Graph Convolutional Neural Networks for ADME Prediction in Drug Discovery

Patrick Hop

1Kapirta Technologies, San Francisco, CA, USA

Abstract
ADME in-silico methods have grown increasingly powerful over the past twenty years, driven by advances in machine learning and the abundance of high-quality training data generated by laboratory automation. Meanwhile, in the technology industry, deep-learning has taken off, driven by advances in topology design, computation, and data. The key premise of these methods is that the model is able to pass gradients back into the feature structure, engineering its own problem-specific representation for the data. Graph Convolutional Networks (GC-DNNs), a variation of neural fingerprints, allow for true deep-learning in the chemistry domain. We use this new approach to build human plasma protein binding, lipophilicty, and humans clearance models that significantly outperform random-forests and support-vector-regression.

1 Introduction
Drug discovery and development continues to suffer from high attrition - various methods to improve the overall efficiency of the whole process are being pursued. Hence, in silico models for potency, ADME, and toxicology have been implemented in the drug discovery realm. Predictive accuracy of these methods is now high enough that it can be used to inform chemical synthesis, and exploration of chemical space in drug discovery programs Khojasteh [2011]. These methods typically rely on circular / ECFP4 fingerprints Rogers [2010] to featurize the molecular graph, and traditional machine-learning methods such as random-forests, LASSOs, and support-vector-regression for modeling.

The key premise of deep-learning is that featurization is performed by the model itself, replacing bespoke, domain-specific human insight with sophisticated graphical topologies and vast amount of computation. Recent research has shown that this methodology can be applied to in-silco models. Dahl [2014] showed the deep-learning on ECFP4 fingerprints outperform random-forests on Merck’s internal chemical data, which was shown again by Kearnes [2017] on internal ADME data from Vertex Pharmaceuticals. While this strategy has led to performance gains, it’s not in our view true deep-learning, which should be able to pass gradients back into the feature representation. The solution to this are Graph Convolutional Neural Networks which replace the circular fingerprint representation with a sub-differentiable, learnable fingerprint that has been shown to be significantly more performant in the MoleculeNet benchmark Ramsundar [2017].

2 Graph Convolutional Neural Networks
Molecular graphs can be of arbitrary shape and size, which is troublesome for traditional machine learning systems that are designed to operate on fixed-shape input tensors. This has historically been solved with
the computation of fixed-shape chemical fingerprints (figure 1), of which circular fingerprints are considered state of the art. Graph-Convolutional Neural Networks differ from traditional molecular machine-learning methods in that they replace the fingerprinting stage with a sub-differentiable, convolutional neural network that can learn machine-optimized features directly from the molecular graph to suit the particular problem at hand.

There are three types of layers used to featurize the molecular graph: graph-convolution \( h_{\text{conv}} \), graph-pool \( h_{\text{pool}} \), and graph-gather \( h_{\text{gather}} \) Altae-Tran [2016]. These operations preserve sub-differentiability, ensuring that gradients can flow through and optimize the learnable weights. While seemingly complicated, these operations are somewhat analogous to the canonicalization, hashing, and indexing of traditional circular fingerprints. It can be show that these new learnable fingerprints approach their traditional counterparts when initialized with large, random weights Duvenaud [2015].

Let \( G \) be a graph, \( V \) be the set of all vertices (atoms) in that graph, \( E \) let be the set of edges (bonds), and \( \sigma \) be the activation function. We can then formulate the following graph primitives Altae-Tran [2016]:

\[
h_{\text{conv}}(v) = \sigma \left( \sum_{(u,v) \in E} W^{\text{deg}(v)} u + U^{\text{deg}(v)} v + b^{\text{deg}(v)} \right)
\]

\[
h_{\text{pool}}(v) = \max \{ \max_{(u,v) \in E} u, v \}
\]

\[
h_{\text{gather}}(G) = \sum_{v \in V} v
\]

3 Experiments

Some of the larger ADME datasets released by AstraZeneca on ChEMBL were selected, despite these being approximately an order of magnitude smaller than the datasets found in a large biotech Dahl [2014] and are far smaller than what would be typically used for training a molecular machine-learning system. In light of this, readers should focus on the stratification of performance among methods, not the absolute numbers. A variety of models are used to as baselines to compare Graph-Convolutional Neural Networks (GC-DNNs) to. Support Vector Regression (SVR) is used to represent the max-margin models that are quite popular in industry, and random-forests (100 trees) are used as they represent a strong classical machine-learning baseline Sheridan [2016]. The Pyrimidal Networks (P-DNNs) Kearnes [2017] used use hidden layers of size (384, 196) which have be shown to be more performant than other configurations and represent in our view the limit of classical feature representations.

Measuring the performance of these models is deceptively difficult, because it has been shown that compounds sampled from the same time-slices significantly correlate, and are not representative of data
that the model will see in a future, production environment. In this setting, a random-split validation set is no longer an adequate measure of future performance, and should be replaced by a time-split validation set Aliagas [2015]. Since none of the AstraZenica datasets have temporal information, dissimilarity of the validation and training sets is enforced by splitting on the Bemis-Murcko scaffolds Bermis [1996].

3.1 Clearance

Clearance is the rate at which the human body removes circulating, unbound drug from the blood Lombardo [2014]. This is one of the key parameters to assess in drug discovery because low clearance will result in a longer residence time and a lower dose Khojasteh [2011]. In drug discovery, this property is assessed by measuring the metabolic stability of drugs in either human liver microsomes or hepatocytes Aliagas [2015]. This dataset includes 1102 examples of intrinsic clearance measured in human liver microsomes (µL/min/mg protein) following incubation at 37C.

3.2 Human Plasma Protein Binding (PPB)

Human Plasma Protein Binding assays measures the proportion of drug that is bound reversibly to proteins such as albumin and alpha-acid glycoprotein in the blood. Knowing the amount that is unbound/free is critical because it is that amount that can diffuse or actively taken up into tissues to exert its pharmacological activity or the liver to be cleared Khojasteh [2011]. PPB is measured as % bound to plasma by equilibrium dialysis after being incubated with whole human plasma at 37C for >5hrs. This measure is transformed using log(1 – bound), so we can capture a measure more informative to scientists. The dataset contains measurements for 1614 compounds.

<table>
<thead>
<tr>
<th></th>
<th>Splitting</th>
<th>Train $R^2$</th>
<th>Val $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>Random</td>
<td>.987</td>
<td>.315</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Random</td>
<td>.938</td>
<td>.239</td>
</tr>
<tr>
<td>P-DNN</td>
<td>Random</td>
<td>.985</td>
<td>.319</td>
</tr>
<tr>
<td>GC-DNN</td>
<td>Random</td>
<td>.888</td>
<td>.317</td>
</tr>
<tr>
<td>SVR</td>
<td>Scaffold</td>
<td>.987</td>
<td>.049</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Scaffold</td>
<td>.944</td>
<td>.108</td>
</tr>
<tr>
<td>P-DNN</td>
<td>Scaffold</td>
<td>.606</td>
<td>.275</td>
</tr>
<tr>
<td>GC-DNN</td>
<td>Scaffold</td>
<td>.721</td>
<td>.319</td>
</tr>
</tbody>
</table>

Table 1: Performance on Clearance Dataset. Note that all non deep-learning methods virtually fail to generalize at all on the scaffold-split validation set.

<table>
<thead>
<tr>
<th></th>
<th>Splitting</th>
<th>Train $R^2$</th>
<th>Val $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>Random</td>
<td>.967</td>
<td>.304</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Random</td>
<td>.937</td>
<td>.160</td>
</tr>
<tr>
<td>P-DNN</td>
<td>Random</td>
<td>.981</td>
<td>.372</td>
</tr>
<tr>
<td>GC-DNN</td>
<td>Random</td>
<td>.877</td>
<td>.317</td>
</tr>
<tr>
<td>SVR</td>
<td>Scaffold</td>
<td>.961</td>
<td>.199</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Scaffold</td>
<td>.937</td>
<td>.172</td>
</tr>
<tr>
<td>P-DNN</td>
<td>Scaffold</td>
<td>.978</td>
<td>.262</td>
</tr>
<tr>
<td>GC-DNN</td>
<td>Scaffold</td>
<td>.741</td>
<td>.404</td>
</tr>
</tbody>
</table>

Table 2: Performance on PPB dataset
3.3 Lipophilicity

Lipophilicity is a compound’s affinity for fats and oils. More formally, we use logd (pH7.4) which is captured experimentally using the octan-1 / water distribution coefficient measured by the shake flask method. This is an important measure for scientists to have, because drugs that are not lipophilic enough frequently have poor permeability Khojasteh [2011]. Alternatively, highly lipophilic compounds are usually encumbered by low solubility, high clearance, high plasma protein binding, etc. Indeed, most drug discovery projects have a sweet spot for lipophilicity Khojasteh [2011]. This dataset is the largest of the three at 4200 compounds.

4 Conclusion

We’ve shown the dominance of Graph Convolutional Neural Networks (GC-DNNs) across the MoleculeNet benchmark extends to the ADME domain, beating out not only commonly used methods such as Support Vector Regression and Random Forests, but even the more sophisticated pyramidal deep-neural networks. This is due to GC-DNNs ability to machine optimize it’s feature representation directly from the molecular graph.

Future work includes scaling up to industrial scale datasets that are easily an order of magnitude larger than the ones used in this paper, and experimenting with multi-task GC-DNNs. Together, these improvements could potentially could push performance even higher, perhaps even to a level where they could compete
with the physical experiments themselves for some assays Kramer [2016].

Overall, it’s increasingly clear that deep-learning for drug discovery has become increasingly mature. Recent advances in potency models Ramsundar [2017] Wallach [2014], graph-convolutions Altae-Tran [2016], and one-shot-learning Altae-Tran [2016] demonstrate performance far beyond reach of traditional methods. Looking forward, it’s our view that drug-discovery programs that adopt this new technology will be well rewarded with shorter development cycles and decreased costs by exchanging some physical experiments and bespoke human insight in favor of artificial intelligence tools.

Acknowledgements

I would like to thank the DeepChem community and Bharath Ramsundar for their helpful discussions and excellent implementations of some of the algorithms described in this paper. The ADME benchmark code for the paper will be open-sourced into the DeepChem library, enabling interested parties to be able to reproduce our results.

References


G. Dahl. Multi-task Neural Networks for QSAR Predictions. 2014.

D. Duvenaud. Convolutional Networks on Graphs for Learning Molecular Fingerprints. 2015.


