Effect of RDKit Fingerprinting Preprocessing Strategies on Predictive Accuracy of Lipophilicity in Artificial Neural Networks

Christopher Fu

Challenges in drug discovery, lead identification

- Need to find lead compound with desired properties
  - These properties are all experimentally determined
- Can screen through existing compound libraries with known and tested properties
  - Limited to existing compounds
- Can synthesize and test new molecules and experimentally determine pharmacokinetic properties
  - High-throughput screening (HTS) allows us to rapidly assay a series of compounds
  - Still expensive and time consuming

Lipophilicity

- Key property I will focus on - Lipophilicity (LogP)
  - Strongly contributes to ADME properties of compound
  - Key property in QSAR
  - A drug has to pass through a multitude of cellular membranes.
  - Log P - the concentration ratio of the compound dissolved between an organic and aqueous solution where the organic solvent is commonly 1-octanol (Liu, Testa, & Fahr, 2010)
  - Experimentally determination is time consuming and requires skilled operators.

$$\log P = \log \frac{C_{organic}}{C_{aqueous}}$$

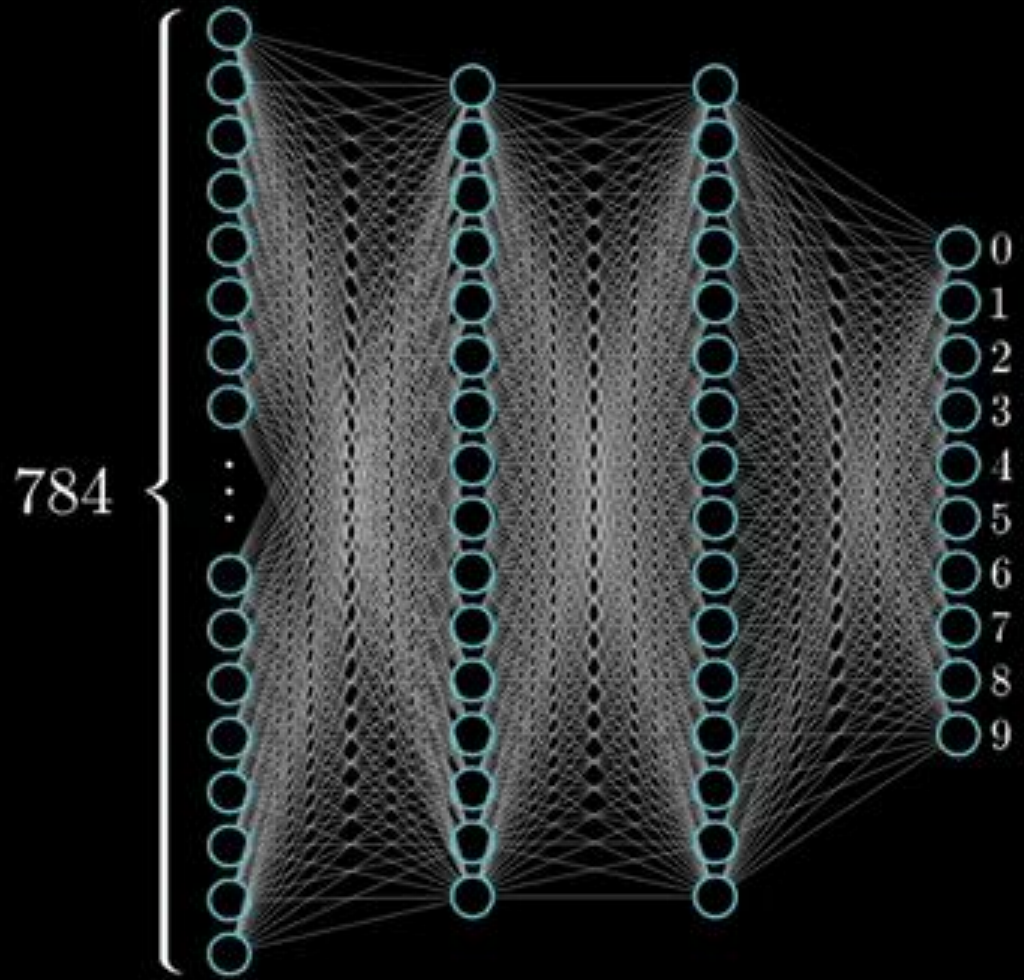
Predicting properties instead of experimentally determining them

- What if we could predict properties based on structure instead of determining them experimentally?
  - Save time, costs, and possibly increase speed of drug discovery
- Complex problem - Requires a tool that can abstract high level concepts from raw data
  - Example - Algorithm that can determine which parts of a molecule's structure contribute most to its reactivity with a target

Deep Learning as a predictive tool

- Strong at representation learning - Abstracting high level concepts from raw data
  - Have been shown to be strong in pattern detection and natural data processing
  - Abstracts using layers - each layers abstracking from the layer before it
- Example: Simple Neural Network as a image number classifier
  - Input layer would be a raw pixel array from the image.
  - First layer could be a layer to detect edges and areas of high contrast;
  - Hidden layers could detect combinations of features in the previous layer (shapes)
  - Last layer could detect shapes in the image that correspond strongly with a number class

- Quick note: Deep learning and Machine learning will never replace domain knowledge. Your jobs are safe



(Sanderson, 2017)

Considerations for Neural Network Design

- Determine structure of NN to use
  - VERY DIFFICULT
  - Active field of research
  - Some heuristics exist, but many people use guess and check
- Tune hyperparameters
  - Properties that affect how a NN acts
  - Learning rate, regularization
  - Some heuristics exist
- How will the data be preprocces and fed into the network?
  - Important consideration for this work as we are working with molecular data

How to encode/preprocess molecules?

- Want to encode molecule as a vector that accurately and uniquely identifies the molecule
- Descriptor extraction
  - Convert molecule into a vector of descriptors
- Fingerprinting
  - Commonly used in molecular similarity searches
  - Hash molecule using specific functions to convert molecule into a vector
  - Resulting vectors are a “Fingerprint” of the initial molecule
  - Similar molecules generate a similar fingerprint
  - Because of how a hash function works - resulting fingerprint may result in a loss of molecular information
    - Curse of dimensionality

(Landrum, 2012)

Open source tools used in this research

- Tensorflow
  - Machine and Deep learning framework
- Keras
  - Library that utilizes tensorflow as a backend. Allows for quick prototyping and testing of deep learning structures
- RDKit
  - Extensive chemoinformatics library
  - I will use for generating fingerprints from molecular SMILES
  - Fingerprints that will be tested include: Daylight-like, atom pairs, topological torsions, Morgan algorithm, Estate, Avalon bit based, ErG, RDKit

Goal

The aims of this proposed research are twofold:

- Construct an Artificial Neural Network (ANN) regressor to predict logP values given molecular structure of a molecule
    - Determine what structure and hyperparameters lead to acceptable prediction accuracies - Trial and Error similar to Devillers (1998)
- Explore and compare the use of different fingerprint encodings provided in RDKit on the accuracy of the model.
    - Fingerprints that will be tested include: Daylight-like, atom pairs, topological torsions, Morgan algorithm, Estate, Avalon bit based, ErG, RDKit

Impact

- Insight on which representation could be used in lipophilicity screening
- Use of accessible tools for “DIY Drug Discovery” among researchers and hobbyists

Proposed research

- “Lipophilicity_Dataset_-_logD7_4_of_1_130_Compounds” dataset published by Wang (2015).
  - contains the ID, Simplified molecular-input line-entry system (SMILE), and Log(D)7.4 of 1130 molecules
- Neural Network Architecture - Start with a structure similar to Devillers (1998), use trial and error to fine tune
  - Same strategy for hyperparameters
- Preprocess molecules using fingerprinting strategies provided in RDKit
  - Daylight-like, atom pairs, topological torsions, Morgan algorithm, Estate, Avalon bit based, ErG, RDKit
- Train and test accuracies of the model using each fingerprint strategy, increasing the fingerprinting length each time.
- Accuracy versus fingerprint length curves will be generated for each fingerprinting strategy.
  - The best performing strategy will be selected by comparing area under the curves.

References

Andreeva, E. P., & Raevsky, O. A. (2009). Lipophilicity of organic compounds calculated using structural similarity and molecular physicochemical descriptors. *Pharmaceutical Chemistry Journal, 43*(5), 258-262. doi:10.1007/s11094-009-0280-5

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. Drug Discovery Today, 20(3), 318-331. doi:10.1016/j.drudis.2014.10.012

An overview of the RDKit. (n.d.). Retrieved from http://www.rdkit.org/docs/Overview.html

Bahmani, A., Saaidpour, S., & Rostami, A. (2017). A Simple, Robust and Efficient Computational Method for n-Octanol/Water Partition Coefficients of Substituted Aromatic Drugs. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-05964-z

Dashtbozorgi, Z., & Golmohammadi, H. (2010). Quantitative structure-property relationship modeling of water-to-wet butyl acetate partition coefficient of 76 organic solutes using multiple linear regression and artificial neural network. *Journal of Separation Science, 33*(23-24), 3800-3810. doi:10.1002/jssc.201000448

References

Devillers, J., Domine, D., Guillon, C., & Karcher, W. (1998). Simulating Lipophilicity of Organic Molecules with a Back-Propagation Neural Network. *Journal of Pharmaceutical Sciences, 87*(9), 1086-1090. doi:10.1021/js980101j

Landrum, G. (2012). Fingerprints in the RDKit. Retrieved from http://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

LeCun, Yann & Bengio, Y & Hinton, Geoffrey. (2015). Deep Learning. Nature. 521. 436-44. 10.1038/nature14539.

Liu, X., Testa, B., & Fahr, A. (2010). Lipophilicity and Its Relationship with Passive Drug Permeation. *Pharmaceutical Research, 28*(5), 962-977. doi:10.1007/s11095-010-0303-7

Nielsen, M. A. (2015). Using neural nets to recognize handwritten digits. In *Neural Networks and Deep Learning*. Retrieved from http://neuralnetworksanddeeplearning.com/chap1.html

Ng, A. (n.d.). Deep Learning. Retrieved from http://cs229.stanford.edu/notes/cs229-notes-deep_learning.pdf

References

Sanderson, G. (2017, October 05). But what *is* a Neural Network? | Chapter 1, deep learning. Retrieved from https://www.youtube.com/watch?v=aircAruvnKk&t

Yang, S., Lu, W., Gu, T., Yan, L., & Li, G. (2009). QSPR Study of n-Octanol/Water Partition Coefficient of Some Aromatic Compounds Using Support Vector Regression. *QSAR & Combinatorial Science, 28*(2), 175-182. doi:10.1002/qsar.200810025

Wang, J-B., D-S. Cao, M-F. Zhu, Y-H. Yun, N. Xiao, Y-Z. Liang (2015). In silico evaluation of logD7.4 and comparison with other prediction methods. Journal of Chemometrics, 29(7), 389-398.