CISUC
CENTRO DE INFORMÁTICA E SISTEMAS DA UNIVERSIDADE DE COIMBRA

EXPLORING A SIAMESE NEURAL NETWORK ARCHITECTURE FOR DRUG DISCOVERY

Luis H. M. Torres [1] · uc2015241578@student.uc.pt
Bernardete Ribeiro [1] · bribeiro@dei.uc.pt
Joel P. Arrais [1] · jpa@dei.uc.pt

[1] **CISUC, DEI, University of Coimbra**
D4 - Deep Drug Discovery and Deployment Project
https://www.cisuc.uc.pt/projects/show/259

Vision

The application of deep neural networks is an important asset to significantly increase the predictive power when inferring the properties and activities of small-molecules and those of their pharmacological analogs. However, in the traditional drug discovery process, the lead-optimization step is, inherently, a low-data problem, which makes it difficult to find analogous molecules with the desired pharmacotherapeutic potential and reduced patient risks, making it impossible to obtain accurate predictions for candidate drug analogs.

Using an one-shot learning strategy, we only need one instance per class for the network's training and a small amount of data and computational resources to build an accurate model.

Ultimately, we demonstrate how a one-shot-based Siamese Neural Network allow us to outperform the state-of-the-art models in the accurate and reliable prediction of pharmacological analogs for drug discovery applications.

Objectives

- Use of deep learning methods to tackle the challenges identified in the lead optimization step of drug discovery tasks;
- Face the low-data problem of the lead optimization step of drug discovery tasks using an one-shot learning strategy;
- Use of a Siamese Neural Network architecture to achieve strong results on one-shot classification tasks based on a similarity score between two input molecules;
- Evaluate the performance of a one-shot Siamese neural network and compare it with the state-of-the-art machine learning and deep learning approaches;
- Optimization of the prediction of drug analogs with increased biological activity based on a reduced set of candidate molecules.

Methodology

One-Shot Siamese Neural Network Model

A Siamese Neural Network built upon two parallel and identical convolutional neural networks is introduced as the proposed model approach. This network is compatible with a set of pairs of compounds provided for training. The proposed model accepts different pairs of molecules and learns a similarity function, which returns a similarity score between two input molecules based on a reduced set of candidate molecules. Thus, according to the learned similarity rule, the network predicts a similarity score in one-shot. Tox21 was the dataset used to extract SMILES (Simplified Molecular Input Line Entry System) for compound representation and encoding (one-hot encoding).

The model architecture that maximizes performance is described in Figure 1.

The output feature map of the last convolutional layer is flattened into a single vector that serves as an input to a fully connected layer with 1024 units. This layer learns a similarity function between two feature vectors by applying a distance metric to the learned feature map. It is followed by a dense layer that computes the absolute difference between the two output feature vectors. This value serves as input to a sigmoid function in the last layer which condenses the prediction into a probability value between 0 and 1. The predicted similarity score is given by,

$$score = sigmoid\ (\sum_i |\ v^i_{1,l} - v^i_{2,l}\ |)$$

v_1 and v_2 are the output feature vectors of the last convolutional layer of each Siamese twin, l the index representing the dense layer, i the index in each output feature vector and *sigmoid* the activation function. This similarity measure is a probability, assuming a value between 0 and 1. If *score* is equal to 1, the probability of both compounds belonging to the same class is maximum. If *score = 0*, this probability is minimum according to the learnt similarity rule.

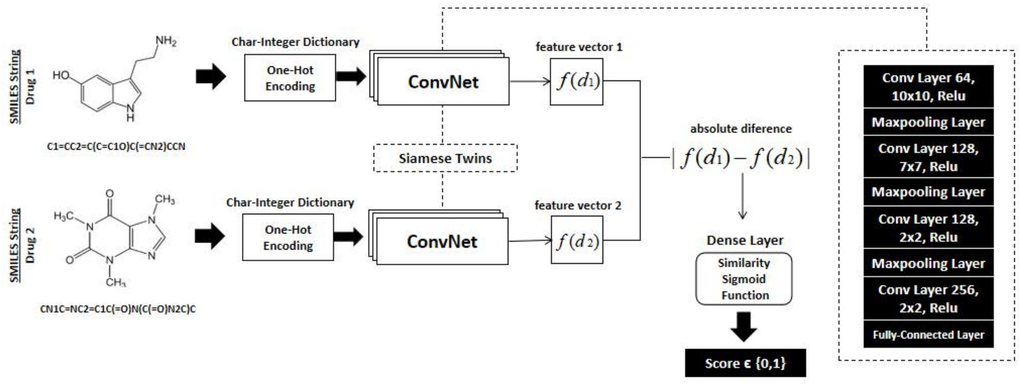

Figure 1. One-Shot Siamese Neural Network model architecture.

N-way One Shot Learning Classification

The reduced amount of biological data available for training led us to adopt a new strategy to predict novel compounds using the proposed model approach. A one-shot classification strategy is applied to demonstrate the discriminating of the similarity rule (Figure 2). The data for classification is organized in pairs, one example from a support set *S* and the other from a test set *T*. The support set consists of set of molecules representing each class selected at random whenever a one-shot learning task is performed. The support set has compounds representing each one of the compound classes and the test set has the test molecule of unknown class provided for classification.

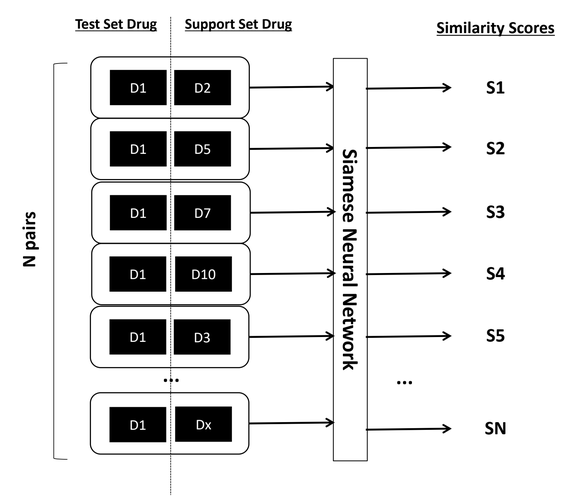

Figure 2. *N*-way One-Shot Learning Classification using a Siamese Neural Network.

Note that for every pair of input twins, our model generates a similarity score between 0 and 1 in one-shot. Therefore, to evaluate whether the model is really able to recognize similar molecules and distinguish dissimilar ones, we use an *N*-way one shot learning strategy. The test molecule is compared to *N* different ones and only one of those matches the original input. Thus, we get *N* different similarity scores $\{score_1, ..., score_N\}$ denoting the similarity between the test molecules and those on the support set. The pairs are organized so that the first pair is a pair of instances of the same class, with the remaining pairs formed by compounds of different classes. If the pair of compounds of the same class gets the maximum similarity score, the model prediction is correct.

$$pred(d_i, S) = \arg\max(score(d_i, d_j))\ \ , d_j \in S, d_i \in T$$

The prediction *pred* corresponds to the pair (d_i, d_j) that returns the highest similarity score $score(d_i, d_j)$ in a one-shot trial with d_i the test molecule and d_j the support set molecule. This process is repeated across multiple trials, the model accuracy being determined as the percentage of correct predictions.

$$accuracy(\%) = \frac{number\ of\ correct\ predictions}{number\ of\ trials\ per\ one-shot\ task}$$

	N					
	2	3	4	5	7	10
Siamese Neural Network (validation)	**94%**	**90%**	**84%**	**78%**	**70%**	**65%**
Siamese Neural Network (training)	**95%**	**92%**	**86%**	**84%**	**72%**	**70%**
KNN	70%	55%	49%	43%	36%	30%
Naive Model	61%	43%	34%	31%	22%	19%
SVM	56%	42%	30%	24%	16%	12%
Random Forest	71%	58%	60%	44%	34%	20%
Multi-Layer Perceptron	76%	60%	36%	34%	22%	13%
Convolutional Neural Network	81%	70%	58%	46%	41%	39%

Table 1. *N*-way One-shot Classification Accuracy Results.

Results and Conclusions

In this paper, we propose a model able to predict novel compounds according to a degree of similarity between molecules. To measure the model performance, the accuracy was determined using a *N*-way one-shot learning strategy described previously. The comparison of a given complex model with a set of simpler base models is a common strategy when assessing performance. Therefore, it was crucial to compare the proposed model with traditional machine learning approaches and simpler deep learning methods (Table 1).

In this study, we validate the potential of learning a similarity metric capable of classifying a set of compound pairs using a one-shot Siamese deep neural network. The high discriminating power of the learnt features extracted from a reduced set of candidate compounds is a property that results from the combination of a one-shot learning approach and a model built as a set of parallel convolutional neural networks.

References

[1] Altae-Tran, H., Ramsundar, B., Pappu, A. S., \& Pande, V. (2017). Low Data Drug Discovery with One-Shot Learning. ACS Central Science.
[2] Lake, B. M., Salakhutdinov, R., Gross, J., \& Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In {Proceedings of the 33rd Annual Conference of the Cognitive Science Society}.
[3] Salakhutdinov, R. R., Tenenbaum, J. B., \& Torralba, A. (2012). One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. JMLR Workshop and Conference Proceedings.
[4] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, “DeepTox: Toxic-ity prediction using deep learning,”Frontiers in Environmental Science, 2016.
[5] Koch, G. Siamese neural networks for one-shot image recognition. Ph.D. thesis, University of Toronto, 2015.
[6] N. R. C. Monteiro, B. Ribeiro, and J. Arrais, “Drug-Target Interaction Prediction: End-to-End Deep Learning Approach,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.

UNIVERSIDADE D COIMBRA
FCT Fundação para a Ciência e a Tecnologia
Cofinanciado por: COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA