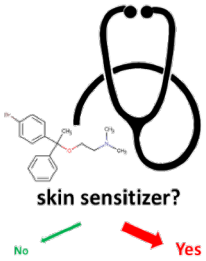
Natural Product-Likeness



NP-Scout

Identification and visualization of natural product-likeness

Skin Sensitization



Skin Doctor

Prediction of skin sensitization potential