TeachOpenCADD

Open Source Teaching Platform for Computer-Aided Drug Design

RDKit UGM 2019

Dominique Sydow
Jaime Rodríguez-Guerra
VolkamerLab@Charité

Open resources

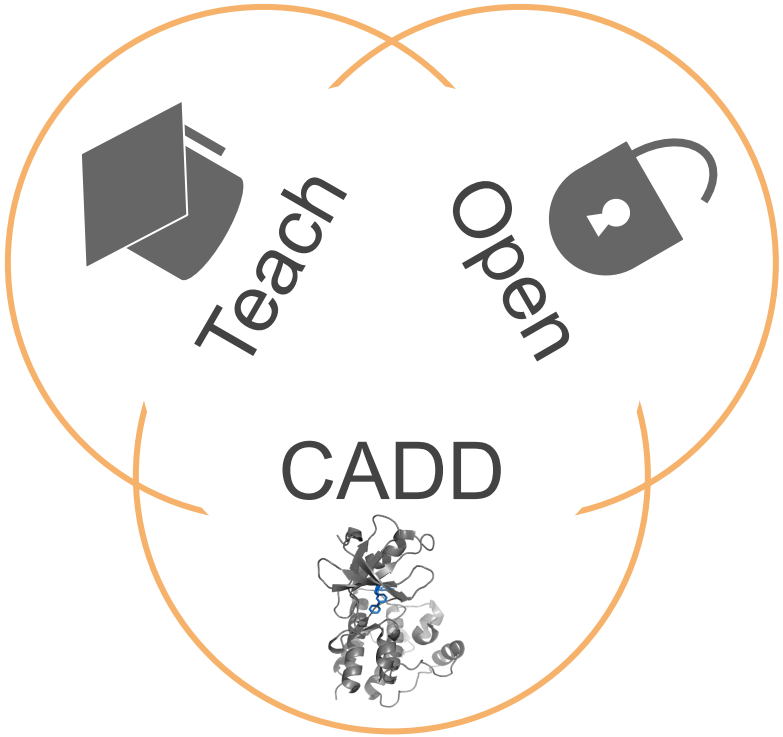
… are more and more available and used

- Databases
- Software/packages
- Code version control and deposition platforms


… allow for **transparent**, **reproducible** and **reusable** research

TeachOpenCADD objectives

- Collect useful open resources
- Show usage in drug design
- Provide starting point
  - Teach common tasks in drug design
  - Address both coders or drug designers



Sydow D, Morger A, Driller M, Volkamer A. TeachOpenCADD: A teaching platform for computer-aided drug design using open source packages and data. *JChem,* **2019** (https://doi.org/10.1186/s13321-019-0351-x).
Sydow D and Wichmann M, Rodríguez-Guerra J, Goldmann D, Landrum G, Volkamer A. TeachOpenCADD-KNIME: A teaching platform for computer-aided drug design using KNIME workflows. *JCIM*, **2019** (http://dx.doi.org/10.1021/acs.jcim.9b00662).

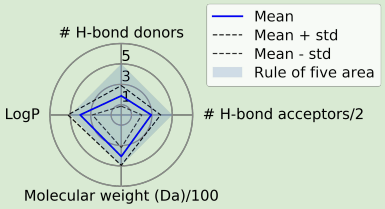
TeachOpenCADD: Topic overview

1 Data acquisition

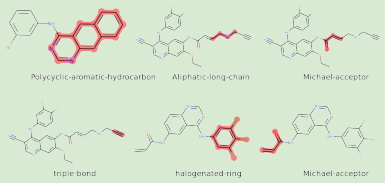
ChEMBL

	molecule_chembl_id	units	value	molecule_structures
0	CHEMBL68920	nM	41	Cc1cc(C)c(C=C2/C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c23)[nH]1
1	CHEMBL69960	nM	170	Cc1cc[nH]c1C=C2/C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c23)C(=O)N5CCOCC5
2	CHEMBL137635	nM	9300	CN(c1ccccc1)c2ncnc3ccc(NN=NCc4ccccn4)cc23
3	CHEMBL306988	nM	500000	CC(=C(C#N)C#N)c1ccc(NC(=O)CCC(=O)O)cc1
4	CHEMBL66879	nM	3e+06	OC(=O)/C=C/c1ccc(O)cc1

2 Molecular filtering: ADME criteria



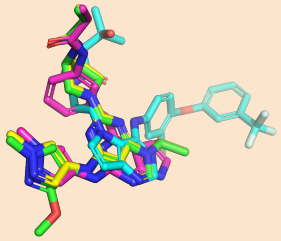
3 Molecular filtering: Unwanted substruct.



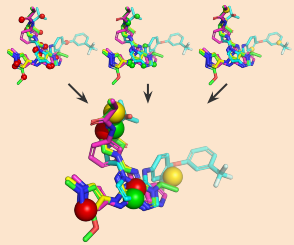
Query target (EGFR)

8 Data acquisition

RCSB PDB PROTEIN DATA BANK

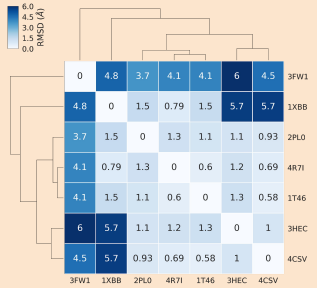


9 Ligand-based ensemble pharmacophores

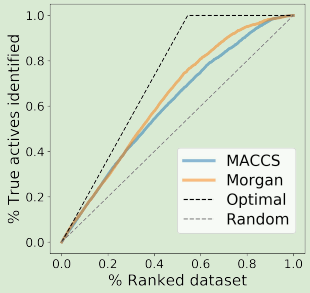


Query ligand (Imatinib)

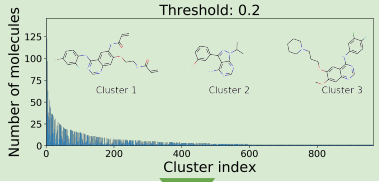
10 Binding site comparison



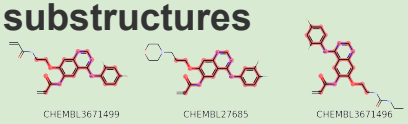
4 Ligand-based screening: Compound similarity



5 Compound clustering



6 Maximum common substructures



7 Ligand-based screening: Machine learning

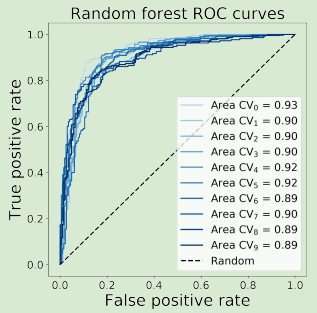


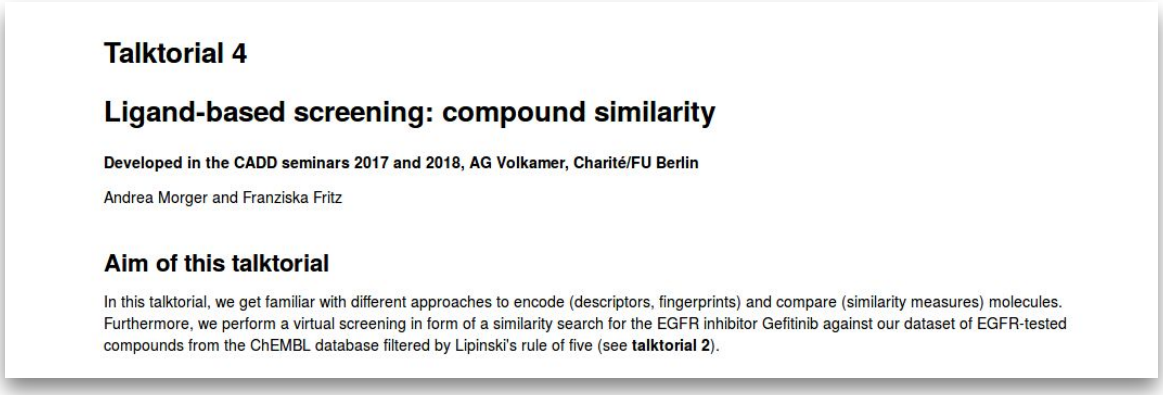
Figure adapted from:
https://doi.org/10.1186/s13321-019-0351-x

TeachOpenCADD: Material format

- **Coding**-based
- Interactive computing across many programming languages
  - TeachOpenCADD Jupyter Notebooks: **Python**
- All-in-one
  - Narrative text
  - Live code
  - Visualizations (also interactively using `ipywidgets`)

- **GUI**-based
- Interactive workflows = connected **nodes**
  - Small pre-implemented code units
  - Standardized functionalities
  - Configurable with individual settings
- All-in-one
  - Create data workflows
  - Execute selected steps
  - Check intermediate and final results, models and views

TeachOpenCADD: Jupyter Notebooks



Talktorial 4

Ligand-based screening: compound similarity

Developed in the CADD seminars 2017 and 2018, AG Volkamer, Charité/FU Berlin

Andrea Morger and Franziska Fritz

Aim of this talktorial

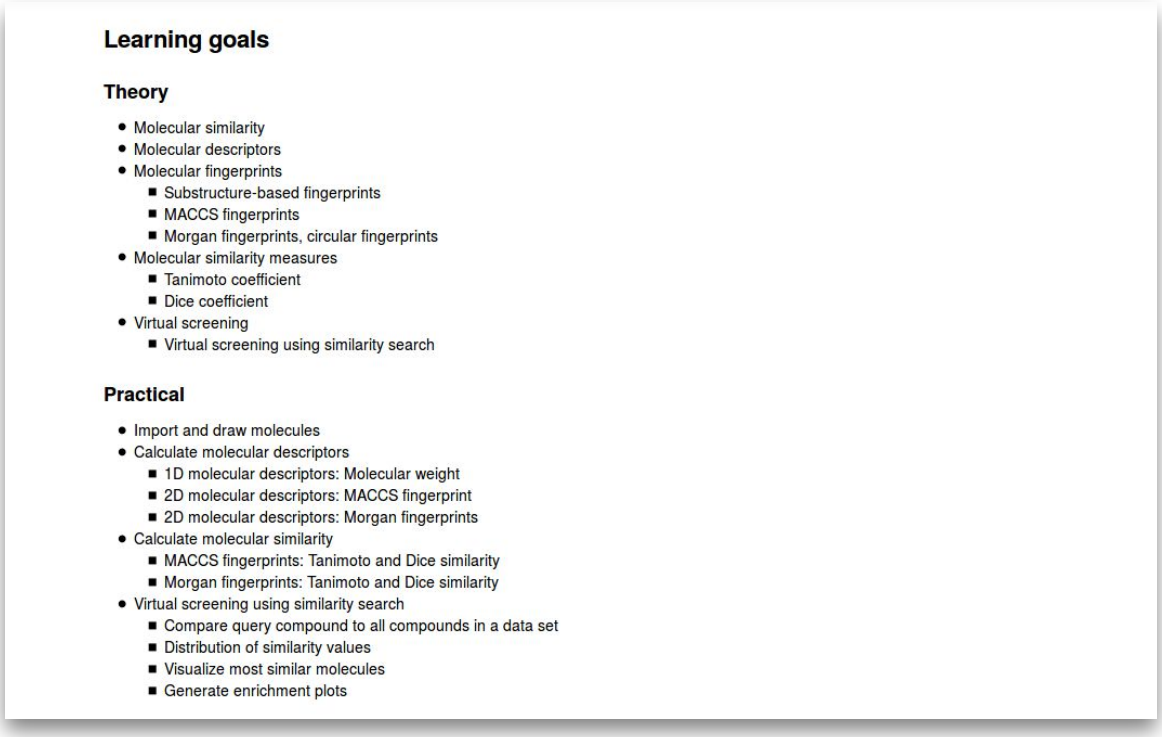
In this talktorial, we get familiar with different approaches to encode (descriptors, fingerprints) and compare (similarity measures) molecules. Furthermore, we perform a virtual screening in form of a similarity search for the EGFR inhibitor Gefitinib against our dataset of EGFR-tested compounds from the ChEMBL database filtered by Lipinski's rule of five (see **talktorial 2**).

Talktorial (talk + tutorial) sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

https://github.com/volkamerlab/teachopencadd

TeachOpenCADD: Jupyter Notebooks



Learning goals

Theory

- Molecular similarity
- Molecular descriptors
- Molecular fingerprints
  - Substructure-based fingerprints
  - MACCS fingerprints
  - Morgan fingerprints, circular fingerprints
- Molecular similarity measures
  - Tanimoto coefficient
  - Dice coefficient
- Virtual screening
  - Virtual screening using similarity search

Practical

- Import and draw molecules
- Calculate molecular descriptors
  - 1D molecular descriptors: Molecular weight
  - 2D molecular descriptors: MACCS fingerprint
  - 2D molecular descriptors: Morgan fingerprints
- Calculate molecular similarity
  - MACCS fingerprints: Tanimoto and Dice similarity
  - Morgan fingerprints: Tanimoto and Dice similarity
- Virtual screening using similarity search
  - Compare query compound to all compounds in a data set
  - Distribution of similarity values
  - Visualize most similar molecules
  - Generate enrichment plots

Talktorial
(talk + tutorial)
sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

https://github.com/volkamerlab/teachopencadd

TeachOpenCADD: Jupyter Notebooks

References

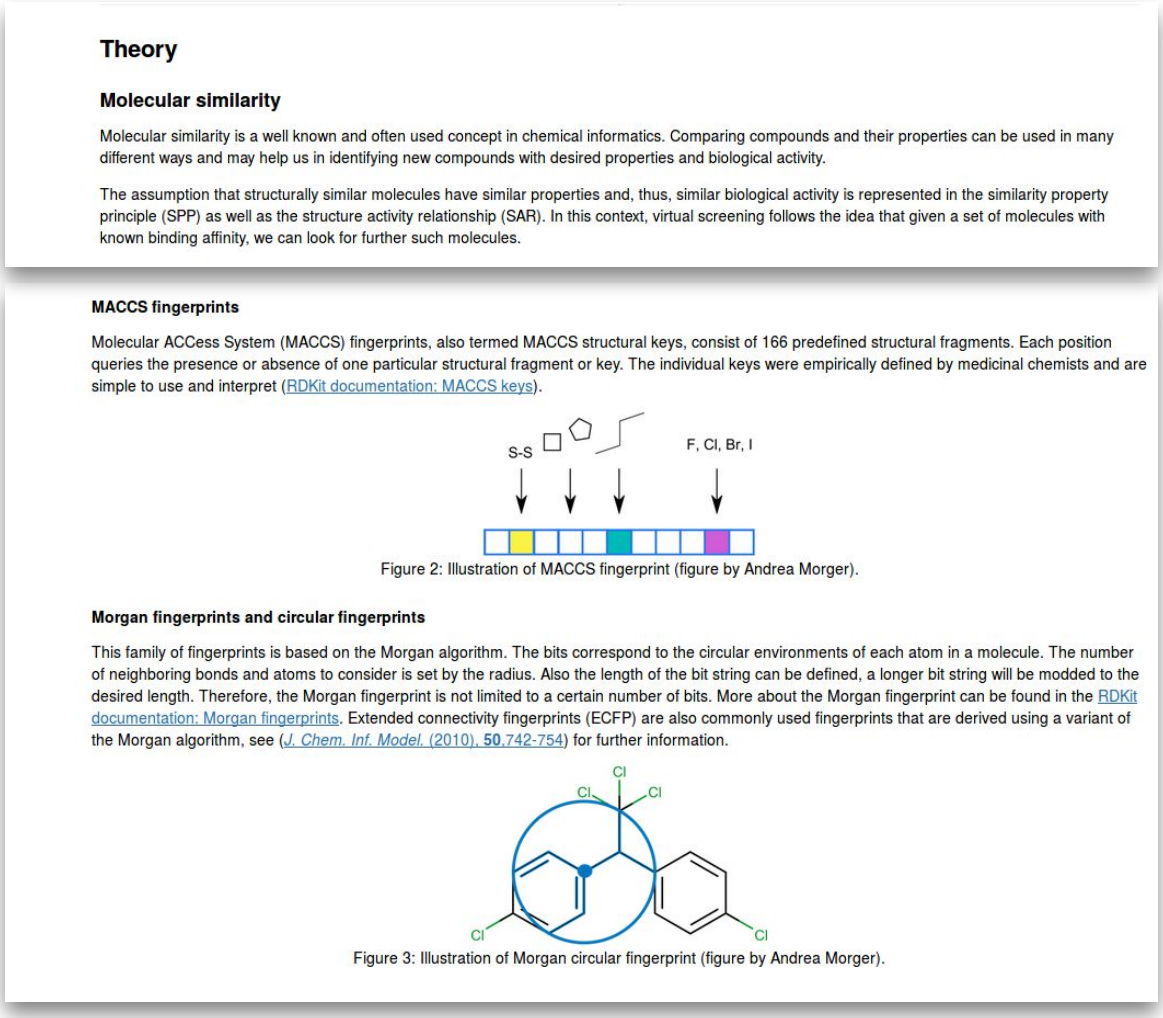
- Review on "Molecular similarity in medicinal chemistry" (*J. Med. Chem.* (2014), **57**, 3186-3204)
- Morgan fingerprints with RDKit (RDKit tutorial on Morgan fingerprints)
- ECFP - extended-connectivity fingerprints (*J. Chem. Inf. Model.* (2010), **50**,742-754)
- Chemical space (*ACS Chem. Neurosci.* (2012), **19**, 649-57)
- List of molecular descriptors in RDKit (RDKit documentation: Descriptors)
- List of fingerprints in RDKit (RDKit documentation: Fingerprints)
- Enrichment plots (Applied Chemoinformatics, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, (2018), **1**, 313-31)

Talktorial (talk + tutorial) sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

https://github.com/volkamerlab/teachopencadd

TeachOpenCADD: Jupyter Notebooks



Theory

Molecular similarity

Molecular similarity is a well known and often used concept in chemical informatics. Comparing compounds and their properties can be used in many different ways and may help us in identifying new compounds with desired properties and biological activity.

The assumption that structurally similar molecules have similar properties and, thus, similar biological activity is represented in the similarity property principle (SPP) as well as the structure activity relationship (SAR). In this context, virtual screening follows the idea that given a set of molecules with known binding affinity, we can look for further such molecules.

MACCS fingerprints

Molecular ACCess System (MACCS) fingerprints, also termed MACCS structural keys, consist of 166 predefined structural fragments. Each position queries the presence or absence of one particular structural fragment or key. The individual keys were empirically defined by medicinal chemists and are simple to use and interpret (RDKit documentation: MACCS keys).

Figure 2: Illustration of MACCS fingerprint (figure by Andrea Morger).

Morgan fingerprints and circular fingerprints

This family of fingerprints is based on the Morgan algorithm. The bits correspond to the circular environments of each atom in a molecule. The number of neighboring bonds and atoms to consider is set by the radius. Also the length of the bit string can be defined, a longer bit string will be modded to the desired length. Therefore, the Morgan fingerprint is not limited to a certain number of bits. More about the Morgan fingerprint can be found in the RDKit documentation: Morgan fingerprints. Extended connectivity fingerprints (ECFP) are also commonly used fingerprints that are derived using a variant of the Morgan algorithm, see (*J. Chem. Inf. Model.* (2010), **50**,742-754) for further information.

Figure 3: Illustration of Morgan circular fingerprint (figure by Andrea Morger).

Talktorial (talk + tutorial) sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

TeachOpenCADD: Jupyter Notebooks

MACCS fingerprints: Tanimoto similarity

In [18]:
```
# Add similarity scores to DataFrame
sim_df['tanimoto_MACCS'] = DataStructs.BulkTanimotoSimilarity(maccs_fp1,maccs_fp_list)
```

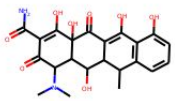
In [19]:
```
# DataFrame sorted by Tanimoto similarity of MACCS fingerprints
sim_df_sorted_t_ma = sim_df.copy()
sim_df_sorted_t_ma.sort_values(['tanimoto_MACCS'], ascending=False, inplace=True)
sim_df_sorted_t_ma
```

Out[19]:

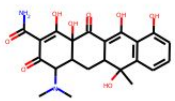
	name	smiles	tanimoto_MACCS
0	Doxycycline	CC1C2C(C3C(C(=O)C(=C(C3(C(=O)C2=C(C4=C1C=CC=C4O)O)O)O)C(=O)N)N(C)C)O	1.000000
6	Tetracycline	CC1(C2CC3C(C(=O)C(=C(C3(C(=O)C2=C(C4=C1C=CC=C4O)O)O)O)C(=O)N)N(C)C)O	0.928571
1	Amoxicilline	CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=C(C=C3)O)N)C(=O)O)C	0.590909
7	Hemi-cycline D	CC1C(CC(=O)C2=C1C=CC=C2O)C(=O)O	0.403509
2	Furosemide	C1=COC(=C1)CNC2=CC(=C(C=C2C(=O)O)S(=O)(=O)N)Cl	0.321839
4	Hydrochlorothiazide	C1NC2=CC(=C(C=C2S(=O)(=O)N1)S(=O)(=O)N)Cl	0.306818
5	Isotretinoine	CC1=C(C(CCC1)(C)C)C=CC(=CC=CC(=CC(=O)O)C)C	0.288136
3	Glycol dilaurate	CCCCCCCCCCCC(=O)OCCOC(=O)CCCCCCCCCCC	0.149254

In [20]:
```
# Draw molecules ranked by Tanimoto similarity of MACCS fingerprints
draw_ranked_molecules(sim_df_sorted_t_ma, "tanimoto_MACCS")
```

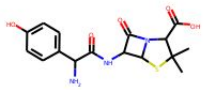
Out[20]:



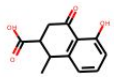
Query: Doxycycline (1.0)



#1: Tetracycline (0.93)



#2: Amoxicilline (0.59)

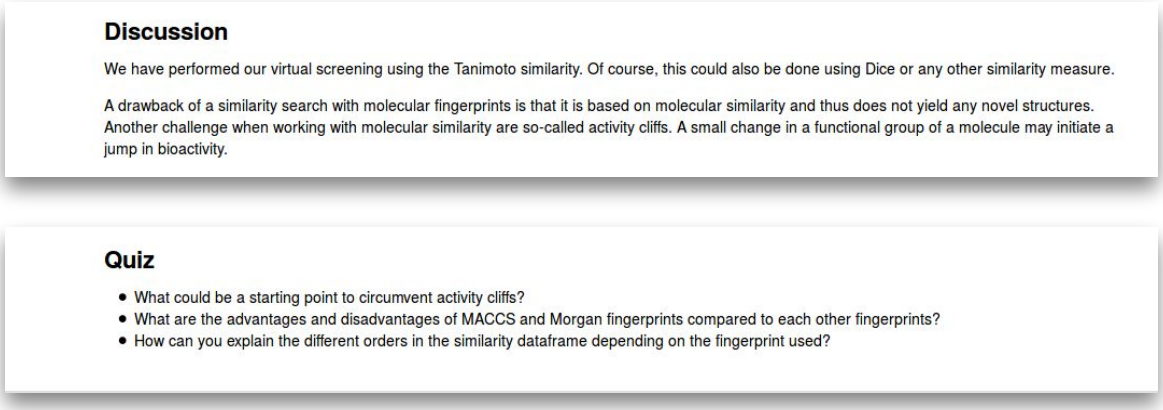


#3: Hemi-cycline D (0.4)

Talktorial (talk + tutorial) sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

TeachOpenCADD: Jupyter Notebooks



Discussion

We have performed our virtual screening using the Tanimoto similarity. Of course, this could also be done using Dice or any other similarity measure.

A drawback of a similarity search with molecular fingerprints is that it is based on molecular similarity and thus does not yield any novel structures. Another challenge when working with molecular similarity are so-called activity cliffs. A small change in a functional group of a molecule may initiate a jump in bioactivity.

Quiz

- What could be a starting point to circumvent activity cliffs?
- What are the advantages and disadvantages of MACCS and Morgan fingerprints compared to each other fingerprints?
- How can you explain the different orders in the similarity dataframe depending on the fingerprint used?

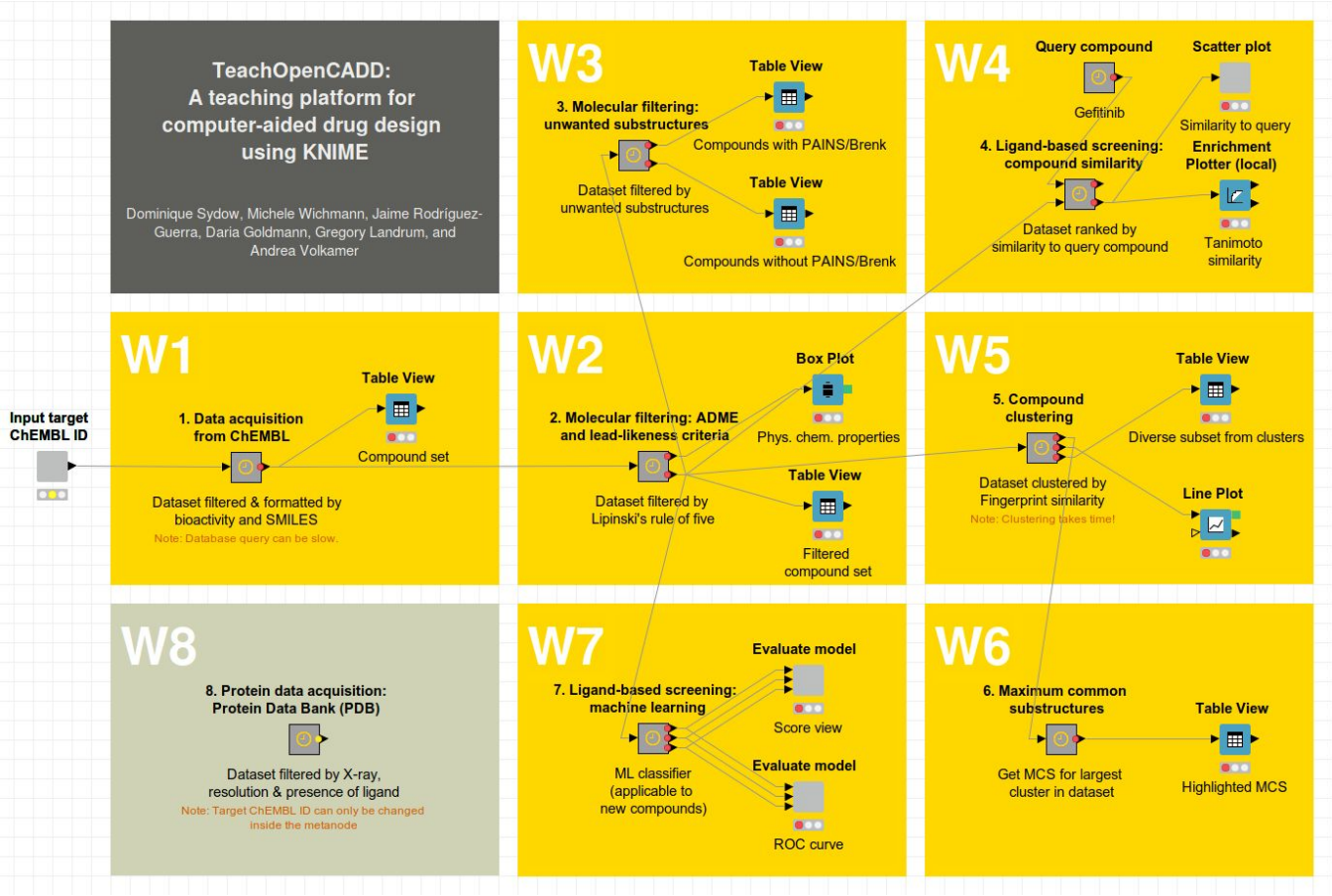
Talktorial (talk + tutorial) sections

- Aim of this talktorial
- Learning goals
- References
- Theory
- Practical
- Discussion
- Quiz

https://github.com/volkamerlab/teachopencadd

TeachOpenCADD: KNIME workflows

Topic overview

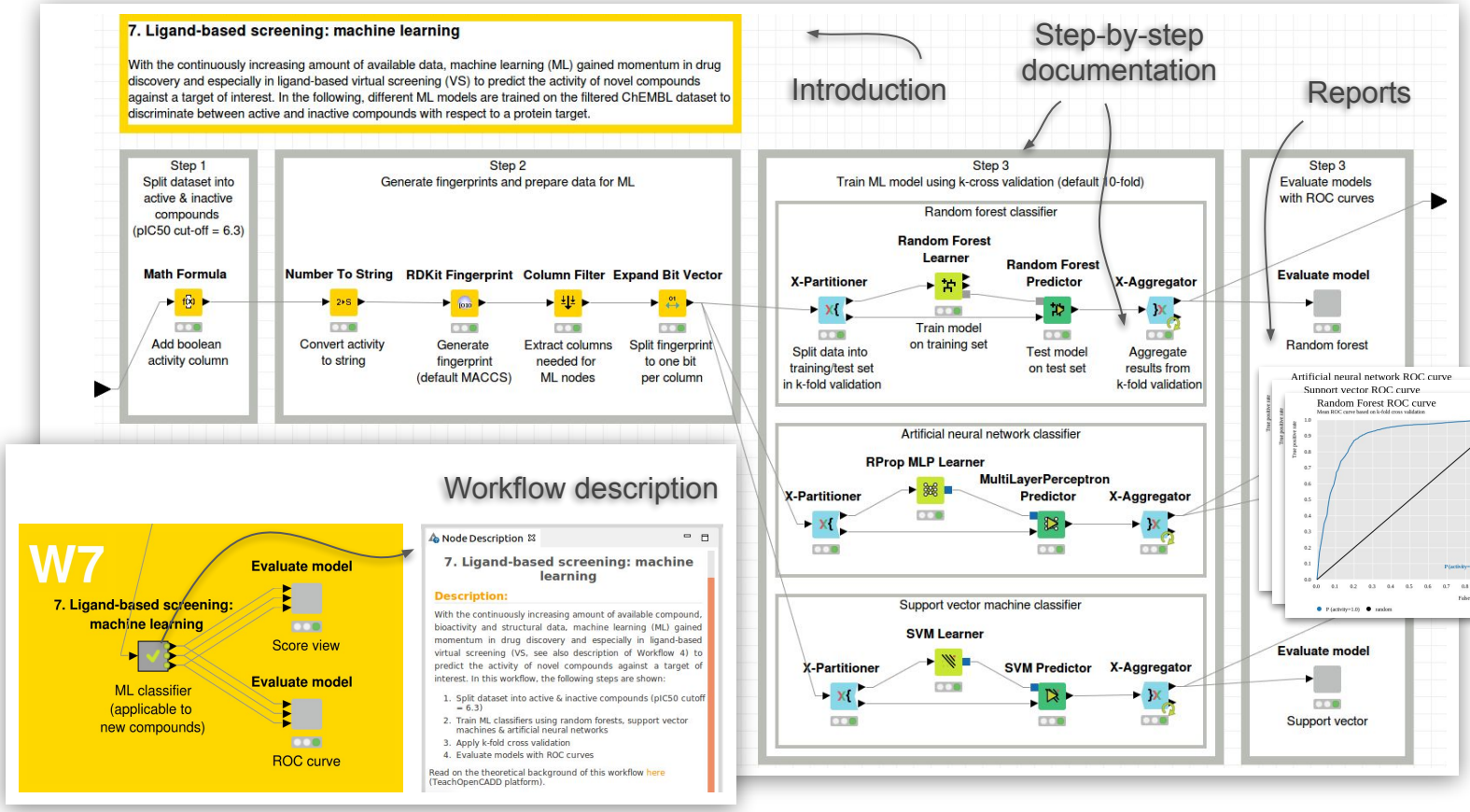


https://hub.knime.com/volkamerlab/space/TeachOpenCADD/

Figure taken from:
http://dx.doi.org/10.1021/acs.jcim.9b00662

TeachOpenCADD: KNIME workflows

Composition of workflow W7 (Machine learning)

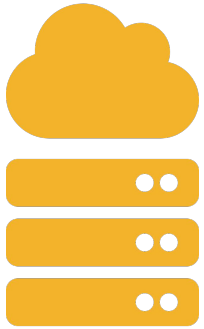


https://hub.knime.com/volkamerlab/space/TeachOpenCADD/

Figure taken from:
http://dx.doi.org/10.1021/acs.jcim.9b00662

TeachOpenCADD: New topics

New Jupyter Notebook by Jaime Rodríguez-Guerra



- Topic 11: Structure-based CADD using online APIs/servers
  - Querying KLIFS & PubChem for potential kinase inhibitors
  - Docking the candidates against the target
  - Assessing the results and comparing against known data
- https://github.com/volkamerlab/TeachOpenCADD/tree/master/talktorials/11_online_apis

Join us at the hackathon

- Work on new topics: contribute.volkamerlab.org
- Update/extend existing topics
- Learn by using the TeachOpenCADD material

Acknowledgements

Volkamer Lab
Andrea Volkamer
Dominique Sydow
Andrea Morger
Maximilian Driller
Michele Wichmann
Jaime Rodríguez-Guerra
Talia Kimber

Student CADD course 2017 & 2018
(Freie Universität Berlin / Charité)

Greg Landrum (RDKit, KNIME)
Daria Goldmann (KNIME)

John Chodera (MSKCC)