Lecture 14: Biophysical Modeling

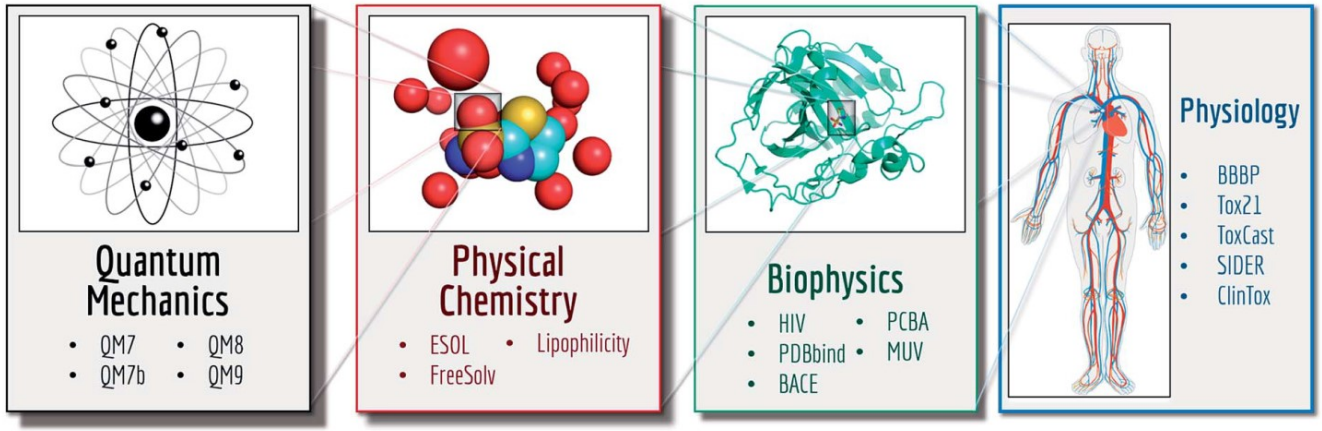
Credit

Slides are partially based on the material in Ramsundar, Bharath; Eastman, Peter; Walters, Patrick; Pande, Vijay. Deep Learning for the Life Sciences, Chapter 5.

https://medium.com/@stefan.schroedl/machine-learning-for-drug-discovery-in-a-nutshell-part-ii-24f90d5963d9

MoleculeNet

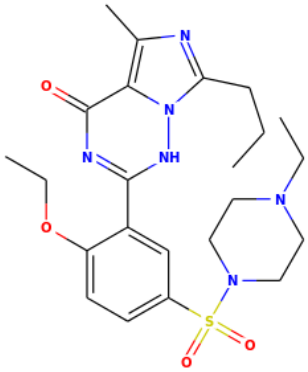
- A large set of dataset useful for molecular machine learning, included with DeepChem.
- Data from over 70,000 compounds
- Integrated with DeepChem package



Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

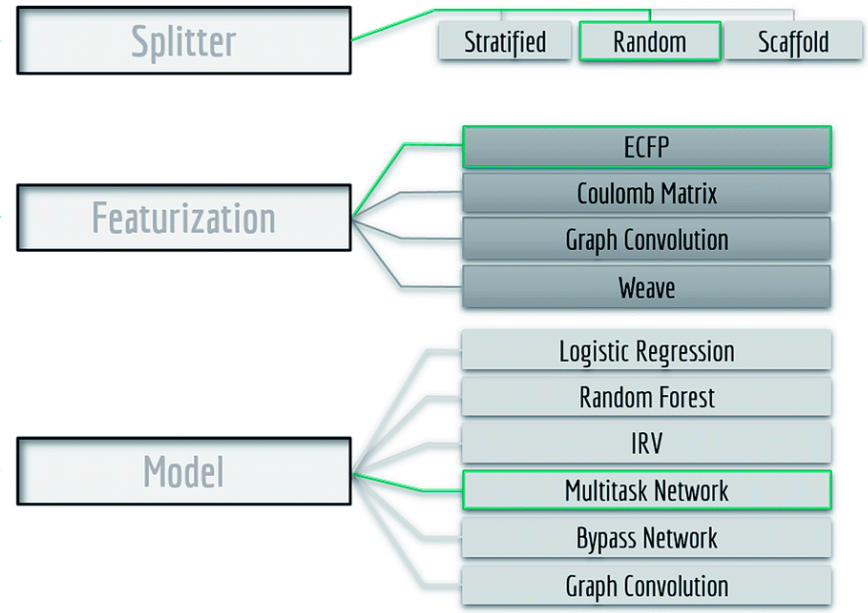
RDKit

- RdKit is a popular Cheminformatics Python library for computing features and molecular representation.



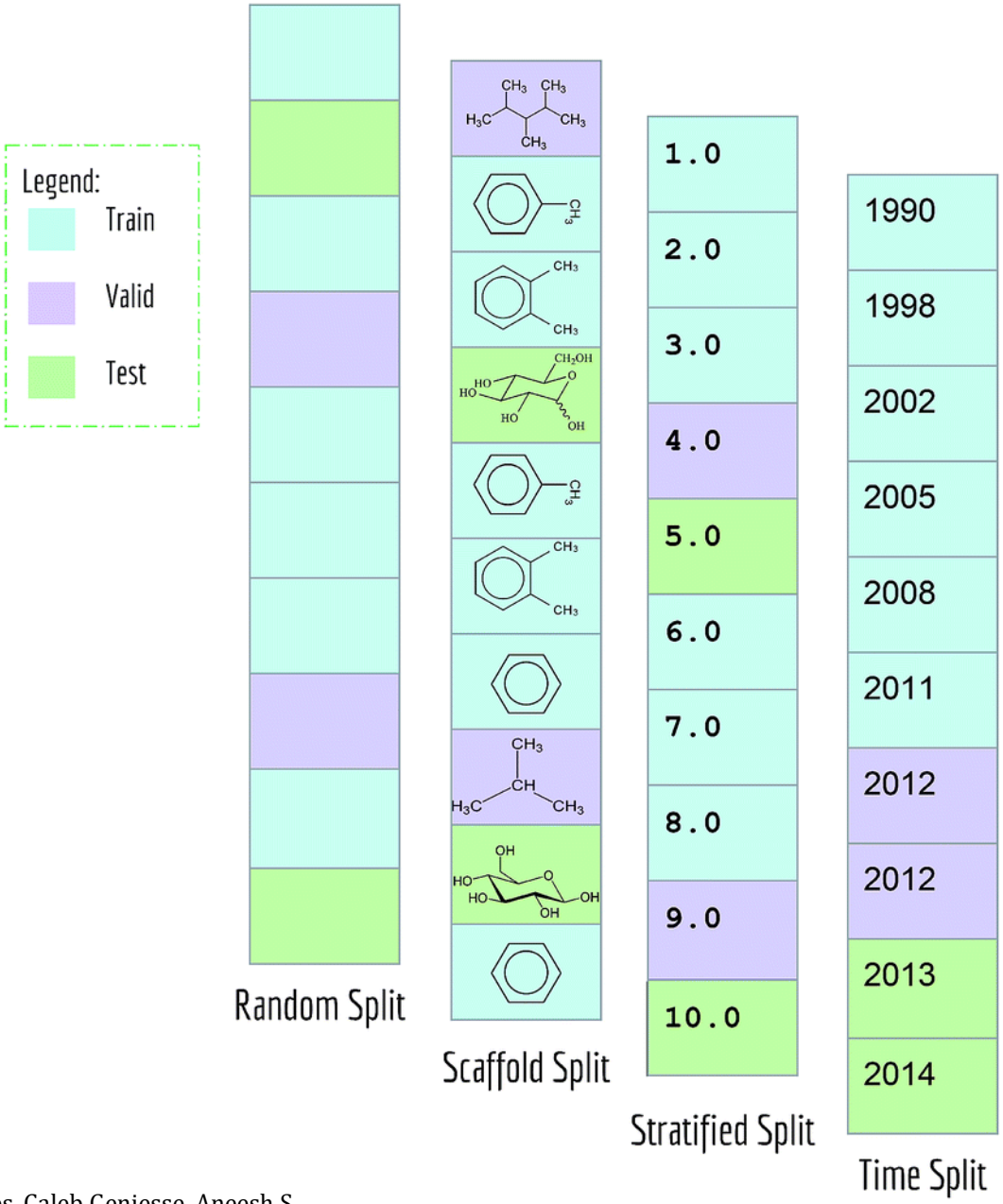
Coding using DeepChem

```python
"""
Script that trains multitask models on Tox21 dataset.
"""
from __future__ import print_function
from __future__ import division
from __future__ import unicode_literals

import numpy as np
import deepchem as dc

np.random.seed(123)

# Load Tox21 dataset
n_features = 1024
tox21_tasks, tox21_datasets, transformers = dc.molnet.load_tox21(split="random",
                                                                  featurizer="ECFP")

train_dataset, valid_dataset, test_dataset = tox21_datasets

# Define metric
metric = dc.metrics.Metric(dc.metrics.roc_auc_score, np.mean)
# Define model
model = dc.models.TensorflowMultiTaskClassifier(
    len(tox21_tasks), n_features, layer_sizes=[1000], dropouts=[.25],
    learning_rate=0.001, batch_size=50)

# Fit trained model
model.fit(train_dataset)
model.save()

print("Evaluating model")
train_scores = model.evaluate(train_dataset, [metric], transformers)
valid_scores = model.evaluate(valid_dataset, [metric], transformers)

print("Train scores")
print(train_scores)

print("Validation scores")
print(valid_scores)
```



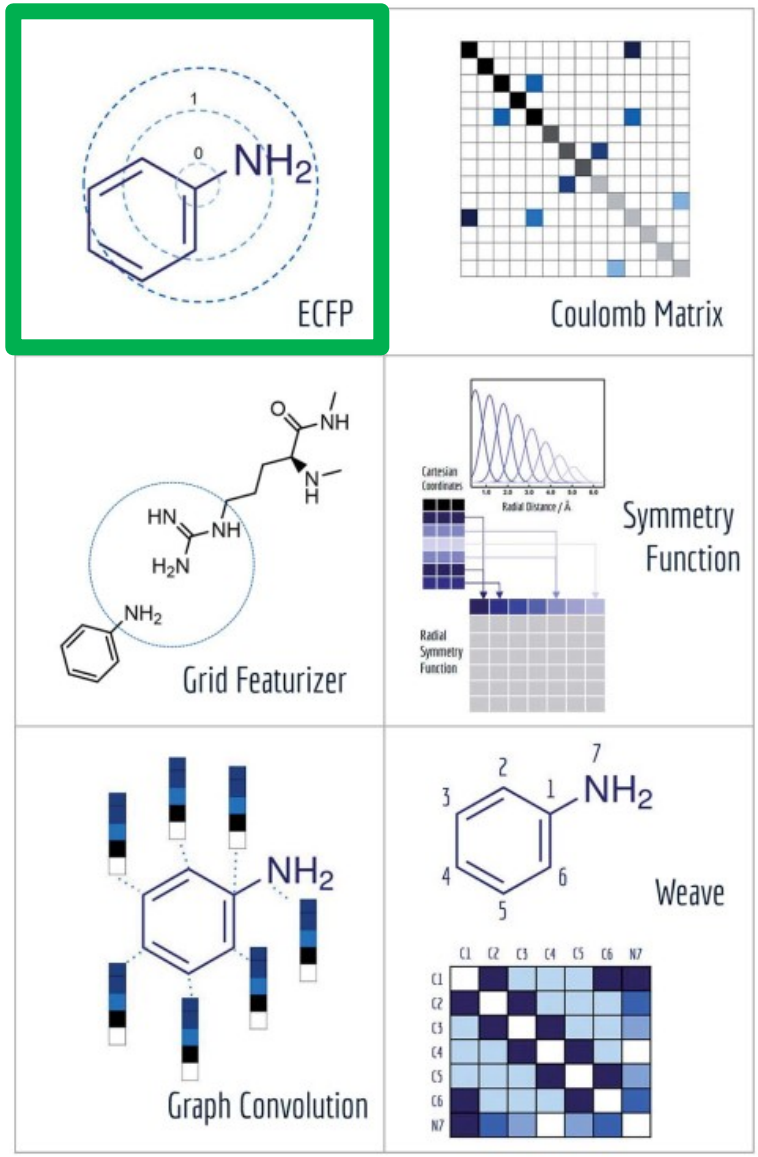
Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Splitting Method



Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

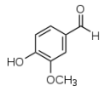
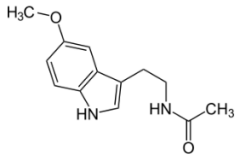
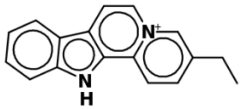
Featurization Methods in MoleculeNet

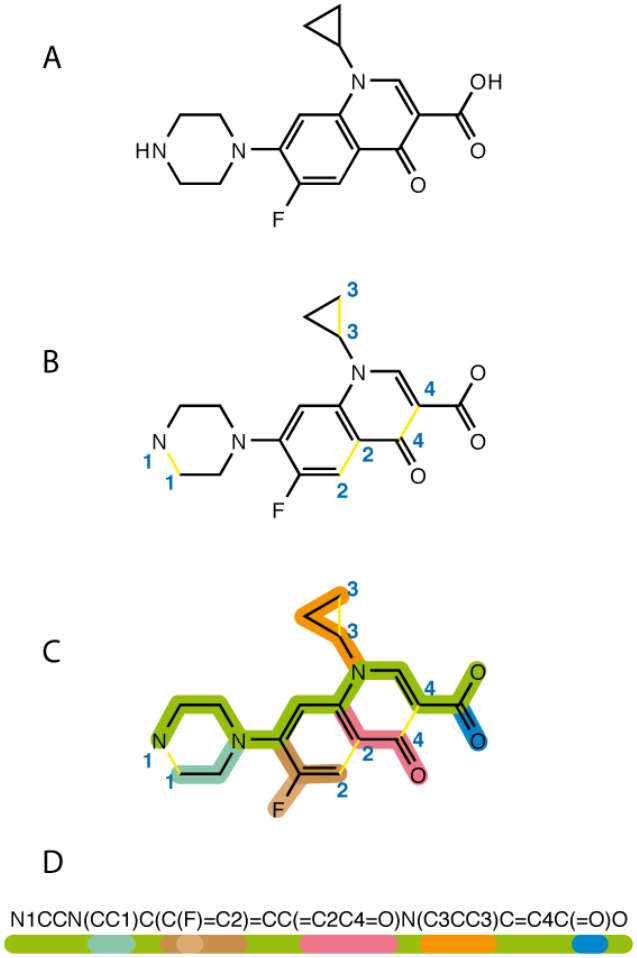


Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

SMILES

- The simplified molecular-input line-entry system (SMILES)
- A valence model of a molecule

Structure	SMILES Formula
N≡N	`N#N`
CH_3–N=C=O	`CN=C=O`
$Cu^{2+}SO_4^{2-}$	`[Cu+2].[O-]S(=O)(=O)[O-]`
	`O=Cc1ccc(O)c(OC)c1` `COCc1cc(C=O)ccc1O`
	`CC(=O)NCCC1=CNc2c1cc(OC)cc2` `CC(=O)NCCc1c[nH]c2ccc(OC)cc12`
	`CCc(c1)ccc2[n+]1ccc3c2[nH]c4c3cccc4` `CCc1c[n+]2ccc3c4ccccc4[nH]c3c2cc1`



SMILES

- It can helpful when representing a molecule, but not sufficient for machine learning tasks.
  - Lack electronic charges or topological features.
- A core challenge in molecular machine learning is to effectively encode molecules into fixed-length vectors.
- MoleculeNet contains implementation of six featurization methods.

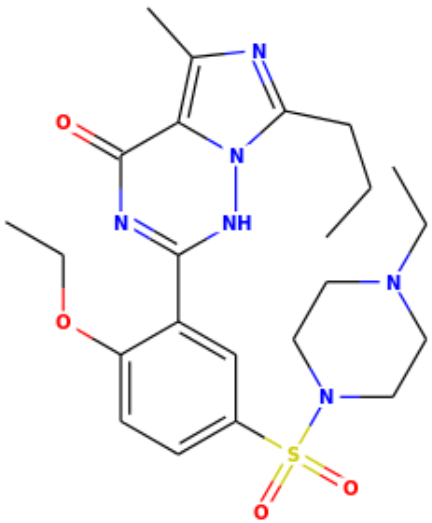
Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

1-D Feature Representations

- These representation are a collection of experimental and calculated molecular properties.
- They do not take into account the structure and bonds.
- Often, they are used for simple classification.
- In some cases, they work well.
  - E.g. the partition coefficient (ratio of solubility of two different substances)

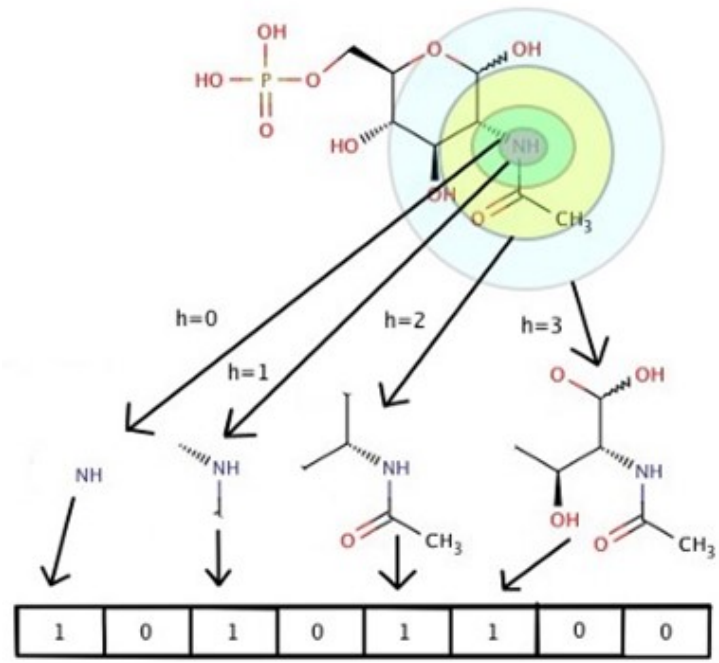
2-D Feature Descriptors

- The 2-D descriptors take into account the covalent and aromatic bonds, but not the spatial coordinates.



2D Descriptors: Fingerprints

- Fingerprinting is mapping variable-size molecular structures to a fixed-size vector (e.g. 1024 bits).
- An iterative approach
  - Diameter increasing

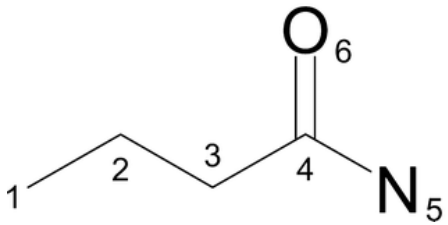


2D Descriptors: Fingerprints

- There are several fingerprinting methods, including circular fingerprint methods.
  - Each atom is examined along with its neighbors at a distance of 1, 2, ...
  - A function of atom properties (e.g., atom type) and its immediate neighborhood is computed.
  - Resulting function is hashed into a bit vector.
- A popular circular method is the extended circular fingerprints (ECFPx, x is the maximum diameter).

Fingerprints Assignment

- The initial assignment is done based on daylight atomic invariant rule:
  - the number of immediate neighbors who are “heavy” (non-hydrogen) atoms,
  - the valence minus the number of hydrogens,
  - the atomic number,
  - the atomic mass,
  - the atomic charge,
  - and the number of attached hydrogens (both implicit and explicit).
- These values are hashed into a single 32-bit integer value.

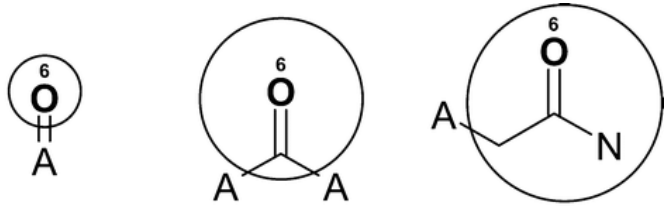


1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

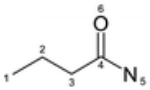
David Rogers and Mathew Hahn. Journal of Chemical Information and Modeling 2010 50 (5), 742-754

Fingerprint Iteration

- The iterative updating process generates features that represent each atom within larger and larger circular substructural neighborhoods.
- Conceptually, as the process is repeated, the feature denoted by an atom identifier represents an atom-centered substructure of increasing size.



Iteration 0 Iteration 1 Iteration 2

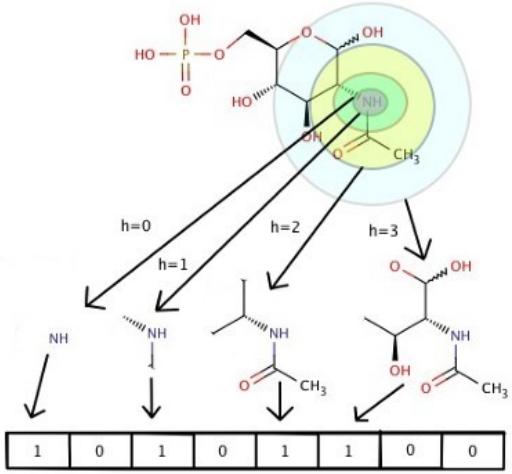


> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>
734603939	734603939	734603939	734603939
1559650422	1559650422	1559650422	1559650422
-1100000244	-1100000244	-1100000244	-1100000244
1572579716	1572579716	1572579716	1572579716
-1074141656	-1074141656	-1074141656	-1074141656
	863188371	863188371	863188371
	-1793471910	-1793471910	-1793471910
	-1789102870	-1789102870	-1789102870
	-1708545601	-1708545601	-1708545601
	-932108170	-932108170	-932108170
	2099970318	2099970318	2099970318
		-87618679	-87618679
		1112638790	1112638790
		-627599602	-627599602

David Rogers and Mathew Hahn. Journal of Chemical Information and Modeling 2010 50 (5), 742-754

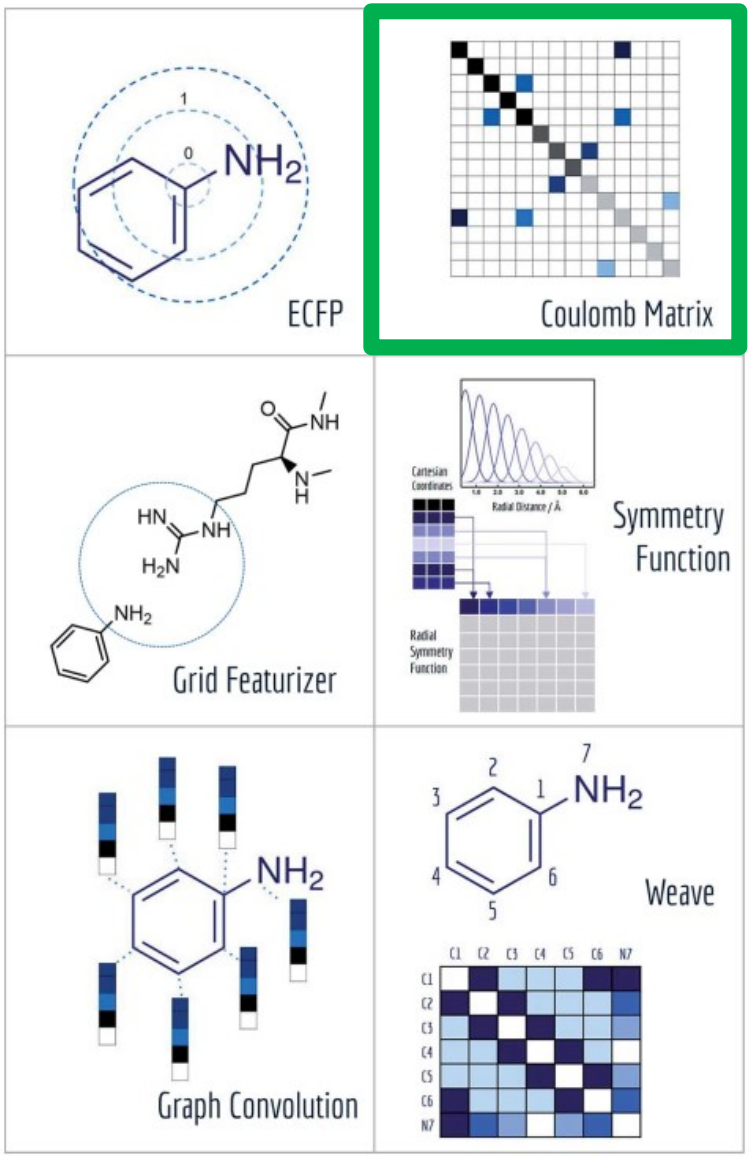
ECFP Application

- Similarity search can be easily done by comparing two bit vectors in an efficient manner.
- The representation is not unique, two completely different molecules can have the same representation.



Read more here

Featurization Methods in MoleculeNet



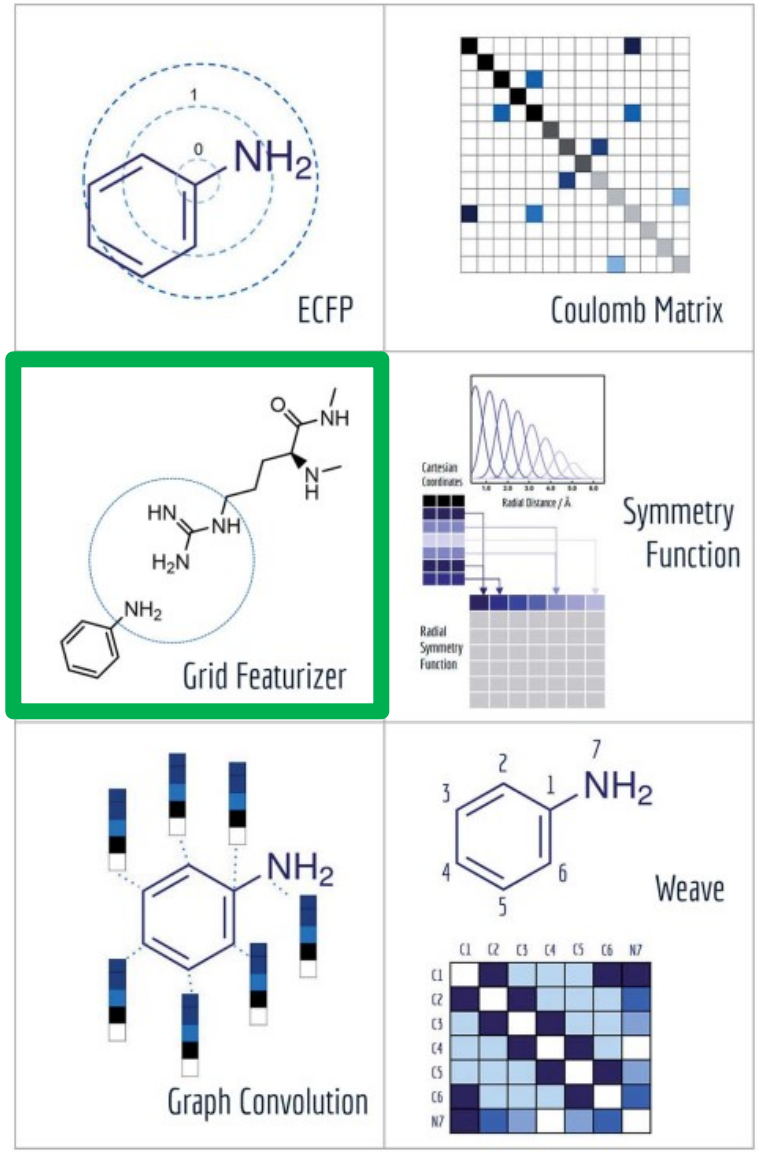
Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Coulomb Matrix

- Constructed using the nuclear charges and distances.
- A matrix M is constructed using the following equation.
  - Z is the nuclear charge, and R refers to Cartesian coordinates.

$$M_{IJ} = \begin{cases} 0.5 Z_I^{\,2.4} & \text{for } I = J \\ \dfrac{Z_I Z_J}{\left| \boldsymbol{R}_I - \boldsymbol{R}_J \right|} & \text{for } I \neq J \end{cases}$$

Featurization Methods in MoleculeNet

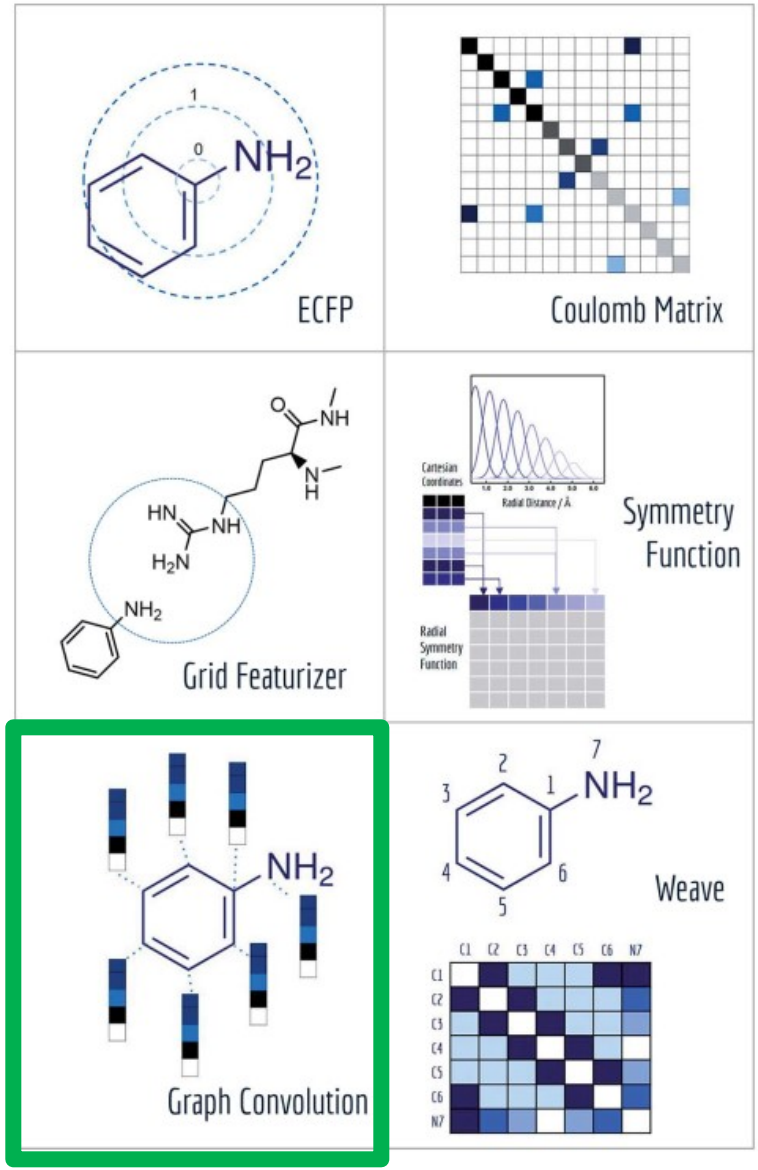


Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Grid Featurization

- DeepChem has a grid featurization tool based on *RdKitGridFeaturizer().*
- It searches for presence of chemical interactions and constructs a feature vector that contains the counts of such interactions.
  - Hydrogen bonds,
  - Salt bridges between amino acids,
  - Pi-stacking between aromatic rings.

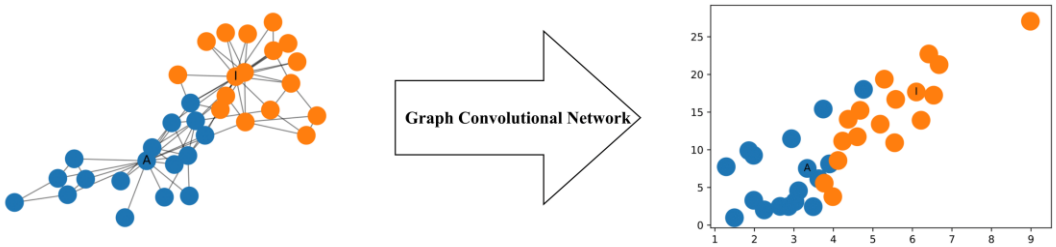
Featurization Methods in MoleculeNet



Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Graph Convolutional Networks

- A neural network operating on graphs with input:
  - X: a matrix of [nodes × node_features]
  - A: adjacency matrix of [nodes × nodes]



Description and Image based on this Link

X is per-atom features of each graph node

A is the atom adjacency matrix

H is the feature map for all atoms

Graph Convolutional Neural Network (GCNN)

1 bond length away

$$H^{(1)} = ReLU\left(W^{(1)} \cdot A \cdot X\right)$$

2 bond lengths away

$$H^{(2)} = ReLU\left(W^{(2)} \cdot A \cdot H^{(1)}\right)$$

$$\vdots$$

K bond lengths away

$$H^{(K)} = ReLU\left(W^{(K)} \cdot A \cdot H^{(K-1)}\right)$$

Sum over per atom features

$x^{(NN)}$ is convolutional "fingerprint" of entire molecule

$$x^{(NN)} = \sum_{atoms} H^{(K)}$$

$$h^{(1)} = ReLU\left(W^{(1)} \cdot x^{(NN)}\right)$$

$$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$$

$$\vdots$$

K fully connected layers

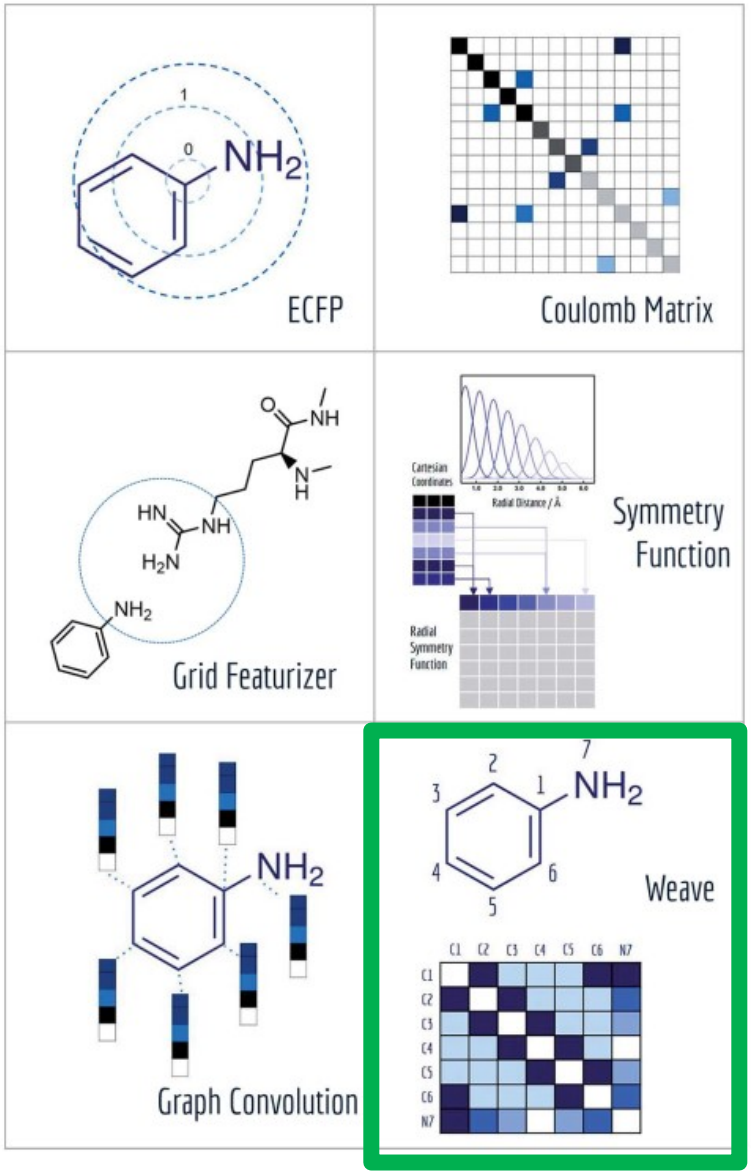
$$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$$

Credit: https://mit6874.github.io/assets/slides/6.874-lecture-14-2019.pdf

Graph Convolutions

- Convolutions in graphs are more challenging, we are dealing with abstract concepts, with no notion of sliding up or down the image!
- Two types of graph convolutions
  - Spatial
  - Spectral

Featurization Methods in MoleculeNet



Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Weave

- Similar to graph convolutions, the weave featurization encodes both local chemical environment and connectivity of atoms in a molecule.
- Difference:
  - Atomic feature vectors are exactly the same.
  - More detailed pair features instead of neighbor listing.
  - The weave featurization calculates a feature vector for each pair of atoms in the molecule, including bond properties (if directly connected), graph distance and ring info, forming a feature matrix.

Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9, no. 2 (2018): 513-530.

Applications

Predicting Protein Structures

- Homology modeling
  - If two proteins are homolog (near relatives), they probably have similar structures.
  - This works well for the overall shape, but often gets the details wrong.
- Physical Modeling
  - Using knowledge of physics laws to predict possible conformation
  - Computationally expensive
  - Will often predict the right structure, but not always.

Protein Binding

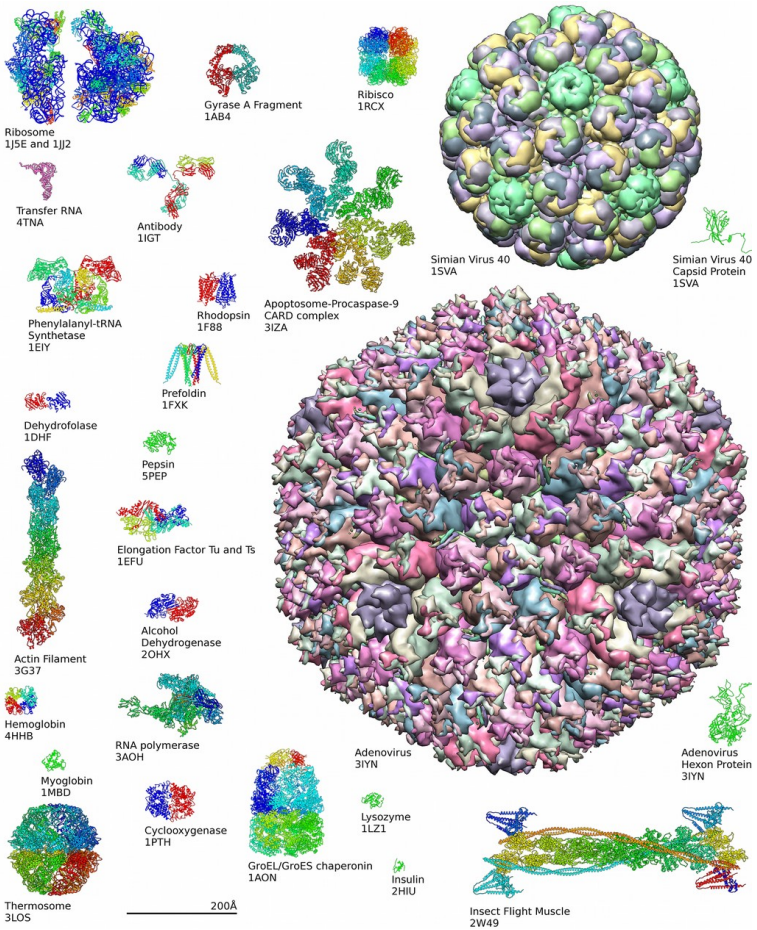
- Protein binding is very important.
  - E.g. Signaling transduction
- It involves lots of very specific details.
  - E.g. changing a few atoms can determine if a molecule binds to a protein or not.

Biophysical Featurization

1. Grid Featurization
   - It explicitly searches a 3D structure for the presence of critical physical interactions such as hydrogen bonds or salt bridges.
   - We can rely on a wealth of known facts.
   - Yet, bound by known physics.
2. Atomic Featurization
   - It provides a processed representation of the 3D positions and identities of all atoms.
   - It must learn to identify critical physical interactions,
   - Yet, feasible to detect new patterns of interesting behavior

Protein Data Bank (PDB)

- The primary repository for known protein structures
- It contains over 142,000 structures.
  - Far less than what we want.



PDBind Dataset

- The PDBind dataset contains a large number of biomolecular crystal structures.
- About 15,000 complex structures, each annotated with the binding affinity measure.
- We will look at the problem of predicting the binding affinity in protein-ligand complexes.
- Design of better machine learning models to accurately predict the thermodynamic behavior of these systems is still an open problem.

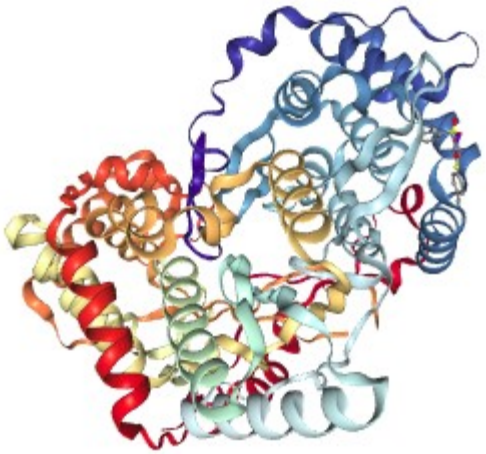
PDB Files

- Protein structures are stored in PDB files.
  - Text files containing description of the atoms in the structure and their positions.
  - Often malformed, as experiments fail to have adequate resolution to completely specify a portion of the protein.
  - PDB files can be difficult to understand, so we use visualization packages.
    - NGLview

NGLview

```
SEQRES  27 B  570  PHE THR GLU ALA MET THR ARG TYR SER ALA PRO PRO GLY
SEQRES  28 B  570  ASP PRO PRO GLN PRO GLU TYR ASP LEU GLU LEU ILE THR
SEQRES  29 B  570  SER CYS SER SER ASN VAL SER VAL ALA HIS ASP ALA SER
SEQRES  30 B  570  GLY LYS ARG VAL TYR TYR LEU THR ARG ASP PRO THR THR
SEQRES  31 B  570  PRO LEU ALA ARG ALA ALA TRP GLU THR ALA ARG HIS THR
SEQRES  32 B  570  PRO VAL ASN SER TRP LEU GLY ASN ILE ILE MET TYR ALA
SEQRES  33 B  570  PRO THR LEU TRP ALA ARG MET ILE LEU MET THR HIS PHE
SEQRES  34 B  570  PHE SER ILE LEU LEU ALA GLN GLU GLN LEU GLU LYS ALA
SEQRES  35 B  570  LEU ASP CYS GLN ILE TYR GLY ALA CYS TYR SER ILE GLU
SEQRES  36 B  570  PRO LEU ASP LEU PRO GLN ILE ILE GLU ARG LEU HIS GLY
SEQRES  37 B  570  LEU SER ALA PHE SER LEU HIS SER TYR SER PRO GLY GLU
SEQRES  38 B  570  ILE ASN ARG VAL ALA SER CYS LEU ARG LYS LEU GLY VAL
SEQRES  39 B  570  PRO PRO LEU ARG VAL TRP ARG HIS ARG ALA ARG SER VAL
SEQRES  40 B  570  ARG ALA ARG LEU LEU SER GLN GLY GLY ARG ALA ALA THR
SEQRES  41 B  570  CYS GLY LYS TYR LEU PHE ASN TRP ALA VAL LYS THR LYS
SEQRES  42 B  570  LEU LYS LEU THR PRO ILE PRO ALA ALA SER GLN LEU ASP
SEQRES  43 B  570  LEU SER GLY TRP PHE VAL ALA GLY TYR SER GLY GLY ASP
SEQRES  44 B  570  ILE TYR HIS SER LEU SER ARG ALA ARG PRO ARG
HET    CCT  A1001      27
HET    CCT  B2001      27
HETNAM     CCT 5-(4-CYANOPHENYL)-3-{[(2-METHYLPHENYL)
HETNAM   2 CCT  SULFONYL]AMINO}THIOPHENE-2-CARBOXYLIC ACID
FORMUL   3  CCT    2(C19 H14 N2 O4 S2)
FORMUL   5  HOH   *754(H2 O)
HELIX    1   1 LEU A   26  LEU A   31  1                                   6
HELIX    2   2 HIS A   33  ASN A   35  5                                   3
HELIX    3   3 THR A   41  ARG A   43  5                                   3
HELIX    4   4 SER A   44  THR A   53  1                                  10
HELIX    5   5 ASP A   61  SER A   76  1                                  16
HELIX    6   6 SER A   84  LEU A   91  1                                   8
HELIX    7   7 GLY A  104  ASN A  110  1                                   7
HELIX    8   8 SER A  112  ASP A  129  1                                  18
HELIX    9   9 ASP A  164  GLY A  188  1                                  25
HELIX   10  10 SER A  189  TYR A  195  5                                   7
HELIX   11  11 SER A  196  SER A  210  1                                  15
```

PDB File



Other Visualization Tools

- Note that there are other tools used in professional drug discovery.
  - VMD, PyMOL, Chimera
- NGLview is nicely integrated into Jupyter and is open source.