Ligand-based virtual screening for discovery of JAK2 inhibitors

Roman Zubatiuk,[1] Omar Aljawfi,[1] Third Author, Benjamin Garlington,[1] and Robin Ghosh[1]
[1]Jackson State University, Jackson MS roman@icnanotox.org

Abstract*– Various machine learning techniques were applied to dataset of over 15,000 molecules with experimentally measured IC50 values against JAK2 kinase targets. The purpose of the work was to develop a QSAR model for prediction of IC50 values. Using a set of 918 2D molecular descriptors, the best prediction performance was archived with feed-forward deep neural network. RMSD on target pIC50 values was estimated as 0.48 using 5-fold cross-validation of the trained model. It was demonstrated, that Random Forest can archive RMSD of 0.54 using less than 50 descriptors. Developed QSAR models can be used for in silico screening of potential JAK2 inhibitors.

Keywords—SQAR, JAK2 inhibitors, IC50, Neural Networks

I. Introduction

Abnormal activity of tyrosine-protein kinase Janus kinase 2 (JAK2) is related to both chronic and acute forms of eosinophilic, lymphoblastic and myeloid leukemia [1]. The V617F JAK2 gene mutation results in the production of a JAK2 protein that is constantly turned on (constitutively activated), which, in essential thrombocythemia, leads to the overproduction of abnormal blood cells called megakaryocytes. Because platelets are formed from megakaryocytes, the overproduction of megakaryocytes results in an increased number of platelets. Excess platelets can cause abnormal blood clotting (thrombosis), which leads to many signs and symptoms of essential thrombocythemia. Somatic mutations in the JAK2 gene are associated with polycythemia vera, a disorder characterized by uncontrolled blood cell production. The V617F mutation is found in approximately 96 percent of people with polycythemia vera. Somatic JAK2 gene mutations are also associated with primary myelofibrosis, a condition in which bone marrow is replaced by scar tissue (fibrosis) [2].

Up to date, only one approved JAK2 inhibitor exists, with positive, but not curative effects in myeloproliferative neoplasms, and promising effects in autoimmune diseases and cancer [3]. Several other drug candidates are in various stages of clinical trials [4].

One of the important steps for the drug discovery process is a search for the molecules, which are potent to demonstrate high activity towards specific biological target, and possess required physical, chemical and biological properties, like solubility, lipophilicity, toxicity, etc. Such screening in large scale performed in silico with Qualitative Structure Activity Relationship (QSAR) approach, which is based on machine learned models for prediction of specific activity based on molecular structure.

The aim of the project is to discover potential efficient inhibitors of JAK2 based on ligand-based virtual screening approach using a database experimentally measured inhibition activities for about 15,000 drug-like compounds.

II. Computational Details

A. Descriptors generation

Set of 2D molecular descriptors was calculated using PaDEL [5] (1444 descriptors) and RDKit [6] (196 descriptors) software packages. A combined set of 1640 descriptors for the dataset of approximately 15k molecules was further cleaned using following procedure:

- Removed 321 features with low relative variance, defined as Std / (Max – Min). The criteria for dropping feature was chosen 0.05 for discrete features and 0.01 for continuous features.
- Removed 401 features which show correlation with R2 > 0.95 with any other feature within the dataset using greedy algorithm.

Final set of 918 features was used for model selection and for regression analysis.

B. Machine Learning Software Libraries

Data storage, preprocessing and manipulation was performed using Pandas 0.19.2 Python library [7]. For the of machine learning techniques Scikit-learn [8] library was used, except for Neural Networks, for which Keras [9] implementation with Theano [10] backend was used. Seaborn library [11] was used for plotting.

III. Training Dataset

The dataset of 15349 JAK2 inhibitor molecules along with experimentally measures activities (IC50) was compiled from public ChEMBL [12] and PubChem [13], and commercial Kinase Knowledgebase databases (KKB [14]). Within the dataset, pIC50 values range from 2.0 to 11.4 with mean and median of 7.0 and standard deviation of 1.2. The distribution of pIC50 values within the dataset is shown at Fig. 1.

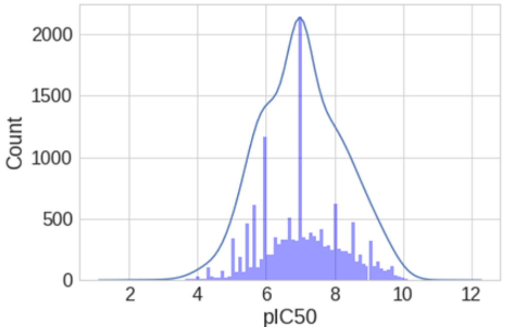

Figure 1. Histogram of experimentally measured pIC50 values within the training dataset.

Our main goal is to build a model to predict IC50 against JAK2 target based on 2D molecular structure. To be generally applicable, the model need to be trained against diverse dataset. We have estimated diversity of the molecules within the dataset using topological RDKit fingerprints. This type of fingerprint enumerates occurrence of substructures within a molecule with up to 7 bonds involved, and encodes the result as 2048-bit vector. The bit vectors *Vi* and *Vj* could be further compared using Tanimoto distance metric:

$$Tani\left(V_i, V_j\right) = \frac{V_i \cdot V_j}{|V_i| + |V_j| - V_i \cdot V_j}$$

At the Fig. 2 distribution of pairwise Tanimoto similarities are shown, which are defined as *1 – Tani*. Most of the molecule pairs have similarity of 40%, and only a small fraction (0.73 %) of pairwise similarities are above 60%. DBSCAN clustering algorithm, applied with minimum similarity of 90% and minimum cluster size of 20, partitions the dataset into 90 distinct clusters, with 25% of the points located outside of any cluster. Such distribution indicates substantial diversity of the dataset and let as expect good generalization of the derived IC50 prediction models.t

IV. RESULTS AND DISCUSSION

A. *Linear Models*

Linear regression models represent one of the most easily interpretable and simple machine learning techniques. Simple L1 (Lasso) and L2 (Ridge) regularization models could be used to identify the most important features for the model. We have applied both models to the dataset using all 918 features generated by RDKit and PaDEL. Regularization parameters for Lasso and Ridge regressions were selected using leave-one-out 5-fold cross-validation procedure to 0.0002 and 10, respectively. Cross-validation lead to R2 fit scores of 0.642(13) and 0.637(11), which could be translated into RMS errors of pIC50 prediction of about 0.67 for both methods.

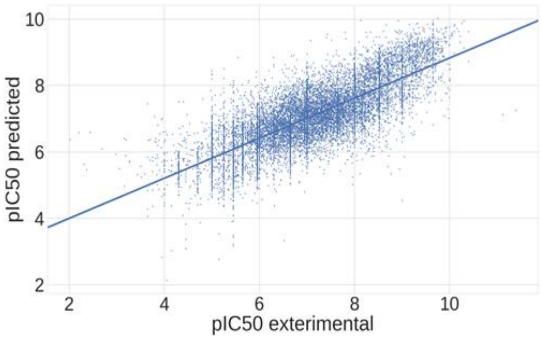

Figure 3. Experimental pIC50 values vs predicted using non-regularized linear regression (R2 = 0.56).

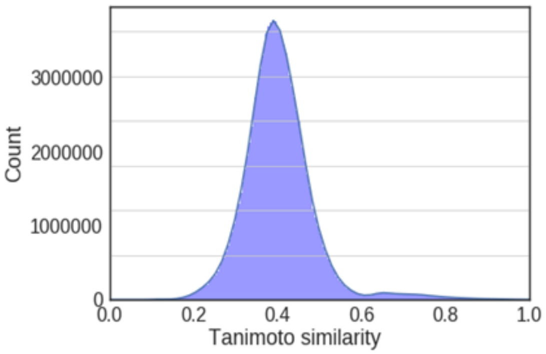

Figure 2. Pairwise Tanimoto similarities for the training dataset.

Both Lasso and Ridge discriminate feature importance and identify 329 (Lasso) and 420 (Ridge) with coefficients above 0.01. Non-regularized linear regressions fitted against these sets of features give R2 5-fold cross-validation scores of 0.564(15) and 0.579(15), respectively. Thus, reducing feature set leads to downgrade of regression models. Fit quality of non-regularized linear model on 329 descriptors is presented on Fig. 3.

Failure of linear models to provide decent regression quality might indicate complicated nature of our target concept. This could be rationalized by the fact, that the JAK2 family of kinases have several binding sites for inhibitor ligands, with different pocket structure. Thus, different factors might descript binding affinity for the different molecules within the dataset.

B. *Recursive Feature Elimination with Random Forest*

Random Forest [15] is an ensemble method, it combines a set of Decision Trees, each is built upon a

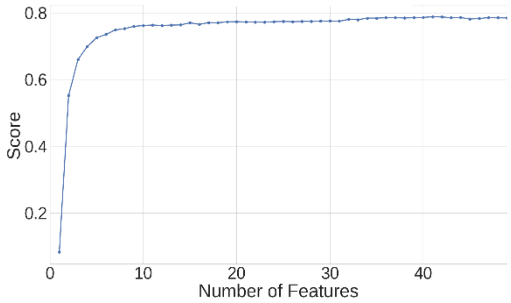

Figure 4. R^2 regression score for recursive feature elimination with Random Forest.

bootstrapped sample of data instances. The output of all trees is combined to produce one final prediction. For regression problems, the output is average of outputs of individual trees. Compared to single Decision Tree, Random Forest has several major advantages. It is resistant to overfitting, can efficiently deal with large number of features, and can learn very complex concepts. For the QSAR application Random Forest usually provides much better classification and regression models, and feature importance.

Application of the Random Forest repressor model to the training dataset with full set of 918 features immediately gives R^2 of the regression model of 0.78, which is significantly better, compared to archivable with linear models.

We have applied Random Forest method to identify the most important features of the molecules within the dataset. This might further lead to better understanding the complexity of the problem and will facilitate rationalization of the QSAR model. We used recursive feature elimination (RFE) method with the following protocol:

- Split dataset into train and test subsets with 0.7:0.3 ratio;
- Fit Random Forest model against train data;
- Record regression R^2 score on test data;
- Remove n percentile of the least important features, with n = 100 / N(features). Thus, removing up to 10% of features on first iteration, and removing a single feature when their number is below 100.
- Repeat loop until single feature left.

R^2 scores for RFE run are shown on Fig. 4. Starting from 918 features, elimination of least important features does not result in drop of regression performance, until about 50 features left (R^2 = 0.79-0.79). Starting from that point, performance decreases slowly, until 10 (R^2 = 0.76) or 4 features left (R^2 = 0.70). Starting from that point sharp performance drop is observed with elimination of every single feature.

Four most important features are path count index MPC9, and Burden modified eigenvalues SpMin1_Bhp, SpMax1_Bhe and SpMax1_Bhv. All these indexes have no direct connection with physical or chemical properties of the molecules, but rather encode topological information from the whole molecule.

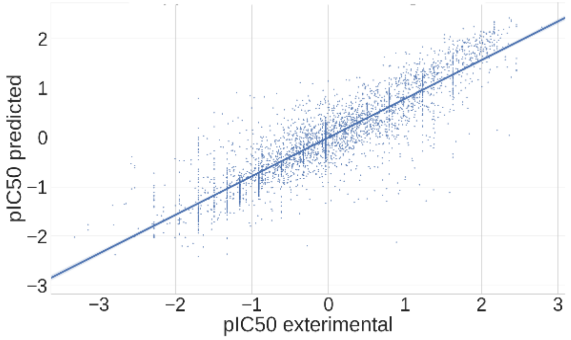

Figure 5. Experimental pIC50 values vs predicted using Support Vector Regression (R^2 = 0.82).

More rigorous evaluation of Random Forest repressor with respect to the number of features using 5-fold cross-validation technique is presented in Table 1. According to these data, number of features close to 42 is optimal for Random Forest to learn decision function for IC50 values for the dataset of about 15k JAK2 inhibitors.

Table 1. Performance of Random Forest repressor with respect of number of features used

Number of features	CV5 R^2 score	Standard deviation
4	0.714	0.007
42	0.793	0.007
99	0.793	0.006
918	0.784	0.006

C. Support Vector Machine Regression

Support Vector Machine (SVM) class of methods use embedding procedure – increasing dimensionality of feature space by application of linear or non-linear kernel functions in order to linearize the problem. With polynomial or Gaussian-type kernels, SVM is very powerful machine learning technique which can learn very complex concepts. However, the algorithm requires matrix inversion operations, thus it is computationally expensive both for training and prediction phases.

Application of Support Vector Regression (SVR) model to the dataset of 15k instances with 918 features could be considered as impractical. Thus, we have applied SVR method to learn pIC50 values from a limited set of 99 features, obtained by RFE procedure with Random Forest technique.

SVR method has several tunable parameters: kernel function, regularization strength C and tube size ε. The data points inside the tube do not introduce additional penalty in cost function. All three parameters were selected using grid search algorithm. Good parameters are found to be: Gaussian kernel, C = 10 and ε = 0.1. With this parameters SVR method with 5-fold cross-validation data selection scheme gives regression R2 score of 0.819(7) (Fig 5). Thus, SVR method shows better performance then Random Forest, when is applied to the same set of features.

D. Feedforward Neural Network

Neural Networks (NN) are the proven method to describe concept of high complexity. They are archived by the vary substantial number of adjustable parameters of network, and complex network topologies. From the other hand, this introduces two major drawbacks for practical applications of neural networks. They generally need large datasets to archive good generalization for target concept, however still prone to overfitting. And there is no universal systematic approach to build a network topology for the precipice problem.

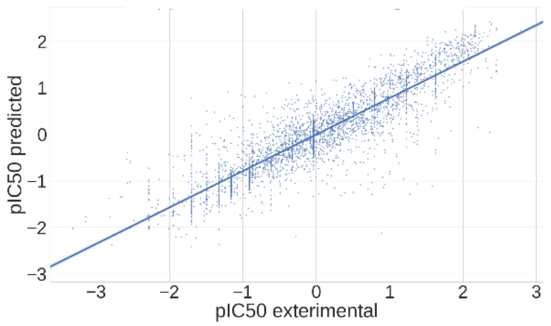

Figure 6. Experimental pIC50 values vs predicted using Feed-Forward Neural Network ($R^2 = 0.84$).

The overfitting problem for neural networks could be solved with several popular regularization approaches:

- Introducing L2 or L1 penalty for neural network weights;
- Introduce sparsity to network by dropping some connections or switching off neutrons randomly during training phase;
- Batch normalization [16].

The former is essentially the normalization of layer inputs for every training minibatch. This have been proven to be a very powerful technique, which allows higher learning rates and better generalization. For our problem, we found that Batch Normalization works much better that dropout or L1/L2 regularization.

After experimenting with different NN parameters, we came up with the network with 1,169,301 trainable parameters and the following architecture:

- Number of hidden layers: 5;
- Layer dimensions: [918, 400, 200, 100, 100, 1];
- Activation function: ReLU;
- Regularization: Batch Normalization;
- Batch size: 16.

The network was trained for 250 epochs using Adam optimization algorithm with scheduled learning rate decrease every 50 epochs by 5 times: from 10-3 to 10-5. During training phase, we have not observed overfitting effect, i.e. raise of cost function for test dataset while optimizing network against training dataset.

With 5-fold cross-validation procedure, trained NN demonstrated R2 regression score of 0.840(11).

V. SUMMARY AND FUTURE WORK

We used several machine learning techniques to build predictive model of IC50 values against JAK2 kinases using set of 2D molecular descriptors for 15020 molecules with reported experimental values of IC50. The best performance we archived is summarized in Table 2. Here, we do not report performance of linear models, because it was found to be sub-optimal.

Table 2. Performance of deferent ML methods for prediction of IC50 values, estimated using 5-fold cross-validation

Number of features	CV5 R^2 score	Standard deviation
4	0.714	0.007
42	0.793	0.007
99	0.793	0.006
918	0.784	0.006

Neural network model provides the best regression score, which is however only marginally better, then Support Vector regression. Z-score between these methods is 1.7. From the other hand, Random Forest archives satisfactory performance with very limited number of features, and provides valuable information about relative feature importance.

Developed QSAR models will further be used for screening of potential active compounds and for the rationalizations of molecule activity towards JAK2 kinases.

III. REFERENCES

[1] Tefferi, A., 2010. Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. *Leukemia*, 24(6), pp.1128-1138.

[2] James, C., 2008. The JAK2V617F mutation in polycythemia vera and other myeloproliferative disorders: one mutation for three diseases*?* *ASH Education Program Book*, 2008(1), pp.69-75.

[3] Verstovsek, S., Mesa, R. A., Gotlib, J., Gupta, V., DiPersio, J. F., Catalano, J. V., & Winton, E. F. (2017). Long-term treatment with ruxolitinib for patients with myelofibrosis: 5-year update from the randomized, double-blind, placebo-controlled, phase 3 COMFORT-I trial. *Journal of Hematology & Oncology*, 10(1), 55.

[4] Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E. and Davies, M., 2017. The ChEMBL database in 2017. *Nucleic acids research*, 45(D1), pp.D945-D954.

[5] Yap, C.W., 2011. PaDEL‐descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry, 32(7), pp.1466-1474.

[6] Landrum, G., 2016. *RDKit: Open-source Cheminformatics*. http://www.rdkit.org

[7] McKinney, W., 2015. Pandas: a Python data analysis library. http://pandas.pydata.org.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), pp.2825-2830.

[9] Chollet, F., 2015. Keras: Deep learning library for Theano and Tensorflow. http://keras.io

[10] Theano Development Team, 2016 Theano: A Python framework for fast computation of mathematical expressions arXiv e-prints, vol. abs/1605.02688, May 2016

[11] M. Waskom, O. Botvinnik, Drewokane, P. Hobson, Y. Halchenko, S. Lukauskas, J. Warmenhoven, J. B. Cole, S. Hoyer, J. Vanderplas, S. Villalba, E. Quintero, M. Martin, A. Miles, K. Meyer, T. Augspurger, T. Yarkoni, P. Bachant, C. Evans, C. Fitzgerald, T. Nagy, E. Ziegler, T. Megies, D. Wehner, S. St-Jean, L. P. Coelho, G. Hitz, A. Lee, Luc Rocher, 2016 Seaborn: v0.7.1 http://seaborn.pydata.org

[12] Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E.

and Davies, M., 2017. The ChEMBL database in 2017. *Nucleic acids research*, 45(D1), pp.D945-D954.

[13] Wang, Y., Bryant, S.H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B.A., Thiessen, P.A., He, S. and Zhang, J., 2017. PubChem BioAssay: 2017 update. *Nucleic Acids Research*, 45(D1), pp.D955-D963.

[14] Eidogen-Sertanty Kinase Knowledgebase (KKB) http://eidogen-sertanty.com/kinasekb.php

[15] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958.

[16] Ioffe, S., Szegedy C. 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML-15).*