Molecular Machine Learning

Bharath Ramsundar

What is Molecular Machine Learning?

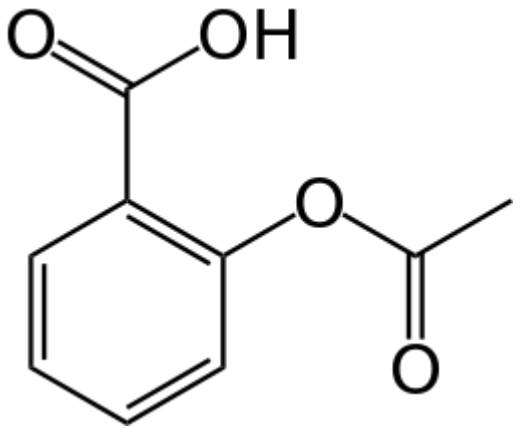
- The biggest applied fields of machine learning today are computer vision, natural language processing, speech processing, and video processing.
- **Molecular ML** is the field attempting to build machine-learning architectures and code-bases for learning to predict properties of molecules
- There are many challenges unique to molecular ML (as opposed to other machine learning fields).

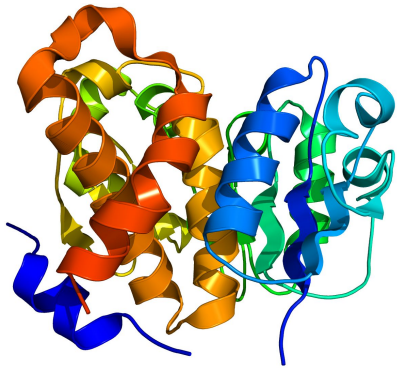
Molecular ML Challenges: Data Scarcity

- Limited data available publicly.
- Not easy to gather data. Most experiments require highly-trained scientists or technicians to perform. (No Mechanical Turking of data).
- Pharmaceutical companies and biotechs have data, but tough to get due to IP concerns.

Molecular ML Challenge: Featurization

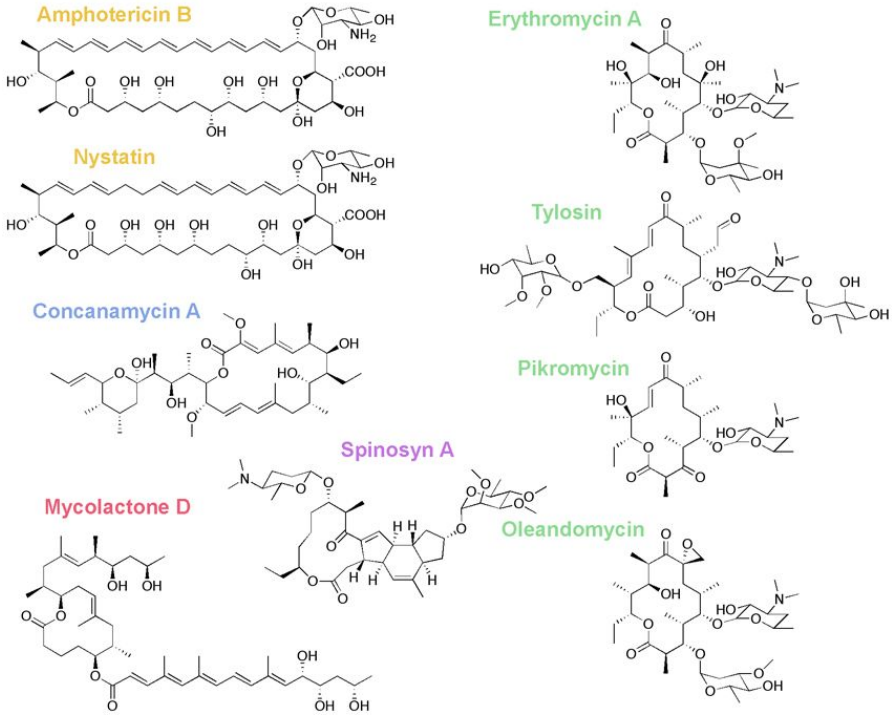
- Molecules come in many sizes and shapes.
- How can a molecule be transformed into a vector/matrix for machine learning?
- Turns out different representations needed for different problems.





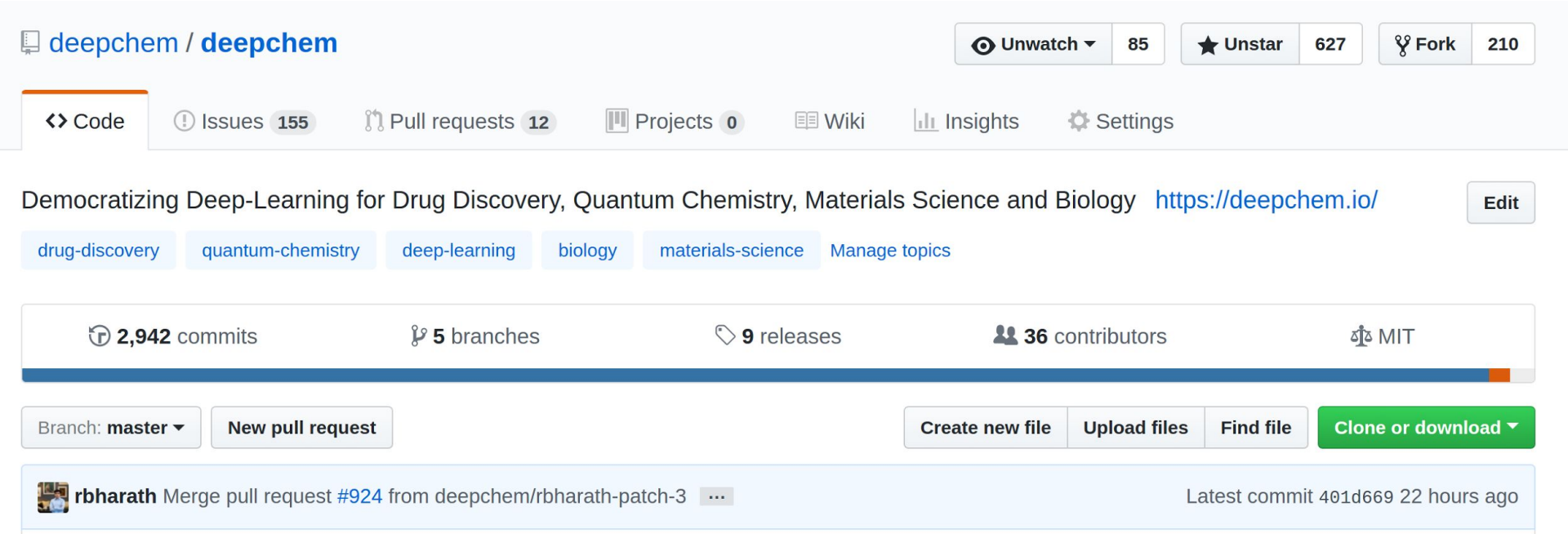
Molecular ML Challenge: Generalization

- Chemical space is vast (10^60 drug-like molecules) and only a small fraction (100 M compounds) actually made.
- So test set for models has to be different from training set! We want to know which new molecules are useful.

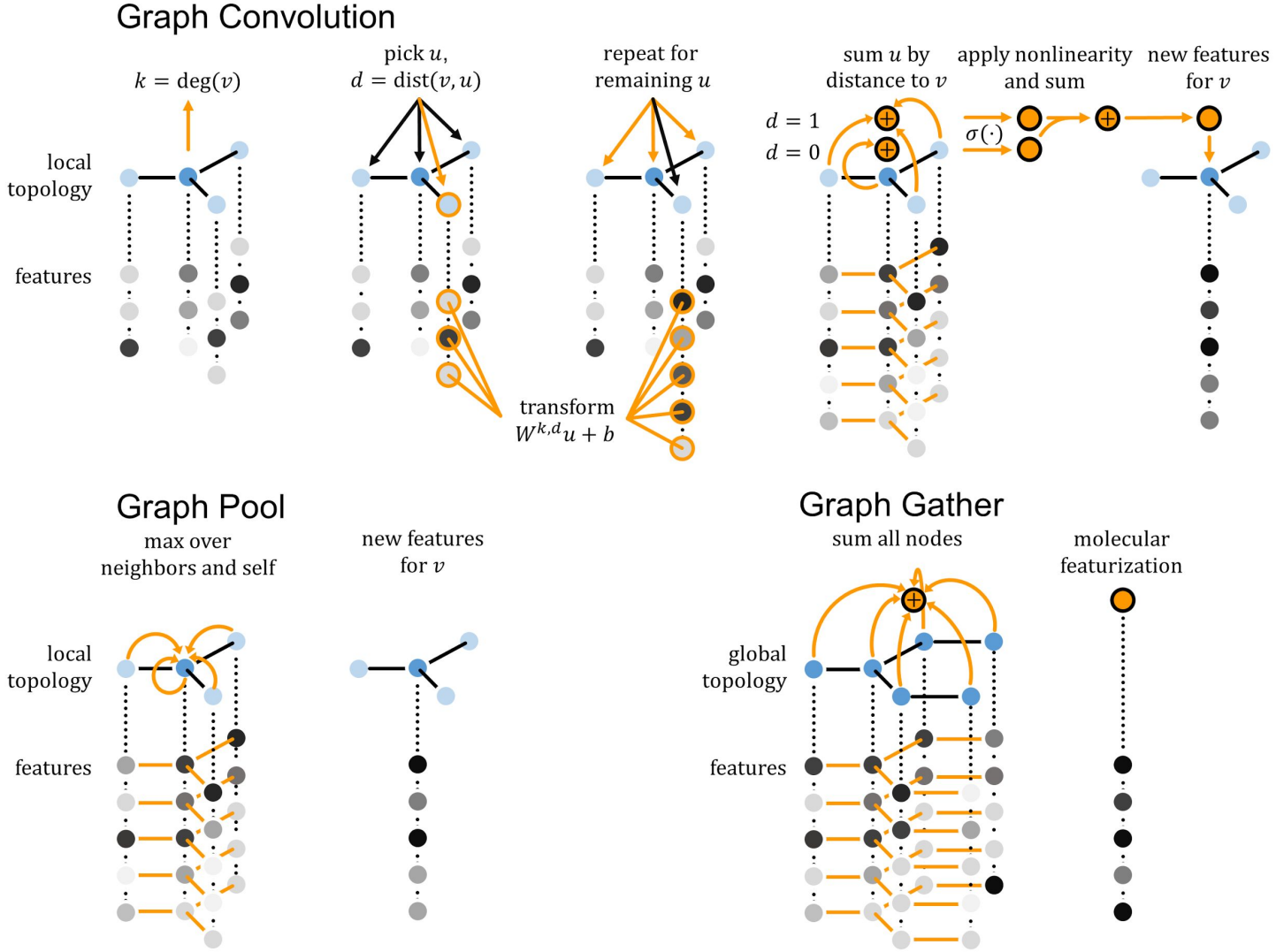


DeepChem: https://deepchem.io

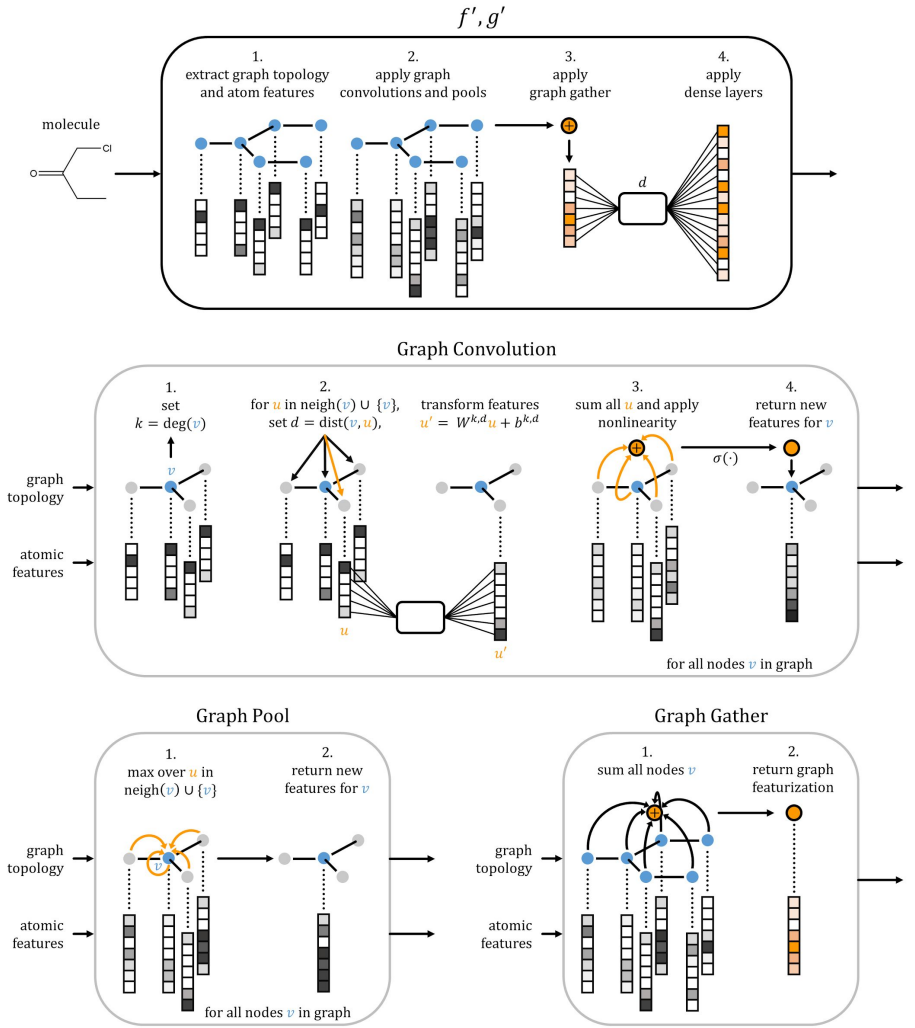
- DeepChem is a molecular machine learning toolchain built on Tensorflow, being used for a variety of projects, including drug-discovery, quantum chemistry, solar cell design, and structural biology.



Graph Convolutions


Graph Convolution
$k = \deg(v)$
local topology
features
pick u, $d = \mathrm{dist}(v, u)$
repeat for remaining u
transform $W^{k,d}u + b$
sum u by distance to v
$d = 1$
$d = 0$
apply nonlinearity and sum
$\sigma(\cdot)$
new features for v
Graph Pool
max over neighbors and self
local topology
features
new features for v
Graph Gather
sum all nodes
global topology
features
molecular featurization

Graph Convolutions Explained Diagrammatically



Low Data Drug Discovery with One Shot Learning

Han Altae-Tran*‡, Bharath Ramsundar*†, Aneesh S. Pappu†, Vijay Pande†

* : equal contribution
†: Stanford University
‡: Massachusetts Institute of Technology

Published in ACS Central Science.

One-Shot: Borrowing Tricks from Babies

Giraffe

?

Google DeepMind's Matching Networks

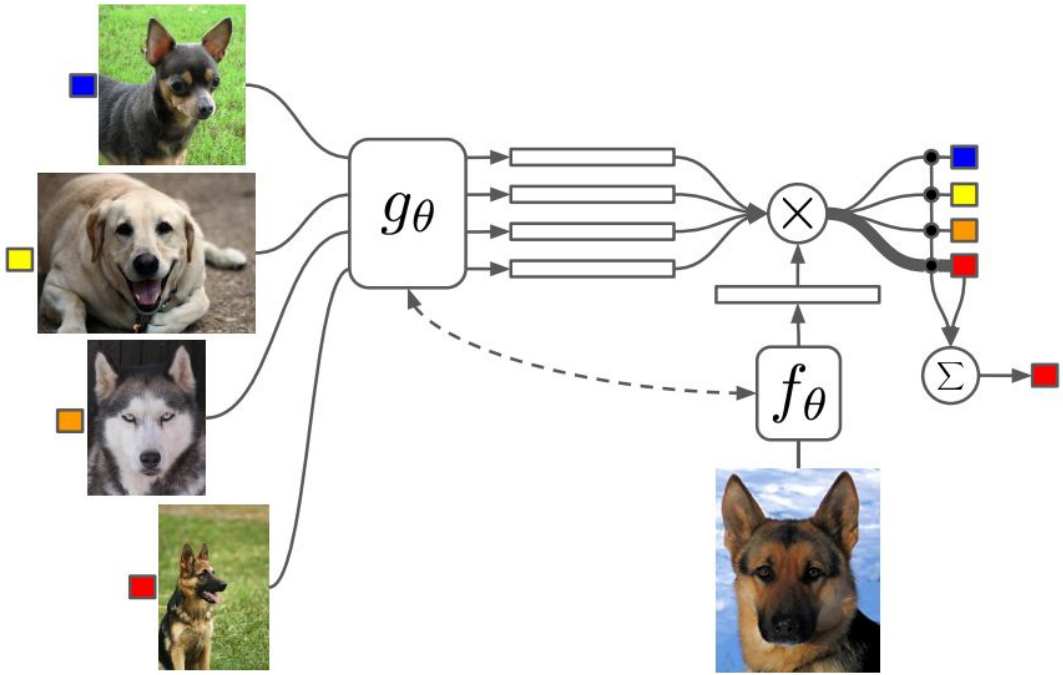


Figure 1: Matching Networks architecture

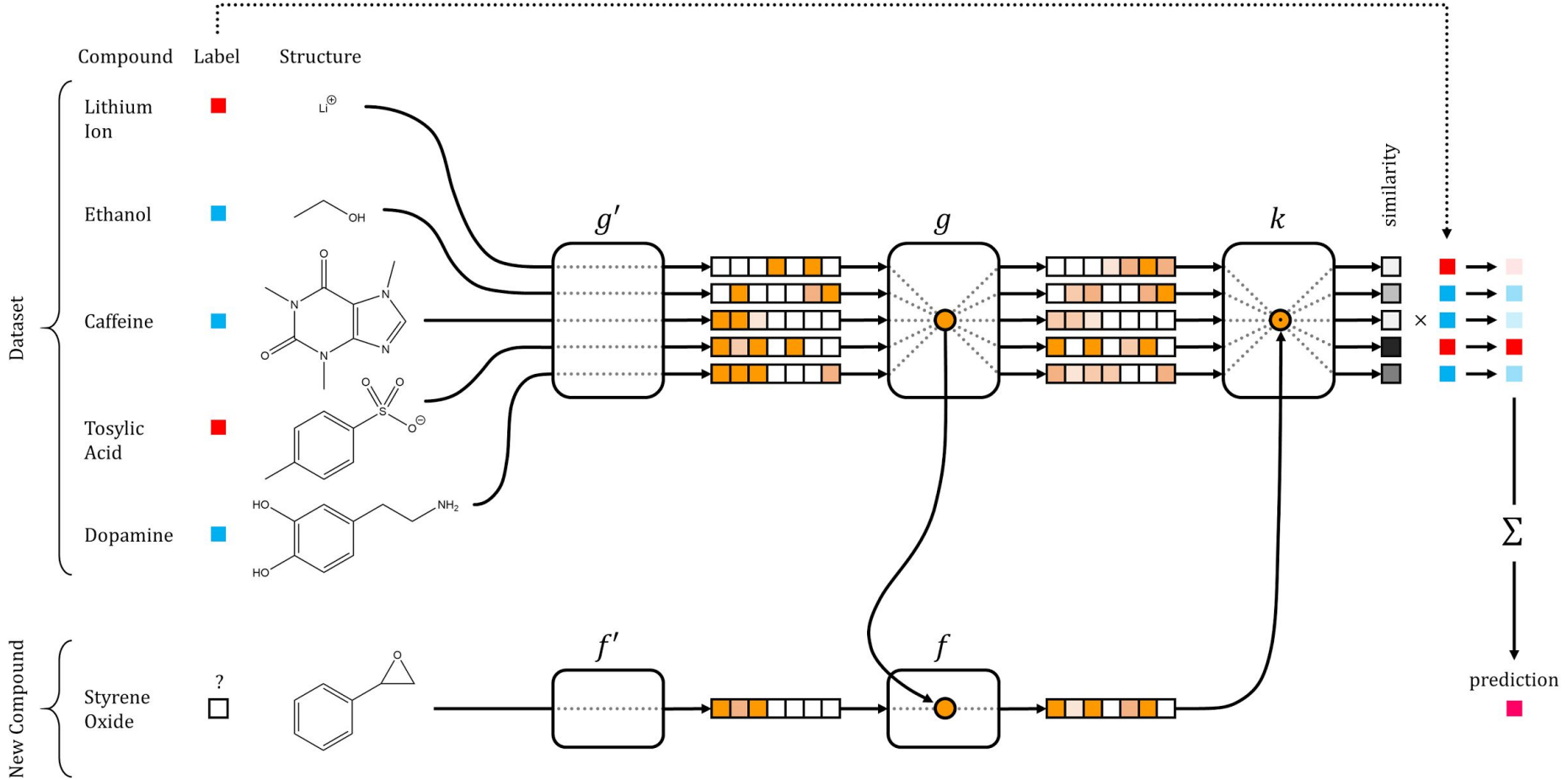

Compound
Label
Structure
Dataset
Lithium Ion
Ethanol
Caffeine
Tosylic Acid
Dopamine
New Compound
Styrene Oxide
?
g′
g
k
f′
f
similarity
Σ
prediction

Table 1: ROC-AUC scores of models on median held-out task for each model on Tox21. Numbers reported are means and standard deviations. Randomness is over the choice of support set; experiment is repeated with 20 support sets. The appendix contains results for all held-out Tox21 tasks. The result with highest mean in each row is highlighted. The notation 10+/10- indicates supports with 10 positive examples and 10 negative examples.

Tox21	RF (100 trees)	Graph Conv	Siamese	AttnLSTM	IterRefLSTM
10+/10-	0.586 ± 0.056	0.648 ± 0.029	0.820 ± 0.003	0.801 ± 0.001	$\mathbf{0.823 \pm 0.002}$
5+/10-	0.573 ± 0.060	0.637 ± 0.061	0.823 ± 0.004	0.753 ± 0.173	$\mathbf{0.830 \pm 0.001}$
1+/10-	0.551 ± 0.067	0.541 ± 0.093	$\mathbf{0.726 \pm 0.173}$	0.549 ± 0.088	0.724 ± 0.008
1+/5-	0.559 ± 0.063	0.595 ± 0.086	0.687 ± 0.210	0.593 ± 0.153	$\mathbf{0.795 \pm 0.005}$
1+/1-	0.535 ± 0.056	0.589 ± 0.068	0.657 ± 0.222	0.507 ± 0.079	$\mathbf{0.827 \pm 0.001}$

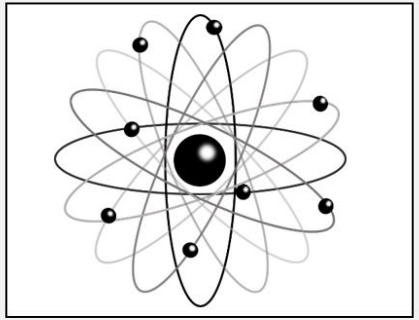
MoleculeNet: https://moleculenet.ai

Zhenqin Wu*†, Bharath Ramsundar*†, Evan N. Feinberg**†, Joseph Gomes**†, Caleb Geniesse†, Evan N. Feinberg†, Aneesh S. Pappu†, Evan N. Feinberg†, Karl Leswing‡, Vijay Pande†

* : equal first authors
**: equal second authors
†: Stanford University
‡: Schrodinger Inc.

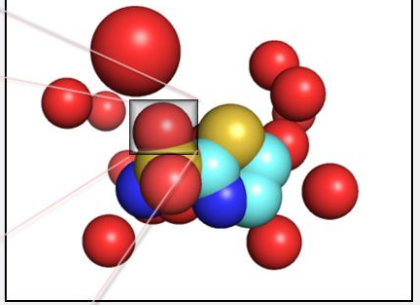
Accepted by Chemical Science.

MoleculeNet Datasets



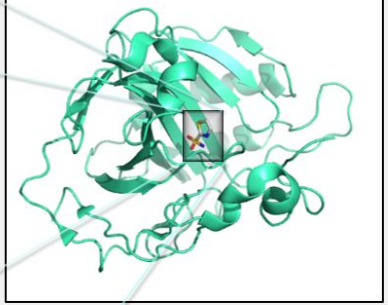
Quantum Mechanics

- QM7
- QM7b
- QM8
- QM9



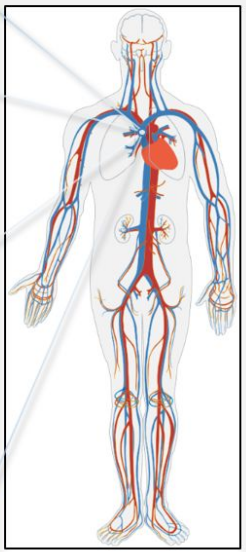
Physical Chemistry

- ESOL
- FreeSolv
- Lipophilicity



Biophysics

- HIV
- PDBbind
- BACE
- PCBA
- MUV

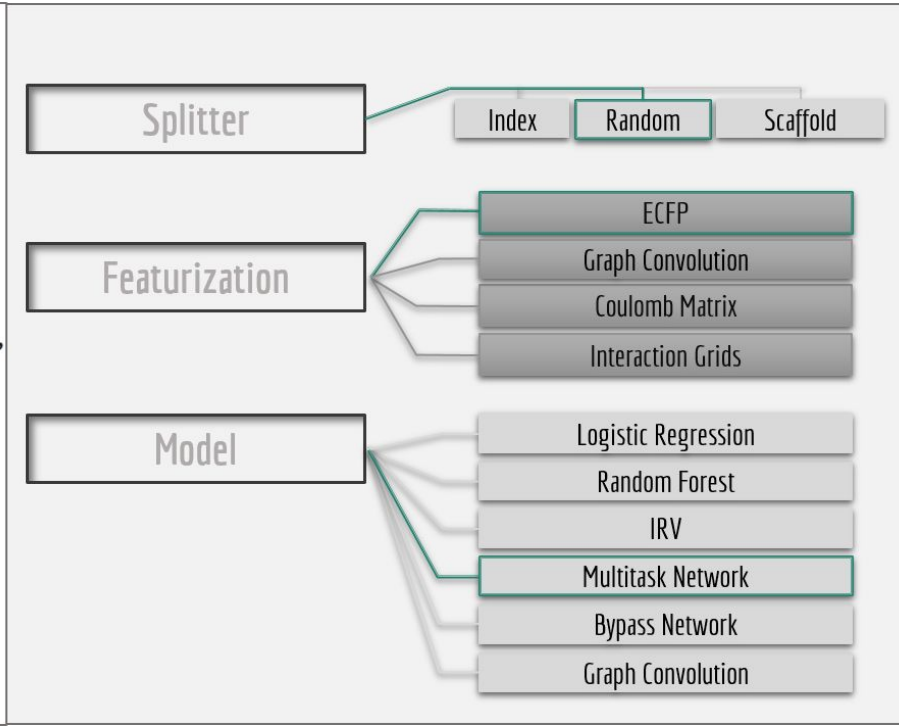


Physiology

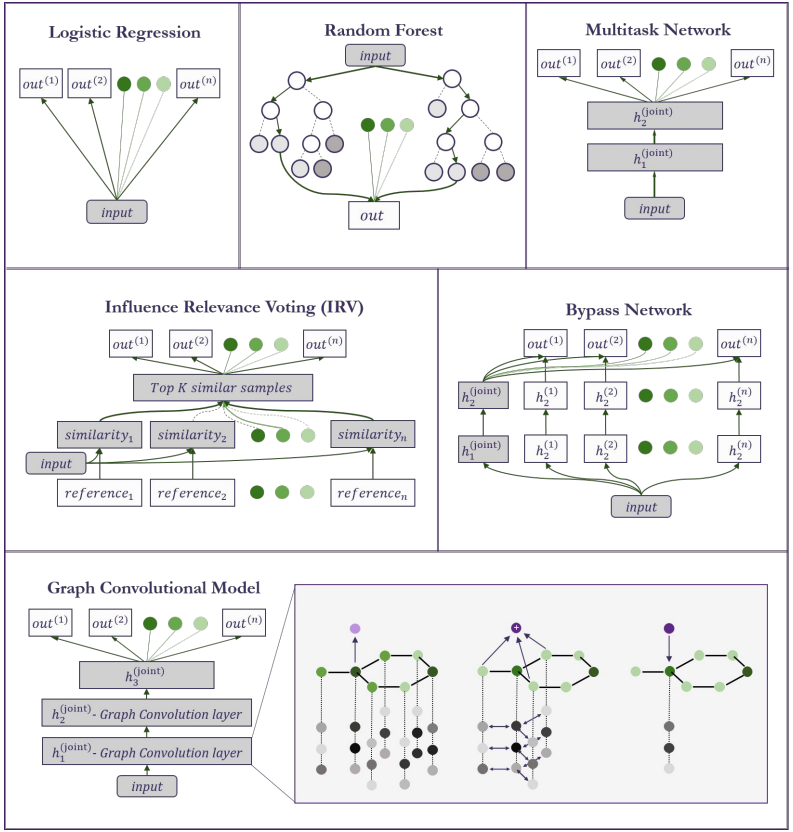
- BBBP
- Tox21
- ToxCast
- SIDER
- ClinTox

MoleculeNet uses DeepChem to Benchmark

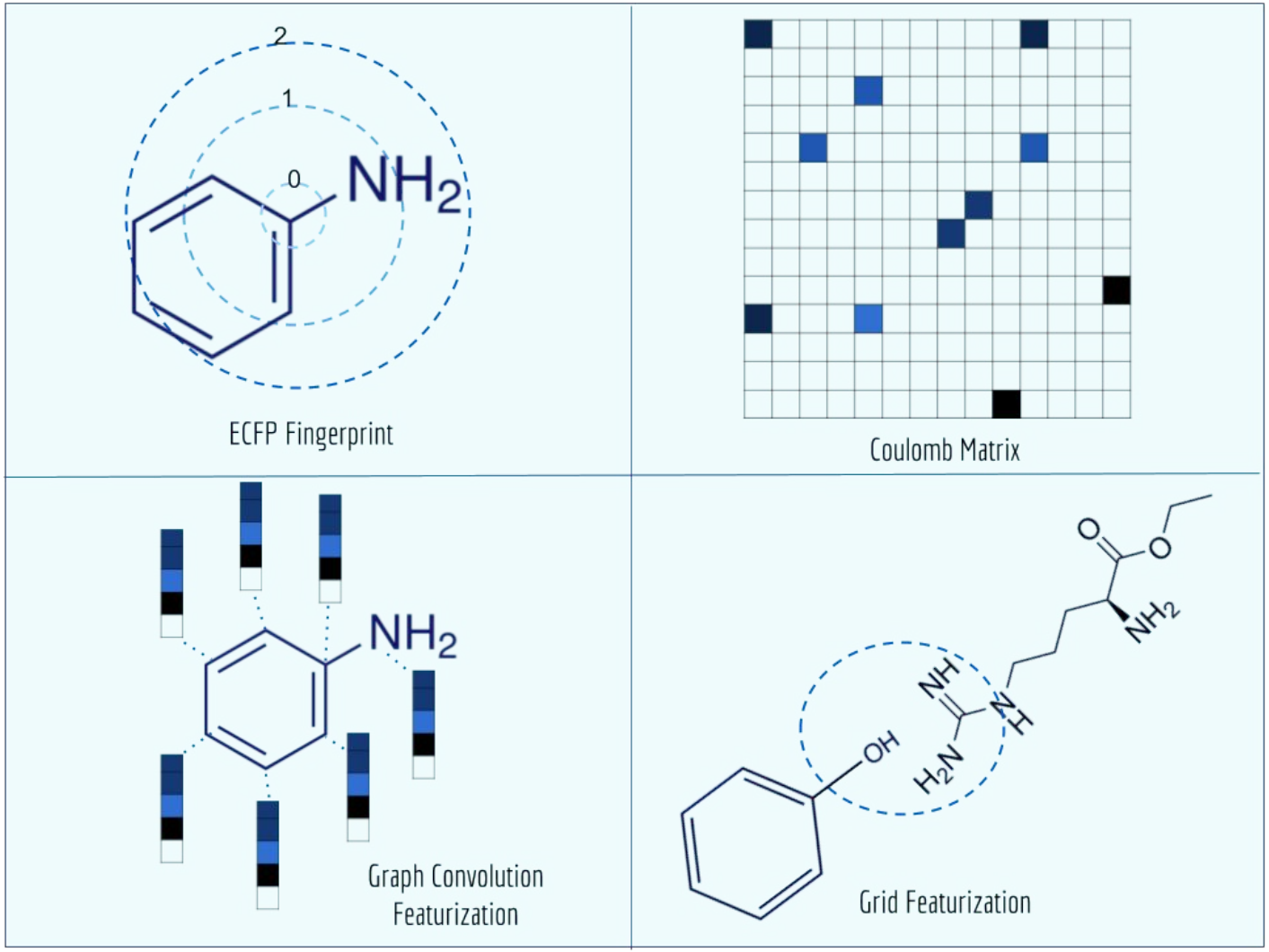
```python
import numpy as np
import deepchem as dc
from clintox_datasets import load_clintox

# Load clintox dataset
n_features = 1024
clintox_tasks, datasets, transformers = load_clintox(split='random',
                                                     featurizer='ECFP')
train_dataset, valid_dataset, test_dataset = datasets
# Defining metrics
metric = dc.metrics.Metric(dc.metrics.roc_auc_score, np.mean)
# Defining models
model = dc.models.TensorflowMultiTaskClassifier(len(clintox_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)
test_scores = model.evaluate(test_dataset, [metric], transformers)
print("Train scores")
print(train_scores)
print("Validation scores")
print(valid_scores)
print("Test scores")
print(test_scores)
```



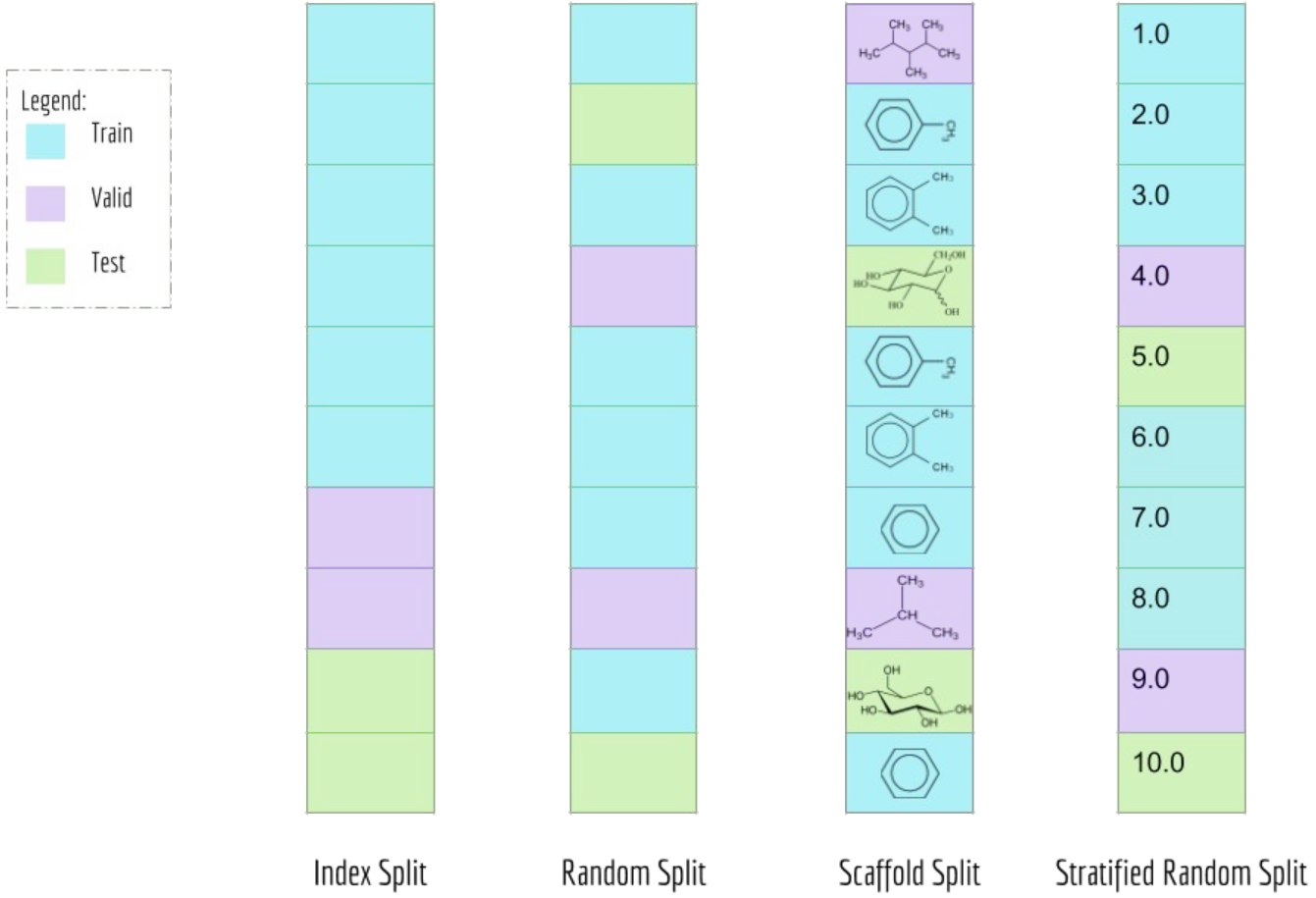
Machine Learning Models Tested for MoleculeNet Benchmarks



MoleculeNet Featurizations

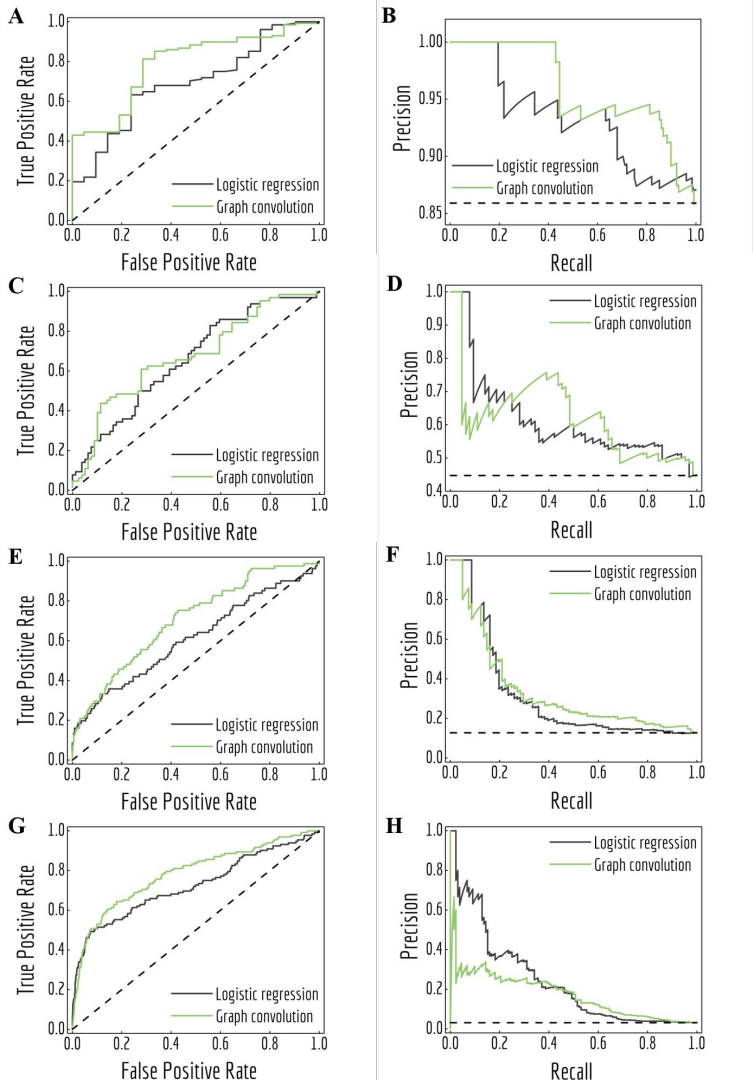


MoleculeNet Tests Generalization with Different Splits



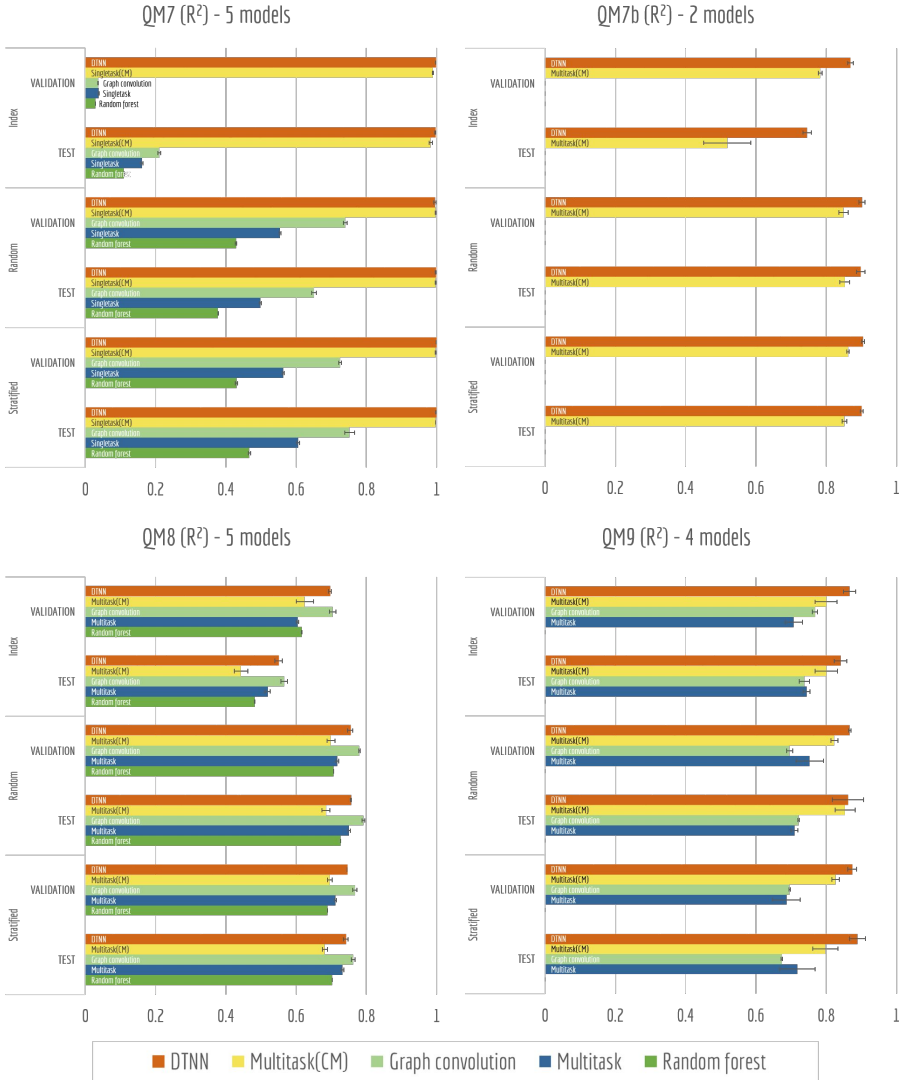
ROC vs PRC

- Most deep drug discovery benchmarked with ROC-AUC
- However, analysis shows that for very imbalanced datasets, ROC and PRC behave differently.
- Deep methods do not currently win on very imbalanced datasets!

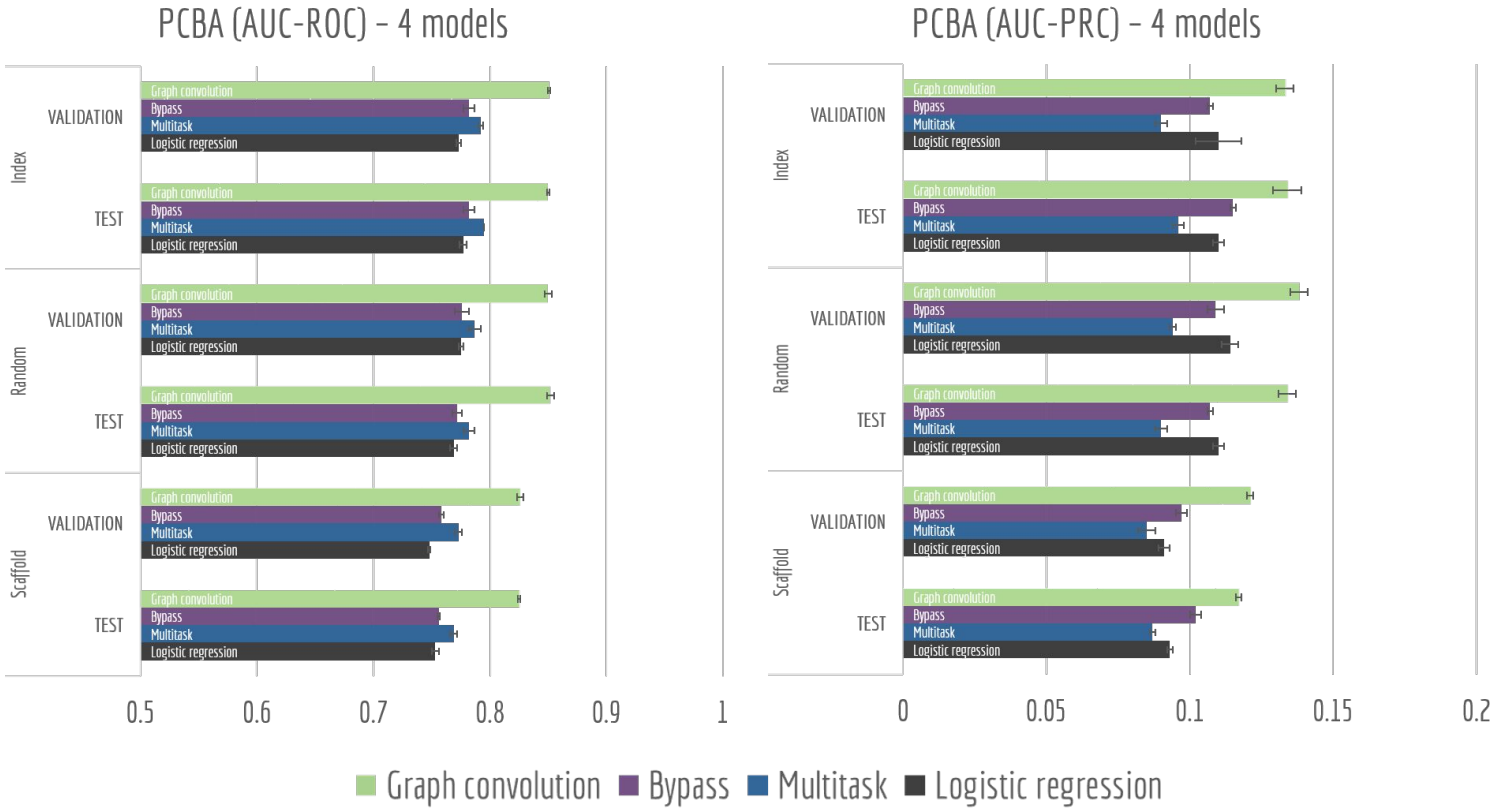


Quantum Datasets

- Deep Learning methods dominate benchmark.
- End-to-end DTNN wins overall, with hand engineered DNN+Coulomb a close second.

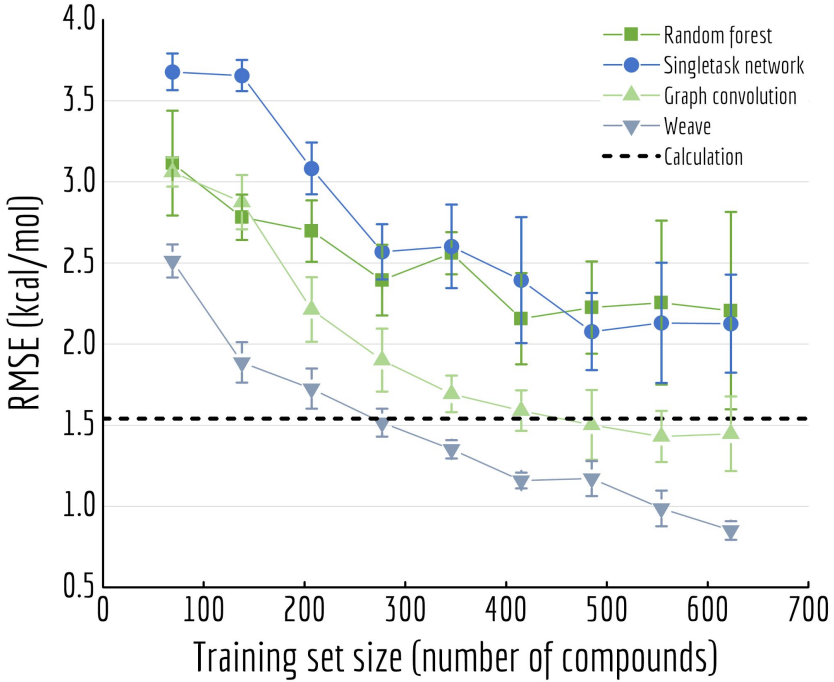


PubChem BioAssay



Molecular ML vs. Physical Computations

- Solubility Data
- Dashed Line measures error of physics-based simulation algorithm.
- 1000 datapoints in set. Plots show that Molecular ML can out-perform physics-based with enough data.



Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity

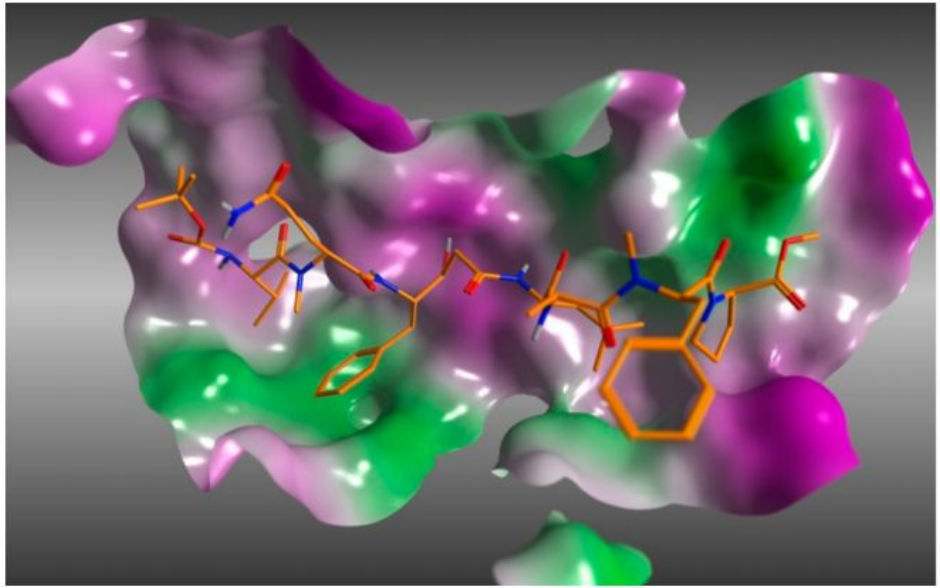
Joseph Gomes*†, Bharath Ramsundar*†, Evan N. Feinberg†, Vijay Pande†

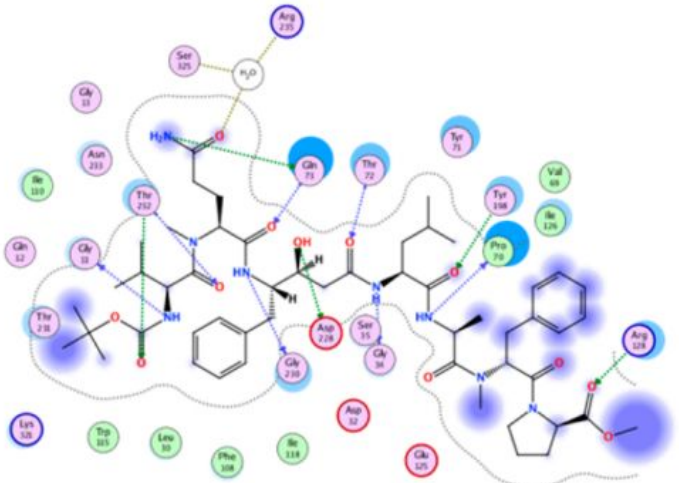
* : equal contribution
†: Stanford University

Currently on Arxiv. Tentatively plan to submit to ACS Central Science.

Example of BACE-1 Binding Pocket Interactions

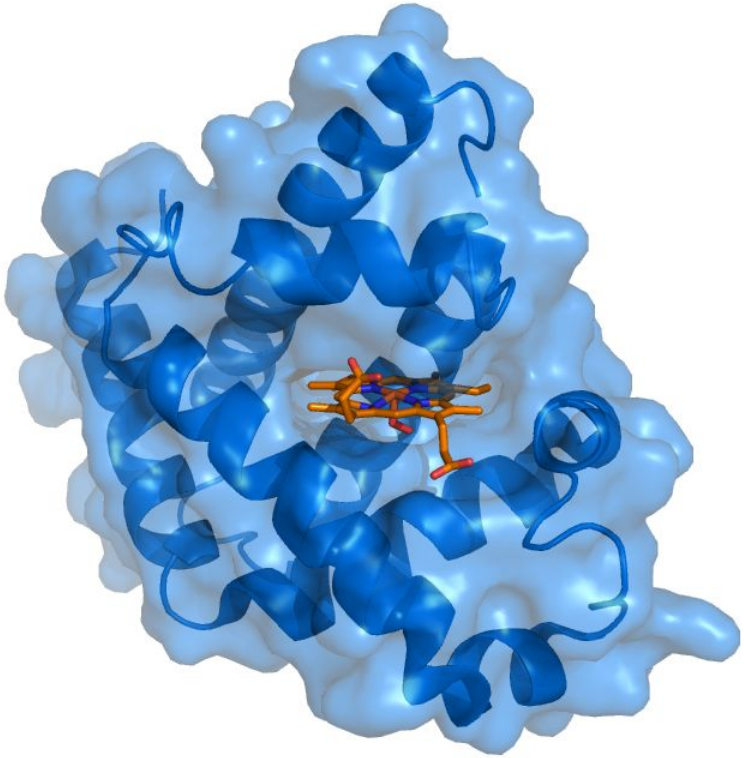
Scheme 1. Depiction of BACE-1 Binding Site (left) Using the Ligand from PDB Code 3UQP along with the Protein–Ligand Interaction (right)



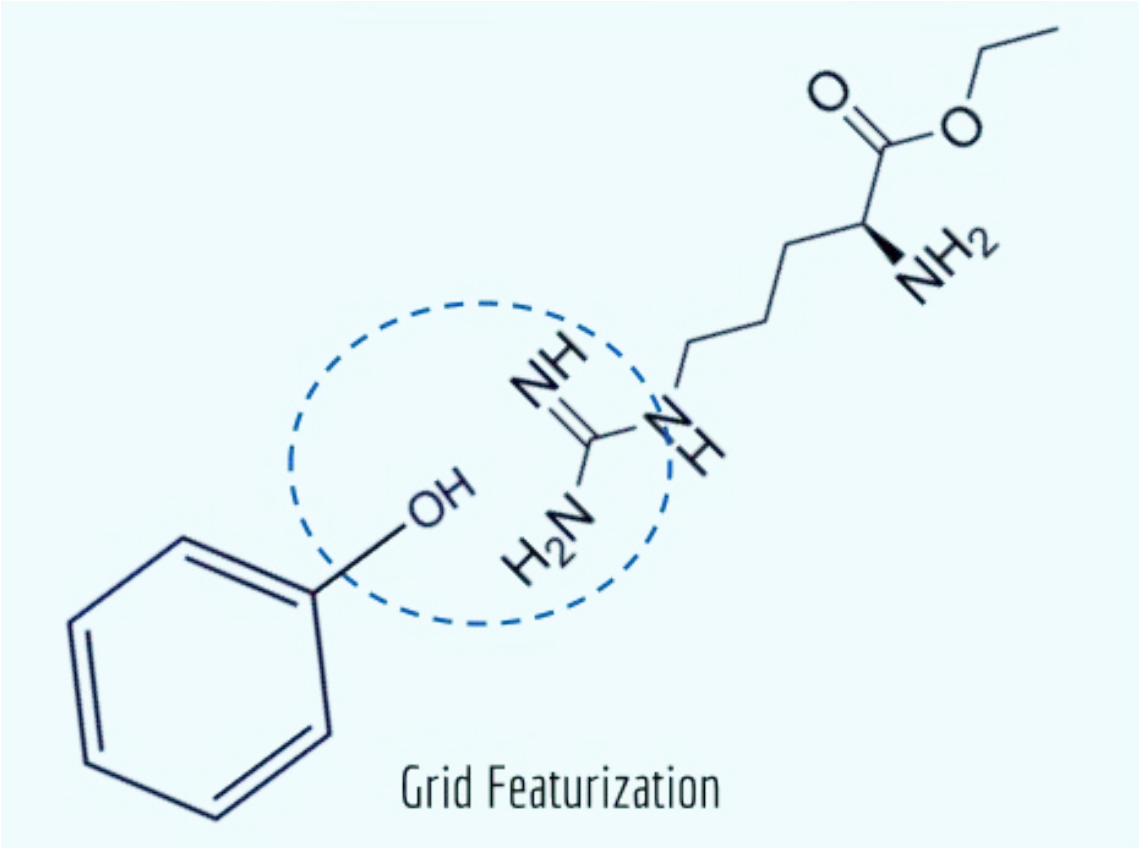


Exploiting 3D Geometry for Drug Discovery

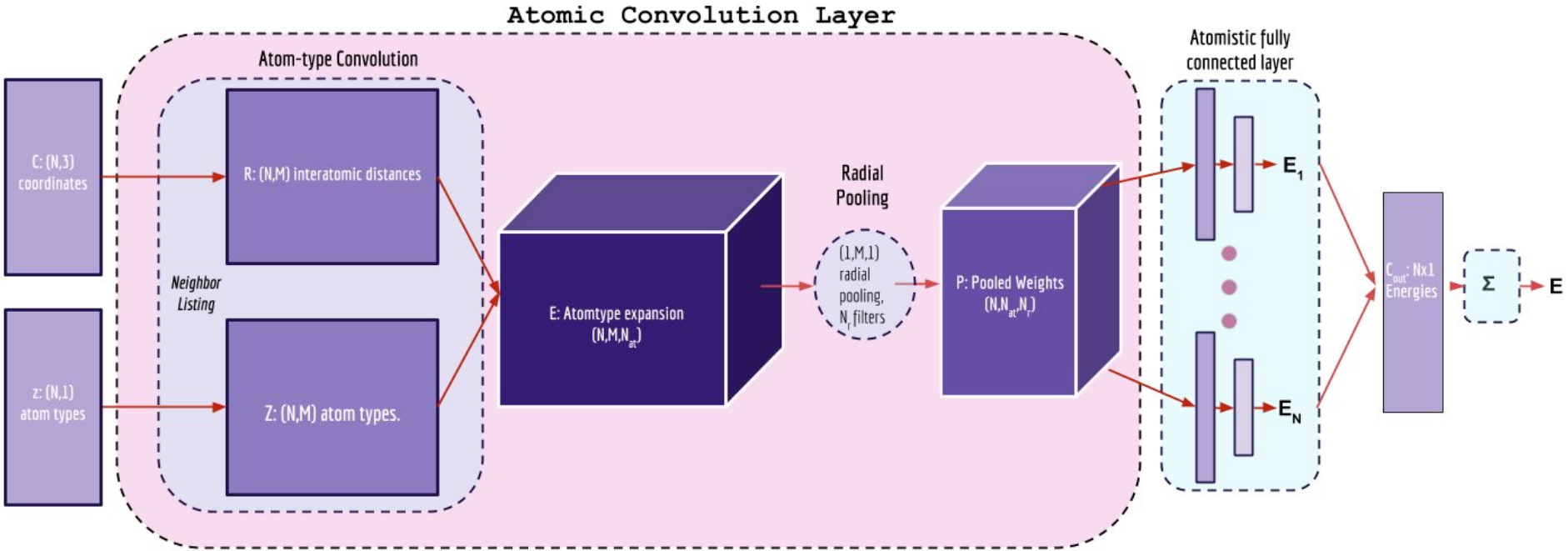
- Learning to predict interaction energies requires modeling the 3D geometry of binding interactions.
- Two approaches: hand tuned and learned featurizations.



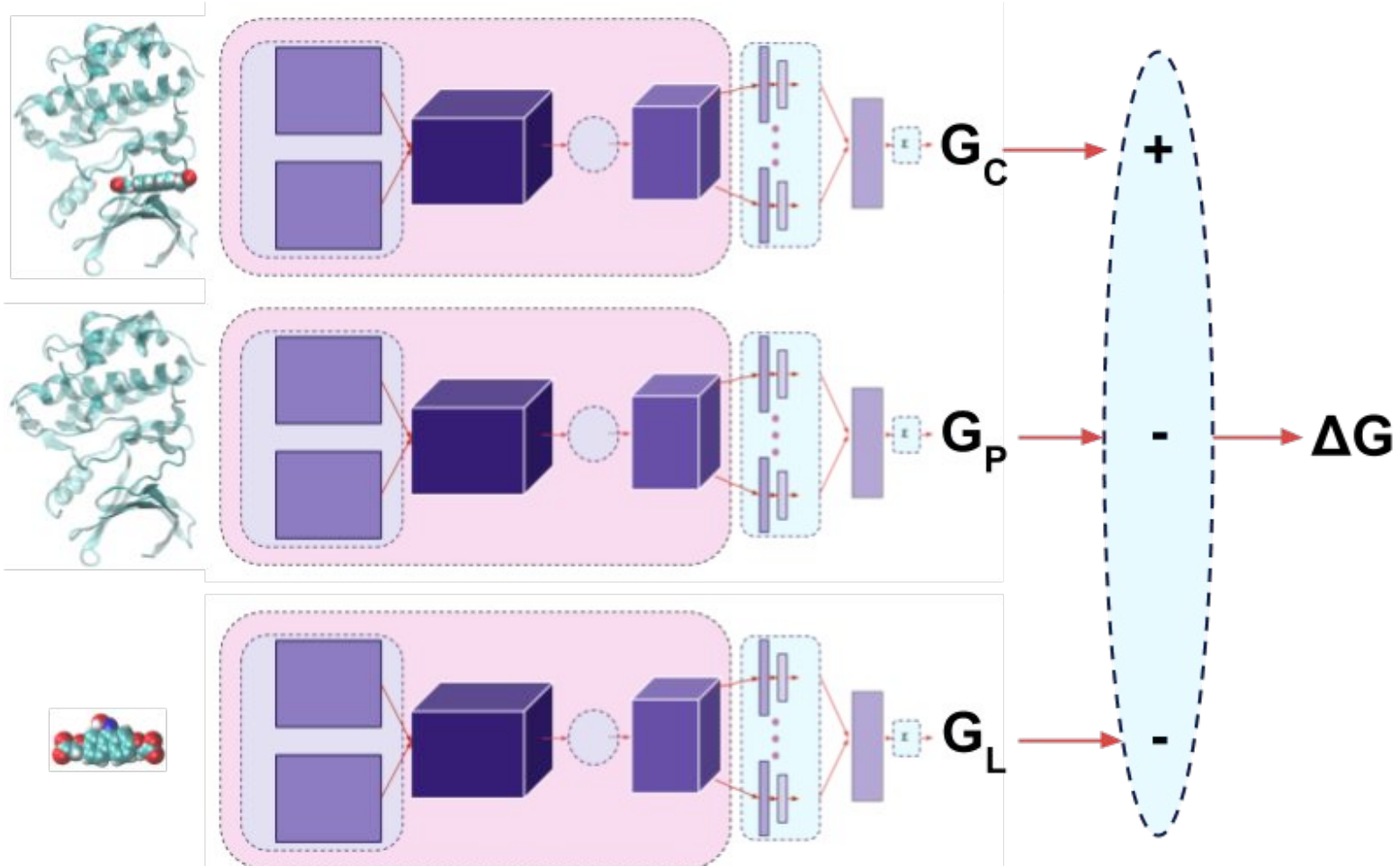
Hand-Designed Grid Featurization



Learnable Atomic Convolutions



Thermodynamic Cycles with End-to-End Training



Results on PDBBind Benchmark

	ACNN		GRID-RF		GRID-NN		GCNN		ECFP-RF		ECFP-NN	
Split	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Random	.705	.508	.962	**.546**	.976	.539	.581	.403	.883	.466	.850	.386
Stratified	.793	.491	.963	**.562**	.983	.494	.635	.410	.883	.477	.840	.462
Scaffold	.752	.267	.964	**.349**	.977	.334	.652	.135	.861	.229	.824	.146
Temporal	.704	**.529**	.963	.486	.978	.455	.596	.315	.875	.401	.827	.362

Table 3. Performance (Pearson R^2) on PDBBind refined train/test sets.

	ACNN		GRID-RF		GRID-NN		GCNN		ECFP-RF		ECFP-NN	
Split	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Random	0.518	0.653	0.240	**0.630**	0.468	1.208	0.598	0.726	0.319	0.661	0.335	0.740
Stratified	0.540	0.686	0.237	**0.650**	0.439	1.109	0.559	0.710	0.319	0.660	0.344	0.678
Scaffold	0.555	0.579	0.245	**0.537**	0.406	1.177	0.557	0.805	0.339	0.749	0.373	0.856
Temporal	0.526	0.668	0.235	**0.666**	0.410	1.184	0.610	0.865	0.321	0.770	0.371	0.798

Table 4. Performance (MUE [kcal/mol]) on PDBBind refined train/test sets.