Using Deep Learning to Predict Toxicity and Lipophilicity from Molecular Fingerprints and 2D Structures

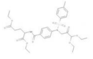
Riccardo Verzeni (rverzeni@stanford.edu), Celia Xinuo Chen (xinuo@stanford.edu)

Poster Presentation youtube link: https://youtu.be/22wDwMKNZ18

Motivation

- Predicting molecular properties of drug-like molecules are crucial in Computer-Aided Drug Discovery
- The conventional machine learning approaches (e.g. QSAR/QSPR[1]) heavily relies on domain specific knowledge for ad-hoc features selection. We used two more general approaches: from standard RDKIT molecular fingerprints and directly from molecular skeletal formula images (a new approach[2]) letting the DL model derive the relevant feature without explicitly including molecular descriptors
- The DL models could be re-purposed to predict different properties, making computer-aided drug design more efficient and less dependent on ad hoc experimentally accumulated data
- We used both approaches: fingerprints and images to predict toxicity and lipophilicity, two very important properties for screening and designing new drugs, and we compared their results

Data

	Toxicity	Lipophilicity
Dataset	National Institutes of Health Tox21dataset[3]	National Cancer Institute Dataset[4]
Data size	~10, 000 molecular structures	~250, 000 molecular structures
Data Pre-processing	We used RDKit library to parse the structure-data into SMILES,and then converted them into fingerprints (binary vector of size 2048) and Skeletal formulas gray-scale images of size 150 x 150 x 1 for the two different approaches	
Input features	• Fingerprints which are high-dimensional binary vector, each entry representing a manually encoded sub-structure of the molecule • Image representations of the molecular structures. Here is a sample: 	
Output features	binary values (1: toxic; 0:non-toxic) of 12 toxicological properties (with missing labels)	numerical value of Log P (theoretically calculated / experimentally measured) with missing labels

Results and Discussion

1. Toxicity

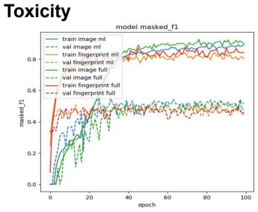


Fig. 1. Performance of four models evaluated by the average f1 score across three toxic properties

- For fully connected networks using fingerprints, the best performing one is the three-layer model, with **0.495 f1-score** and 96.4% accuracy on the test set.
- For convolutional neural networks, the best performing model is the residual network with 12 residual blocks, achieving **0.487 f1-score** and 96.6% accuracy on the test set.
- Considering the dataset is limited and highly imbalanced, both have been reasonable results.
- Overall, the performance of the convolutional neural network is on par with that of the fully connected network using fingerprint inputs.
- Multi-label classification models with missing labels generally perform better than those without, most likely due to the boost of dataset size.
- Methods such as L2, dropout, and batchnorm (for CNN) do not help improve the performance, as they increase the bias more than they reduce the variance.

	F1 Score (average)		Dataset Size		
1.Toxicity models	Train	Test	Train	Val	Test
Image missing-label	0.892	0.487	7742	860	955
Fingerprint missing-label	0.839	0.495			
Image full input	0.911	0.456	6708	746	838
Fingerprint full input	0.864	0.428			
	R^2 value		Dataset Size		
2. Liphophilicity (LogP) models	Train	Test	Train	Val	Test
fingerprints fcnn_6l logP	0.966	0.839	173865	19318	21464
fingerprints fcnn_6l experim_logP (transf learn)	0.985	0.923	2288	572	715
2Dimg incept_resnet_compact_v4 logP	0.963	0.852	173865	19318	21464
2Dimg incept_resnet_compact_v4 experim_logP (transf learn)	0.993	0.964	2288	572	715

2. Lipophilicity
(LogP and experimental LogP)

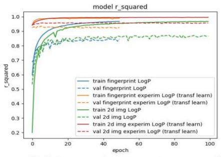


Fig. 2. Comparison of all the best performing models R^2 during training. Fingerprints models have been trained for 50 epochs while the others for 100 epochs.

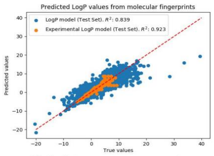


Fig. 3. R^2 of LogP and experimental LogP from fingerprints. N.B: Experimental LogP model is trained using transfer learning from the LogP model.

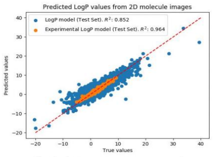


Fig. 4. R^2 of LogP and experimental LogP from 2D molecule images. N.B: Experimental LogP model is trained using transfer learning from the LogP model.

- The best LogP and Experimental LogP predictors for both inputs types are all abundantly above the baseline (R^2: 0), see results table above. Best fully connected NN: 6 layers (hidden layers: 2048, 1024, 256, 512, 128 neurons). Best CNN: see Fig.6 .
- Using the model weights learnt on the theoretical LogP values (~210000 data samples) to perform transfer learning on the model trained on experimentally measured LogP values (~ 3500 data samples) seemed to boost the performances of the experimental LogP predictor, which outperformed the original LogP predictor with both inputs types (Fig.4, 5).
- The CNN model trained on the 2D molecule images, outperformed the fully connected neural networks trained on the molecular fingerprints (Fig.3) suggesting that the features learnt by the convolutional layers were better than the human engineered, albeit general, molecular fingerprints.

Models

- Output layers
  For both architectures the output layers are respectively 1 linear unit for the logP prediction and n sigmoid units for n toxicological properties.
  The respective loss functions are:
  - Mean Square Error (for LogP regression)
  - Binary Cross Entropy (for toxicity classification)
- Fully connected architecture for 1D molecular fingerprints input

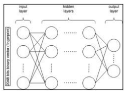


Fig. 5. Fully Connected Neural Network architecture

- CNN architecture for 2D molecule images input

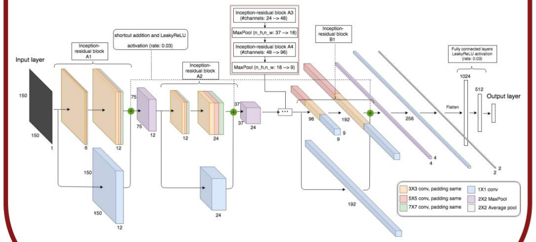

Input layer
150
150
1
Inception-residual block A1
6
12
shortcut addition and LeakyReLU activation (rate: 0.03)
150
150
12
75
75
12
Inception-residual block A2
12
24
24
37
37
24
Inception-residual block A3 (#channels: 24 --> 48)
MaxPool (n_h,n_w: 37 --> 18)
Inception-residual block A4 (#channels: 48 --> 96)
MaxPool (n_h,n_w: 18 --> 9)
Inception-residual block B1
96
192
9
9
192
256
4
4
Flatten
Fully connected layers LeakyReLU activation (rate: 0.03)
1024
512
Output layer
2
2
3X3 conv, padding same
5X5 conv, padding same
7X7 conv, padding same
1X1 conv
2X2 MaxPool
2X2 Average pool

Fig. 6. Convolutional Neural Network architecture inspired by Inception-ResNet[5] and Chemception[2].

Future work

- It would be helpful if we could gather more data to reduce the variance for the toxicity problem. We could also explore more state-of-the-art models and see if we could bring up the result to those of the best-performing papers.
- Considering the promising results on predicting both LogP and experimental LogP value from 2D images it would be interesting to see if possible to improve even more the results with 3D molecular structure inputs and/or trying predicting different properties with regression.

References

[1] URL:https://en.wikipedia.org/wiki/Quantitative_structure-activity_relationship.
[2] Charles Vishnu Abhinav O. Hodas Nathan Baker Nathan B. Goh, Garrett Siegel. Chemception: A deep neural network withminimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. 2017. URL:https://arxiv.org/ftp/arxiv/papers/1706/1706.06689.pdf
[3] URL:https://tripod.nih.gov/tox21/challenge/data.jsp.
[4] URL:https://cactus.nci.nih.gov/download/nci/.
[5] S.; Vanhoucke V.; Alemi A. Szegedy, C.; Ioffe. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016. arXiv:1602.07261.