Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Drug-Target Prediction: an application of Machine Learning

Henrik Otterstedt[1,*], Christian Dallago[1], Burkhard Rost[1]

[1]TU München, Garching

*To whom correspondence should be addressed; ge52keq@mytum.de

Abstract

Experimental drug repurposing is a resource-intensive process. To save resources, computational methods largely based on machine learning techniques are used to find new Drug-Target Interactions (DTI) for already approved drugs. The main challenge when using machine learning for this purpose is the availability of negative DTI to train on. In this work, negatives were artificially created based on heuristics which take the quality and extent of experimental annotation for drugs in drugbank into consideration. Training different machine learning models, the support vector machine (SVM) achieved an area under the ROC curve (AUC) of 0.753 ± 0.006, which taking the difference in computational resources into consideration compares well to the AUC of 0.886 ± 0.010 Wang et al.s' network-based state-of-the-art algorithm archives.

Supplementary information: The source code, as well as the SeqVec embeddings of the proteins used are available at https://github.com/HenrikOtterstedt/drug-target-prediction

1 Introduction

Experimental drug development is usually a high cost and time-consuming procedure. The process of bringing a new drug to the market can take up to 10 years and require an investment of a billion Dollars (Institute of Medicine, 2014). There is a lack of reliable ways to predict drug targets, which can be observed by the low clinical target validation success rate (Emig et al., 2013). Therefore, new procedures are required to accurately predict drug targets for diseases (Rask-Andersen et al., 2011). Furthermore, accurate DTI prediction may aid the re-use of existing drugs (which have already passed many of the hurdles faced by initial clinical trial phases) for new uses. This process is also dubbed drug repurposing (Gilbert et al., 2003). There are a multitude of methods used to repurpose drugs: Previous research highlights how a combination of LASSO and a Deep neural network (You et al., 2019) outperforms simpler types of machine learning algorithms. Another class of methods for drug-target interactions (DTI) prediction are graph-based approaches (ie.: trying to complete graphs based / graph completion challenges), which successfully and accurately predict previously unknown DTIs (e.g. for traditional Chinese medicine (Wang et al., 2019)). A different approach was taken by Keiser et al. who proposed a chemoinformatic way of finding new DTIs by using the similarity ensemble approach to compare the similarities between a drug and a proteins ligand and predicting according to the ligand similarities (Keiser et al., 2009). Further approaches use reported drug side effects as the distance measurements within their network (Takarabe et al., 2012). This project aims to take a first glance at simple machine learning methods and how they perform within the task of predicting new DTIs.

2 Material & Methods

The dataset used for training and evaluating the machine learning implementations is based on drugbank (Wishart et al., 2018). It is an extensive database providing information about drugs used in clinical settings and their respective protein targets. Overall, drugbank stores 21,473 DTIs. For machine learning applications this data presents one major challenge: it has no negative DTIs available. In this work, negative DTIs were thus generated.

Generation of negative DTIs. The space of possible DTIs (negative and positive ones) is spanned by the combination of all drugs (in our particular case we are interested only in small-compounds) and all proteins reported as targets in the database. These amount to 11,334 small molecules and 2593 proteins, for a combinatorial space of 29,389,062 possible DTI. Negatives cannot simply be set difference of the combinatorial space of all DTIs and the recorded positives ones, as this would almost certainly lead to the inclusion of interactions that are marked as negatives but in reality, do exist. To overcome this limitation, we generated the negatives only using well-annotated small molecules. The intuition here is that drugs which are better annotated have had more research done about them. Following this logic the more research there has been on a drug the more likely it is that all existing DTI for this drug have been found. Since drugbank does not include an indication of the quality of annotation, we used two heuristics based on the number of known interactions. We decided that drugs which have between 3 and 10 noted interactions are to be considered “well annotated”. This was done on the basis that a drug with less than three interactions has an increased likelihood of not being extensively annotated or researched. On the other hand, drugs having more than ten interactions were eliminated based on the high possibility of them binding to more proteins than the ones reported, thus possibly being party hubs. In the end this leaves us with 1,161 drugs and 2593 proteins, for a space of 3,010,473 possible DTIs. Removing the positive recorded DTIs (995), we generated 21,331 negative interactions to achieve a ratio of 1:1 for positives and negatives.

For the one-class classifier a separate redundancy reduced dataset was created. The proteins and drugs within it are reduced to 50% similarity. To achieve this for the drugs we employed the Tanimoto index (Bajusz et al., 2015). This leaves us with a total of 1558 drugs. The proteins were reduced using CD-HIT (Fu et al., 2012) resulting in 3798. The drugs and proteins then were examined as to existing DTI between them. 4568 were found and subsequently used to train on.

Feature extraction. Small molecules and proteins had to be brought into a suitable form for machine learning. We achieved this by converting the SMILES (simplified molecular input line entry specification) representing the small molecules into a topological fingerprint via RDKit (Landrum, 2016). Bitvectors of length 1024 thus represent the small molecules. This process can be seen in Figure 1.

Proteins were converted to 1024 dimensional vectors by using SeqVec (Heinziger et al., 2019).

The drug and protein vectors where thus of equal length / same numbers of features (1024), and the concatenation of small molecule and SeqVec was fed to various ML devices.

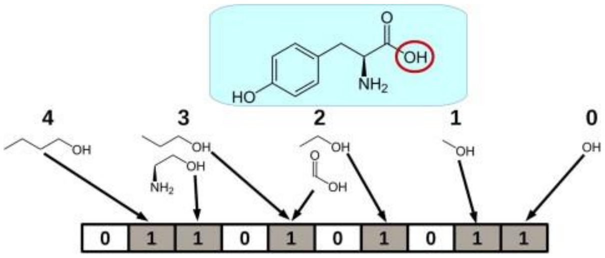


Fig. 1: The process of constructing a chemical fingerprint for a molecule (Cereto-Massagué et al., 2015)

ML devices. In the context of this project, various ML devices where tested on the aforementioned input features. Models were implemented using Scikit-learn (Pedregosa et al., 2011). Models tried include: a Multi-layer Perceptron (MLP), an artificial neural network (ANN) with one hidden layer, which consists of 75 neurons and uses a constant learning rate as and ReLu (Nair et al., 2010) as well as Adam (Kingma et al., 2014) as the activation function and the solver, respectively.

The second model tested was a one-class classifier. This model was implemented using the Support Vector Classifier (SVC), which uses the concept of Support Vector Machines (SVM) (Cortes, 1995). The classifier only trained with positive interactions, but for evaluation purposes the test set has a ratio of 1:1 for positive and negative samples.

Lastly, a SVC was also trained using the entirety of the data. Both SVCs used default parameters.

Data was fitted using 5-fold cross-validation. This cross-validation was kept for all further implementations as well.

Evaluation criteria. To evaluate the performance of any given machine learning we chose to use the confusion matrix, precision, recall, accuracy, the F1-score and the Matthews correlation coefficient (MCC), which can be computed using the data in the confusion matrix. Each scores formula can be seen below.

The shown performance was calculated on a randomly sampled 20% of the available data.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

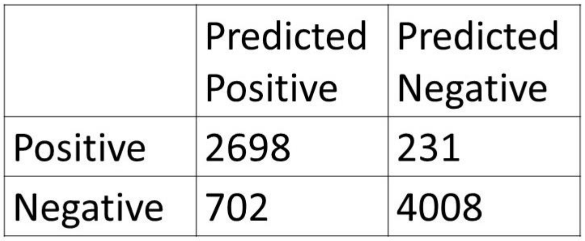
$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1-score} = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

3 Results & Discussion

The first performance to take a look at are the ones achieved by the MLP. When looking at the results, displayed as confusion matrices in Figure 2, it springs to mind that the MLP, with only 231 falsely predicted negatives was very good at predicting whether a given interaction is negative or not. Overall it has good accuracy (at 0.877 ± 0.003). The precision, which shows how sure the ANN is in its positive predictions, lies at 0.921 ± 0.005. This score is excellent, especially when the recall at 0.793 ± 0.006 still lies reasonably high. This is reflected in the 0.852 ± 0.004 for the F1-Score, which is the harmonic mean of precision and recall. The MCC at 0.755 ± 3.33e-16 shows that there is some substance to the predictions. The Area Under the ROC (AUC), a widely used metric, is 0.946 ± 0.002 All these scores paint a picture of an ANN that performed way above expectations and even better than some state-of-the-art algorithms, e.g. several SVMs proposed by Yu et

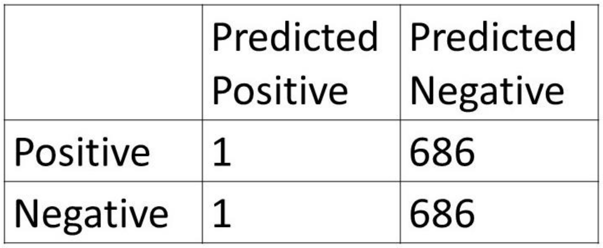


	Predicted Positive	Predicted Negative
Positive	2698	231
Negative	702	4008

Fig. 2: Confusion matrix showing the performance achieved by the MLP

al. that achieved an AUC of 0.905 (Yu et al., 2012). This prompted further inspection. We started by reviewing all of the cases the MLP predicted samples to be negative. These samples seemed to all originate from the same individual drugs. Through further inspection, we discovered that the drugs with interactions that were predicted negative and the ones which were predicted positive are mutually exclusive, which shouldn't happen as every drug within the dataset at least has one positive DTI. This means that the MLP had discovered the pattern by which the negative samples were generated and bases its predictions on it, rather than using the chemical information to determine if a given drug binds to the protein. The MLP performs well in theory but does not do so in practice as purely biological data does not include the artificial patter, which our heuristics for the creation of negative DTI introduced and which the predictions are heavily biased towards. The MLP seems to not be compatible with our intuition on how to create negative DTI to train on. The best way to train an MLP with generated data seems to be to mirror the distribution for positive DTI for negative DTI as well (You et al., 2019). An even further improvement for the quality of predictions would be to base the negative DTI in reality and therefor use experimentally annotated negative DTI as data.

The aforementioned problem might be solved by using a one-class SVM, which only trains on positive samples. Testing used a ratio of 1:1 for positive as well as negative interactions.

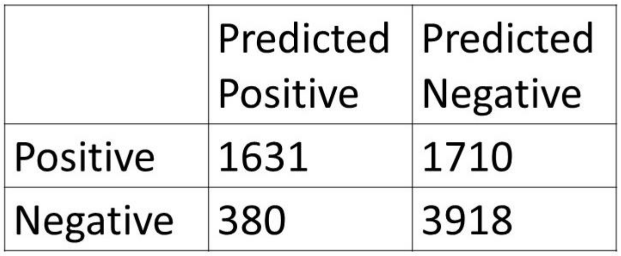


	Predicted Positive	Predicted Negative
Positive	1	686
Negative	1	686

Fig. 3: Confusion matrix displaying the one-class SVCs' performance

The results in the form of a confusion matrix can be seen in Figure 3. The most salient figure about the one-class SVM are predicted negatives, which are only 2. The performance evaluating scores of precision at 0.5, the recall at 0.001, the accuracy with a score of 0.5, the F1-score at 0.002 and the MCC at 0.0. The nature of an SVM can explain these results. SVMs use hyperplanes to separate classes. In the particular case of a one-class SVM, the hyperplane is set to enclose all of the training data and then make predictions based on the distance in hyperspace between the sample which is to be predicted and the enclosed area. In this case, it so happened that most of the data within the test set fell into an area that was not included by the SVM.

As the last machine learning device, we chose to employ an SVM trained with all the data. The resulting confusion matrix is displayed in Figure 4.



	Predicted Positive	Predicted Negative
Positive	1631	1710
Negative	380	3918

Fig. 4: Confusion matrix resulting from the SVMs' performance

The precision of 0.811 ± 0.008 is almost as good as the one achieved by the faulty ANN. Whereas the recall, describing the amount of correctly identified positives, at 0.488 ± 0.008 has dropped down significantly compared to the faulty MLP. The accuracy with 0.726 ± 0.005 and the f1-score at 0.609 ± 0.007 are still up to par. The MCC at 0.450 ± 0.001 suggests a lower correlation between the prediction and reality than the one shown by the MLP. These results are still better than chance.

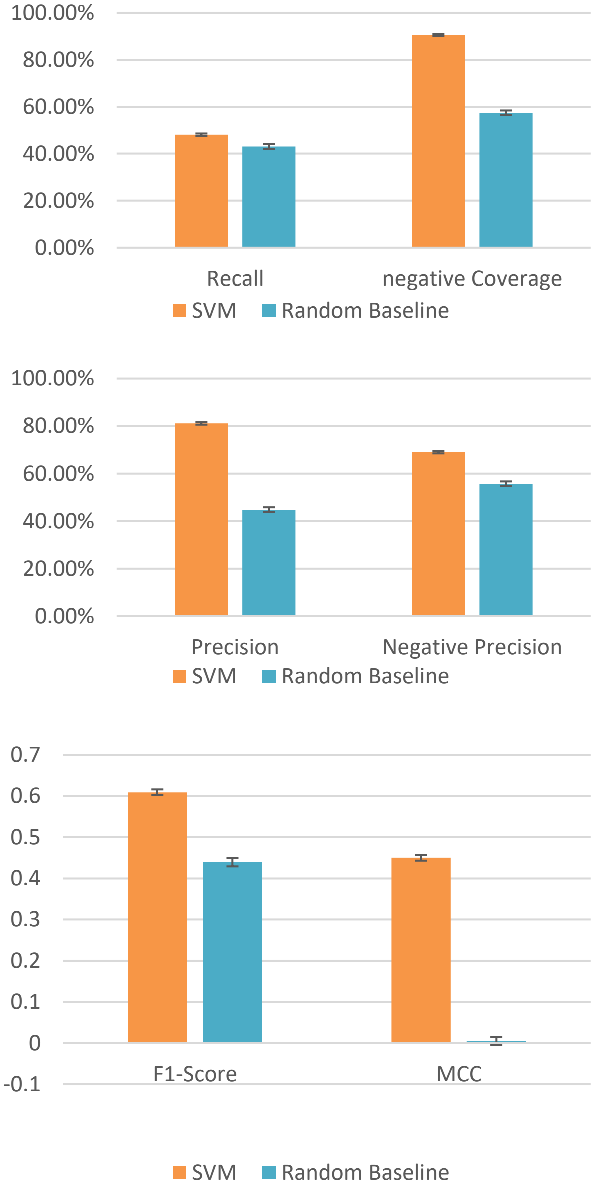


Fig. 5: Comparison between the SVM and a Random Baseline
A: Recall and negative Coverage
B: Precision and negative Precision
C: F1-Score and MCC

Figure 5 displays how these scores compare to baseline predictions, which makes its predictions entirely on the class distributions. From these comparisons we gather that all metrics, but the recall, do outperform the baseline. The measurements of negative coverage and negative precision are also shown in Figure 5.

$$\text{Negative Coverage} = \frac{TN}{TN+FP} \quad (6)$$

$$\text{Negative Precision} = \frac{TN}{TN+FN} \quad (7)$$

They are computed as shown and describe how well the SVM manages to identify negative samples. This score is high as well. A further investigation into the detailed results demonstrates that the issue appearing within the predictions of the ANN does not repeat within the SVMs' predictions. While a basic form of the trend (aka SOME of the drugs being mutually exclusive between positive and negative predictions still exist) it is not as extreme as in the ANNs' predictions. About 70% of all data points partly consisting of one of the drugs used to create the arbitrary negative DTIs were still predicted negative. In contrast to the ANN, the SVM managed to correctly identify some DTIs including these drugs, which stem from drugbank and so are true positives.

Lastly, we want to see how our machine learning devices perform in comparison to one state-of-the-art algorithm. This is done by comparing it to a network-based algorithm, which was developed by Wang et al. (Wang et al., 2019). This method also leverages data from drugbank [cit], and employs a combination of a Least absolute shrinkage and selection operator (LASSO) [cit] and a deep neural network (DNN) (You et al., 2019). The comparison is made based on the area under the ROC curve (AUC), which shows how much the model is capable of distinguishing between classes. Figure 6 shows the results of this comparison. It clearly shows that the LASSO-DNN closely outperforms the homogenous network-based prediction. Both are vastly outperformed by the faulty ANN, whereas both SVMs are worse. The ANNs' superb performance can be disregarded The SVM trained with the entirety of the dataset performed 0.13 worse than state of the art, but does not need as many computational resources to train.

Wang et al.	You et al.	MLP	One-class-classifier	SVC
0.886 ± 0.010	0.89	0.946 ± 0.002	0.237 ± 0.012	0.753 ± 0.006

Fig. 6. Model performance comparison. AUCs of different state-of-the-art algorithms and models proposed in this paper

4 Conclusion

The biggest limitation of machine learning based drug-target prediction lies with the availability of negative data. When arbitrarily generating negative data, it is crucial to make sure that downstream predictions don't leverage artificially introduced distinctions in the data to perform predictions.. This pattern was indeed present in the ML devices tested in this work. We discovered that in our case, an SVM is less likely to stick to a pattern introduced in the data than an ANN. Data-driven error in DTI prediction might be prevented by using a different approach in constructing the negative samples. One alternative approach is proposed by You (You et al., 2019). In their work, You et al. eliminate the most likely unnoticed positive interactions and mirror the distribution of targets for the negatives as well seemed to work well for a similar approach (Endres et al., 2020). Even more relevant would be experimentally validated negative samples.

References

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics, 7. https://doi.org/10.1186/s13321-015-0069-3

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G.(2015). Molecular fingerprint similarity search in virtual screening. *Methods, 71,* 58-63. https://doi.org/10.1016/j.ymeth.2014.07.005

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., & Bessarabova, M. (2013). Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLOS ONE*, *8*(4), e60618. https://doi.org/10.1371/journal.pone.0060618

Endres L., Dallago, C., & Rost, B. (2020). Drug Target Prediction using a Artificial Neural Network and a Random Forrest Classifier. Unpublished report

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England), 28(23), 3150-3152. https://doi.org/10.1093/bioinformatics/bts565

Gilbert, J., Henske, P., Singh, A. (2003). Rebuilding big pharma's business model. Vivo, Bus. Med. Rep. 21, 73-80

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, *20*(1), 723. https://doi.org/10.1186/s12859-019-3220-8

Institute of Medicine. (2014). *Drug Repurposing: Workshop Summary.*Washington, DC: The National Academies Press. https://doi.org/10.17226/18731

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., … Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature*, *462*(7270), 175–181. https://doi.org/10.1038/nature08506

Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations

Landrum, G. (2016). RDKit: Open-Source Chminformatics Software. https://rdkit.org

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Bolzmann Machines. 8.

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rask-Andersen M, Almen MS, Schioth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10: 579–590.

Takarabe, M., Kotera, M., Nishimura, Y., Goto, S., & Yamanishi, Y. (2012). Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, *28*(18), i611–i618. https://doi.org/10.1093/bioinformatics/bts413

Wang, Z., Lin, H.-H., Linghu, K., Huang, R.-Y., Li, G., Zuo, H., … Hu, Y. (2019). Novel Compound-Target Interactions Prediction for the Herbal Formula Hua-Yu-Qiang-Shen-Tong-Bi-Fang. *Chemical and Pharmaceutical Bulletin,* 67(8), 778-785. https:/doi.org/10.1248/cpb.c18-00808

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. https://doi.org/10.1093/nar/gkx1037.

You, J., McLeod, R. D., & Hu, P. (2019). Predicting drug-target interaction network using deep learning model. *Computational Biology and Chemistry*, *80*, 90-101. https://doi.org/10.1016/j.compbiochem.2019.03.016

Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li,X., Zhou, W., Wang, W., & Wang, Y. (2012). A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data. PLoS ONE, 7(5). https://doi.org/10.1371/journal.pone.0037608