Virtual Screening Assisted by Siamese Neural Networks

Alan Santos and **Duncan Ruiz**
Pontifical Catholic University of Rio Grande do Sul
6681 Ipiranga Avenue, Faculty of Informatics
Porto Alegre, Brazil

Abstract

High-throughput virtual screening relies on scoring functions to evaluate binding affinity between ligand and receptor. Although useful for identification of new potential drugs, imperfections in these scoring functions can lead to incorrect classification of small molecules. In this context, non-parametric machine-learning approaches can identify implicit binding interactions that can be hard to model explicitly. We present an approach to distinguish between ligands and decoys using energy-based models with siamese neural networks. Taking as inputs 3D biochemical property grids from ligand and receptor, it is computed the compatibility between them. We show that this model outperforms other machine-learning approaches in a Fully Flexible Receptor model of InhA-NADH complex.

1 Introduction

High-throughput virtual screening can be used to identify small molecules that effectively bind to a drug target. In virtual screening, thousands or even millions of chemical compounds are tested for a potential binding target using computer programs. Computational approaches include algorithms that *dock* small molecules into a biological target's binding site and *score* their binding affinity (Ain et al. 2015). Molecular docking and associated scoring functions can enrich the pool of candidate inhibitors. However, their accuracy is not enough to characterize a single ligand (Durrant and McCammon 2011).

Scoring functions evaluate the conformation of a molecule as docked to the target's binding site. These functions are fast mathematical methods that approximate physical interactions between ligand and receptor in order to predict binding affinity. However, imperfections in these functions can lead to a poor prediction performance (Ain et al. 2015), causing incorrect selection of non-binding compounds over true active ligands.

In this context, machine-learning models can exploit available data about molecular recognition, increasing the efficiency of ligand classification tasks. These models could determine when and how molecular binding occurs, learning to identify implicit binding interactions (Ain et al. 2015). Besides, the study of the learned parameters can clarify some internal aspects involved in molecular recognition.

The solution presented in this work focus on *learning a ligand classification model*, which can be used as a scoring function. The proposed model calculates the compatibility between ligand and receptor using 3D grids of biochemical properties. These grids are processed by a siamese neural network (SNN), which extract a vector of features of each grid. *Compatibility* is calculated based on the norm of the sum of these vectors. Besides, we show that the model learned to identify important molecular interactions, like hydrogen bonds.

2 Related Work

In drug discovery, molecular binding affinity is estimated based on theory-inspired energy functions, which may include parameters fitted to experimental or simulation data. However, molecular recognition involves implicit binding interactions, which are hard to model explicitly (Ain et al. 2015). These binding interactions could be captured by non-parametric machine-learning techniques, improving the performance of scoring functions (Ain et al. 2015).

Recently, deep learning models are achieving outstanding results in different complex tasks. Deep learning is part of a family of machine learning algorithms that attempt to learn high-level representations from raw data by using deep graph with multiple processing units (Schmidhuber 2015). In recent years, different deep learning approaches were applied to drug discovery, like deep convolutional networks for bioactivity prediction (Wallach, Dzamba, and Heifets 2015) and target prediction (Unterthiner et al. 2014). Also, ensembles of neural networks were used to predict binding affinity (Durrant and McCammon 2011) or aqueous solubility for drug-like molecules (Lusci, Pollastri, and Baldi 2013). Wang (Wang et al. 2014) used a pairwise input neural network to predict target-ligand interactions. In this architecture, ligand and target are represented as n-dimensional vectors, which are linearly combined and submitted to a neural network.

Our approach is different from these methods: it learns a function that maps 3D grids of biochemical properties into vectors in low-dimensional space and computes the *compatibility* between input grids using these vectors. The function that maps grids into vectors uses a *Convolutional Neural Network (CNN)* to encode each 3D grid. The training of this

model is similar to approach described in (Hadsell, Chopra, and LeCun 2006).

3 Design and Implementation

In context of data classification, probabilistic models assign a normalized probability to every possible configuration of input features. *Energy-based models* (EBM) assign an unnormalized energy to those configurations. According to Lecun (LeCun et al. 2006), learning, in these models, consists in finding weights W that associate low energy to compatible samples and high energy to incompatible ones. Using EBMs, the estimation of normalized distributions over input space is not necessary, which is an advantage over probabilistic models (LeCun et al. 2006).

A trainable compatibility metric can be seen as an energy function $E_W(X_1, X_2)$ that computes the energy of a pair of inputs, X_1 and X_2. In drug discovery context, it is possible to compare a small molecule and a receptor in order to evaluate the binding affinity between them. Due the differences of interactions that results in high-affinity binding and low-affinity binding, a trained machine-learning model could learn to distinguish between high-affinity and low-affinity molecules, given a receptor.

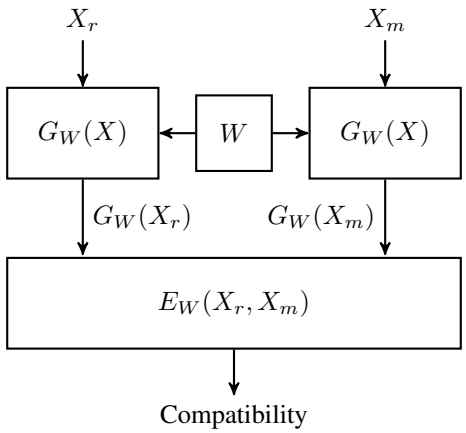

Figure 1: Diagram of the SNN to calculate $E(X_r, X_m)$

In this context, our approach builds a trainable system that maps 3D grids of biochemical properties to vectors in low-dimensional space, where *compatibility* between small molecule and receptor can be easily calculated. The learning consists in training two identical *CNN*s that share the same parametrization W - a *Siamese Architecture* (Hadsell, Chopra, and LeCun 2006). This architecture is shown in Figure 1, where X_r and X_m refers to grids of biochemical properties of receptor and small molecule, respectively. $G_W(X)$ refers to function that translate X_r and X_m into low-dimensional vectors. Using these vectors, the merge layer computes $E_W(X_r, X_m)$, the function that calculate the *compatibility* between X_r and X_m, i.e., the output of the proposed model.

Convolutional Neural Networks

As shown in Figure 1, the siamese architecture uses two identical *CNN*s that shares the same parameterization W. CNNs are trainable models specialized in identification of raw-data local correlations, and they can learn low-level features and combine them to create high-level representations. In the proposed model, CNNs are used to convert 3D raw grids into low-dimensional vectors.

In context of molecular recognition, interactions between atoms are predominantly local (Bissantz, Kuhn, and Stahl 2010). Once convolutional layers of CNNs can exploit local correlations due to its locally connectivity pattern between layers, these layers can be used to detect important interactions between atoms. Due to this behavior, the application of *CNN*s is appropriated (Wallach, Dzamba, and Heifets 2015).

Merge Layer

The merge layer is responsible for computing the output of the model, i.e., the compatibility between X_r and X_m. Once X_r and X_m must match complementary biochemical properties to achieve high affinity binding (Hildebrandt et al. 2007), the sum of representation vectors, $G_W(X_r)$ and $G_W(X_m)$, must be as close to 0 as possible. So, the calculus of compatibility can be designed as the sum of two vectors, which must be close to 0 if X_r and X_m are compatible, and $+\infty$, otherwise. In the proposed model, the energy E_W is defined as the norm of the sum of vectors $G_W(X_r)$ and $G_W(X_m)$, as shown in Equation 1.

$$E_W(X_r, X_m) = \|G_W(X_r) + G_W(X_m)\| \tag{1}$$

Loss function

In order to learn the optimal parameterization W, we use the loss function described in (Hadsell, Chopra, and LeCun 2006), shown in Equation 2. The y refers to label assignment of pair ($y = 0$ if pair is compatible, $y = 1$, otherwise) and the constant m is the margin, which defines a radius around $G_W(X_r)$ where dissimilar pairs contribute to loss functions. The loss function is composed by a sum of two terms: the first term is the partial loss function for compatible pairs and the second, for incompatible pairs. The minimization of this function decreases the energy of compatible pairs and increases the energy of incompatible ones.

$$\begin{aligned} L(W, Y, X_r, X_m) =& (1 - Y)\frac{1}{2}(E_W)^2 + \\ & Y\frac{1}{2}\{max(0, m - E_W)\} \end{aligned} \tag{2}$$

4 Experiments

The experiments presented in this section show the results obtained by our model, compared to a deep convolutional neural network model (DCNN), similar to approach described in (Wallach, Dzamba, and Heifets 2015).

Dataset and Data encoding

We demonstrate the application of the SNN on a Fully Flexible Receptor (FFR) model containing 19.5 nanoseconds of

molecular dynamics simulation of InhA-NADH complex of *Mycobacterium tuberculosis* (PDB ID: 1ENY) (Gargano, Costa, and de Souza 2007). In this model, there is one conformation for each 20 picoseconds.

To generate compatible pairs, we docked inhibitors from 19 crystallographic structures of InhA, obtained from RCSB PDB (Berman et al. 2000), to every conformation in FFR model, using Autodock 4.2 (Morris et al. 2009). Similarly, to create incompatible pairs, we selected 19 most similar decoys in InhA subset of DUD-E (Mysinger et al. 2012), which were also docked in FFR model. DUD-E is a benchmark of virtual screening studies and contains ligands and decoys for biological drug targets. Decoys are selected based on similar physical properties but different chemical structures from ligands.

After docking, we split this data set into training set and testing set. The training set consists of docked conformations of 30 small molecules (15 inhibitors and 15 most similar decoys), each conformation paired with its own FFR snapshot. The docked conformations of 8 remaining small molecules (4 inhibitors and 4 decoys) were used in testing set.

Once electrostatic interactions plays an important role in molecular recognition (Hildebrandt et al. 2007), we chose electrostatic potential as the biochemical property to 3D grids used as model input. Electrostatic potential grids were generated by APBS (Automatic Poisson Boltzmann Solver) (Baker et al. 2001), using Poisson Boltzmann Equation. All grids were adjusted to fit inside a 25 $\mathring{\text{A}}$ $\times$25 $\mathring{\text{A}}$ $\times$25 $\mathring{\text{A}}$ cube, centered at the center of mass, in case of small molecule grids, and covering the binding site, in case of protein grids.

SNN architecture

A series of experiments, using the dataset, were executed to determine the best sub-net architecture. In this work, we only describe the best-performing architecture. C_x denotes convolutional layers with ReLU activation and F_x refers to fully connected layer using linear activation, where x is the layer index.

C_1 Filters: 7; Kernel Size: $5 \times 5 \times 5$; Parameters: 882;
C_2 Filters: 7; Kernel Size: $5 \times 5 \times 5$; Parameters: 6132;
C_3 Filters: 7; Kernel Size: $5 \times 5 \times 5$; Parameters: 6132;
C_4 Filters: 7; Kernel Size: $3 \times 3 \times 3$; Parameters: 1330;
F_5 Number of units: 32; Parameters: 28032;
F_6 Number of units: 16; Parameters: 528;
F_7 Number of units: 2; Parameters: 34;

Results

In order to evaluate the performance of the proposed SNN, we compare it with a DCNN similar to approach described in (Wallach, Dzamba, and Heifets 2015). To avoid the need of positioning each small molecule inside binding site to use both models, the input of DCNN is a single 3D electrostatic potential grid of ligand, and the same grid is used in the input pairs of SNN. The best performing architecture of DCNN consists of 3 convolutional layers followed by 2 hidden layers with 256 and 128 units, respectively, and 2-way softmax.

Table 1: Statistics of performance

	Accuracy	Precision	Sensitivity	AUC
SNN	85.01%	0.85	0.81	0.92
DCNN	72.71%	0.70	0.76	0.81

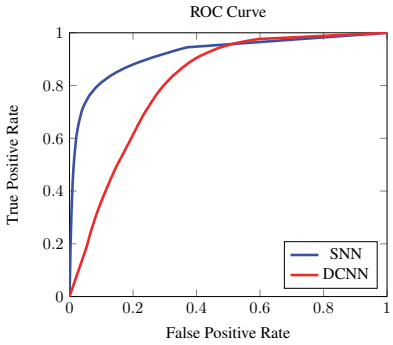

Figure 2: Receiver operating characteristic curves of both tested models.

To allow direct comparison between SNN and DCNN performance, we employed the *logistic function* to normalize the output of SNN. The output of DCNN varies between $(0, 1)$. However, the calculated compatibility of SNN varies between $(0, +\infty)$. To address this difference, we set a threshold of compatibility which is the decision boundary of proposed model. In this case, we define the decision boundary at $m/2$, where m is the margin in equation 2. The probability of classification as active (p_{active}) is calculated according equation 3, where E_W refers to energy calculated by SNN. Using this equation, if E_W is lesser than $m/2$, the small molecule is classified as active ligand. Otherwise, this ligand is classified as decoy.

$$p_{active} = \frac{1}{1 + e^{(E_W - \frac{m}{2})}} \tag{3}$$

In our tests, the area under receiver operating characteristic curve (AUC) of our model was 0.92, outstanding the DCNN, which achieves 0.81 AUC. Besides, the precision and sensitivity of our model were both higher, which denotes that our model is better on ranking ligands. The results are summarized in table 1. Figure 2 presents detailed AUC curves for SNN and DCNN.

Understanding what model learned

In a siamese architecture, the model learns to map input data to a low-dimensional space, where comparison between elements is simple. However, it is necessary to understand how the model combines input features to calculate the compatibility between pairs.

To achieve better understanding about $G_W(X)$, firstly, we identified important interactions involved in molecular recognition between a docked conformation of inhibitor 8PC (8PC400 from PDB ID: 3FNE) and a snapshot of FFR model. Using Ligplot+ (Laskowski and Swindells 2011), we

observed two hydrogen bonds between atom $O22$ of 8PC and residues ALA197 and ILE201, as shown in Figure 3.

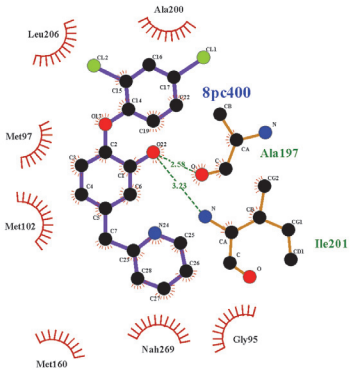

Figure 3: Representation of interactions between small molecule 8PC (PDB ID: 3FNE) and binding site of InhA-NADH complex (PDB ID: 1ENY).

In order to calculate the influence of this interaction in compatibility calculated by SNN, we altered the electrostatic potential grid of 8PC, replacing the negative electrostatic potential at region of atom $O22$ by positive electrostatic potential. The compatibility between this altered grid and the grid of the snapshot of FFR model is 16.75, while the output of SNN feed with original pair was 1.04. This behavior leads to a conclusion that the two hydrogen bonds are very important to compatibility between 8PC and InhA-NADH complex.

5 Conclusion

This work describes a method to calculate compatibility between ligand and receptor using siamese architecture. This *unsupervised* compatibility metric is directly related to affinity binding, where high compatibility (low output energy) means high affinity binding and vice-versa. This metric can be used as scoring function and can be applied to guide a search algorithm in molecular docking techniques. In addition, the learned filters permit to figure out important chemical interactions useful to domain experts. Our tests shown outstanding results on tested dataset achieving 0.92 AUC and outperforming the DCNN model.

References

Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; and Ballester, P. J. 2015. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5(6):405–424.

Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; and McCammon, J. A. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* 98(18):10037–10041.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research* 28(1):235–242.

Bissantz, C.; Kuhn, B.; and Stahl, M. 2010. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry* 53(14):5061–5084.

Durrant, J. D., and McCammon, J. A. 2011. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling* 51(11):2897–2903.

Gargano, F.; Costa, A.; and de Souza, O. N. 2007. Effect of temperature on enzyme structure and function: a molecular dynamics simulation study. In *Proceedings of the 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

Hildebrandt, A.; Blossey, R.; Rjasanow, S.; Kohlbacher, O.; and Lenhof, H.-P. 2007. Electrostatic potentials of proteins in water: a structured continuum approach. *Bioinformatics* 23(2):e99–e103.

Laskowski, R. A., and Swindells, M. B. 2011. Ligplot+: multiple ligand–protein interaction diagrams for drug discovery. *Journal of chemical information and modeling* 51(10):2778–2786.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data* 1.

Lusci, A.; Pollastri, G.; and Baldi, P. 2013. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* 53(7):1563–1575.

Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; and Olson, A. J. 2009. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* 30(16):2785–2791.

Mysinger, M. M.; Carchia, M.; Irwin, J. J.; and Shoichet, B. K. 2012. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry* 55(14):6582–6594.

Schmidhuber, J. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61:85–117.

Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; and Hochreiter, S. 2014. Deep learning as an opportunity in virtual screening. In *Proceedings of the Deep Learning Workshop at NIPS*.

Wallach, I.; Dzamba, M.; and Heifets, A. 2015. Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Wang, C.; Liu, J.; Luo, F.; Tan, Y.; Deng, Z.; and Hu, Q.-N. 2014. Pairwise input neural network for target-ligand interaction prediction. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 67–70. IEEE.