TeachOpenCADD

Teaching Platform for Computer-Aided Drug Design Using Open Source Packages and Data

Dominique Sydow, Maximilian Driller, Andrea Morger, Andrea Volkamer

In silico Toxicology, Institute for Physiology, Universitätsmedizin Berlin, Virchowweg 6, 10117 Berlin

Introduction

Open source programming packages for cheminformatics and structural bioinformatics are powerful tools to build modular, reproducible, and easy-to-share pipelines for computer-aided drug design (CADD). While documentation for such tools is available, only few freely accessible examples teach underlying concepts focused on CADD usage, such as the TDT initiative [1], addressing especially users new to the field. Here, we present TeachOpenCADD, a CADD teaching platform developed from students for students, using ChEMBL [6] and the PDB [7] as well as python tools such as RDKit [2], PyPDB [3], and PyMol [4]. Interactive Jupyter notebooks [5] were developed for central topics, integrating theoretical background and practical code, and are freely available on GitHub: https://github.com/volkamerlab/TeachOpenCADD.

Topics

Illustrated with the example of the epidermal growth factor receptor (EGFR) kinase, we discuss how to acquire data from ChEMBL (T1), filter compounds for drug-likeness (T2), and identify unwanted substructures (T3). Furthermore, we introduce measures for compound similarity, applied for virtual screening (VS) of Gefitinib based on a similarity search (T4) and for compound clustering (T5), including the use of maximum common substructures (MCS) (T6). We also employ machine learning (ML) approaches to build models for predicting active compounds (T7). Lastly, structures are fetched from the PDB (T8), used to generate ligand-based ensemble pharmacophores (T9) and to conduct RMSD-based binding site comparison of Imatinib-binding proteins for off-target prediction (T10).

T1. Data acquisition from ChEMBL

- Get target ChEMBL ID(s) from UniProt ID

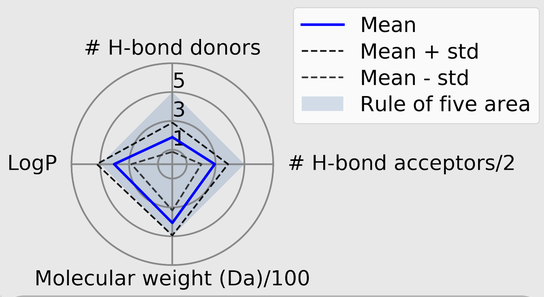
ChEMBL dataset for EGFR
6641 compounds

- Get bioactivity (IC_{50}) & compound (SMILES) data from ChEMBL ID
- Filter out entries with missing values & duplicates (-1601)
- Filter for entries with IC_{50} in molar units (-69)
- Convert bioactivity to unit nM and pIC_{50}

Dataset filtered & formatted
by bioactivity & SMILES
4771 compounds

T2. Molecular filtering: ADME criteria

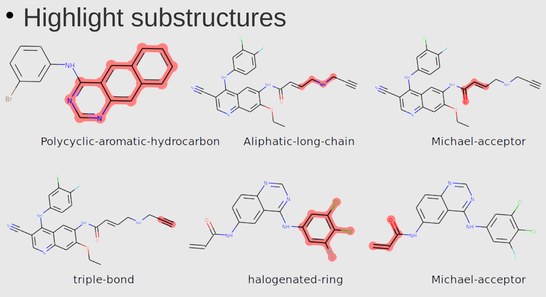
- Calculate 1D chemical properties
- Filter dataset by Lipinski's rule of five (-762)
- Visualize dataset properties with radar plot



Dataset filtered
by Lipinski's rule of five
4009 compounds

T3. Molecular filtering: Unwanted substructures

- Filter out PAINS [8] (-321)
- Filter out unwanted substructures by Brenk *et al.* [9] (-2058)
- Highlight substructures



Dataset filtered
by unwanted substructures
1951 compounds

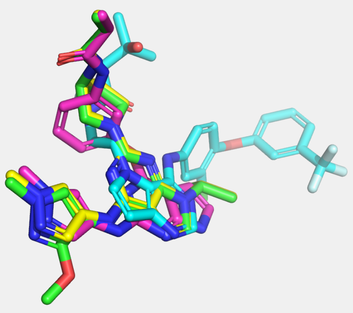
Query target
EGFR (UniProt ID: P00533)

T8. Data acquisition from PDB

- Get PDB IDs from UniProt ID of query target

PDB dataset
179 PDB structures

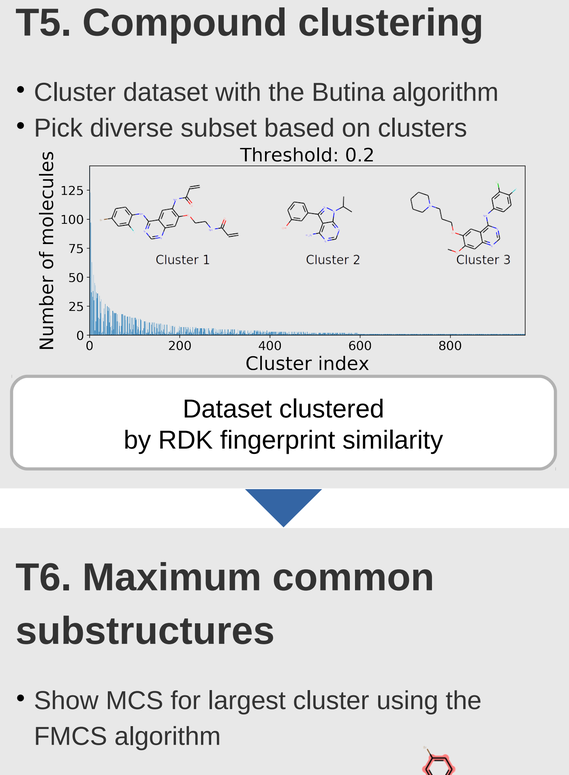
- Filter PDB structures: X-ray, ≤ 3Å resolution & presence of ligand
- Align top 4 protein-ligand structures (by resolution) & save ligands to file



Aligned ligand structures
4 ligands

T9. Ligand-based ensemble pharmacophores

- Extract pharmacophore features: H-bond donors/acceptors & hydrophobic
- Cluster (k-means) each feature type
- Generate ensemble pharmacophore



Ligand-based
ensemble pharmacophore

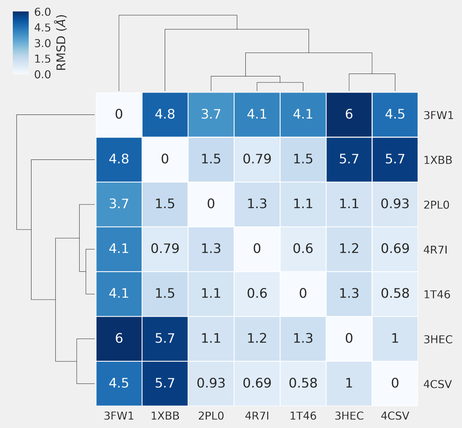
Query ligand
Imatinib (STI)

T10. Binding site comparison

- Get all protein-STI complexes from the PDB

PDB dataset
53 PDB structures

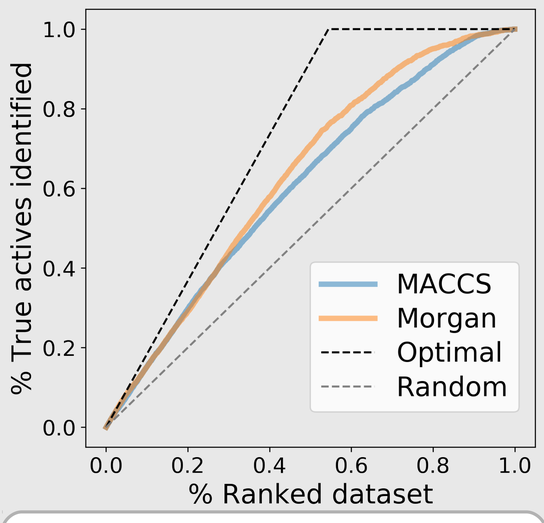
- Filter PDB data set: X-ray & ≤ 3Å resolution
- Align PDB structures & get pairwise RMSD for whole protein & binding site



Pairwise RMSD of aligned binding sites
8 protein structures

T4. Ligand-based screening: Compound similarity

- Calculate MACCS/Morgan fingerprints and Tanimoto/Dice similarity for dataset
- Split dataset into active & inactive compounds (pIC_{50} cutoff > 6.3)
- VS of query Gefitinib against dataset based on a similarity search
- Evaluate screening with enrichment plots



Dataset ranked
by similarity to query compound
Query: Gefitinib

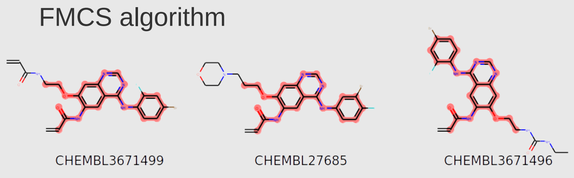
T5. Compound clustering

- Cluster dataset with the Butina algorithm
- Pick diverse subset based on clusters

Dataset clustered
by RDK fingerprint similarity

T6. Maximum common substructures

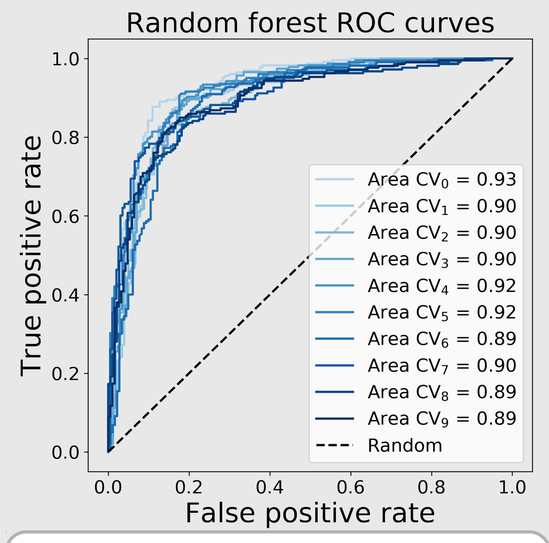
- Show MCS for largest cluster using the FMCS algorithm



Largest cluster of dataset
with MCS

T7. Ligand-based screening: Machine learning

- Split dataset into active & inactive compounds (pIC_{50} cutoff > 6.3)
- Train ML classifiers using random forest, support vector machines & artificial neural networks
- Apply K-fold cross validation (CV)
- Evaluate models with ROC curves



ML classifier
(applicable to new compounds)

Conclusion

With this platform, we aim at introducing interested researches and students to the ease and benefit of using open source cheminformatics tools. Topics will be continuously expanded and are open for contributions from the community. Beyond their teaching purpose, the notebooks can serve as starting point for users' project-directed modifications and extensions.

References

[1] S. Riniker *et al.*, F1000Research, 2017, 6, 1136. [2] http://www.rdkit.org. [3] W. Gilpin, Bioinformatics, 2016, 32, 156-60. [4] The PyMOL Molecular Graphics System, Version 1.8. [5] T. Kluyver *et al.*, IOS Press, 2016, 87-90. [6] A. Gaulton *et al.*, Nucleic Acid Res., 2017, 40, D1100-7. [7] H. Berman *et al.*, Nucleic Acid Res., 2000, 28, 235-42. [8] Brenk *et al.*, Chem. Med. Chem., 2008, 3, 435-44. [9] Baell *et al.*, J. Med. Chem., 2010, 53, 2719-40.