Applying Automated Machine Learning to Drug Discovery

Karl Leswing

Tech Lead Machine Learning

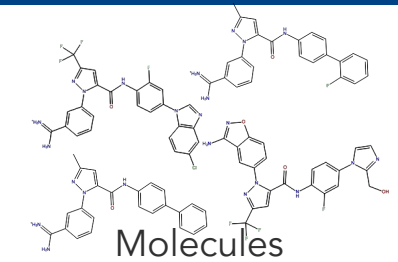
Webinar October 31

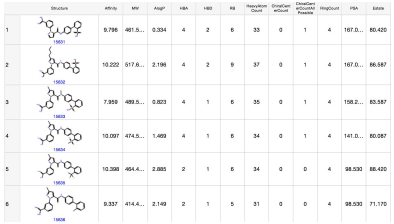
QSAR Modeling in the Chemical Enterprise

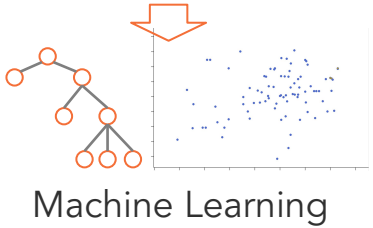
- QSAR models are widely employed for a wide variety of key properties
- Long track record of impact in pharmaceutical projects

Opportunities for improvement

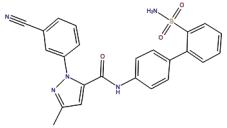
- Creating models is largely an expensive, expert activity
  - No single machine learning method or descriptor set is ideal for all properties
  - Significant human time can be spent experimenting to identify high-performing models
  - Domain experts possess deepest understanding of data to be modeled
- QSAR models can be non-trivial to deploy
- No single machine learning approach is ideal for modeling 10s to 10s of millions of compounds







Property Prediction



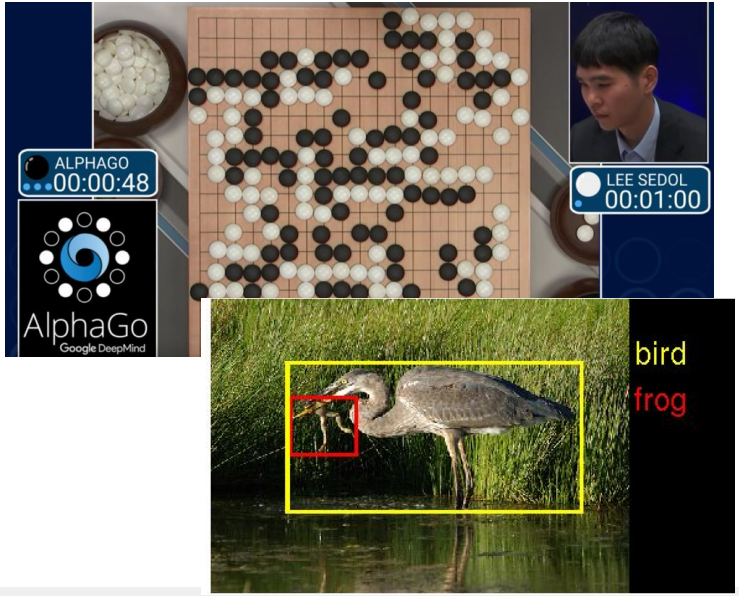
QSAR Model

Democratizing QSAR Modeling with AutoQSAR

- QSAR "expert in a box" to automatically create and validate predictive models
  - Ensure input data adequacy
  - Automated best practices workflow
    - Descriptor generation, feature selection, use of multiple machine learning methods, automated training/test set splits
  - Methods to minimize overfitting
  - Advanced modeling approaches such as consensus methods
  - Assessment of applicability domain
- Easily deploy predictive models
  - Don't need to create scripts to generate descriptors and run machine learning method for each QSAR model
  - Simple command line, desktop and web app deployment

Deep Learning

- Deep learning methods are becoming very popular in image recognition, game playing, and question and answer systems.



More Deep Learning Hype

SECTIONS LATEST POPULAR SEARCH THE NEW YORKER SIGN IN TNY STORE

ANNALS OF MEDICINE APRIL 3, 2017 ISSUE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

By Siddhartha Mukherjee

It's just completely obvious that in five years deep learning is going to do better than radiologists. Hospitals should stop training radiologists now.

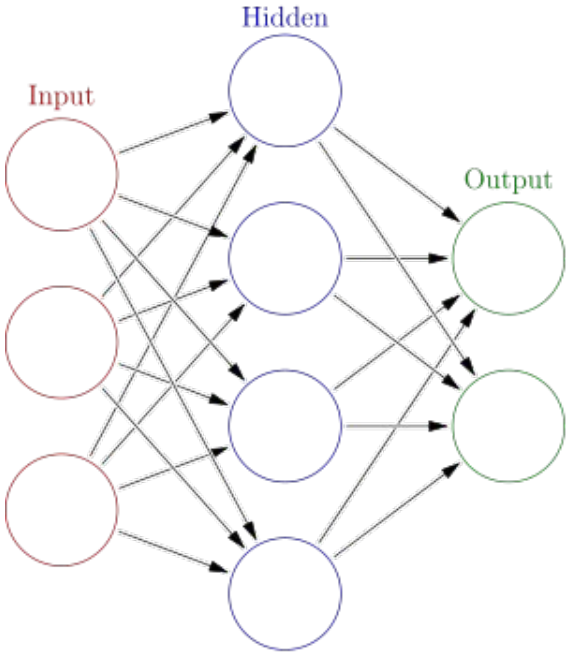
- Geoffery Hinton

Deep Learning

- Lots of excitement to try to use these methods in other contexts
- Should deep learning be used in drug discovery?
- Where does it provide the greatest benefit?

Artificial Neural Network Overview

- Collection of units called neurons (Circles Here)
- Each neuron computes a function over its inputs (real numbers)
- Each neuron and can be connected to multiple outputs
- Trained using back propagation

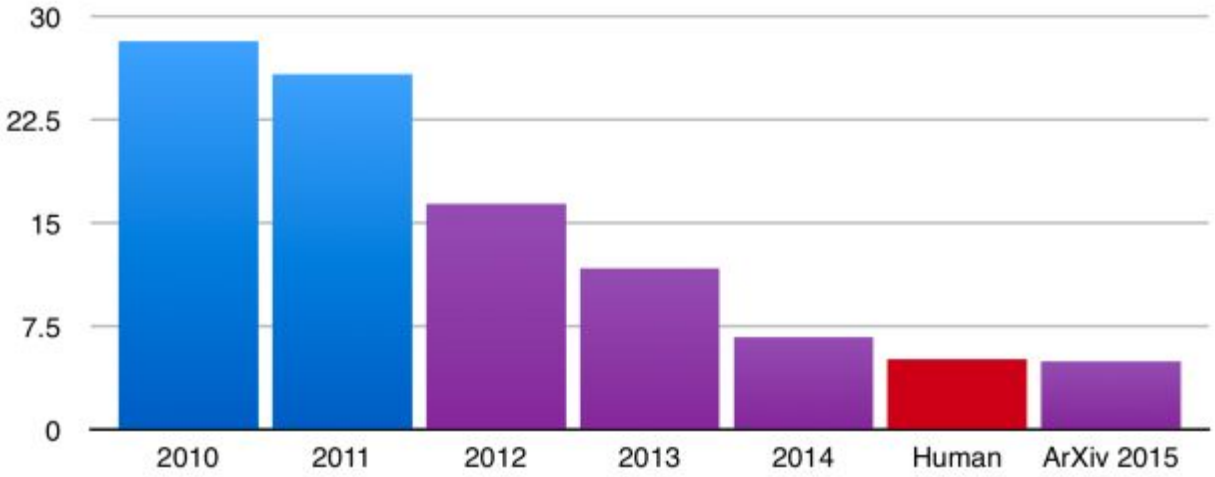


Universal Function Approximation Theorem

- Artificial Neural Networks can represent ANY function
- This does not pan out in practice
  - Limited data and compute power
- Requires us to create data and compute efficient models.

Deep Neural Network Image Classification

ImageNet Large Scale Visual Recognition Challenge Model Accuracy



As of 2015, a 27 layer DNN was more accurate than a human (Stanford student) at sorting 100,000 images into 1,000 different pre-specified categories

Deep Neural Network Image Classification

- The ImageNet classification challenge is *very* difficult:


Ruler


King crab

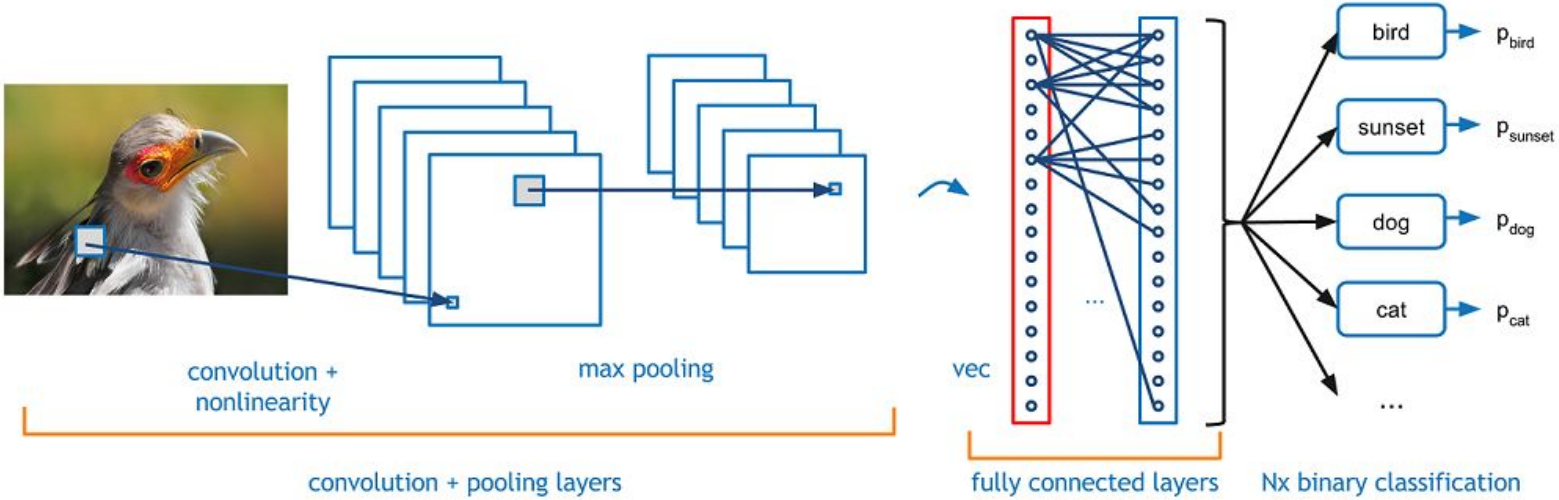

Sidewinder

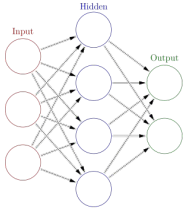
Salt shaker


Reel


Hatchet

Convolutional Neural Networks





Convolution Layer

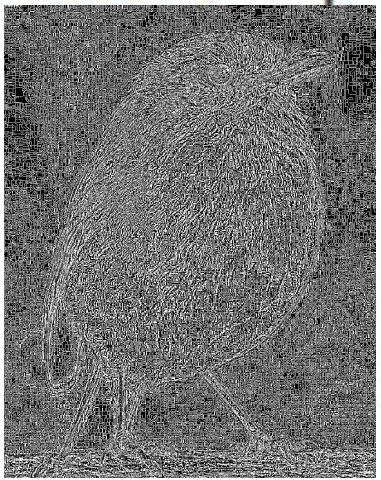
- Slide a learnable mask across the image.

Input image

Convolution Kernel

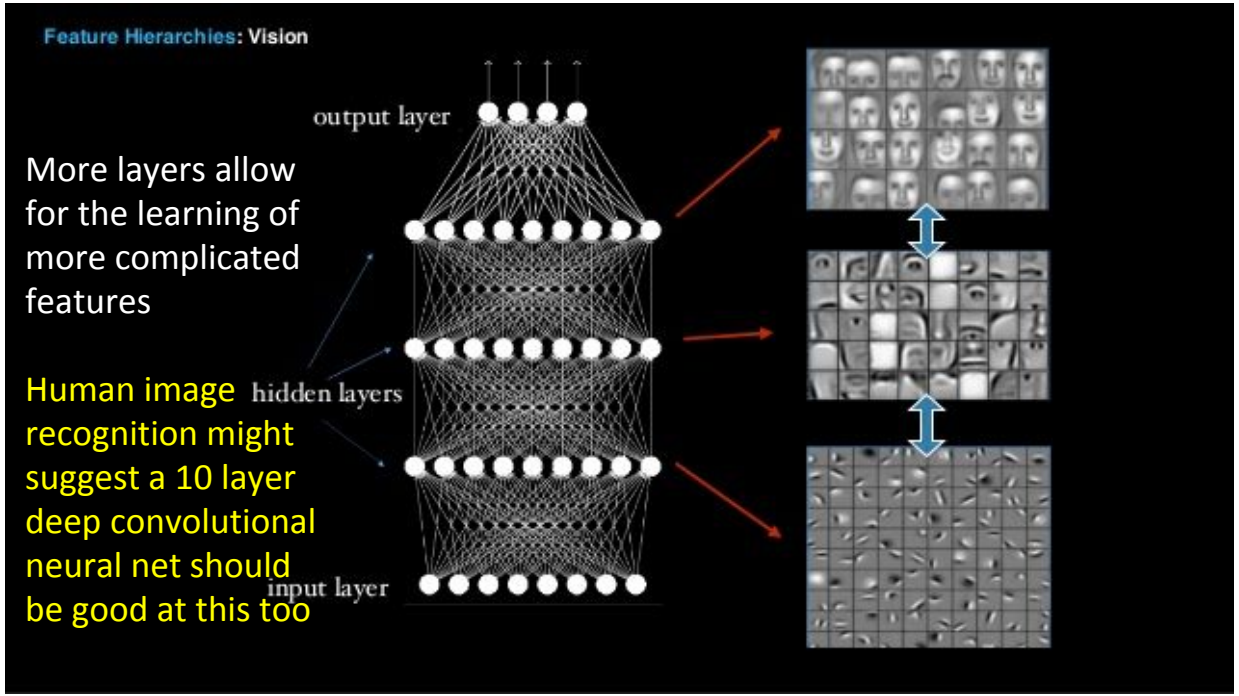
$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Feature map



Deep Neural Network Image Classification

- A unique aspect of Deep Learning is the ability learn new features as the network is trained:



DeepChem

- Started as a Pande group (Vijay Pande Lab) project at Stanford

- Aims to provide a high quality open-source toolchain that democratizes the use of deep-learning in drug discovery, materials science, and quantum chemistry.
  - GPU Enabled Algorithms
  - Built on top of Google TensorFlow

github.com/deepchem/deepchem

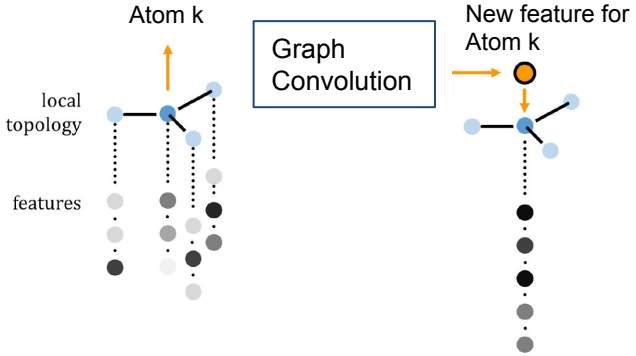
AutoQSAR w/ DeepChem Feature Generation

2D Graphic description of molecules

- Each node represents an atom
- Each edge represents a bond
- Atom features include atoms-type, valences, formal charges, and hybridization

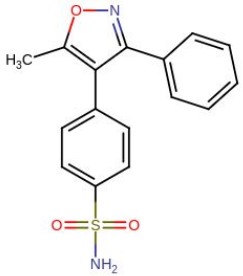
Graph Convolution

- **Automatically learn new local features that suit the endpoint**
- These new features are then converted to molecular feature which is feed to dense neural network for model building

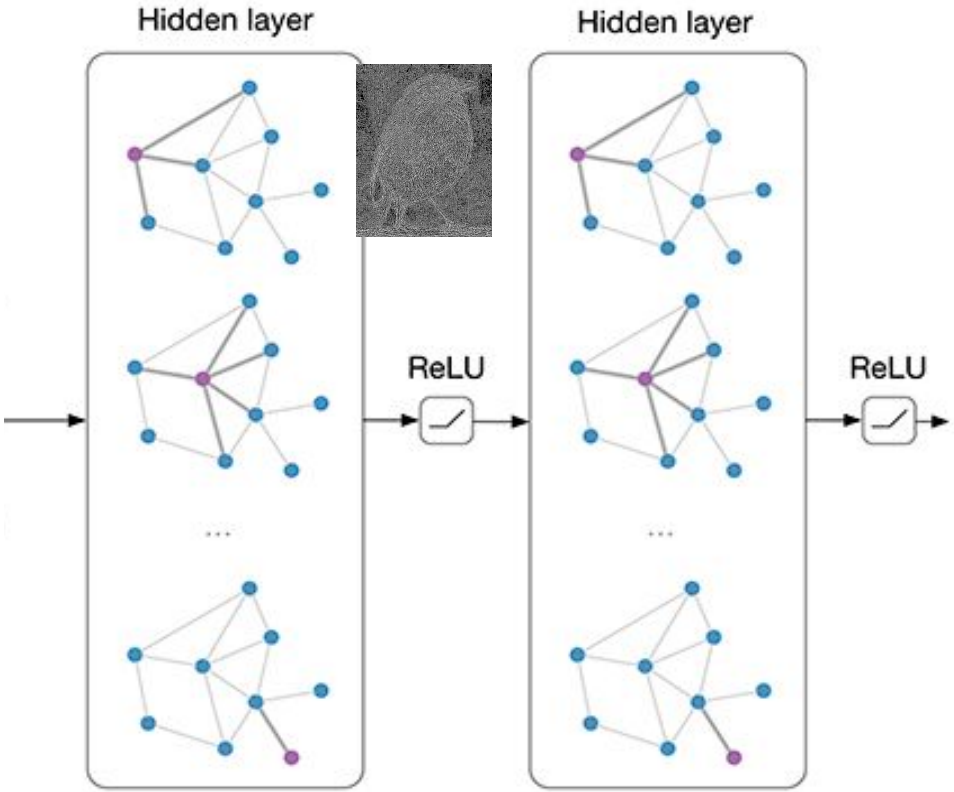


Graph Convolutions

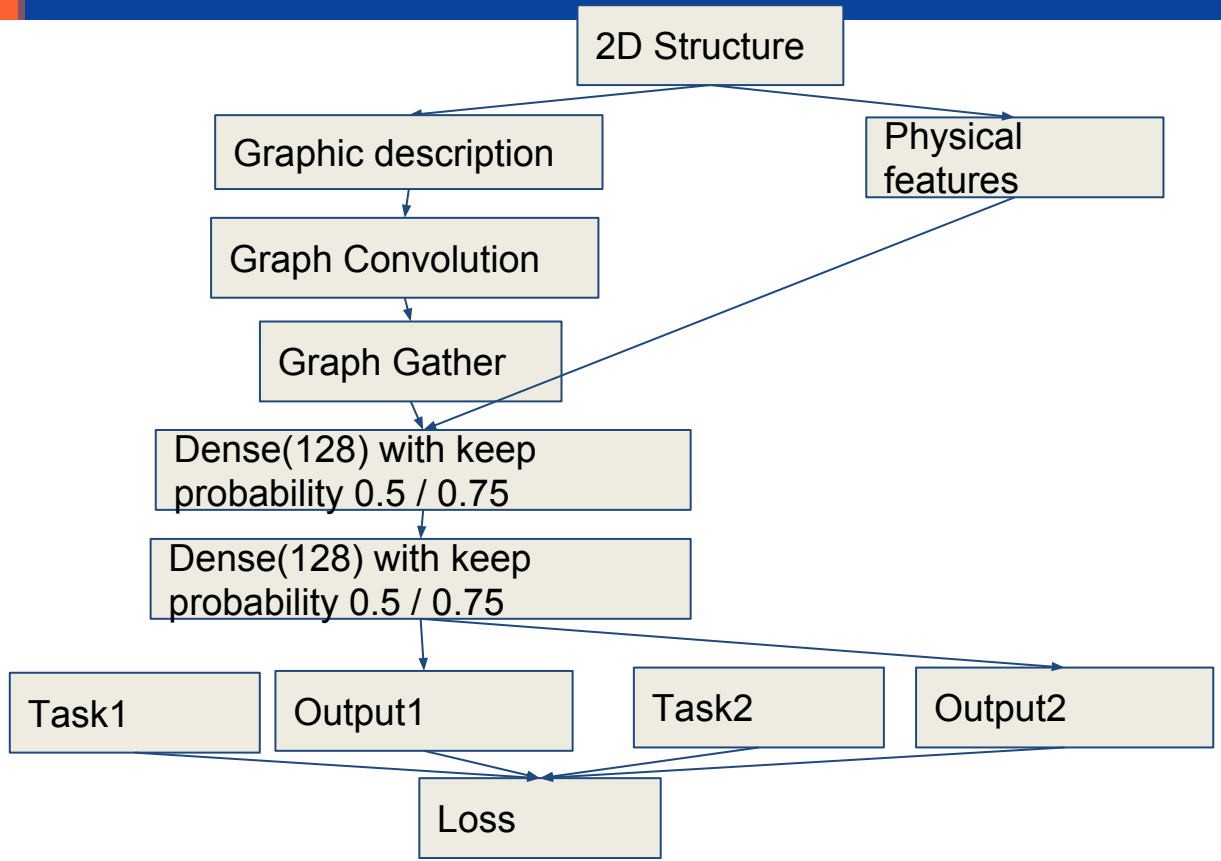
Input



Valdecoxib (pain relief)



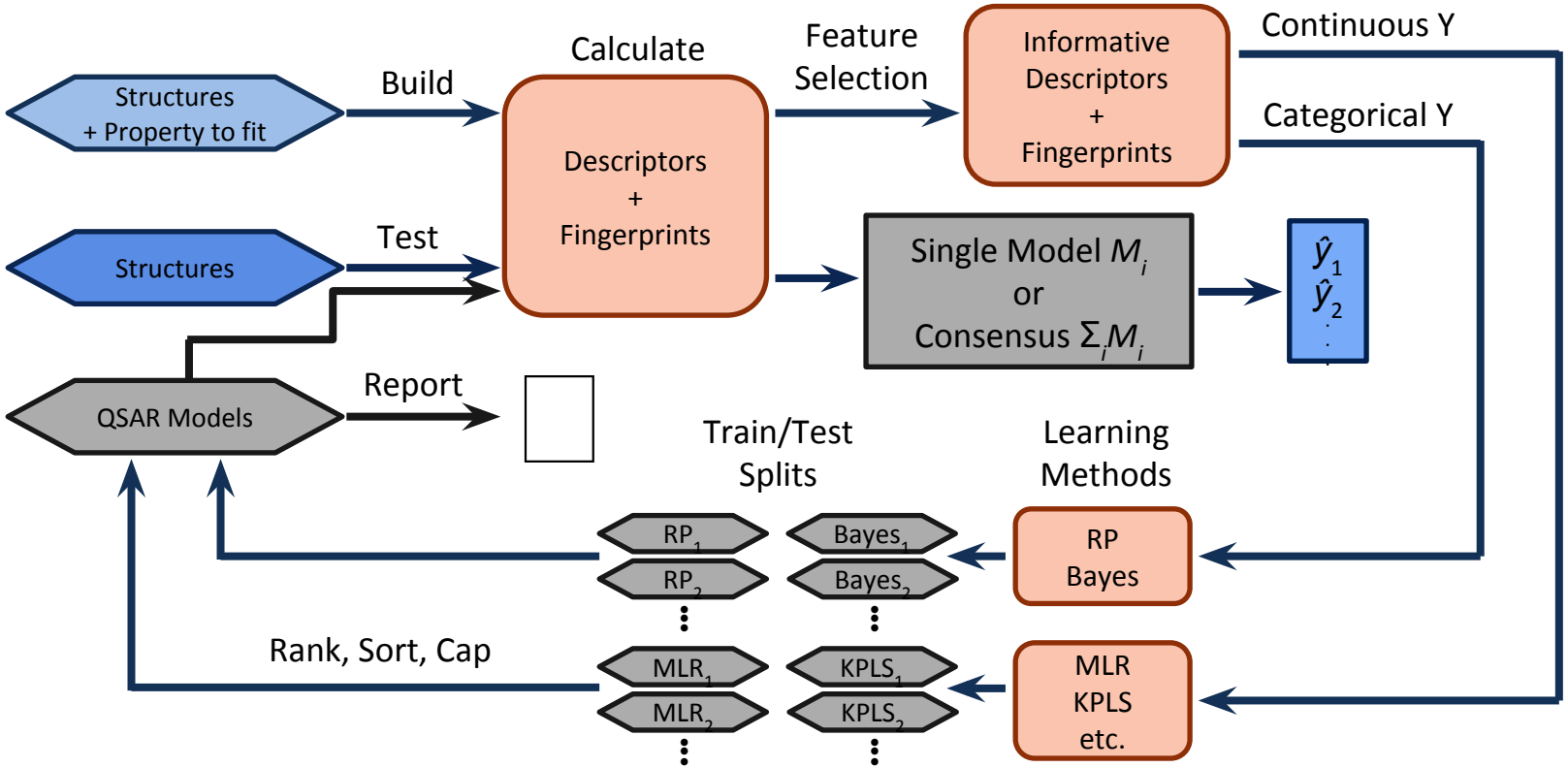
AutoQSAR w/ DeepChem Model Architecture



Model details:

- Physical features are optional
- Training the model by minimizing the loss functions

Traditional AutoQSAR



Results comparison -- Datasets

FUTURE MEDICINAL CHEMISTRY, VOL. 8, NO. 15 | RESEARCH ARTICLE

normal

AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling

Steven L Dixon, Jianxin Duan, Ethan Smith, Christopher D Von Bargen, Woody Sherman & Matthew P Repasky

Published Online: 19 Sep 2016 | https://doi.org/10.4155/fmc-2016-0093

MoleculeNet: A Benchmark for Molecular Machine Learning†

Zhenqin Wu,[a‡] Bharath Ramsundar,[b‡] Evan N. Feinberg,[c¶] Joseph Gomes,[c¶] Caleb Geniesse,[c] Aneesh S. Pappu,[b] Karl Leswing,[d] and Vijay Pande*[a]

- http://moleculenet.ai

Results comparison --- low data applications

Experimental setup:

- All tasks have less than 5000 data points
- 22 regression tasks
- 32 classification tasks
- Comparing with QSAR results from AutoQSAR

Metrics:

- Q^2 and MUE for regression problems
- Area under curve(AUC) for classification problems

Results comparison --- low data applications

Regression dataset description

AutoQSAR reporting publication dataset[1]

- Binding affinity data:
  - Ten IC50 datasets that cover seven different protein targets: Cyclin-dependent kinase 2 (CDK2), Checkpoint kinase 1(Chk1), Factor Xa (FXa), Heat shock protein 90 (Hsp90), Liver X receptor beta (LXR-β), Methionine aminopeptidase 2 (Metap2), and Thrombin
  - Number of ligands per data set ranges from 73 to 203
- Solubility data
  - 1708 data points
- Bioaccumulation
  - Bioconcentration factors (ratio of chemical concentration in fish to the concentration in water)
  - 589 data points

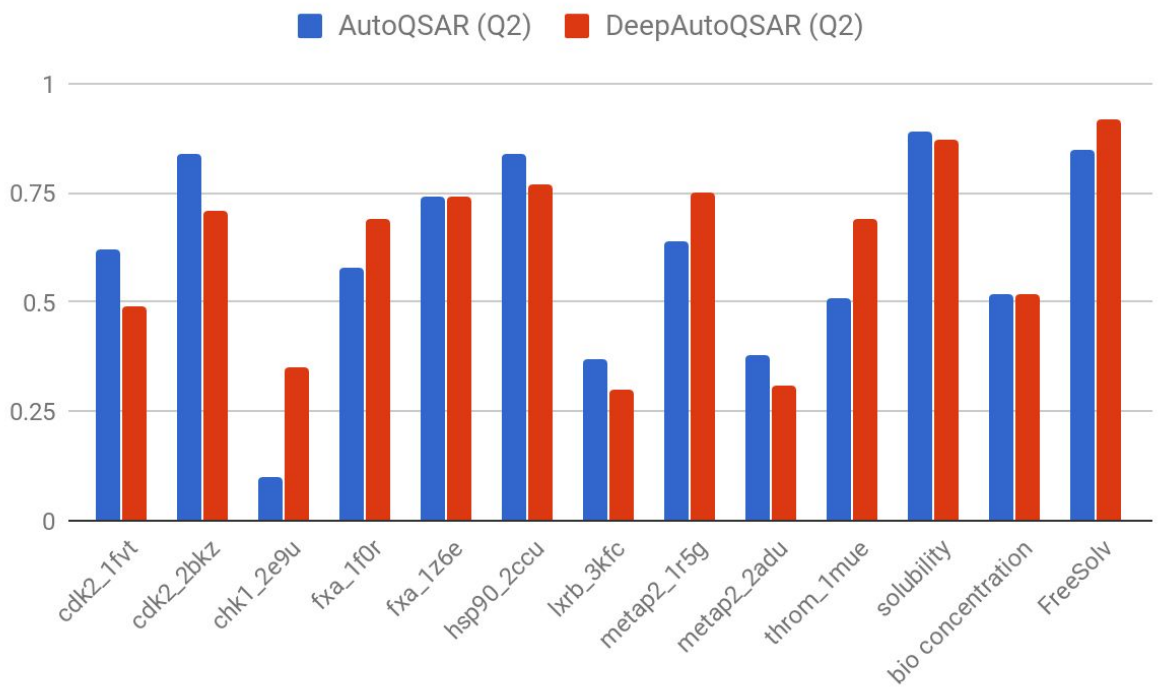
Freesolv[2]

- Solvation free energy
  - 640 data points

1 AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling. *Dixon SL, Duan J, Smith E, Von Bargen CD, Sherman W, Repasky MP Future Med Chem (2016) 8: 1825-1839*
2 FreeSolv: A database of experimental and calculated hydration free energies, with input files. *David L. Mobley and J. Peter Guthrie* J Comput Aided Mol Des. 2014 Jul; 28(7): 711–720.

Results comparison --- low data applications

Regression in Q^2



DeepChem option has similar performance to AutoQSAR in low-data regression tasks

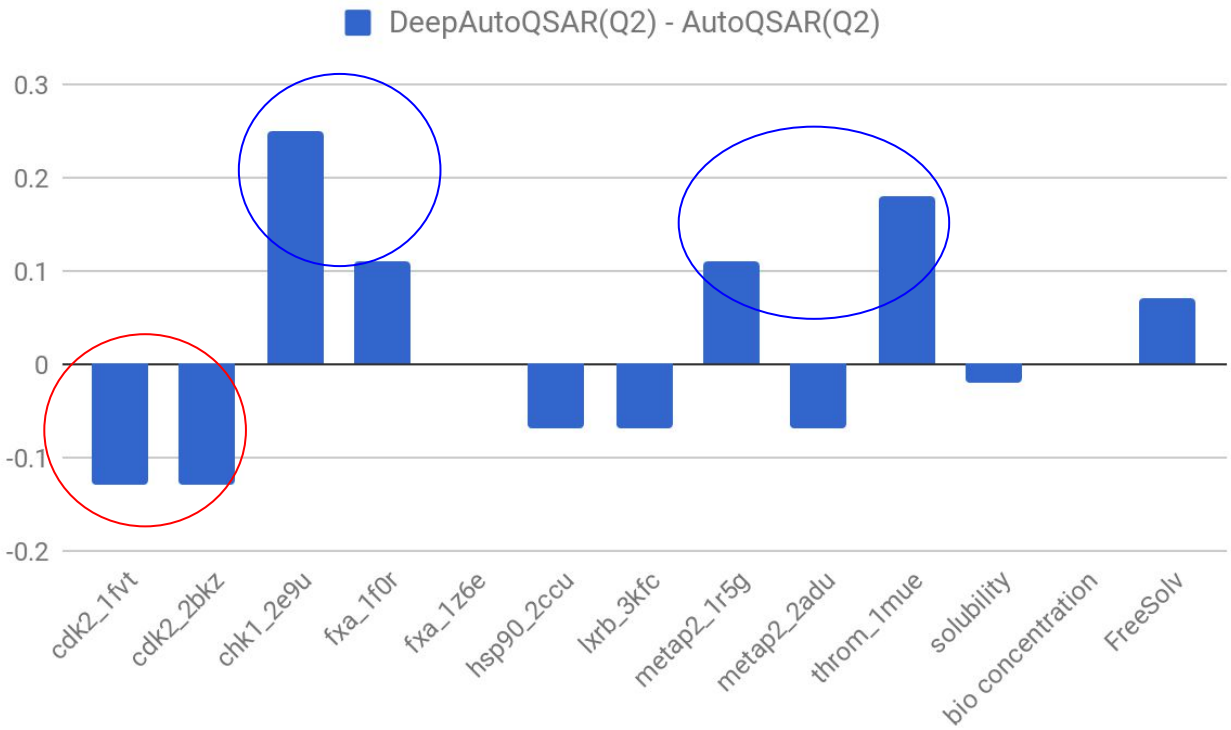
Both methods look better than they should due to random split effects
(A time-split is more reasonable)

Despite over-optimistic performance, random splits allow for head to head comparison with earlier work

Average	AutoQSAR	w/ DeepChem
Weighted by task	0.61± 0.22	0.62 ± 0.20
Weighted by data	0.73 ± 0.20	0.75 ± 0.19

Results comparison --- low data applications

Regression in Q^2

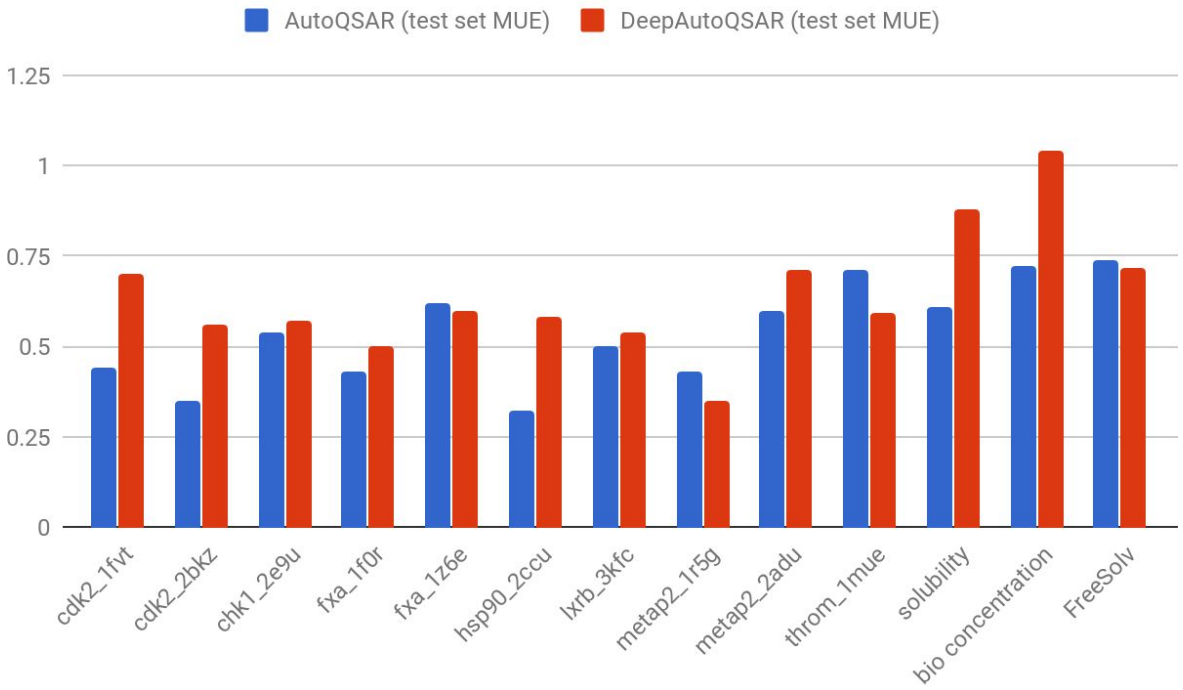


On average, DeepChem option has similar performance as AutoQSAR in regression R^2

Average	AutoQSAR	DeepAutoQAR
Weighted by task	0.61± 0.22	0.62 ± 0.20
Weighted by data	0.73 ± 0.20	0.75 ± 0.19

Results comparison --- low data applications

Regression in MUE (log unit)



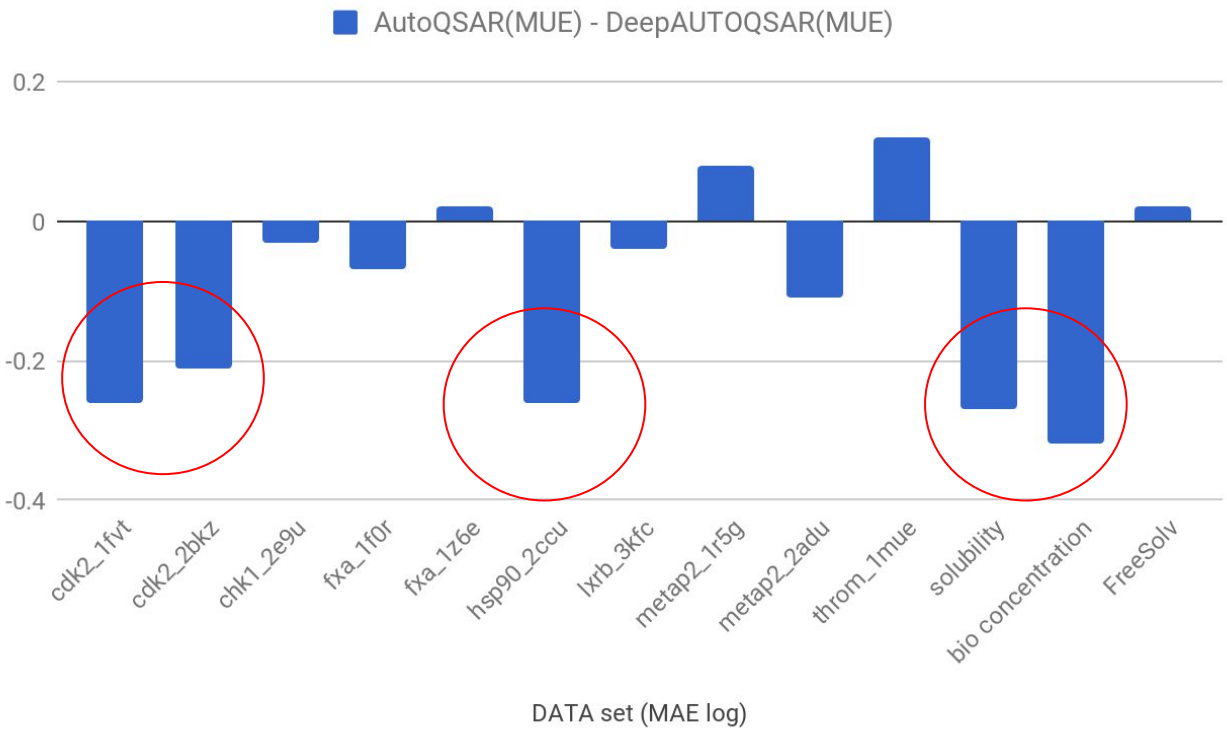
DeepChem option performs slightly worse but within error

Both methods again look better than they should due to random split effects

Average	AutoQSAR	DeepAutoQAR
Weighted by task	0.54 ± 0.13	0.64 ± 0.17
Weighted by data	0.62 ± 0.11	0.78 ± 0.18

Results comparison --- low data applications

Regression in MUE (log unit)



DeepChem option performs slightly worse but within error

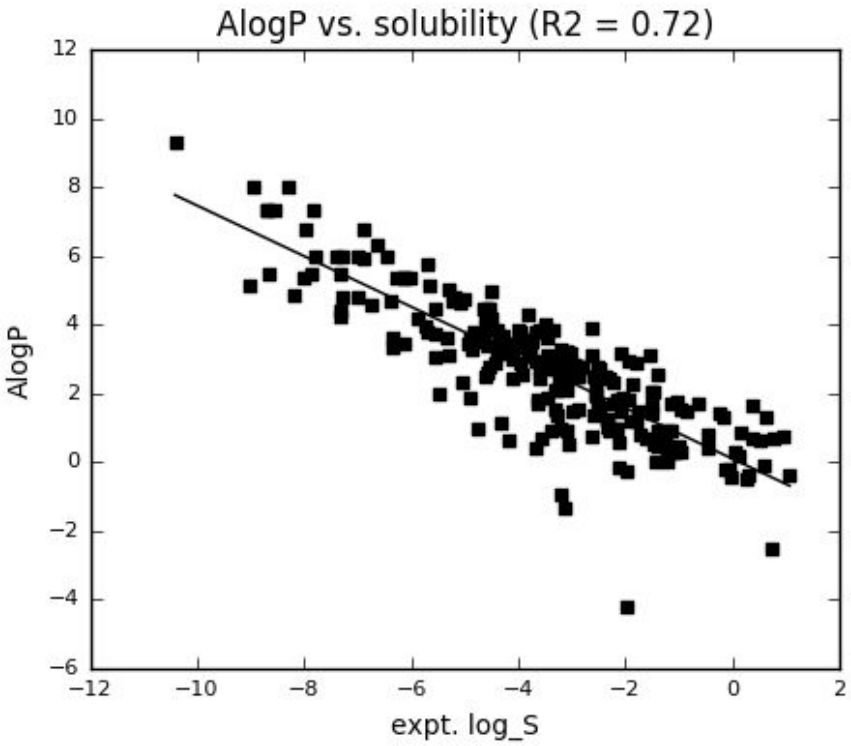
Average	AutoQSAR	DeepAutoQAR
Weighted by task	0.54 ± 0.13	0.64 ± 0.17
Weighted by data	0.62+/-0.11	0.78+/-0.18

Results comparison --- low data applications
Solubility investigation

Why does DeepChem option sometime have worse performance?

Solubility is an illustrative case:

- AlogP shows a good correlation with solubility in this data set
- AlogP is used as an input descriptor for AutoQSAR but not in DeepChem model.
- This gives AutoQSAR an advantage
- New DeepChem model with AlogP has MUE 0.59 comparing with 0.61 from AutoQSAR



Results comparison --- high data applications

Experimental setup:

- All tasks have larger than 5000 data points
- 88 regression tasks
- 30 classification tasks

Metrics:

- Q^2 and MUE for regression problems
- Area under curve(AUC) for classification problems

Training strategies:

- Using 5000 as training set (AutoQSAR scaling limitation)
- 90% as training set

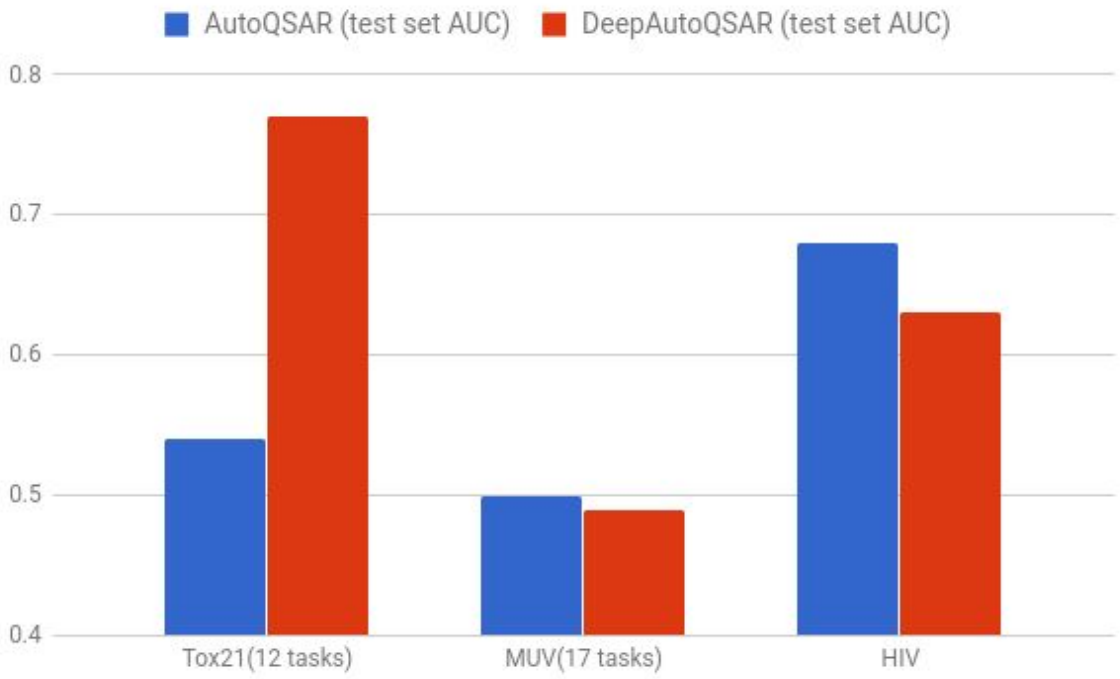
Results comparison --- high data applications

Dataset description

- HIV replication inhibition data:
  - 40426 compounds from Drug Therapeutics Program AIDS Antiviral Screen, which tested the ability to inhibit HIV replication. Results are placed into three categories: confirmed inactive, confirmed active and confirmed moderately active. In this study, confirmed active and confirmed moderately active are combined as one class.
- Toxicity dataset (Tox21 2014):
  - 8014 compounds with quantitative toxicity measurement on 12 different targets.
    - NR-AR,NR-AR-LBD,NR-AhR,NR-Aromatase,NR-ER,NR-ER-LBD,NR-PPAR-gamma,SR-ARE, SR-ATAD5,SR-HSE,SR-MMP,SR-p53
  - On average, each target has ~ 6600 data points
- Virtual screening benchmark dataset:
  - Maximum unbiased validation(MUV) dataset, which contains ~ 90,000 compounds over 17 targets.
  - On average, each target has ~14700 data points

Results comparison --- high data applications

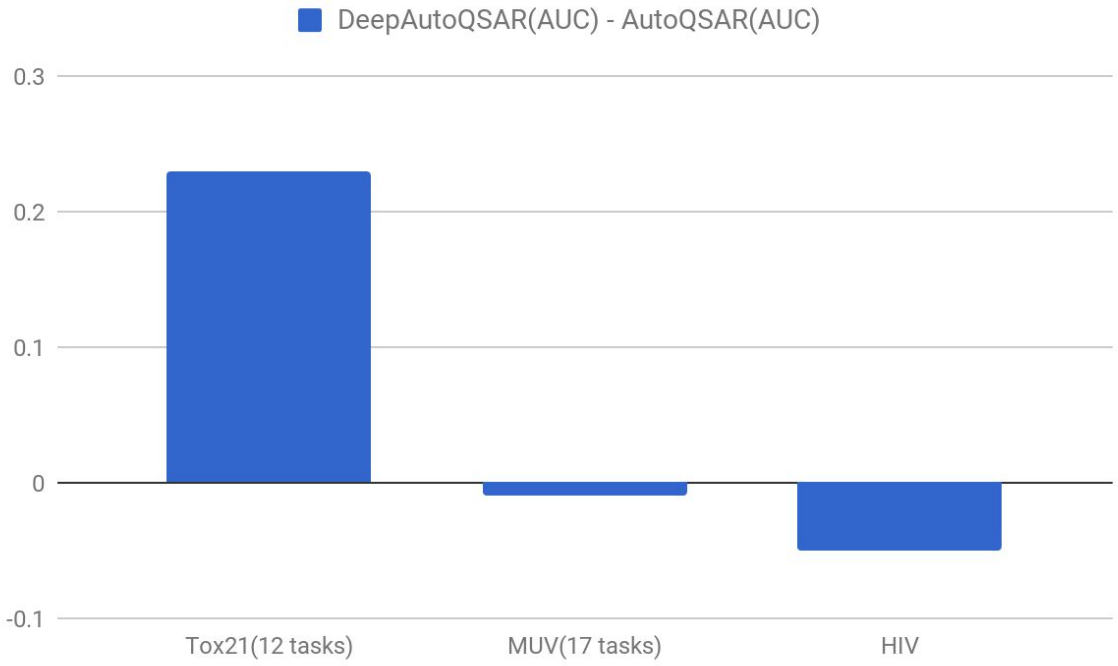
Using 5000 as the training data (classification AUC)



- Apples-to-apples, both methods trained to 5,000 randomly selected points
- DeepChem option performs clearly better in Tox21 dataset
- The other two datasets shows similar performance

Results comparison --- high data applications

Using 5000 as the training data (classification AUC)

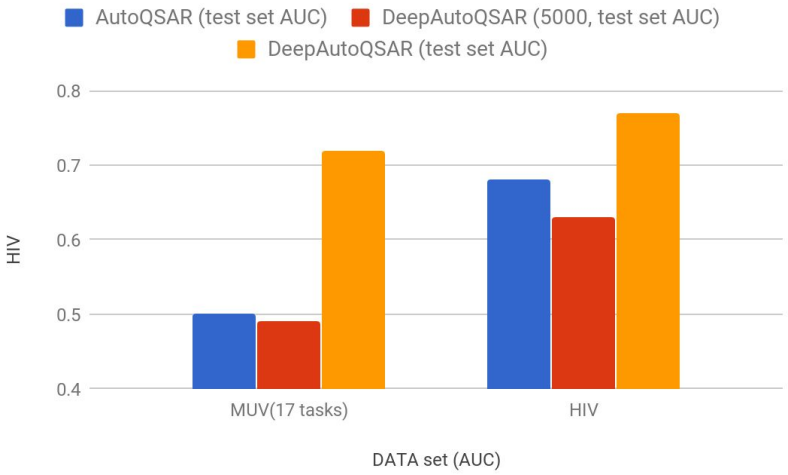


- Apples-to-apples, both methods trained to 5,000 randomly selected points
- DeepChem option performs clearly better in Tox21 dataset
- The other two datasets shows similar performance

Results comparison --- high data applications

Further increase the training data size --- using 90% as training set

DATA set	AutoQSAR (5000 training AUC)	DeepChem (5000 training AUC)	DeepChem (90% training AUC)	Data size (number of targets)
MUV	0.50	0.49	0.72	~14700(17)
HIV	0.68	0.63	0.77	40426



AutoQSAR doesn't scale to training sets over 5000 training data point

DeepChem option can use additional data to obtain much better performance

Generalization of DeepChem option --- Similarity between training and test data set

- Similarity metrics
  - For each cmpd in test set, calculate the max similarity (S_max) this cmpd and all training cmpds
  - Take the average of max similarities S_ave = Mean(S_max)

	Random similarity S_ave	Scaffold similarity S_ave
Selected MUV dataset	0.76	0.65
Selected Tox21 dataset	0.78	0.66

Generalization of DeepChem Option to Novel Scaffolds

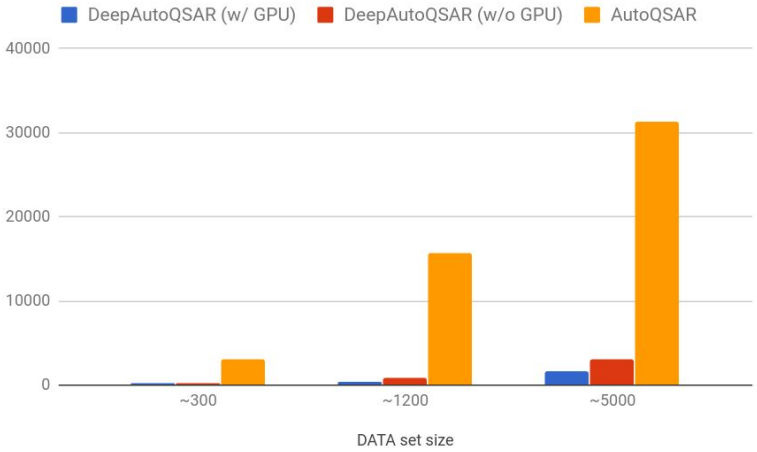
- 5000 training samples
- Classification

Data set (AUC)	% active compounds	DeepChem single task (random)	AutoQSAR (random)	DeepChem single task (scaffold)	AutoQSAR (scaffold)
Tox21	9	0.77	0.54	**0.62**	**0.55**
HIV	3.5	0.63	0.68	**0.60**	**0.50**
MUV	0.2	0.51	0.50	**0.54**	**0.54**

Computational Cost (in seconds)

Improvements to AutoQSAR speed can be made with parallelization

DATA set size	DeepChem (w/ GPU)	DeepChem (w/o GPU)	AutoQSAR
~300	160 s	270 s	3000 s
~1200	440 s	790 s	16000 s
~5000	1600 s	3000 s	31000 s



Even without GPU resources, DeepChem option is significantly faster

Remarks for single task DeepChem comparison

- For low data problems, the DeepChem option performance is comparable to AutoQSAR
  - For data set which there are dominant descriptors AutoQSAR may perform better.
- Even using equivalent training sets (5000 data points) in high-data applications, DeepChem option may have an advantage over AutoQSAR
  - Performs significantly better in Tox21
  - Performs similarly in other two dataset (MUV and HIV)
- The DeepChem option can scale to much larger training sets in high-data applications (200,000), this leads to much better performance in MUV and HIV

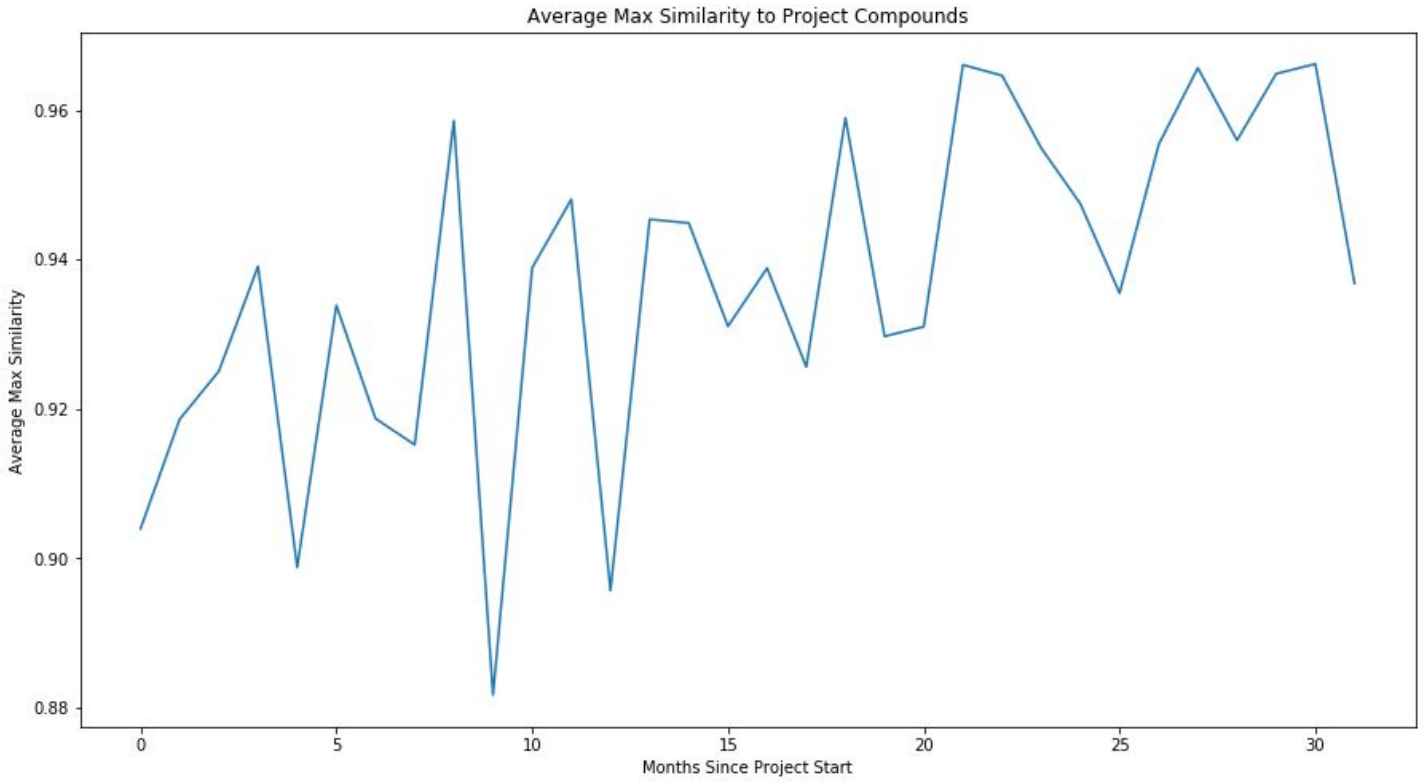
Demo

- Building a Model and Evaluation with DeepChem option

Demo

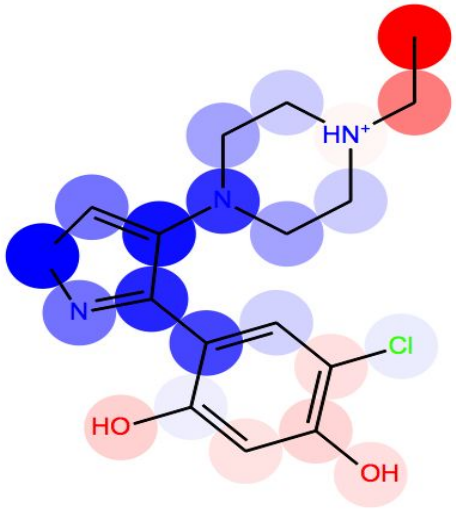
- AutoQSAR model in LiveDesign

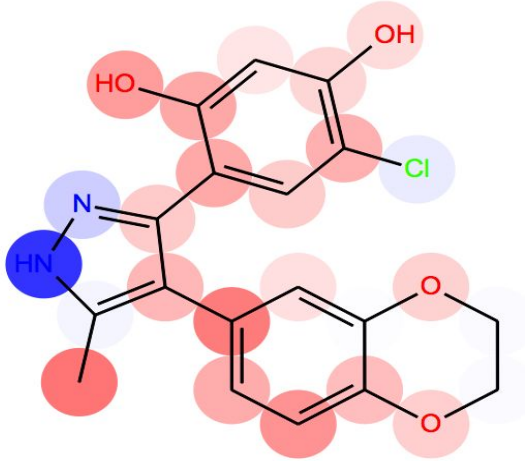
Updating Model Throughout Time



Future Direction

- Adding atom level user descriptors to Deep Learning Models
- LiveDesign Panel for Visualizing Results
- More Robust Splitting Algorithms
- Domain Of Applicability Estimates





Conclusion

- Deep learning methods out-perform existing methods on large datasets
- Deep learning performs within error on smaller datasets as ensembling of traditional methods at lower computational cost.
- Deep learning is not a magic bullet. The improvements in model performance are small to modest over existing state of the art.
- Everyone can run these cutting methods reliability and out of the box.

Q & A

- karl.leswing@schrodinger.com
- https://gitter.im/deepchem/Lobby

SCHRÖDINGER®

Scientific leader in life sciences and materials research