Deep50: Web service for multi-task protein-ligand interaction prediction

***Abstract*—The prediction of drug and protein interactions is crucial for the development of new drugs. We designed and developed a machine learning model, and a coupled web service infrastructure to serve queries about compound activities. The service can handle hundreds of queries real time. We evaluated the predictive performance of the underlying model resulting in a competitive result of mean AUC of 0.930 over different proteins.**

I. Introduction

To find really novel compounds with appropriate physico-chemical properties and pharmacological activity, millions of candidates should be screened. If we are able to reduce this number, we can reach dramatic cost reduction in the early phase of the pharmaceutical development programs.

We believe, that high quality *in-silico* activity prediction service can be especially useful for the case of orphan drug research which now is becoming more attractive area thanks to the legislative easing in the last two decades in the EU (EC 141/2000) and the US (Orphan Drug Act).

Therefore, we propose an easy to use, real-time, on-line prediction service called *Deep50*, which employs a deep factorization based multi-task neural network model.

II. Deep factorization model

The most straightforward way to handle the protein-compound interaction prediction task is to represent all chemical compound with a high-dimensional fingerprint vector, based on its chemical structure. Every element of this vector encodes the occurrence of some substructure, and so the vector is inherently sparse.

However, we know that fingerprints are not covering all aspects of a chemical compound necessary to predict its activity on a target. Most of the times, for example, they do not encode the correct three-dimensional conformation of the compound necessary for the interaction. To overcome this limitation, we propose a *deep factorization model*, which assigns to every compound an embedding vector

$$h_{emb} = \text{mol2vec}(\text{cmpd}), \tag{1}$$

where cmpd is the identifier of the compound, in our case computed from Canonical SMILES code.

This allows us to employ the well known strategy from word embeddings that uses a lookup table to store for every compound its learned embedding vector, which is not explained by the fingerprint [1]. To motivate this approach, let us consider the case of orphan drugs. To make the cost of the clinical testing affordable, we may want to use the strategy of drug repositioning [2][3]. In this case we search in the space of existing compounds that already have results in clinical experiments and bioassays. The use of compound embedding enables efficient transfer of this already existing experimental knowledge by giving the model access to the identity of the compound. This architecture can also be viewed as a nonlinear extension of matrix factorization methods with side information [4][5].

More specifically the deep factorization model is a neural network, with L hidden layers, specified as

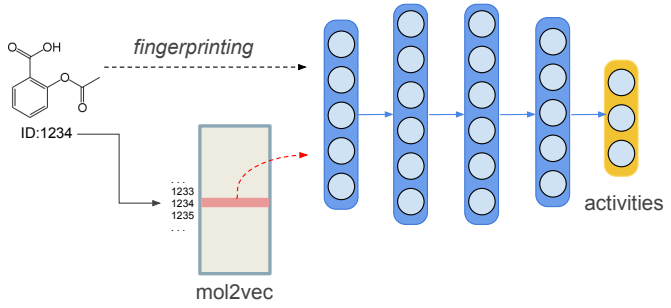

Fig. 1. Architecture of the deep factorization model used by Deep50 for protein-ligand activity prediction.

$$h_1 = \sigma(W_{fp} h_{fp} + W_{emb} h_{emb} + b_0) \tag{2}$$

$$h_{i+1} = \sigma(W_i h_i + b_i), \tag{3}$$

where $i = 1, \ldots, L$, and

$$h_{emb} = \text{mol2vec}(\text{cmpd}) \tag{4}$$

$$h_{fp} = \text{ecfp}(\text{cmpd}) \tag{5}$$

are the vectors from embedding and input fingerprints, respectively. Figure 1 gives an overview of the architecture. The final layer outputs logit scores predicting the activity of the compound for each protein-threshold pair.

The model is trained with Adam optimizer with a fixed learning rate schedule using early stopping on the validation set.

III. Web service

The proposed Deep50 uses TensorFlow for training and inference[6]. We implemented a multi-threaded Flask based stack which exports the trained TensorFlow model as a web service. To implement the client side web interface we used React and Redux technology stack. The service can handle hundreds of parallel queries in real-time. The data preparation steps and the web service architecture is displayed in Figure 2.

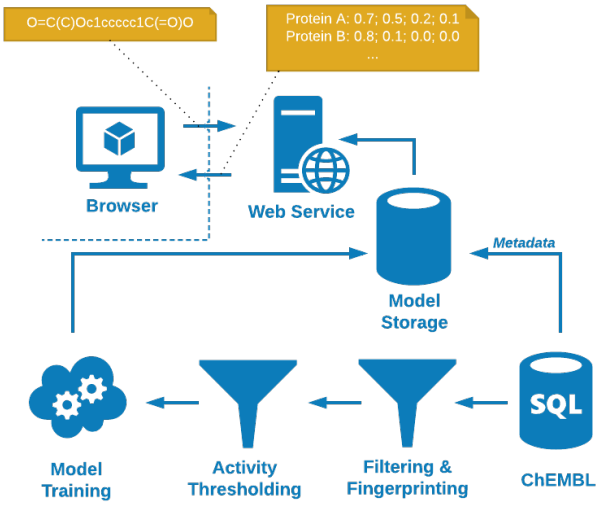

Fig. 2. The architecture of the Deep50 web service.

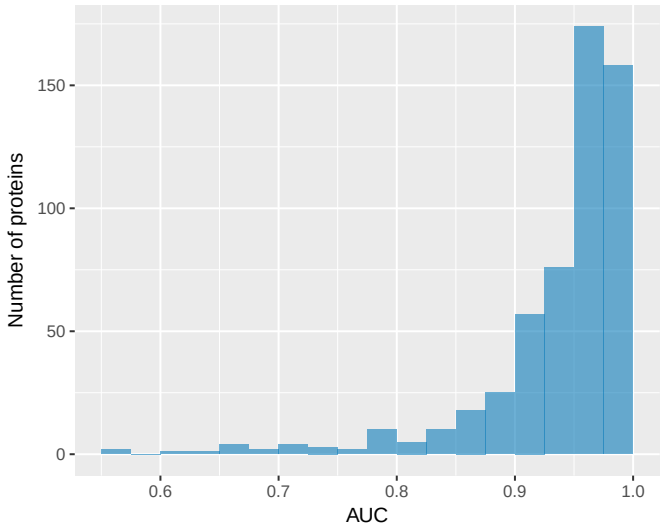

Fig. 3. Histogram of the AUC values on distinct proteins.

IV. Data

In the current version of Deep50 we focus on the interaction measure IC50, which measures the concentration of the chemical compound necessary to inhibit the baseline activity of the protein by 50%. We prepared a dataset from the public bioactivity database ChEMBL, Version 22.1[7]. First we searched for all *Homo sapiens* proteins and removed duplicated measurements. Than we selected the 671 proteins that had at least 100 IC50 measurements, and computed the negative logarithm of the measurement values (pIC50). We set four pIC50 activity cutoffs (5.5, 6.5, 7.5, and 8.5) and defined a binary classification task for each protein-cutoff pair. We transformed the original pIC50 values from ChEMBL to the thresholded values. In ChEMBL there is a relation associated to every value, which can be "greater than", "less than" or "equal" type. In the case of inequalities we set the corresponding binary activity only if their in can be determined unambiguously (the ambiguous values are removed). For extracting the substructural features we used RDKit [8] Morgan Fingerprint with radius 3 [9], which resulted in an 828,000 dimensional sparse vector h_{fp} for each compound.

V. Evaluation

For the evaluation, a test set containing 20% of the observations is chosen at random. We computed AUC for protein-threshold pairs where the test set contained at least 10 active and 10 inactive measurements. We observed average AUC of 0.930. We found the predictive performance comparable but the method significantly more computationally efficient than the matrix factorization based methods [4]. Our model achieves, at least in one of the pIC50 thresholds, an AUC value higher than 0.9 on 465 out of 552 distinct protein targets (note that 119 proteins did not have enough active measurements to accurately evaluate the predictive performance). For the histogram of the observed AUC values see Figure 3.

VI. Future work

One of our further goal is to integrate an error model into Deep50 for the predictions. We aim to extend the well known Platt scaling approach [10] with chemical space coverage information. Our architecture does not depend on a fixed feature representation like the Morgan fingerprint. As the model is fully differentiable, we could combine the proposed compound embedding with the supervised learning of features directly from the chemical structures, like has been presented in a recent work [11].

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[2] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[3] N. Novac, "Challenges and opportunities of drug repositioning," *Trends in pharmacological sciences*, vol. 34, no. 5, pp. 267–272, 2013.

[4] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau, "Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC," *ArXiv e-prints*, Sep. 2015.

[5] A. Arany, J. Simm, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau, "Highly scalable tensor factorization for prediction of drug-protein interaction type," 2015.

[6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[7] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey *et al.*, "The chembl bioactivity database: an update," *Nucleic acids research*, vol. 42, no. D1, pp. D1083–D1090, 2014.

[8] Rdkit: Open-source cheminformatics. [Online]. Available: http://www.rdkit.org

[9] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[10] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[11] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of computer-aided molecular design*, vol. 30, no. 8, pp. 595–608, 2016.