# Machine Learning and Chemistry

QMMM Study Group

Dakota Folmsbee

2018/08/31

# What is Machine Learning?

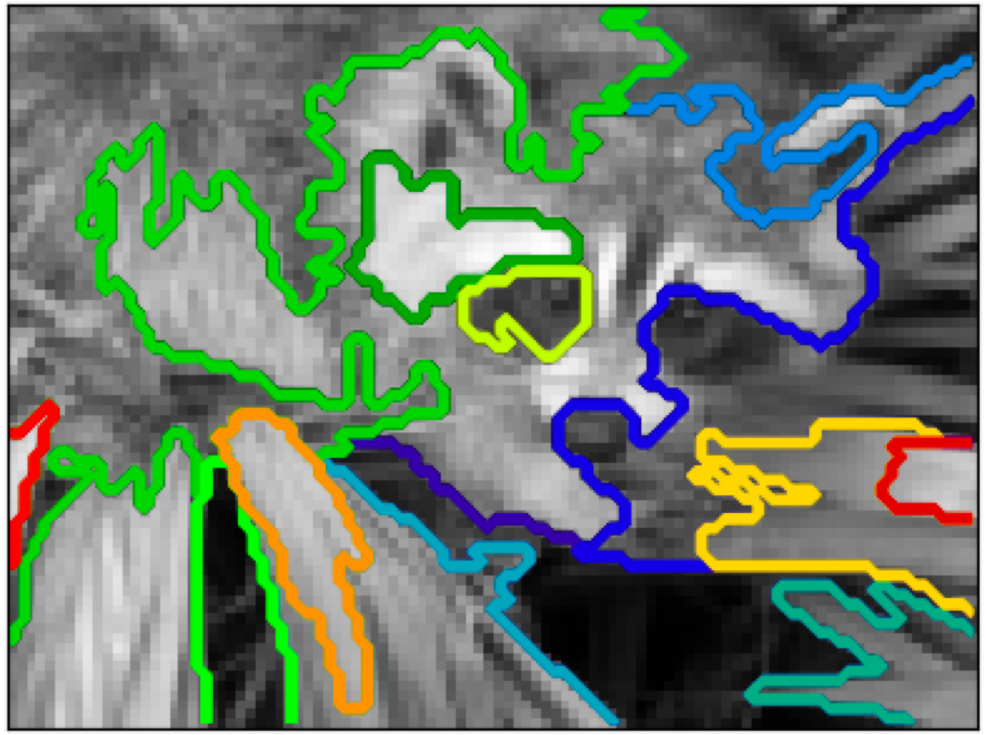
Ability to learn without rules-based programming

Learning:

- Supervised
- Unsupervised
- Reinforcement

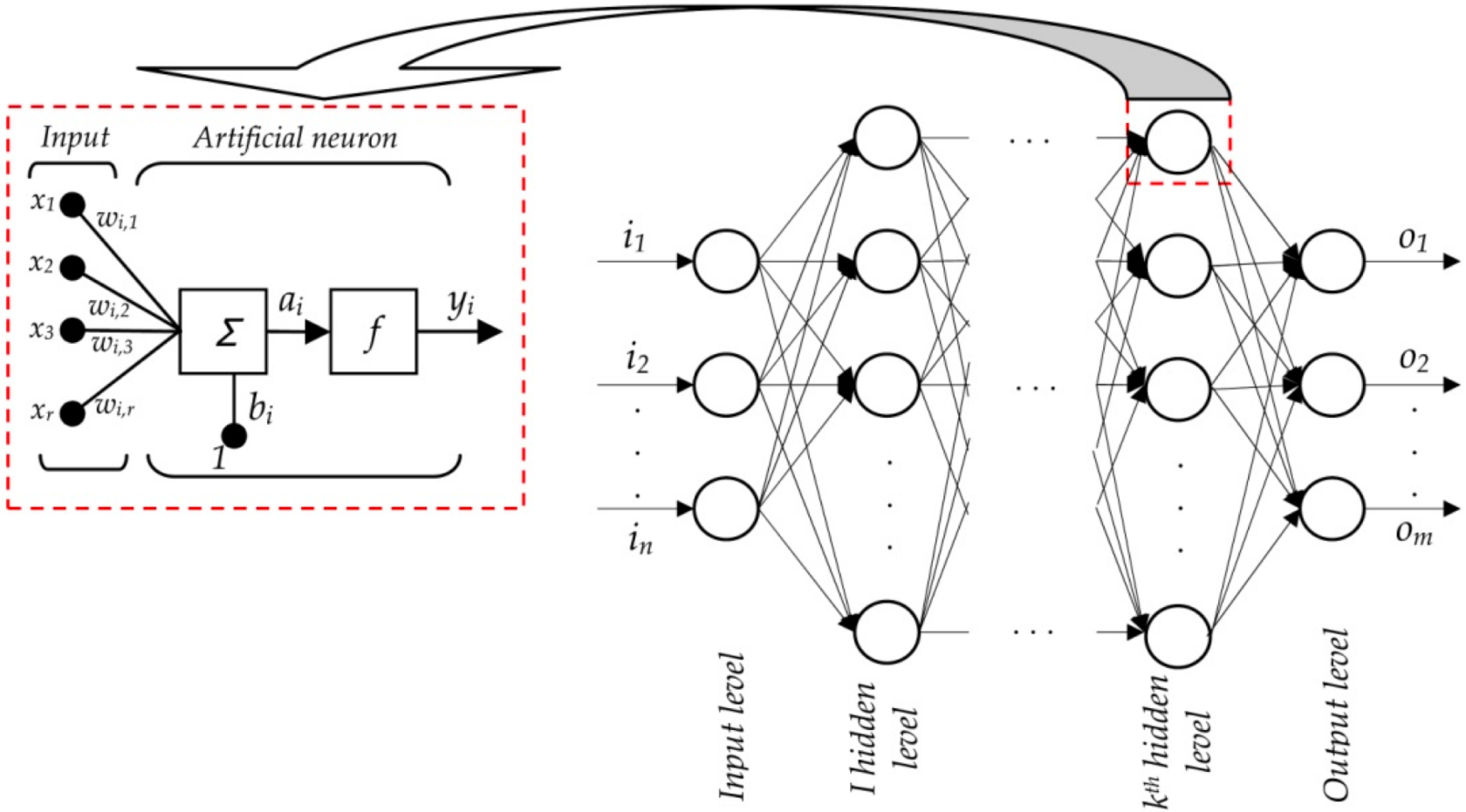
Types:

- Classification
- Regression
- Clustering



http://scikit-learn.org/stable/auto_examples/cluster/plot_face_ward_segmentation.html

# What is Machine Learning?




https://www.analyticsvidhya.com/wp-content/uploads/2016/03/2.-ann-structure.jpg

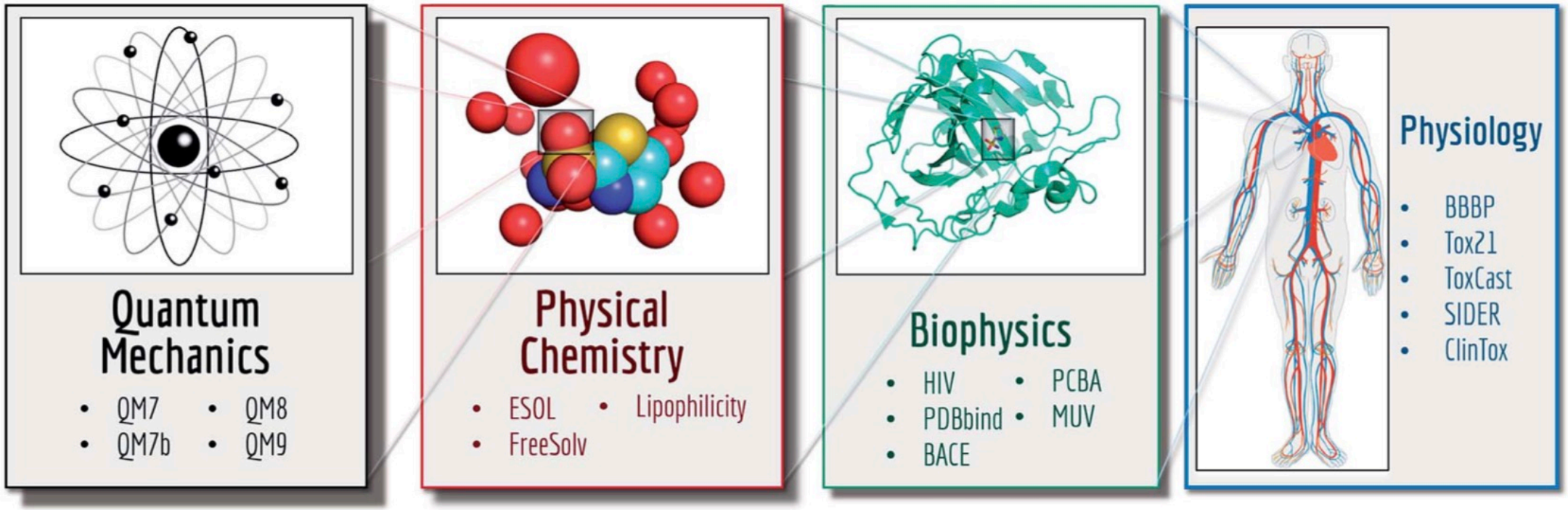
# Why use ML in Chemistry?

- Chemical space ~$10^{60}$ stable molecules
- Time requirement
- Multi-property training
- Multi-level theory training

# What is out there?

- Von Lilenfeld Group
  - Work with QM datasets
  - Bayesian Ridge, Kernel Ridge, and more
  - Coulomb Matrix, Bag of Bonds, BA Representations
- MoleculeNet and DeepChem
  - Benchmarks for machine learning models
  - Implementation of Scikit-Learn and Tensorflow
- And Much More!

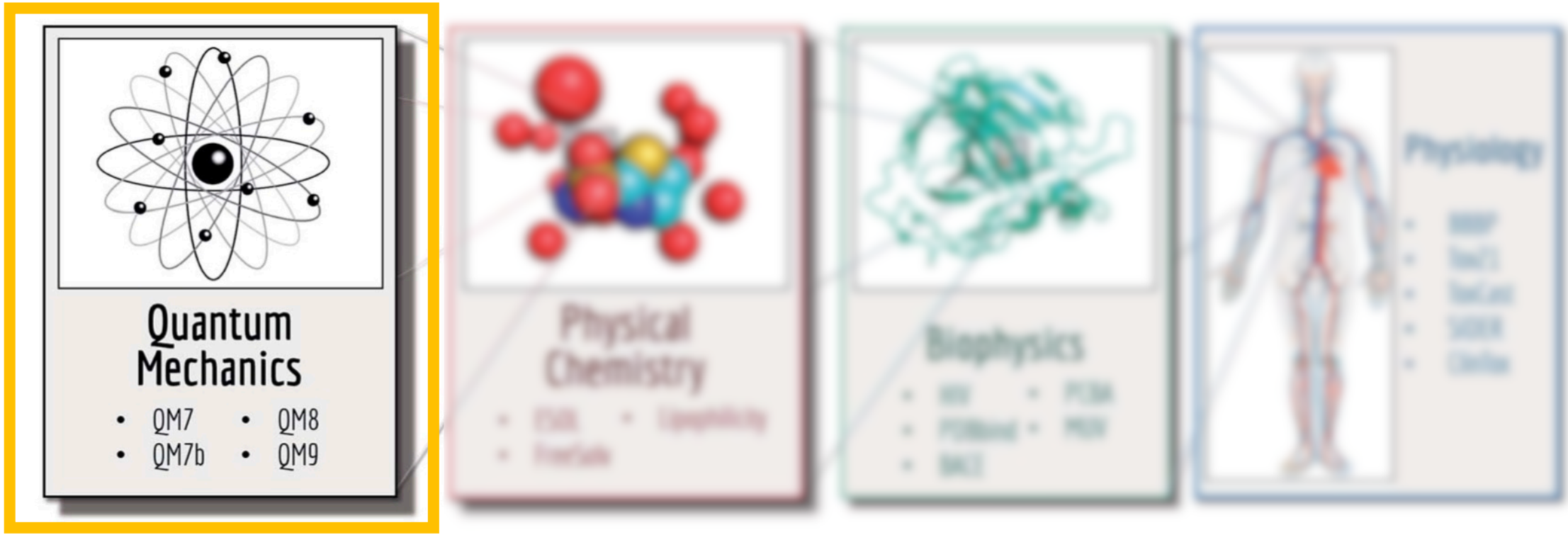
# Datasets






Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

# Datasets






Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

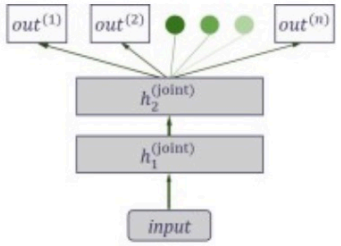
# Current Approaches

- Machine Learning Methods
- Representations
- Benchmarks

# Current Approaches

- **Machine Learning Methods**
  - **Conventional**
  - **Graph**
- Representations
- Benchmarks
- Packages

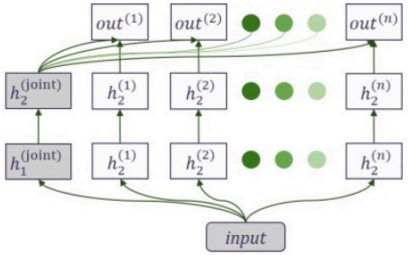
# Conventional Models




## Multitask Network

Standard neural network prediction method designed for multitask settings. Input features are processed through multiple shared fully-connected layers and then fed into separate linear classifiers/regressors for each different task. In case of single task dataset, it will become vanilla neural network model.

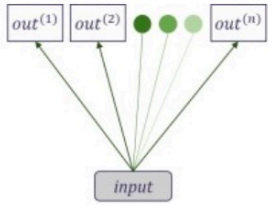
Regression Deterministic Classification




## Bypass Network

A modified version of multitask network designed for uncorrelated tasks. Based on the structure of multitask network, it adds "bypass" layers directly connecting input features and each individual task, hence increasing explanatory power in case of unrelated variations in the sample.

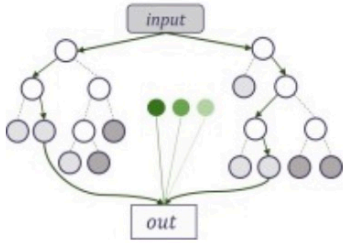
Regression Deterministic Classification




## Logistic Regression

Standard classification model by applying logistic function to weighted linear combination of input features.

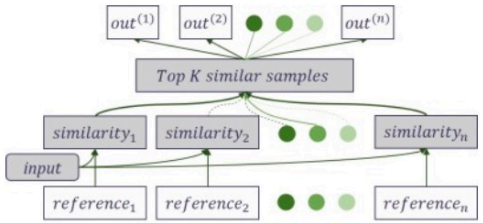
Classification Deterministic




## Random Forest

Standard classification and regression method based on an ensemble of decision trees, each trained on a different subsampled version of the original dataset.

Regression Deterministic Classification




## Influence Relevance Vector

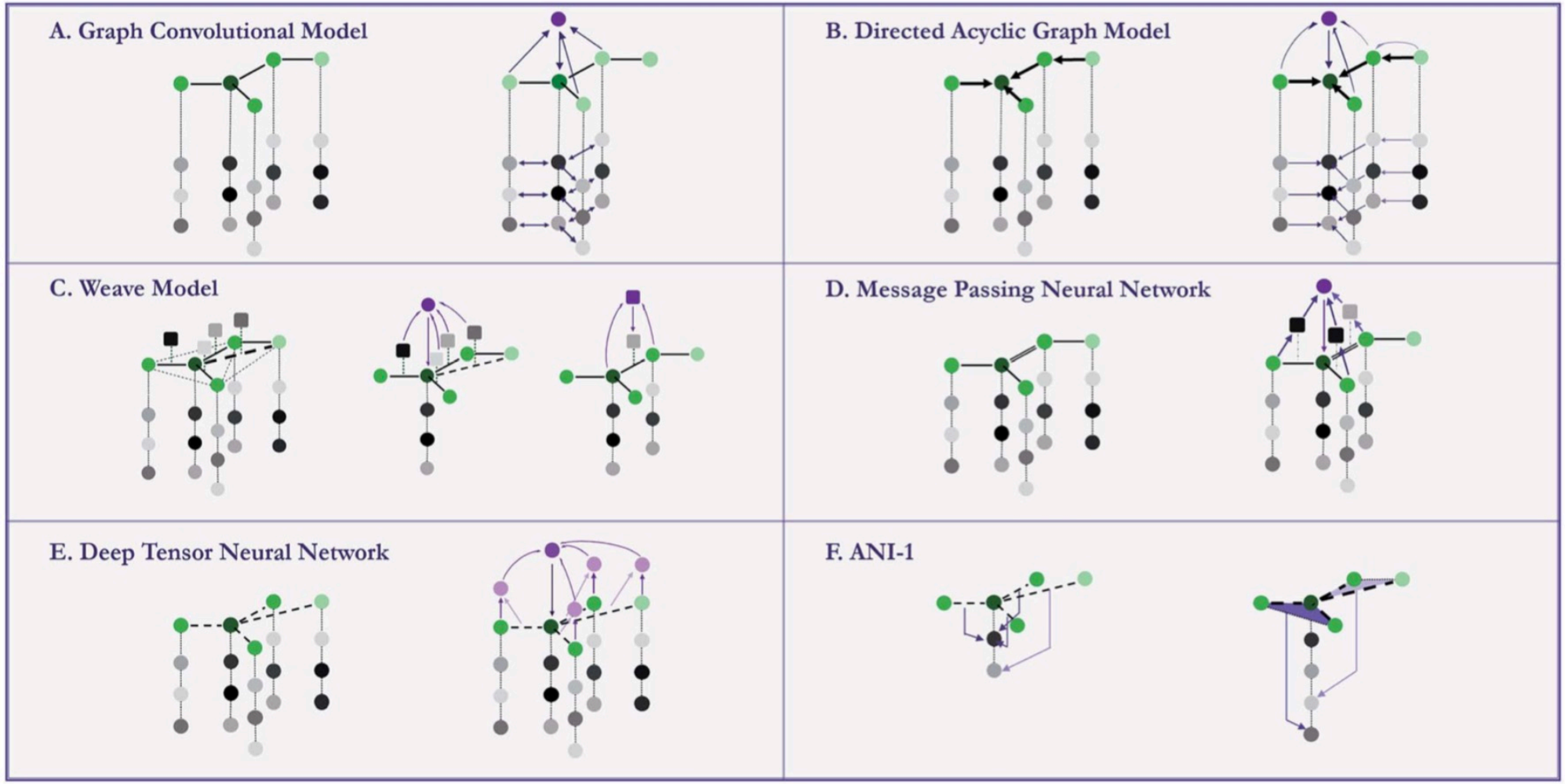
Refined K-nearest neighbour classifier. Using the hypothesis that compounds with similar substructures have similar functionality, it makes prediction by combining labels from the top-K compounds most similar to the sample.

Classification Deterministic



Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

# Graph Models

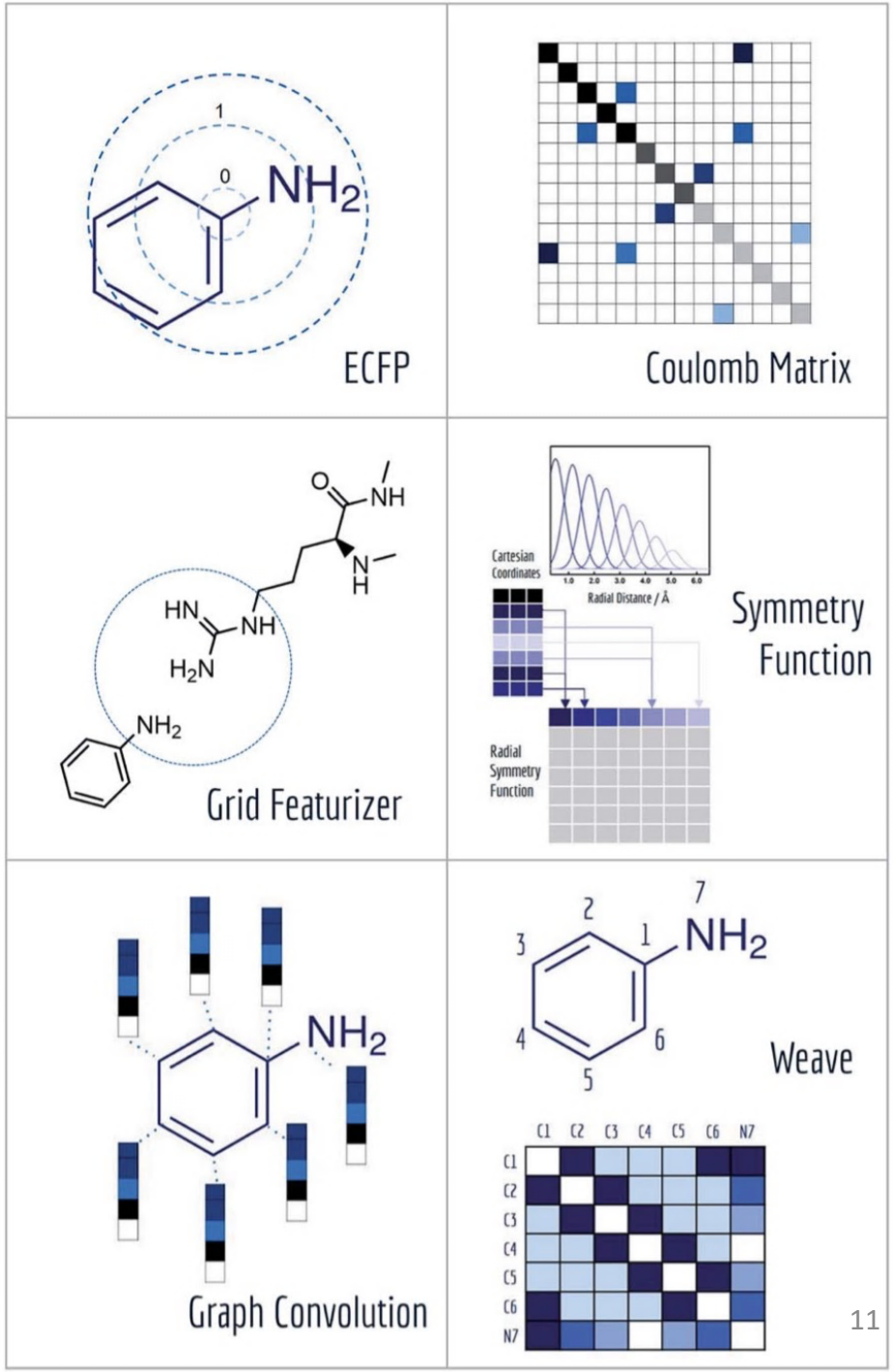





Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

# Current Approaches

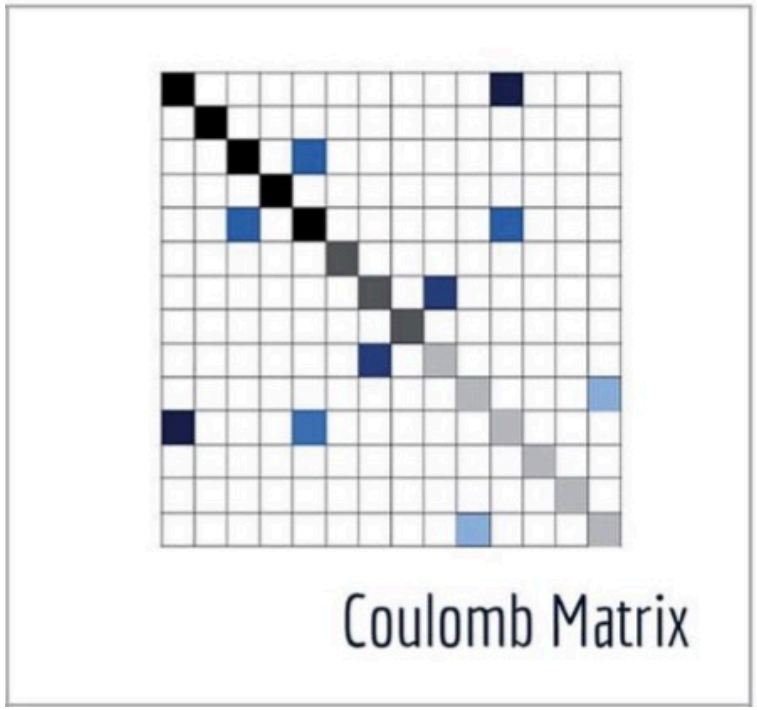
- Machine Learning Methods
- **Representations**
  - **Coulomb Matrix**
  - **Bag of Bonds**
  - **BAML**
  - **HD/HDAD**
  - **Molecular Graphs**
- Benchmarks
- Packages



# Representations

- Coulomb Matrix

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

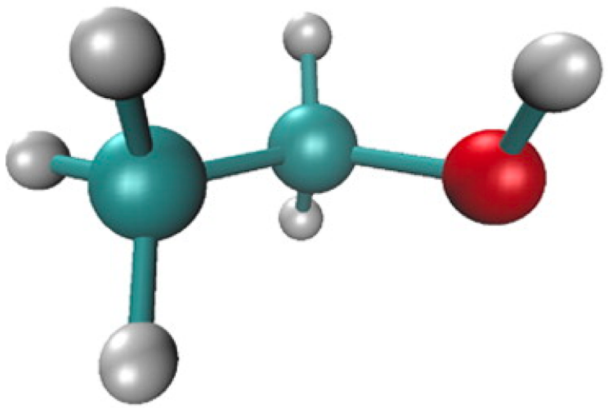





Rupp M., Beigler, Tkatchenko, A., Müller, K., von Lilienfeld, O. A., *Phys. Rev. Lett.* 2012, 108, 58301.
Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

# Representations

- Bag of Bonds



(a)

| | $O$ | $C$ | $C$ | $H$ | $H$ | $H$ | $H$ | $H$ | $H$ |
|---|---|---|---|---|---|---|---|---|---|
| $O$ | $O$ | $OC$ | $OC$ | $OH$ | $OH$ | $OH$ | $OH$ | $OH$ | $OH$ |
| $C$ | $OC$ | $C$ | $CC$ | $CH$ | $CH$ | $CH$ | $CH$ | $CH$ | $CH$ |
| $C$ | $OC$ | $CC$ | $C$ | $CH$ | $CH$ | $CH$ | $CH$ | $CH$ | $CH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $H$ | $HH$ | $HH$ | $HH$ | $HH$ | $HH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $HH$ | $H$ | $HH$ | $HH$ | $HH$ | $HH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $HH$ | $HH$ | $H$ | $HH$ | $HH$ | $HH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $HH$ | $HH$ | $HH$ | $H$ | $HH$ | $HH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $HH$ | $HH$ | $HH$ | $HH$ | $H$ | $HH$ |
| $H$ | $OH$ | $CH$ | $CH$ | $HH$ | $HH$ | $HH$ | $HH$ | $HH$ | $H$ |

(b)

O
C
C
H
H
H
H
H

OC
OC
OH
OH
OH
OH
OH
CC

CH
CH
CH
CH
CH
CH
CH
CH
CH
CH

HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH

(c)

0
O-bag
0
C-bag
0
H-bag
0
OC-bag
0
OH-bag
0
CC-bag
0
CH-bag
0
HH-bag
0

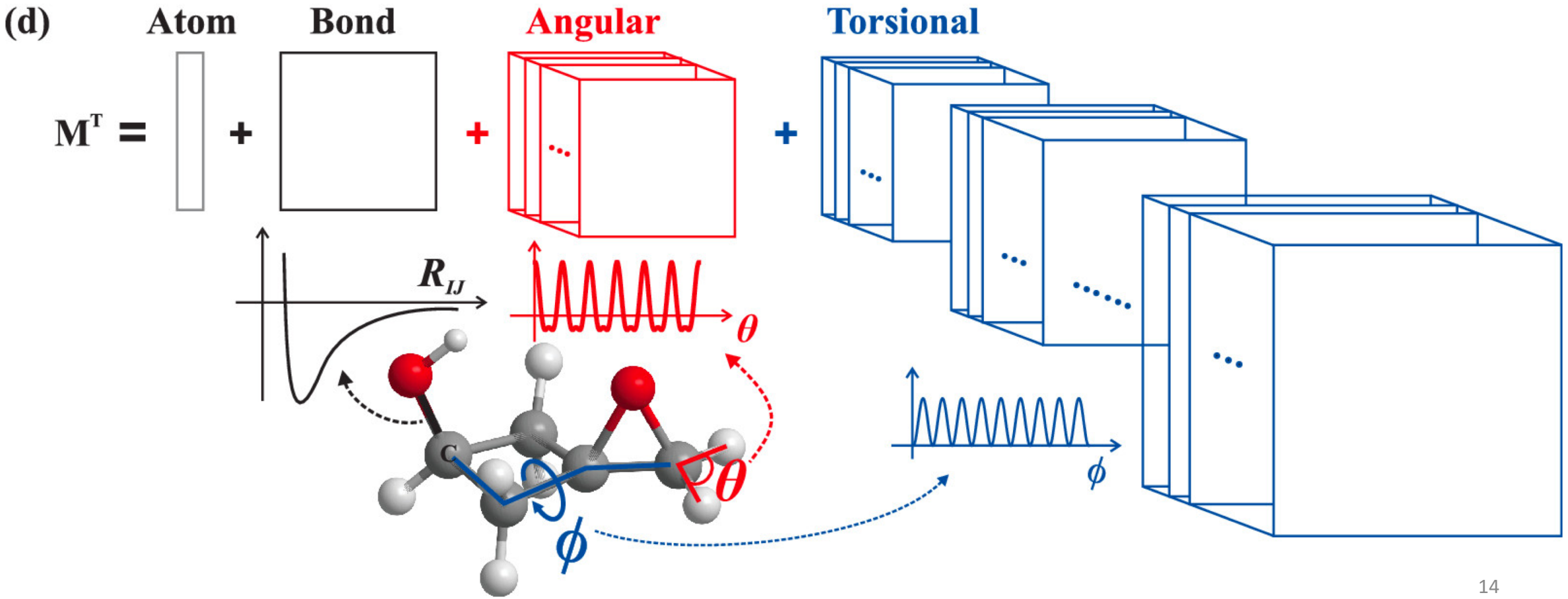
(d)



Hansen K., Beigler, F. Ramakrishnan, R. Pronobis, W. , von Lilienfeld, O. A., Müller, K., Tkatchenko, A., *J. Phys. Chem. Lett.* 2015, 6, 2329.

# Representations

- BAML

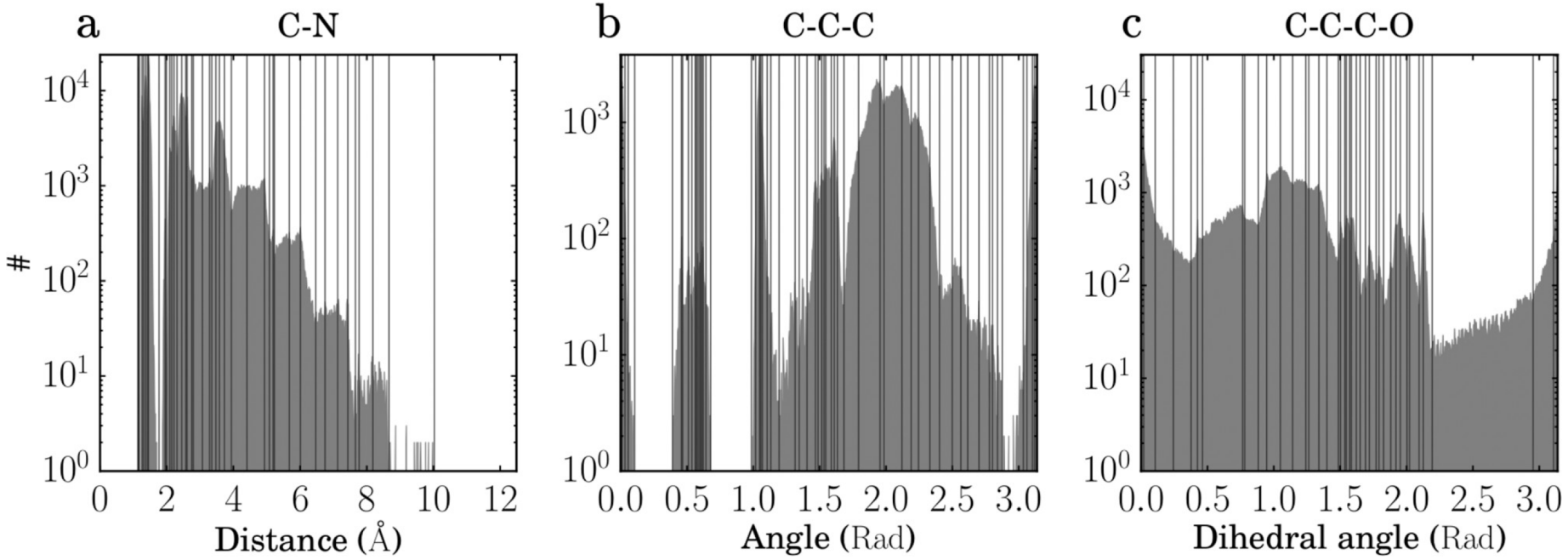





Huang, B., von Lilienfeld, O. A., *J. Chem. Phys* **145**, 161102 (2016).

# Representations

- Histogram of Distances (HD)
- Histogram of Distances, Angles, and Dihedrals (HDAD)

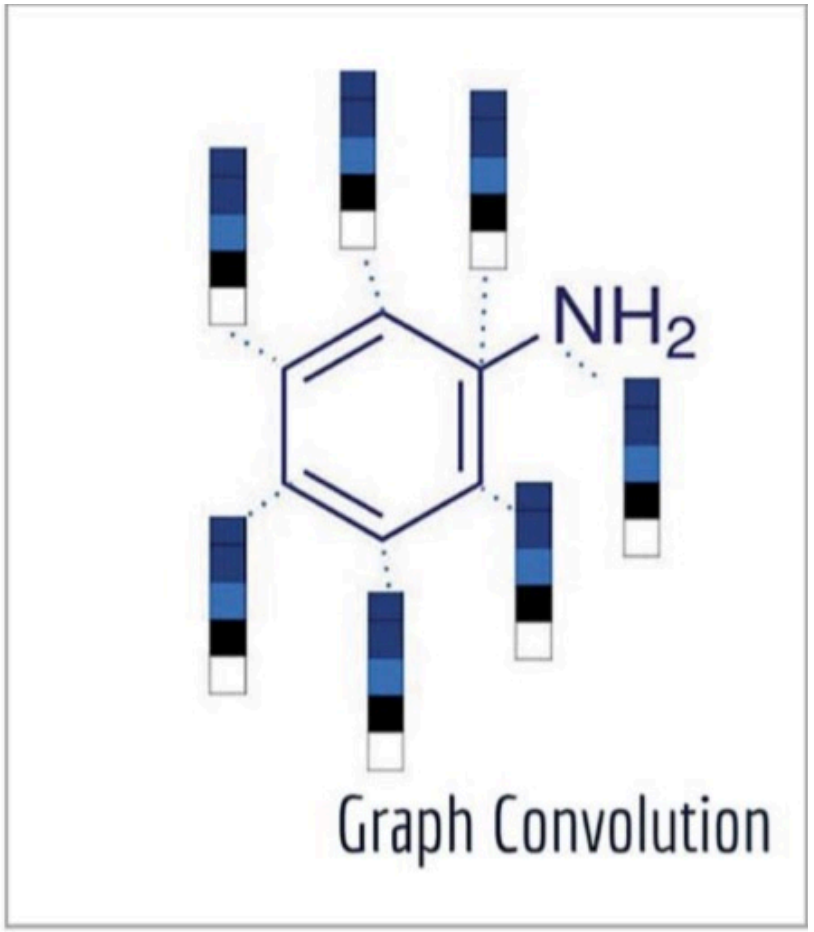





Faber, F., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S., Dahl, G., Vinyals, O., Kearnes, S., Riley, P., von Liliefeld, A., *J. Chem. Theory Comput.* 2017, 13, 5259.

# Representations

- Molecular Graphs
- Contains feature vectors:
  - Local chemical environment
  - Atom types
  - Hybridization types
  - Valence structures






Kearnes, S., McCloskey, K. Berndl, M., Pande, V., *J. Comput. Aided Mol .Des.* 2016, 30, 595-608.

# Current Approaches

- Machine Learning Methods
- Representations
- **Benchmarks**
  - **MoleculeNet**
  - **von Lilienfeld group**
- Packages

# Benchmarks

| Method | Representation | U0 (ev) | $\varepsilon_{HOMO}$ (ev) | $\varepsilon_{LUMO}$ (ev) | $\Delta\varepsilon$ (ev) | $\mu$ (D) | $\alpha$ (bohr$^3$) | ZPVE (ev) | Cv (cal/molK) |
|---|---|---|---|---|---|---|---|---|---|
| KRR† | CM | 0.128 | 0.133 | 0.183 | 0.229 | 0.449 | 0.433 | 0.0048 | 0.118 |
| | BOB | 0.0667 | 0.0948 | 0.122 | 0.148 | 0.423 | 0.298 | 0.00364 | 0.0917 |
| | BAML | 0.0519 | 0.0946 | 0.121 | 0.152 | 0.46 | 0.301 | 0.00331 | 0.082 |
| | HDAD | 0.0251 | 0.0662 | 0.0842 | 0.107 | 0.334 | 0.175 | 0.00191 | 0.0441 |
| | HD | 0.0644 | 0.0874 | 0.113 | 0.143 | 0.364 | 0.299 | 0.00316 | 0.0844 |
| GG† | MG | 0.0421 | 0.0567 | 0.0628 | 0.0877 | 0.247 | 0.161 | 0.00431 | 0.0837 |
| GC† | MG | 0.15 | 0.0549 | 0.062 | 0.0869 | 0.101 | 0.232 | 0.00966 | 0.097 |
| Multitask‡ | CM | 0.0984 | 0.00506 | 0.00645 | 0.0086 | 0.519 | 0.85 | 0.00207 | 0.39 |
| GC‡ | MG | 0.1479 | 0.00716 | 0.00921 | 0.0112 | 0.583 | 1.37 | 0.00299 | 0.65 |
| DTNN‡ | MG | 0.1054 | 0.00388 | 0.00513 | 0.0066 | 0.244 | 0.95 | 0.00172* | 0.27 |
| MPNN‡ | MG | 0.0889 | 0.00541 | 0.00623 | 0.0082 | 0.358 | 0.89 | 0.00216 | 0.42 |

† Faber, F., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S., Dahl, G., Vinyals, O., Kearnes, S., Riley, P., von Liliefeld, A., *J. Chem. Theory Comput.* 2017, 13, 5259.
‡ Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V., *Chem. Sci.* 2018, 9, 527.

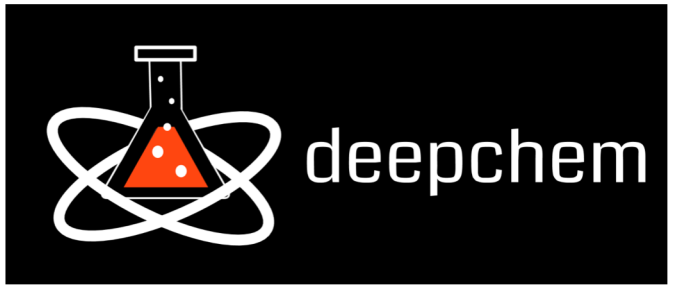
# Current Approaches

- Machine Learning Methods
- Representations
- Benchmarks
- **Packages**
  - **Methods**
  - **Representations**

# Machine Learning Libraries

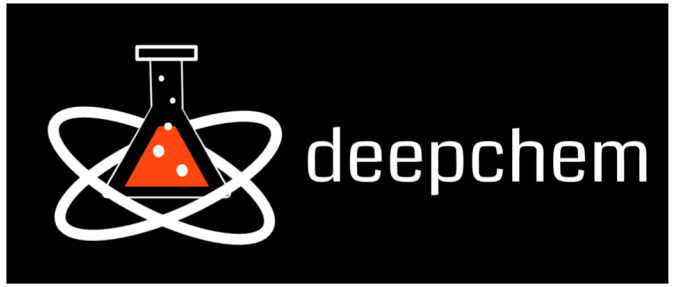
- Scikit-Learn
  - http://scikit-learn.org/stable/
- Tensorflow
  - https://www.tensorflow.org/
- DeepChem
  - https://deepchem.io/
  - https://github.com/deepchem/deepchem





# Representation Libraries

- DeepChem
  - https://deepchem.io/
  - https://github.com/deepchem/deepchem




- QML
  - http://www.qmlcode.org/
  - https://github.com/qmlcode/tutorial

- MolML
  - https://github.com/crcollins/molml

# Find this interesting?

Want to recreate representations from the literature?

Check out chemreps and help Amanda, Shiv, and I do just that!

https://github.com/dlf57/chemreps

You are welcome to fork, submit issues, and add feature requests!

# Useful Resources

- **MoleculeNet** http://moleculenet.ai/
- **MoleculeNet: a benchmark for molecular machine learning**
  Wu, Zhenqin and Ramsundar, Bharath and Feinberg, EvanÂ N. and Gomes, Joseph and Geniesse, Caleb and Pappu, Aneesh S. and Leswing, Karl and Pande, Vijay
  *Chem. Sci.* **2018** *9* (2), 513-530
  DOI: 10.1039/C7SC02664A

- **Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error**
  Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld
  *Journal of Chemical Theory and Computation* **2017** *13* (11), 5255-5264
  DOI: 10.1021/acs.jctc.7b00577