# Aiding Drug Design with Deep Neural Networks

**Thomas Unterthiner**[*], **Andreas Mayr**[*], **Günter Klambauer**
RISC Software GmbH & Institute of Bioinformatics
Johannes Kepler University Linz, Austria

**Marvin Steijaert**
OpenAnalytics, Belgium

**Jörg K. Wegner, Hugo Ceulemans**
Johnson & Johnson
Pharmaceutical Research & Development

**Sepp Hochreiter**
Institute of Bioinformatics
Johannes Kepler University Linz, Austria

## Abstract

An important computational tool in drug design is target prediction where either for a given chemical structure the interacting biomolecules (e.g. proteins) must be identified. Chemical structures interact with different biomolecules if they have similar 3D structure. Thus, the outputs of the prediction are highly interdependent from each other. Furthermore, we have partially labelled molecules since not all training molecules are measured of being active on each biomolecule.

The Merck Kaggle challenge on chemical compound activity was won by Hinton's group with deep networks. This indicates the high potential of deep learning in drug design and attracted the attention of big pharma. However, the unrealistically small scale of the Kaggle dataset does not allow to assess the value of deep learning in drug target prediction if applied to in-house data of pharmaceutical companies. Even a publicly available drug activity data base like ChEMBL is magnitudes larger than the Kaggle dataset. ChEMBL has 13 M compound descriptors, 1.3 M compounds, and 5 k drug targets, compared to the Kaggle dataset with 11 k descriptors, 164 k compounds, and 15 drug targets.

On the ChEMBL database, we compared the performance of deep learning to seven target prediction methods, including two commercial predictors, three predictors deployed by pharma, and machine learning methods that we could scale to this dataset. Deep learning outperformed all other methods with respect to the area under ROC curve and was significantly better than all commercial products. Deep learning surpassed the threshold to make virtual compound screening possible and has the potential to become a standard tool in industrial drug design.

## 1 Introduction

The pharmaceutical industry is currently challenged to increase the efficiency of drug development, since every year fewer drugs reach the market [1]. Machine learning methods could exploit a wealth of measurements that were accumulated by pharma companies and, thereby, offer Big Pharma alternatives.

The first step of a drug design pipeline is to identify a biomolecular *target* upon which a potential drug can act, e.g. a protein whose activity can be modified by a compound to achieve a beneficial therapeutic effect. Predicting these target-interactions using computational approaches is an important tool in modern drug design.

Applying target prediction in a realistic setting involves predicting several hundrets or thousands of outputs at the same time, some of which might be highly correlated. The correlation stems from the

[*]These authors contributed equally to this work

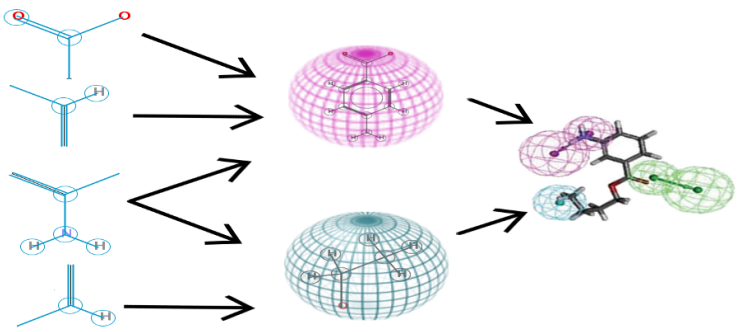

Figure 1: Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

fact that some targets are structurally very similar to each other. For most targets, only have partially labelled data is available, because compounds are typically only measured on a small set of targets.

In the Merck Kaggle challenge, Deep Learning showed promise for these kind of data. We assess the applicability and performance of deep networks at target prediction and compare them to state-of-the-art as well as commercial target prediction methods. Toward this goal we compiled a benchmark data set from ChEMBL, a database which resembles in-house databases of Big Pharma, though it still is considerably smaller. The Kaggle challenge comprised 15 targets, 164,024 compounds, and 11,081 features, while our ChEMBL benchmark contains more than 1,200 targets, 1.3 M compounds with 13 M ECFP12 features. This dataset serves to assess not only the raw performance, but also whether the methods scale to pharma in-house data.

Deep learning architectures seem to be well suited for target prediction because they (1) allow for multi-task learning [2] and (2) automatically construct complex features [3], which for target prediction are assumed to resemble pharmacophore descriptors. First, multiple target learning has two advantages: (a) it naturally allows for multi-label information and therefore can utilize relations between targets; (b) it allows to share hidden unit representations among prediction tasks. The latter item is particularly important as for some targets very few measurements are available, therefore single target prediction may fail to construct an effective representation. In contrast, deep networks exploit representations learned across different tasks and can boost the performance on tasks with few training examples. Secondly, deep networks provide hierarchical representations of a compound, where higher levels represent more complex concepts [4]. In pharmaceutical research complex representations of compounds have a long tradition: A major goal of drug design is the identification of pharmacophores, [5] which are the sets of steric and electronic properties that together enable an interaction with a target. These properties include hydrophobic regions, aromatic rings, electron acceptors or donors, which in turn can be described by substructures yielding these properties. Deep networks with ECFP12 fingerprints (chemical substructures) are ideally suited to represent properties in their first layer and in turn form pharmacophores in higher layers, as seen in Figure 1. The potential of deep learning is to find novel pharmacophores or representations of comparable complexity.

## 2 Experiments

### 2.1 Dataset

We compiled a target prediction benchmark dataset out of the ChEMBL database [6], a manually curated database of bioactivity measurements, which aims to centrally store the high-quality measurements of other chemistry resources.

We extracted all pharmaceutically relevant measurements from ChEMBL. Target measurements are reported in ChEMBL as continuous values, however for a classification task we require binary labels. We thus rely on explicit activity comments where provided, and defined a threshold otherwise.

This yielded 2,103,018 measurements distributed across 5,069 targets and 743,336 compounds. Additionally there are 415,527 measurements which exhibit a very weak signal. These are not used for testing as their signal is no reliable but they may still be a valuable enhancement of the training set. In order to make sure each target was realistically learnable, we discarded all targets with less than 15 samples per label, leaving 1,230 targets.

ChEMBL stores compounds as connected graphs of atoms, which we transformed into a high-dimensional binary representation using Extended Connectivity FingerPrints (ECFP12) [7] features.Each feature/fingerprint denotes the presence or absence of a certain chemical substructure. This yielded a total of 13,558,545 sparse features.

It is important that compounds which share a scaffold are not shared across training and test set, in order to guarantee that our dataset reflects the challenges of the daily drug development reality. As already mentioned, the value of virtual screening is determined by the ability to find new scaffolds with target activity. Thus, we clustered compounds using single linkage clustering to guarantee a minimal distance between training and test set. Clustering yielded 400,000 clusters which were partitioned into three folds of approximately equal size for cross-validation.

The performance of a classifier is evaluated by the AUC (area under the ROC curve) separately for each target. We report the mean AUC for each method.

### 2.2 Methods

#### 2.2.1 Deep Neural Network

Our network consists of one or multiple layers of ReLU hidden units [8, 9], followed by one layer of 1,230 sigmoid output units, one for each molecular target or classification task.

Using all the 7 M inputs for the deep net were infeasible on our hardware, therefore we removed features that were present in less than 100 compounds. 43,340 input features were kept. We stored the weight parameters on a single GPU with 12 GB RAM and used mini-batches of 1,024 samples for stochastic gradient descent learning. Since storing our input data in dense format requires about 5 TB of disk space, we used a sparse storage format. However, it proved to be faster to upload a mini-batch in sparse format to the GPU and then convert it to dense format instead of using sparse matrix multiplication. Overall, training a network takes between 3 to 4 days.

#### 2.2.2 Multi-Task Learning for Deep Networks

Each single training sample contributed only to a few of these tasks. Thus output units that were not active during a training sample were masked out during backpropagation.

Additionally we weighted the output-layer deltas coming from each output by the number of compounds that have been measured on the associated target. This ensures that across the whole training set each target has the same amount of influence on the hidden representation.

#### 2.2.3 Other Methods

We compare Deep Learning to the following Machine Learning methods that are used in target prediction, namely, Support Vector Machines, Binary Kernel Discrimination [10], Logistic Regression, $k$-nearest neighbour. We also re-implemented the following commercial products: a Parzen-Rosenblatt KDE-based approach [11],the Pipeline Pilot Bayesian Classifiers (a Naive Bayes statistics based approach) [12] and the Similarity Ensemble Approach (SEA) [13].

### 2.3 Results

Table 1 shows the mean AUC values across 1,230 targets for each of the classifiers we used. The deep neural network significantly outperformed its competitors, including two commercial methods with respect to the area under ROC curve (AUC) averaged over the prediction tasks, i.e. targets. Other well-established machine learning methods that could be scaled to the data set, such as SVMs, also performed better than the commercial methods.

Table 1: Performance of target prediction methods in terms of mean AUC across targets. The first column gives the method, the second column the AUC value, and the third column the $p$-value of a paired Wilcoxon test with the The alternative hypothesis that the deep neural network has on average a larger AUC than the other method.

| **Method** | **AUC** | $p$**-value** |
|---|---|---|
| Deep network | **0.830** | |
| SVM | 0.816 | 1.0e-07 |
| BKD | 0.803 | 1.9e-67 |
| Logistic Regression | 0.796 | 6.0e-53 |
| k-NN | 0.775 | 2.5e-142 |
| Pipeline Pilot Bayesian Classifier | 0.755 | 5.4e-116 |
| Parzen-Rosenblatt | 0.730 | 1.8e-153 |
| SEA | 0.699 | 1.8e-173 |

The neural net achieves an AUC $\geq 0.8$ on 813 out of the 1,230 targets, or $\approx 66\,\%$ of the time. The median AUC lies at 0.8588. On 12 targets we achieve perfect prediction accuracy (AUC = 1.0). This is in stark contrast to current commercial solutions, where the median AUC lies below $0.8$. Allmost all methods suffered from severe outliers. Of the methods that achieved an average AUC of over 0.8, the Deep Network has the least severe outliers. We hypothesize that the network could leverage its shared hidden representation to predict tasks which are difficult to solve when tackled in isolation.

## References


[1] J. Arrowsmith, "Trial watch: phase III and submission failures: 2007–2010," *Nature Reviews Drug Discovery*, vol. 10, no. 2, pp. 87–87, 2011.

[2] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, p. 41–75, 1997.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives.," *IEEE Trans Pattern Anal Mach Intell*, Feb 2013.

[4] Y. Bengio, "Deep learning of representations: Looking forward," in *Proceedings of the First International Conference on Statistical Language and Speech Processing*, SLSP'13, (Berlin, Heidelberg), pp. 1–37, Springer-Verlag, 2013.

[5] L. Kier, *Molecular orbital theory in drug research*. Medicinal chemistry, Academic Press, 1971.

[6] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, pp. gkr777–D1107, Sept. 2011.

[7] D. Rogers and M. Hahn, "Extended-connectivity fingerprints.," *Journal of Chemical Information and Modeling*, vol. 50, pp. 742–754, May 2010.

[8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

[9] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, pp. 315–323, 2011.

[10] G. Harper, J. Bradshaw, J. C. Gittins, D. V. S. Green, and A. R. Leach, "Prediction of biological activity for high-throughput screening using binary kernel discrimination," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 5, pp. 1295–1300, 2001.

[11] R. Lowe, H. Y. Mussa, F. Nigsch, R. C. Glen, and J. B. Mitchell, "Predicting the mechanism of phospholipidosis," *Journal of Cheminformatics*, vol. 4, no. 1, p. 2, 2012.

[12] X. Xia, E. G. Maliski, P. Gallant, and D. Rogers, "Classification of Kinase Inhibitors Using a Bayesian Model," *Journal of Medicinal Chemistry*, vol. 47, pp. 4463–4470, Aug. 2004.

[13] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnology*, vol. 25, pp. 197–206, Feb. 2007.