Bioinformatics, 2021
Advance Access Publication Date: 31/01/2021

Investigation of the modelling of the molecular structure-activity relationship of the AMPK alpha subunit

Ruzette A.[1], Guoshuo Y.[1], Bu T.[1], Menezes K.[1]

[1] Master in Bioinformatics, Faculty of Bioscience Engineering, KU Leuven

Abstract

Motivation: Drug discovery for diseases is a long, expensive process where only a few candidates are able to reach the testing stage. The goal of the project was to utilize publicly available data of the ligand activity of the AMPK subunit to test various Machine Learning methods to determine their predictive ability to whether the given molecule reacts with the AMPK subunit or not. This is done by transforming molecular information into the format of their molecular descriptors and then testing out the models. Comparison of their accuracy using the full set of descriptors versus a reduced set was done to establish differences in performance.

Results: The publicly available data seemed poor in classification in active/non-active status of the molecules. The models confirmed the same, having a low-test accuracy. A robust pipeline for processing data to testing various methods was developed using the outlined strategy using ensemble methods and Artificial Neural Networks (ANNs), which were combined with LASSO to obtain a reduced set of variables. Further trials with annotated data could yield promise for further testing if the datasets used are significantly larger than the ones tested out in this project.

Availability: scripts provided through GitHub - https://github.com/AntoineRuzy/bioactivity-structure-classification

Contact: antoinealexism.ruzettechevalier@student.kuleuven.be

1 Introduction and motivations

The traditional drug discovery and development procedures are time-consuming, lasting approximately 12-15 years and costing $1-2 billion dollars, from the initial discovery stage to the market. At the same time, the success rate is extremely low and only around 10% of drug candidates can reach human clinical trials [1]. Recent years, the largely increased types and the number of available biological and disease-related data sets grows the opportunities and interests for pharmaceutical industries to apply machine learning algorithms on drug discovery, accelerate the development, and reduce costs [2]. Compared with physical models that are used in drug discovery, for instance, quantum chemistry approaches and molecular dynamics simulations, machine learning is highly efficient, less computationally-intensive, and easier to scale and apply on large data sets [3]. Besides elementary algorithms, applications of advanced machine algorithms such as deep learning have landed on the ground and proved their ability to construct potent models for pharmaceutical companies [2].

In our project, we focus on the in-silico screening step of drug discovery. Multiple machine learning algorithms are applied based on the chemical structures of ligand molecules to identify the bioactivity level with our target: AMPK alpha subunits. Besides obtaining a prediction model for potential molecular candidates, we can also gain insights on the association between the chemical structure and the biological activity, which is instructive to understand the underneath physicochemical properties and how to improve the binding affinity by optimizing its chemical structure for drug development or drug design in the future [3].

AMPK alpha subunit

AMP-activated protein kinase (AMPK) is a phylogenetically conserved fuel-sensing enzyme, that plays a role in cellular energy homeostasis. It is a heterotrimeric protein made up of a catalytic alpha subunit and regulatory beta and gamma subunits. Each of these three subunits takes on a specific role in both the stability and activity of AMPK.

The α subunit has two isoforms, α1 and α2, and both contain conventional serine/threonine kinase domains at the N-terminus, which is the phosphorylation site of many important upstream kinases like LBK1 and the calmodulin-dependent kinase kinases, CaMKK, and so on. The LKB1 had previously been identified as the product of a umor suppressor gene (Alessi, et al., 2006), and AMPK was its first downstream target to be identified [4]. The activity of AMPK is also closely related to diseases like obesity, Alzheimer's disease, cardiovascular disease, diabetes, and so on. Phosphorylation of Thr172 at the α subunit increases AMPK activity by 2-3 orders of magnitude. The conventional serine/threonine kinase domains locate at the N-terminus in α subunit and the kinase activity is increased >100-fold by phosphorylation of a conserved threonine residue which is usually referred as Thr-172 due to its position in the original rat sequence [5], it is a sequence segment also critically involved in the regulation of many other kinases [6] .

Through phosphorylation of Thr-172, AMPK regulates the activity of a number of key metabolic enzymes. It protects cells from stresses that cause ATP depletion by switching off ATP-consuming biosynthetic pathways. So, we try to focus one alpha subunit in this project and generate a tool for identifying early-stage bioactive candidate for it.

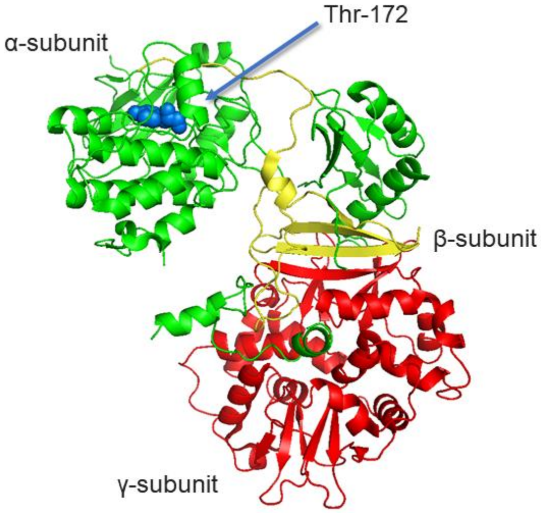

Figure 1: AMPK structure (5Å resolution from PyMol)

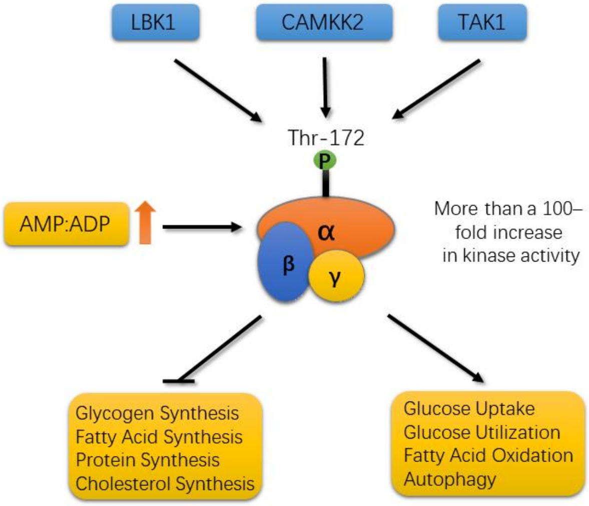

Figure 2: Representation of the effect of AMPK phosphorylation in regulation of energy balance

2 Methods

2.1 Publicly available data sets from CHEMBL and PubChem

A key driver in the project was the use of only publicly available data. Two data sets were retrieved from CHEMBL and PubChem. They both are large, authoritative, open-access drug discovery databases which contain biological activities of compounds. From each source we obtained two data sets of alpha1 and alpha2 separately. We took KD value, which is a quantitative measurement of antibody affinity as our criteria to distinguish the activity state of each compound. Data pre-processing steps like deleting duplication, combining the data of two resources and filtering our missing value were implemented. 1875 descriptors of each compound were calculated using the *PaDELPy* package. Though the initial data sets were sufficiently large, after the steps above the amount of available data had been reduced a lot (initially with 3834 compounds to 242 compounds in the end). This is mainly caused by the duplication of compounds sharing the same cid (Compound ID number), which may be due to the same compound being tested multiple times in the same experiment or against a panel of enzymes with different experiment settings. For these duplications caused by compounds sharing the same cid we use the average KD value of them.

2.2 Molecular Descriptors

The general idea of the project is to investigate the prediction of the bioactivity class of molecules using their respective molecular descriptors. The bioactivity class represents the global biological activity that a molecule can have with a certain target i.e., AMPK. It is encoded as a binary feature with the following levels: *active* or *inactive*. Molecular descriptors are computed from the SMILES string of a molecule. On one hand, SMILES stands for Simplified Molecular-Input Line Entry System and is a linear string representation of the structure, namely of the connectivity and the chirality of any molecule. On the other hand, molecular descriptors are mathematical representations of the molecular properties. It enables scientists to describe the chemical, physical and structural information of molecules through a quantifiable framework. Practically, we computed the molecular descriptors from the SMILES strings using a dedicated python package, named *PaDELPy*, a python wrapper to use the *PaDELDescriptor* [7] freeware (released in 2019, latest version: 0.1.11 on the Dec 5, 2021). The latter software computes 1444 1D or 2D descriptors, 431 3D descriptors and 12 types of fingerprints, for a total of 1875 descriptors.

2.3 Challenges

As illustrated in the previous section, retrieving a large amount of publicly available data is a challenge in itself. Data scarcity imposes another challenge when working with a feature space of about 1875 predictors, that is high-dimensionality. Indeed, the framework we sit in is characterized by a number of observations that is drastically lower than the number of features. We retrieved approx. 250 molecules while we computed 1875 features.

It is in the intrinsic nature of molecular descriptors to be correlated between themselves. Actually, one can say that most of the properties of a molecule are correlated. For example, the number of heavy atoms (nHeavy) is obviously correlated with the number of oxygen (nO), carbon (nC) or nitrogen atoms (nN).

2.4 Exploratory Analysis

Dimension reduction methods

Firstly, both PCA and tSNE resulted in disappointing patterns (see Figure 2). Indeed, the partitioning between active and inactive ligands does not leave any opportunity of correctly clustering the two classes of molecule. It seems that the two classes are indistinguishable based on the linear combinations used as features by the PCA and tSNE.

Feature space reduction using LASSO

Secondly, to reduce the size of the features space, the molecular descriptor set has undergone a LASSO importance selection. It resulted in the selection of 13 descriptors among the 1875, a drastic dimensional reduction of 99.3%. The selected descriptors are: [’AATS5i’, ’ATSC5p’, ’GATS4m’, ’GATS2i’, ’nHBint6’, ’nHeteroRing’, ’n5HeteroRing’, ’nF9HeteroRing’, ’nT5HeteroRing’, ’geomShape’, ’RDF70m’, ’P1p’, ’E3s’].

Correlations

Lastly, we know that molecular descriptors are computed from a single molecular structure, where the structural patterns are intrinsically correlated with each other. Multicollinearity is a typical challenge in life sciences.

Molecular descriptors are no exceptions to the rule. Among the correlations between the above selected features, none of them are extremely high or low as LASSO already did an impressive pruning. Among the correlations between features and the target (i.e., activity class), no extreme values were reported, ranging from -0.2 to 0.2 approximately. The latter might indicate that molecular descriptors are not suited to predict the activity class of molecules.

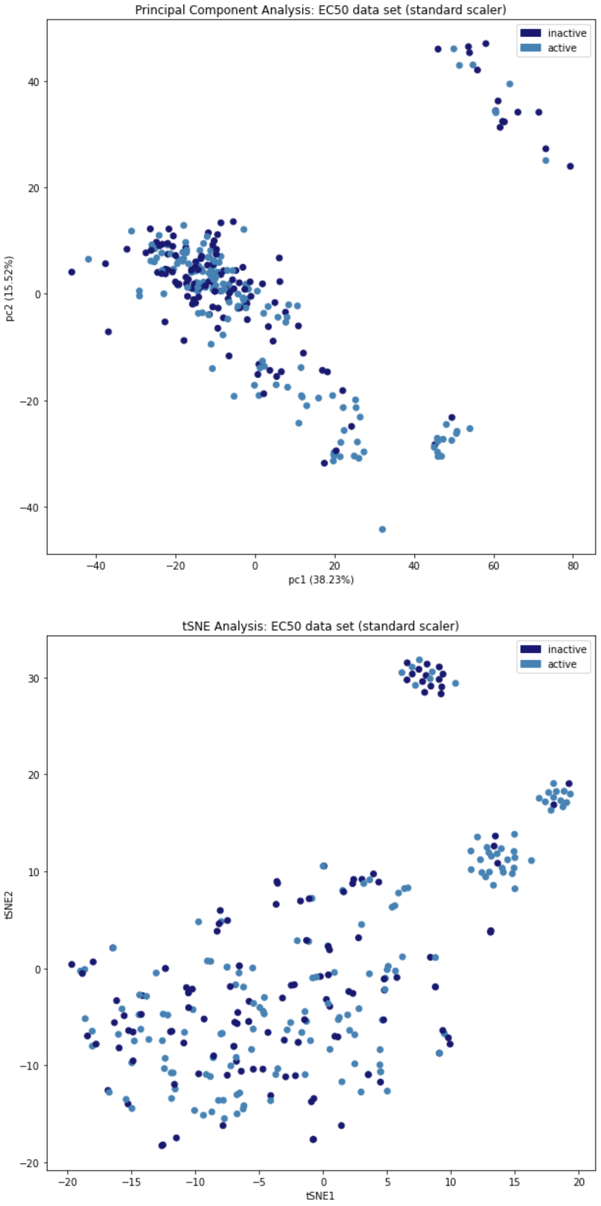

Figure 3: PCA and tSNE dimension reduction analysis applied on the EC50 (balanced) data set

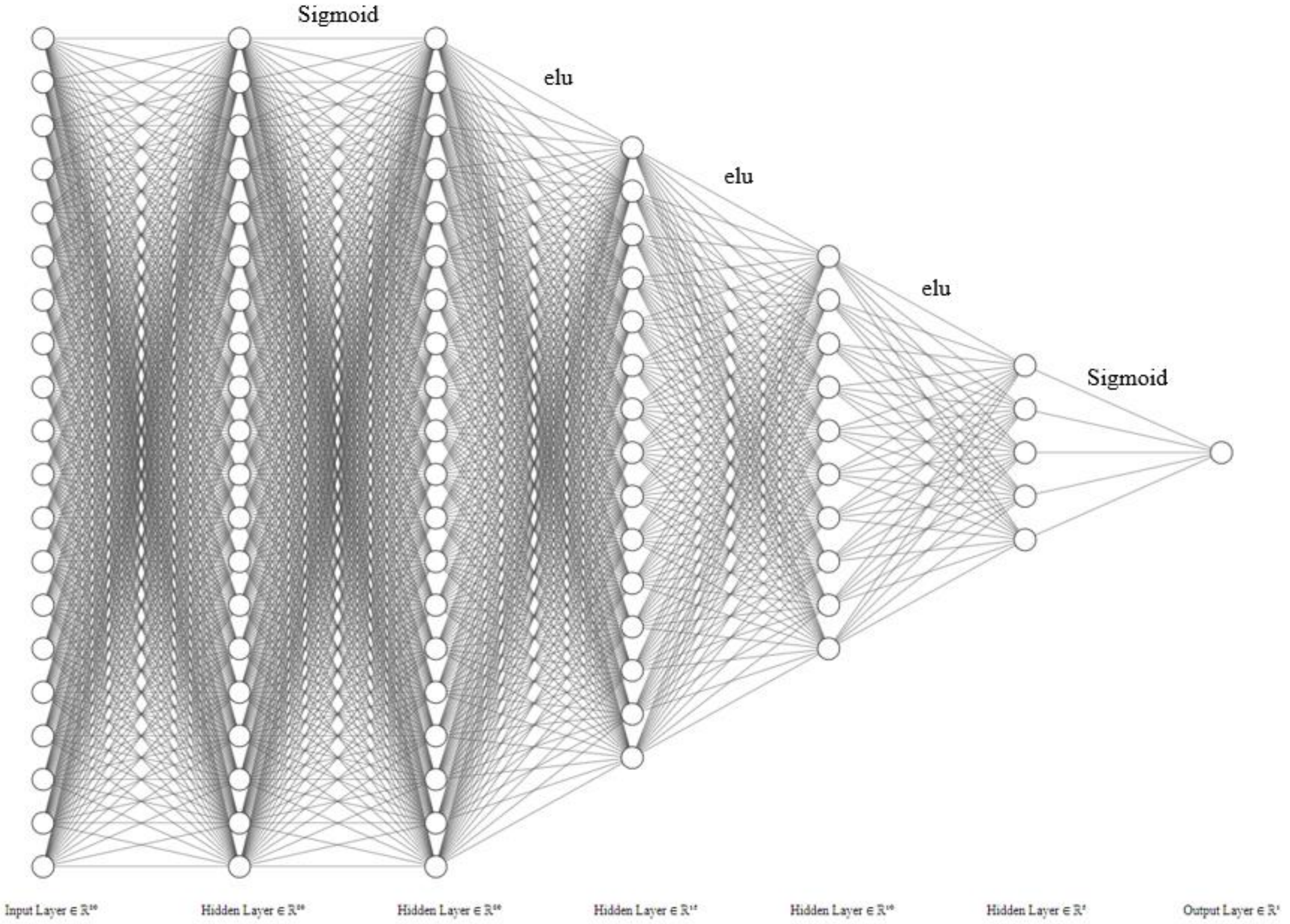

Figure 4: Architecture of the Artificial Neural Network model. Scale: 1:4 - for 25 neurons on the figure, there exist 100 in the actual model.

2.5 Data modelling

2.5.1 Tree based modelling

Decision Tree Classifier

To have a basic visualization before trying out tree methods, a decision tree classifier on the best subset of the data (as visualizing a tree using the entire set of variables would have been too large to view) was visualized. The depth of the trees was compared when allowed to grow fully versus the best tree that could generalize well to unseen data, Grid Search Cross Validation was applied to find the best depth of the tree that would work well on the data and was compared to the fully grown tree and the optimized tree had a slightly improved test accuracy.

Ensemble Methods

Two ensemble methods were tested out, bagging and random forest. Bagging is an ensemble method that fits multiple weak models to multiple samples of the data, combining the predictions to form one single prediction that is stronger than a single well-tuned decision tree.

Random Forests is an improvement over bagging that uses less features, thereby de-correlating the trees which improves the predictive power. The best results were obtained using grid search cross validation to find the best estimators. This was applied on both the whole feature space as well as the LASSO best subset feature space and the results obtained were compared and tabulated.

2.5.2 Artificial Neural Network

The second method we investigate to model the activity-structure relationship of molecules is an artificial neural network. It was used to predict the bioactivity class from the molecular descriptors set. The data set was split in training and test sets using a 75/25 proportion. The training set was further split in training and validation sets using a 70/30 proportion.

Architecture

In order to build the ANN model, we used an architecture that has proven to be working on similar activity-molecular descriptors data in a research project from *Drewe et al* [8].

The input layer is made of 100 neurons (See Figure 4), uses a truncated normal distribution as a weight initializer and a null constant value for the bias initializer. The hidden layers are built in such a way that their number of neurons decreases by 25 at each new layer. The output layer is made of a single neuron and predicts the binary output using a sigmoid activation function on top.

Hyperparameters optimization

The tuning of the parameters was performed using a brute force approach by fitting the model on data for a total of 150 parameters combination. We used the test accuracy as an optimality metric. It returned the following parameters as optimal: learning rate of 0.001, dropout rate of 0.0, 3 hidden layers and batch size of 128.

Furthermore, the model was trained for 300 epochs. The latter was chosen by plotting the train and test accuracy over the number of epochs and looking for a sufficient number of epochs so that the accuracies are stable. In order to ensure an unbiased estimation of the training accuracy, it was computed using an unknown set, namely the validation one.

Combination of ANN and LASSO features selection

Such as with Random Forest, the model was assessed using the whole feature space and with the LASSO-reduced feature space (see Exploratory Analysis for further explanation about LASSO analysis).

3 Results and discussion

3.1 Tree based modelling

Applied on the whole feature space

The whole feature space while having good training results suffer from overlearning even after optimizing the parameters, test accuracy results were not very promising, the AUC curves between the two methods were visualized below

Applied on the LASSO-reduced feature space

After performing a best subset selection using LASSO, the new parameters yielded were applied on 4 methods:

- A decision tree classifier where the tree was allowed to grow fully
- A decision tree classifier that had been optimized with the best parameters set using gridsearchCV
- The Bagging method
- The Random Forests method
- The two ensemble methods showed a small improvement in classifying the test set accurately and the performance of the 4 methods on the LASSO reduced subset were visualized with AUC curve figure below.

3.2 Artificial Neural Network

After disappointing results on the whole feature space, we decided to investigate the improvement that LASSO could bring. With the same optimized parameters, the model returned a classifying accuracy of 60.3\% and 63\% using, respectively, the whole feature space and the LASSO-reduced feature space. Alongside with accuracy, the area under the ROC curve is higher for the model running on

Method	Training Accuracy (%)	Test Accuracy (%)	Precision	AUC
Bagging	97.2	54.7	0.558	0.50
Random Forests	98.63	54.7	0.553	0.51
Decision Tree (unoptimized, best subset)	98.63	43.8	-	0.43
Decision Tree (optimized, best subset)	66.21	56.1	-	0.68
Bagging (best subset)	95.89	50.68	0.536	0.55
Random Forests (best subset)	98.63	52.05	0.540	0.61

Table 1: Tabular results of the performance of the models.

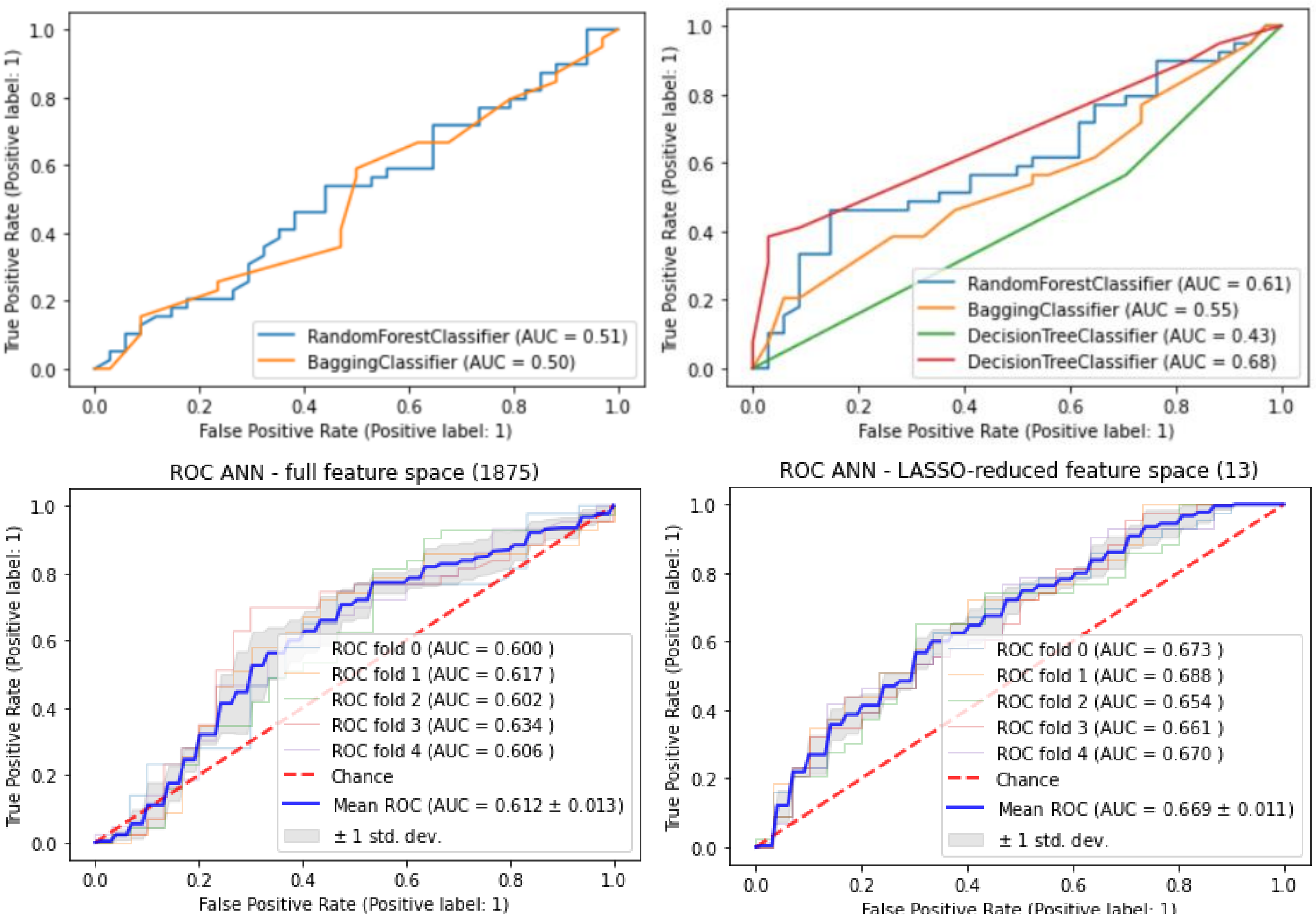

Figure 5: ROC curves for the different data modelling methods

the LASSO-reduced feature space (see Figure X). Overall, we see that the LASSO features reduction improved the performance of the model, but not by a drastic margin. Additionally, evidence of overfitting is displayed by a high training accuracy and a low-test accuracy. One can say that neural network models usually require a gigantic load of data to perform. If we assume the latter to be true, it explains that the lack of data is the major drawback for our model to discriminate between active and inactive molecules.

Discussion

- ANN and Tree Based Models both yielded similar results when applied to publicly available data of AMPK-ligand activity
- The combination of models and LASSO features selection seem to help improve the accuracy by reducing the undoubted overfitting
- The models trained in the present paper have shown their efficacy in modelling the structure-activity of molecules in Jürgen Drewe et al [8]. It is likely that the low accuracy and precision yielded in the present paper is due to poor data. A major improvement would be to apply the models to annotated data from literature.

Code availability

The scripts used for the data analysis in this study are available on GitHub under an open-access license. [https://github.com/AntoineRuzy/bioactivity-structure-classification].

Acknowledgements

We would like to express our gratitude to Dries De Ruysscher, from the Masschelein Lab (KU Leuven), for his guidance and support through this introductory research journey.

Conflict of Interest: none declared.

References

[1] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (March 2010). "How to improve R&D productivity: the pharmaceutical industry's grand challenge". *Nature Reviews. Drug Discovery*. **9** (3): 203–14. doi:10.1038/nrd3078. PMID 20168317. S2CID 1299234.

[2] Vamathevan, J., Clark, D., Czodrowski, P. et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18, 463–477 (2019). https://doi.org/10.1038/s41573-019-0024-5

[3] Lo YC, Rensi ES, Torng W, Altman BR. Machine learning in chemoinformatics and drug discovery. Drug Discovery Today. 23(8): 1538-1546 (2018).
https://doi.org/10.1016/j.drudis.2018.05.010.

[4] LKB1-dependent signaling pathways.Alessi DR, Sakamoto K, Bayascas JRAnnu Rev Biochem. 2006; 75():137-63.

[5] Jiang S, Li T, Ji T, Yi W, Yang Z, Wang S, Yang Y, Gu C. AMPK: Potential Therapeutic Target for Ischemic Stroke. *Theranostics* 2018; 8(16):4535-4551. doi:10.7150/thno.25674. Available from https://www.thno.org/v08p4535.htm

[6] Gowans GJ, Hawley SA, Ross FA, Hardie DG. AMP is a true physiological regulator of AMP-activated protein kinase by both allosteric activation and enhancing net phosph

[7] www.yapcwsoft.com/dd/padeldescriptor/ (PaDELDescriptor software documentation)

[8] Drewe, Jürgen, Ernst Küsters, Felix Hammann, Matthias Kreuter, Philipp Boss, and Verena Schöning. 2021. "Modeling Structure–Activity Relationship of AMPK Activation" Molecules 26, no.21: 6508. https://doi.org/10.3390/molecules26216508