香港中文大學
The Chinese University of Hong Kong

Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models

Lewen Wang

Goh et al., arXiv:1706.06689 (2017)

Outline

1.Background

2.Method

3.Result & Discussion

4.Conclusion

Research Background

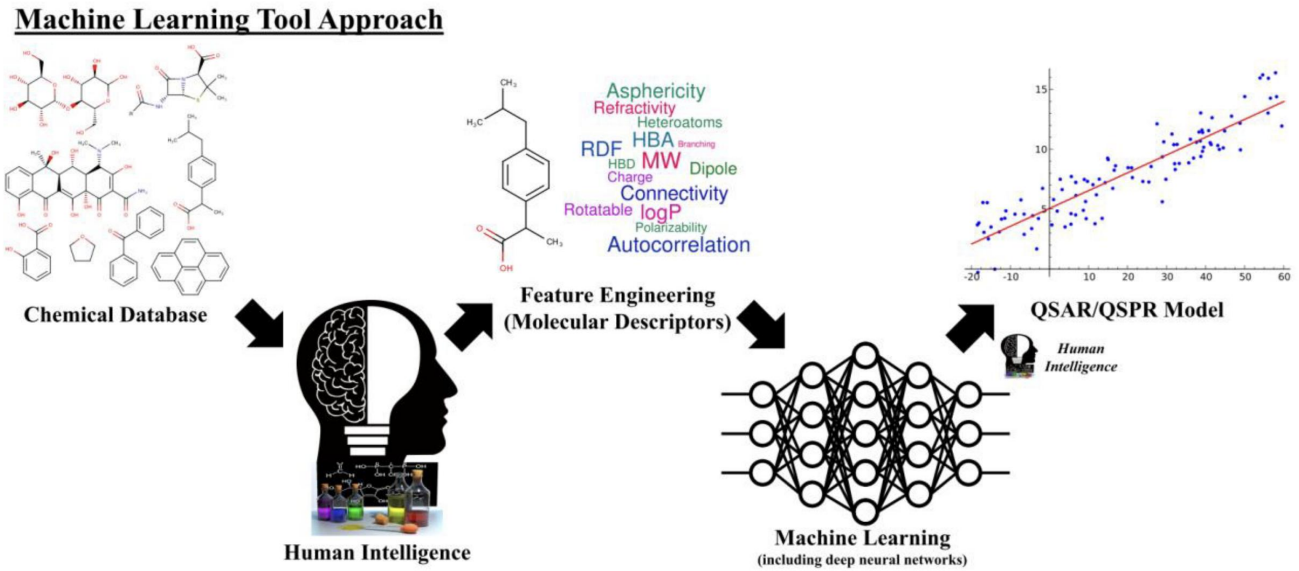
Background

Deep learning entered the computer vision community

DNN models had reached human-level accuracy in image recognition tasks.

- High-energy particle physics
- Astrophysics
- Bioinformatics
- Others

➢ Background



Traditional application approaches of machine learning in chemistry

Human Intelligence

↓

chemistry-specific features

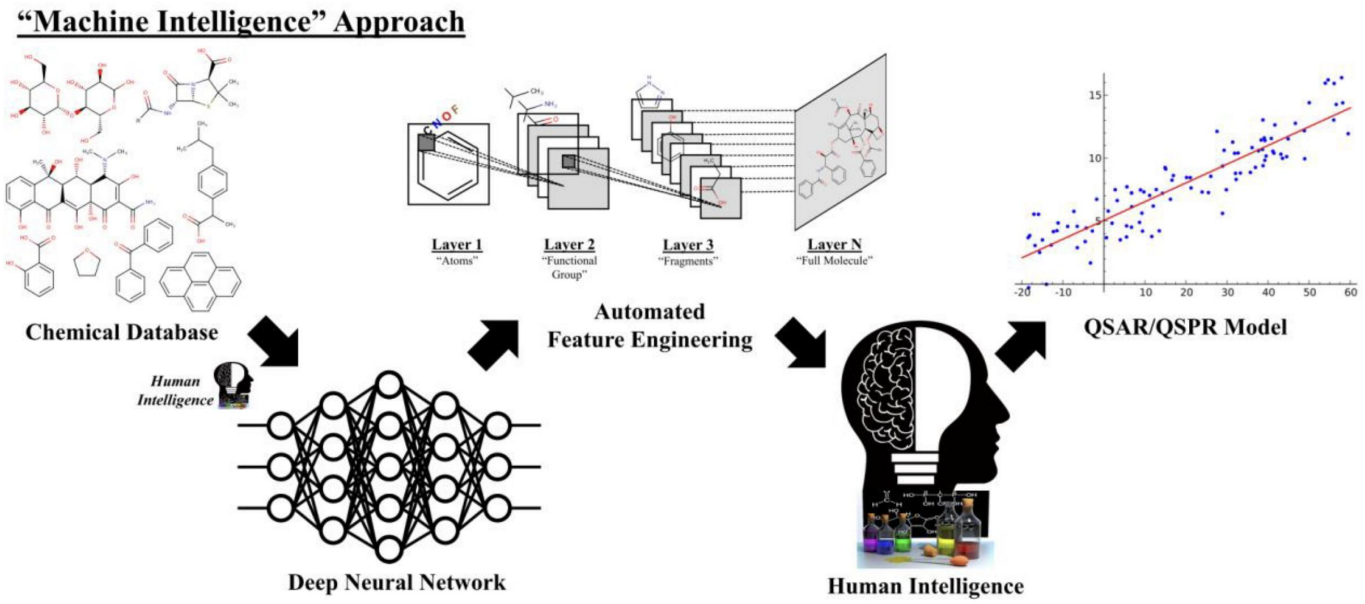
↓

ML algorithm

↓

QSAR/QSPR modeling

Background



Using deep learning model as "machine intelligence"

"Machine Intelligence"

time-consuming feature engineering

• Chemception

- A deep convolutional neural network trained on 2D molecular images.
- Requires only minimal chemical knowledge (high-school level) to generate input.
- Successfully predicts:

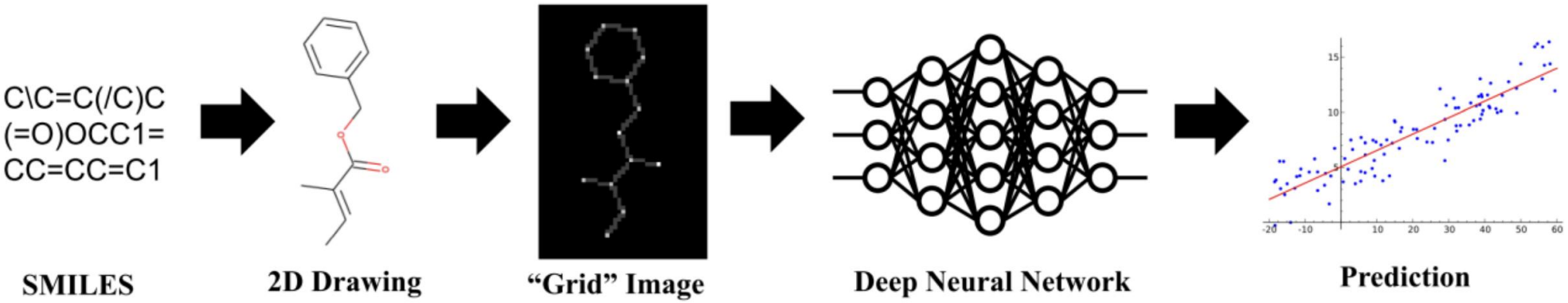
  Toxicity, HIV activity and solvation free energy.
- Achieves performance comparable to traditional QSAR/QSPR models without feature engineering.
- Other advantages:

  ✓ General-purpose architecture.

  ✓ Low computational cost.

Method

➢ Pipeline of Chemception



- SMILES strings were converted to their respective 2D molecular structures, which were then mapped onto an input array used to train the convolutional neural network in a supervised fashion.

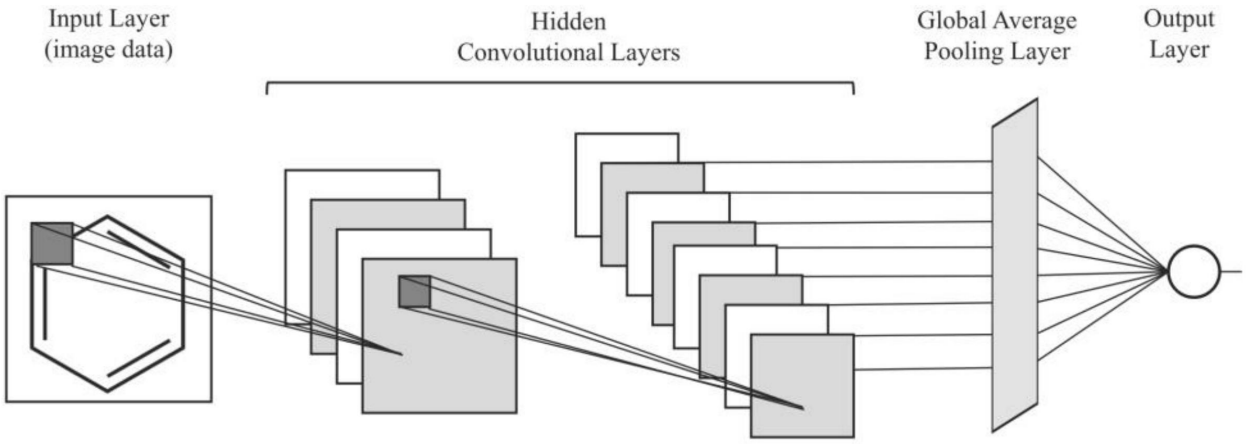
- No additional chemistry-inspired features.

Dataset Processing

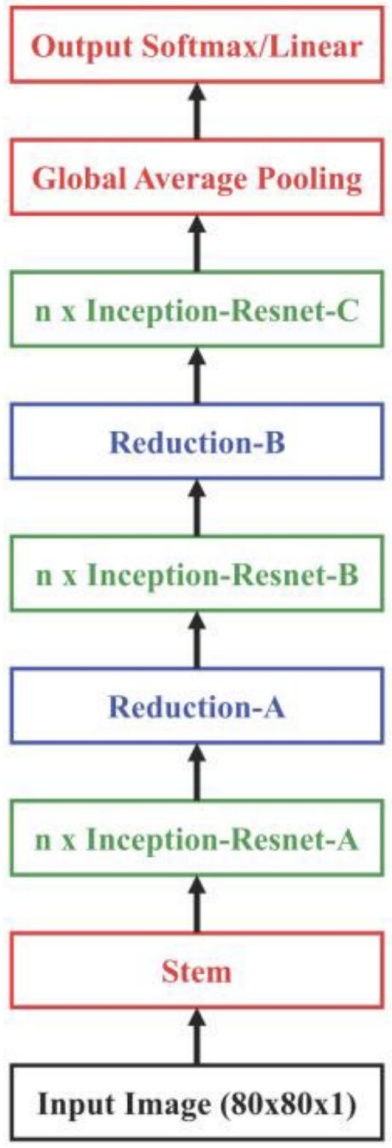
Dataset	Property	Task	Size
Tox21	Physiological: Toxicity	Multi-task binary classification	8014
HIV	Biochemical: Activity	Single-task binary classification	41,193
FreeSolv	Physical: Free energy of solvation	Single-task regression	643

- Used a 5-fold cross validation protocol for training.
- Stratification was enforced for the classification tasks to ensure each subset of the data maintains the same class distribution as the original dataset.
- Oversampled the minority class as ratio of the classes for classification tasks were imbalanced.

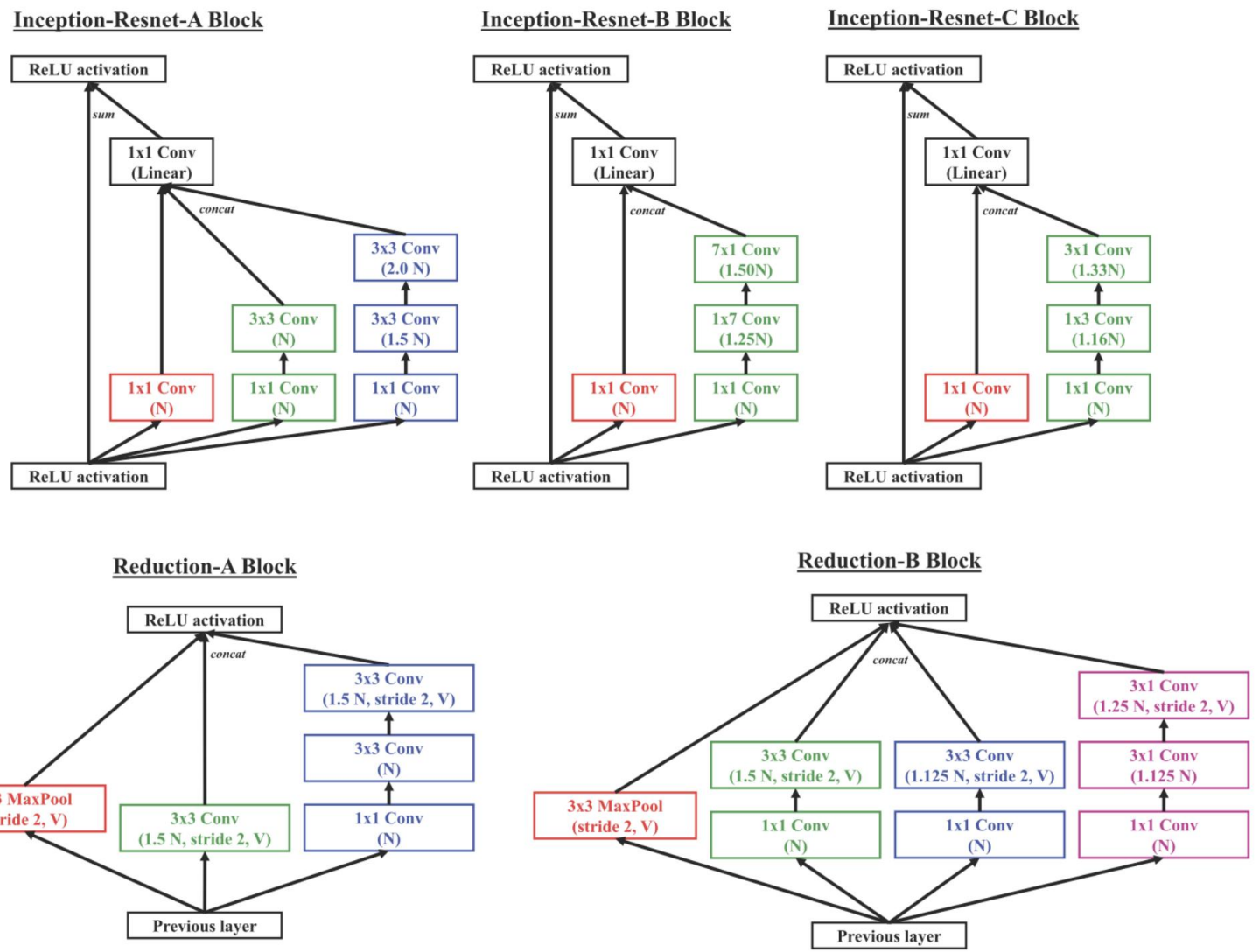
➢ Network Design

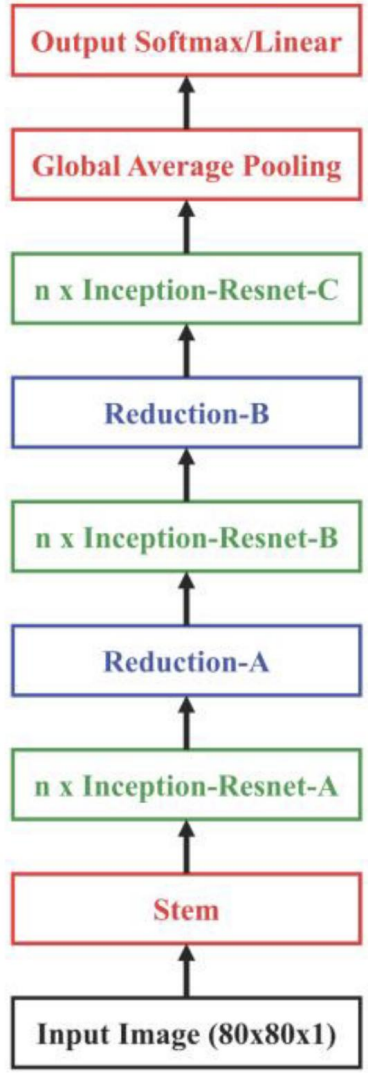


- Developed Chemception based on the Inception-ResNet v2.
- High level architectural design includes 6 segments
- Includes the stem layer, a series of Inception-ResNet blocks and reduction blocks, which is then passed to a global pooling layer that leads directly to the final output layer.
- Classification problems used a softmax layer as the output layer; Regression problem used a linear layer as output layer.



Network Design





➢ Training Protocol

Using 5-folds cross validation protocol and a two-stage protocol.

- Two-stage protocol.

In the first stage, they used the RMSprop algorithm for 50 epochs using the standard settings recommended (learning rate = 10^{-3}, ρ = 0.9, ε = 10^{-8}).

In second stage, they used the stochastic gradient descent (SGD) algorithm with momentum for another 50 epochs, using an initial learning rate of 10^{-3} with an exponential learning rate decay mapped using the following function:

$$\mathrm{lr} = \mathrm{lr}_{\mathrm{ini}} \times \gamma^{\mathrm{epoch}}$$

- For Tox21 and HIV dataset, the evaluation metric is area under the ROC-curve(AUC).
- For the FreeSolv dataset, the evaluation metric is RMSE.

Result & Discussion

➢ Model Optimization Results

	Train AUC			Validation AUC			Test AUC			
nr-ahr	0.825	+/-	0.018	0.779	+/-	0.015	0.800	+/-	0.020	Y
nr-ar	0.843	+/-	0.010	0.797	+/-	0.049	0.757	+/-	0.029	Y
nr-ar-lbd	0.887	+/-	0.034	0.834	+/-	0.046	0.886	+/-	0.014	Y
nr-aromatase	0.801	+/-	0.010	0.759	+/-	0.027	0.799	+/-	0.016	Y
nr-er	0.747	+/-	0.020	0.710	+/-	0.023	0.694	+/-	0.013	Y
nr-er-lbd	0.824	+/-	0.029	0.765	+/-	0.036	0.762	+/-	0.009	Y
nr-ppar-gamma	0.791	+/-	0.038	0.742	+/-	0.025	0.819	+/-	0.015	Y
sr-are	0.724	+/-	0.009	0.702	+/-	0.025	0.654	+/-	0.009	N
sr-atad55	0.841	+/-	0.022	0.759	+/-	0.048	0.776	+/-	0.011	Y
sr-hse	0.776	+/-	0.032	0.732	+/-	0.013	0.717	+/-	0.018	N
sr-mmp	0.791	+/-	0.020	0.759	+/-	0.016	0.755	+/-	0.010	Y
sr-p53	0.844	+/-	0.034	0.782	+/-	0.036	0.776	+/-	0.011	Y
Tox21	**0.808**		**0.044**	**0.760**		**0.035**	**0.766**		**0.058**	

- The individual toxicity measurements in the Tox21 dataset were predicted with validation AUC that ranged from 0.702 to 0.834, with the mean AUC value for the entire Tox21 dataset at 0.760.

- The training protocol was robust and prevented overfitting

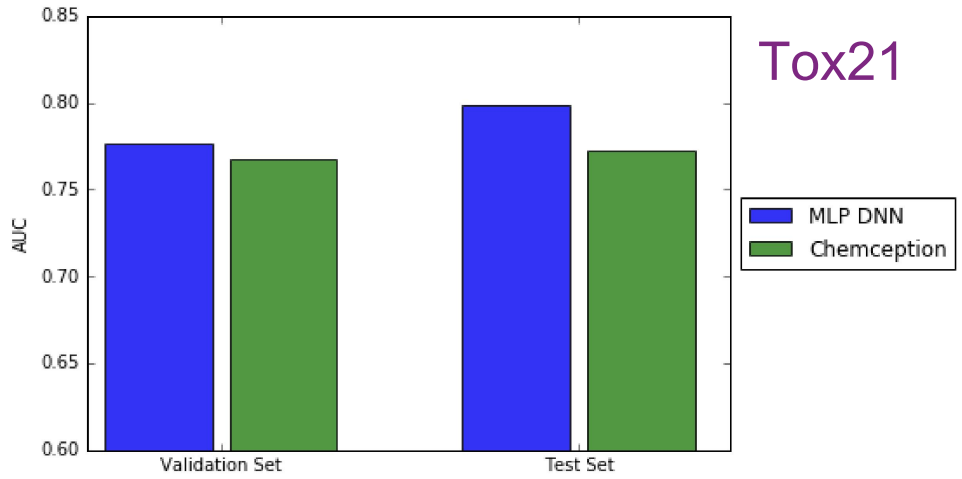
➢ Model Optimization Results

Model	No. of Inception Block per segment	No. of conv filters per block	No. of parameters
Chemception T1	1	32	276,603
Chemception T1_F16	1	16	69,875
Chemception T1_F64	1	64	1,100,967
Chemception T2	2	32	435,516
Chemception T2_F16	2	16	109,808
Chemception T2_F64	2	64	1,735,324
Chemception T3	3	32	594,429
Chemception T3_F16	3	16	149,741
Chemception T3_F64	3	64	2,369,681

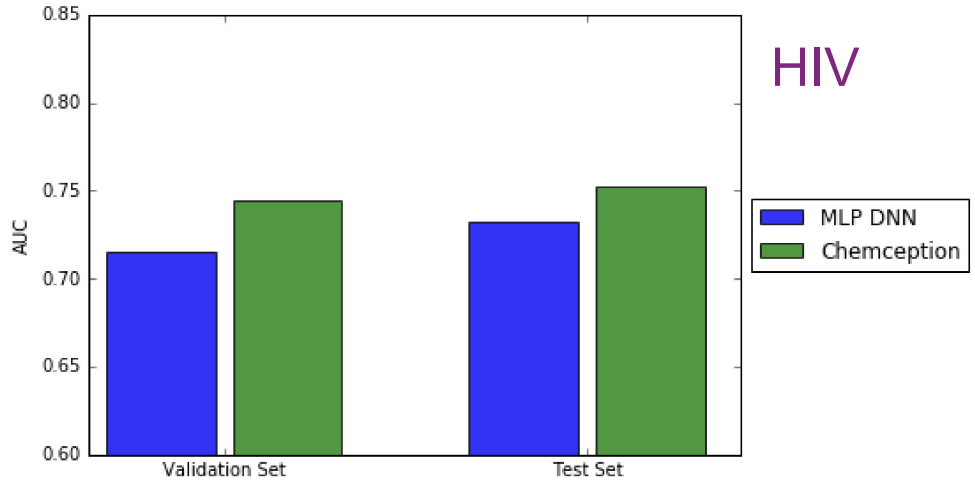
Architecture	Train AUC			Validation AUC			Test AUC		
Chemception_T1_F16	0.810	+/-	0.035	0.758	+/-	0.033	0.766	+/-	0.051
Chemception_T1_F32	0.808	+/-	0.044	0.760	+/-	0.035	0.766	+/-	0.058
Chemception_T1_F64	0.805	+/-	0.043	0.758	+/-	0.034	0.765	+/-	0.055
Chemception_T2_F16	0.805	+/-	0.043	0.760	+/-	0.037	0.769	+/-	0.054
Chemception_T2_F32	0.810	+/-	0.044	0.760	+/-	0.034	0.772	+/-	0.056
Chemception_T2_F64	0.806	+/-	0.047	0.759	+/-	0.033	0.766	+/-	0.055
Chemception_T3_F16	**0.815**	**+/-**	**0.044**	**0.768**	**+/-**	**0.037**	**0.773**	**+/-**	**0.058**
Chemception_T3_F32	0.814	+/-	0.045	0.763	+/-	0.034	0.771	+/-	0.055
Chemception_T3_F64	0.765	+/-	0.046	0.733	+/-	0.039	0.739	+/-	0.052

- For the shorter network depth, increasing the width of the layers provide no statistically significant improvement in the overall performance.
- In deeper network, skinnier network topology performed better.
- A deeper and skinnier Chemception architecture might be advantageous.

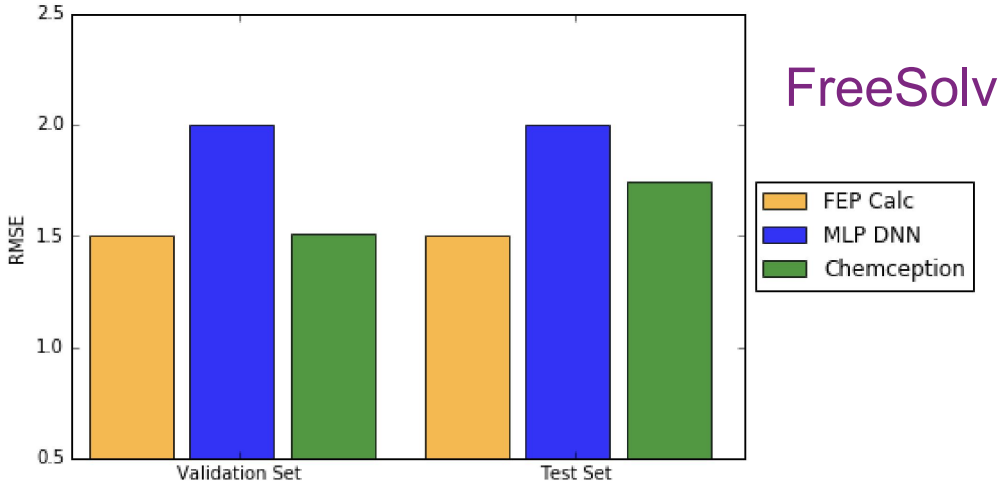
Training Result



Tox21



HIV



FreeSolv

Chemception slightly outperforms in HIV activity and solvation prediction and slightly underperforms in toxicity prediction.

Conclusion

➢ Conclusion

1. Chemception can only use 2D molecular image data for prediction of chemistry properties.

2. Chemception architecture can serve as a general-purpose neural network for learning a range of distinct properties while using a modest training database ranging from only ~600 to ~40,000 compounds.

3. The general accuracy matches or outperforms MLP deep neural networks trained on engineered features.

4. This study demonstrates that deep neural networks can effectively assist or replace human-driven feature engineering in chemistry.

Questions? Comments?

Thank You

Standard Deviation (σ)

A statistical measure that tells you how spread out or dispersed a set of values is from the mean (average).

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

- x_i = each data point
- μ = the mean(average)
- n = total number of data points