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

Ruzette Antoine, Bu Taofeng, Francis Menezes Kevin and Ye Guoshuo

KU Leuven, Faculty of Bioscience, MSc in Bioinformatics, Integrated Bioinformatics Project

Motivations

The traditional drug discovery and development procedures last approx. 12-15 years and cost $1-2 billion dollars, from the initial discovery stage to the market. Meanwhile, the success rate is extremely low and only around 10% of drug candidates can reach human clinical trials [1]. Alongside with the booming quantity of publicly available biological data, machine learning methods arose as fairly simple, highly efficient and scalable candidates to undertake drug-discovery steps in pharmaceutical industries.
In the present project, we focus on ML methods to find drug-like candidates. Multiple machine learning algorithms are applied based on the molecular structures of ligands to identify the bioactivity level with our target: AMPK alpha subunits. Besides, we aim to explore the molecular patterns of AMPK alpha sub-unit ligands, which is crucial to understand the underlying physio-chemical binding properties.

AMPK α subunit structure

AMP-activated protein kinase (AMPK) is a phylogenetically conserved fuel-sensing enzyme and crucial in cellular energy homeostasis, by regulating the activities of a number of key metabolic enzymes. The activity of AMPK is also closely related with diseases, such as obesity, Alzheimer's disease, cardiovascular disease,diabetes. It is a heterotrimeric protein made up of three subunits:

1. α subunit
2. β regulatory subunit
3. γ regulatory subunit

Each of these three subunits takes on a specific role in both the stability and activity of AMPK. In the present project, we focus on the α subunit. The α subunit of AMPK protein is most relevant to the activity. It has two isoforms, $\alpha1$ and $\alpha2$. Both isoforms contain conventional serine/threonine kinase domains at the N-terminus, and their kinase activity is increased >100-fold by phosphorylation of a conserved threonine residue e.g. Thr-172, which is a sequence segment also critically involved in regulation of many other kinases [2]. Once phosphorylation happens, activated AMPK stimulates energy generating processes such as glucose uptake and fatty acid oxidation and decreases energy consuming processes such as protein and lipid synthesis.

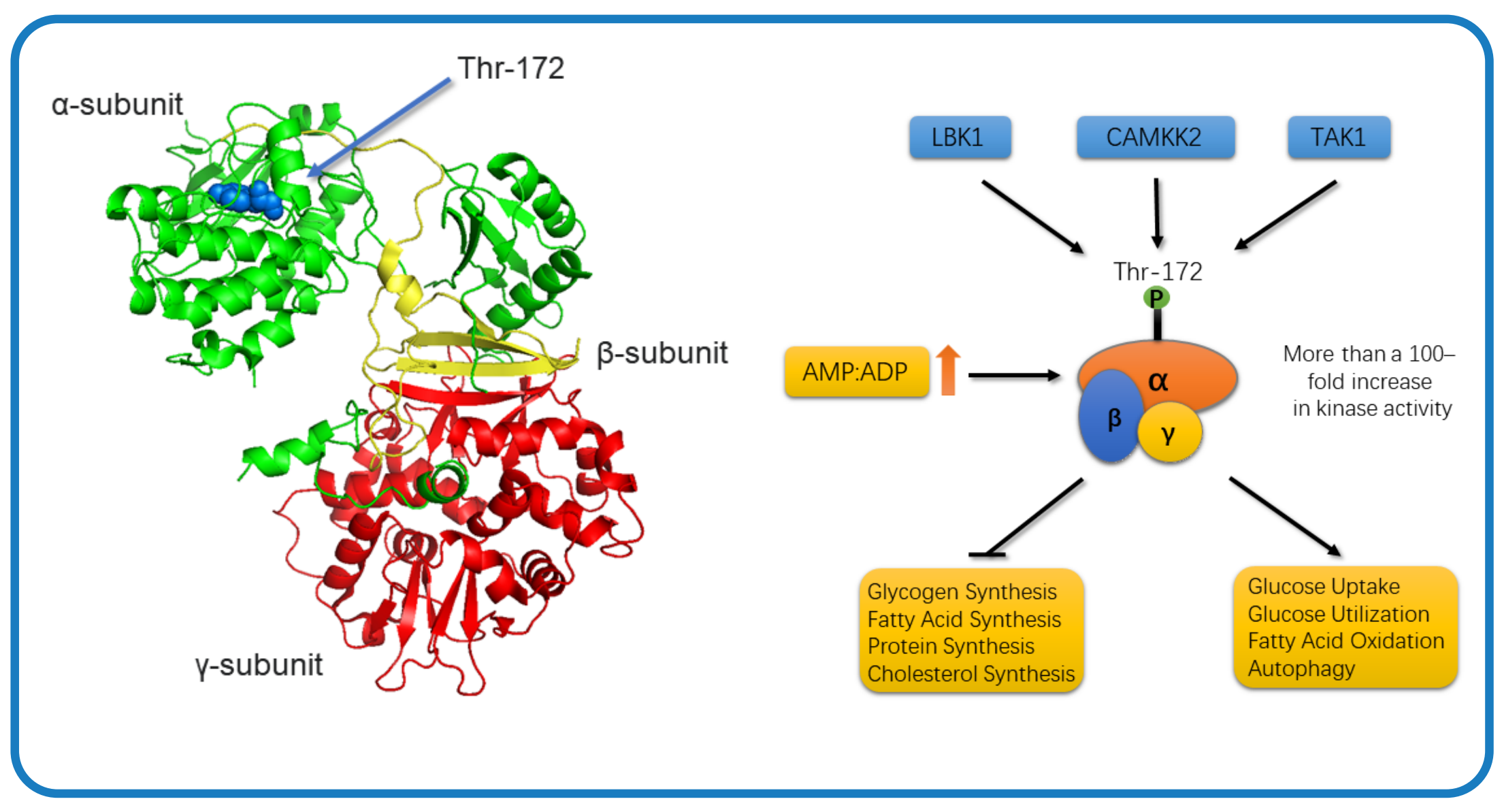

Figure 1: AMPK structure and representation of the effect of AMPK phosphorylation in regulation of energy balance

Data sets

A key driver in the project was the use of only publicly available data. Two data sets were retrieved from CHEMBL and PubChem. One based on the Kd values that is unbalanced with an active to inactive molecules ratio equal to 1.4. Another one based on the EC50 values that is balanced with a similar ratio of almost 0.25. Thus, we decided to go on with the balanced one, namely the EC50 values data set. Using the *PaDELPy* package, we computed 1875 1D, 2D or 3D molecular descriptors and fingerprints.

References

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

[2] 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 phosphorylation". *Cell Metab.* 2013;18(4):556-566. doi:10.1016/j.cmet.2013.08.019.

[4] 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.

Exploratory Analysis

Firstly, both PCA and tSNE resulted in disappointing patterns (see Figure 2). Indeed, the partitioning between active and inactive ligands does not result in a high potential 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. 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']. Lastly, we know that molecular descriptors are computed from a single molecular structure. In such a structure, the structural patterns are intrinsically correlated with each other. Multi-collinearity 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.

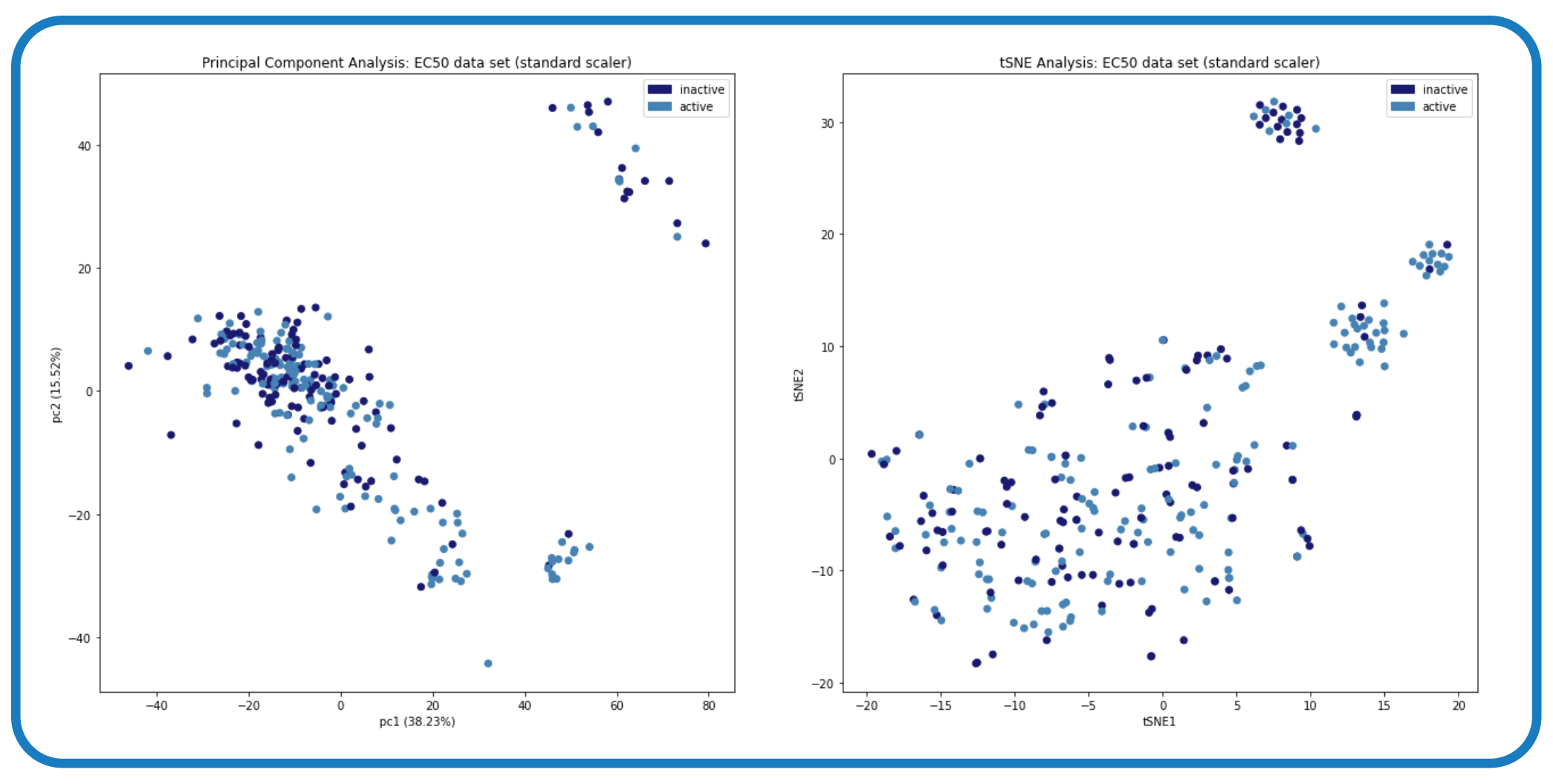

Figure 2: PCA and tSNE applied on the EC50 data set

Random Forest Classifier

Firstly, a single decision tree classifier was performed. When tested out, it had a test accuracy of around 57% when the max depth was limited and an AUC of 0.55. Compared to that, when the tree was allowed to fully grow the accuracy was only 50 percent. A single tree is a weak model (however, very useful for interpretation) thus we will tree to combine the power of several weak models into ensemble models.
Two ensemble methods utilizing decision trees were tested out, namely bagging and random forests. Using the *scikit-learn* library, the two methods were trained on our dataset (with a train-test split of 3:1). The models yielded excellent training results with a classification accuracy of around 99.1%. However, when applied to training data, the precision yielded was only 65.4% for bagging and random forest, using the whole feature space. The area under the ROC curves (see Figure 3) gave a result of around 0.63 and 0.64 with both Bagging and Random Forest.

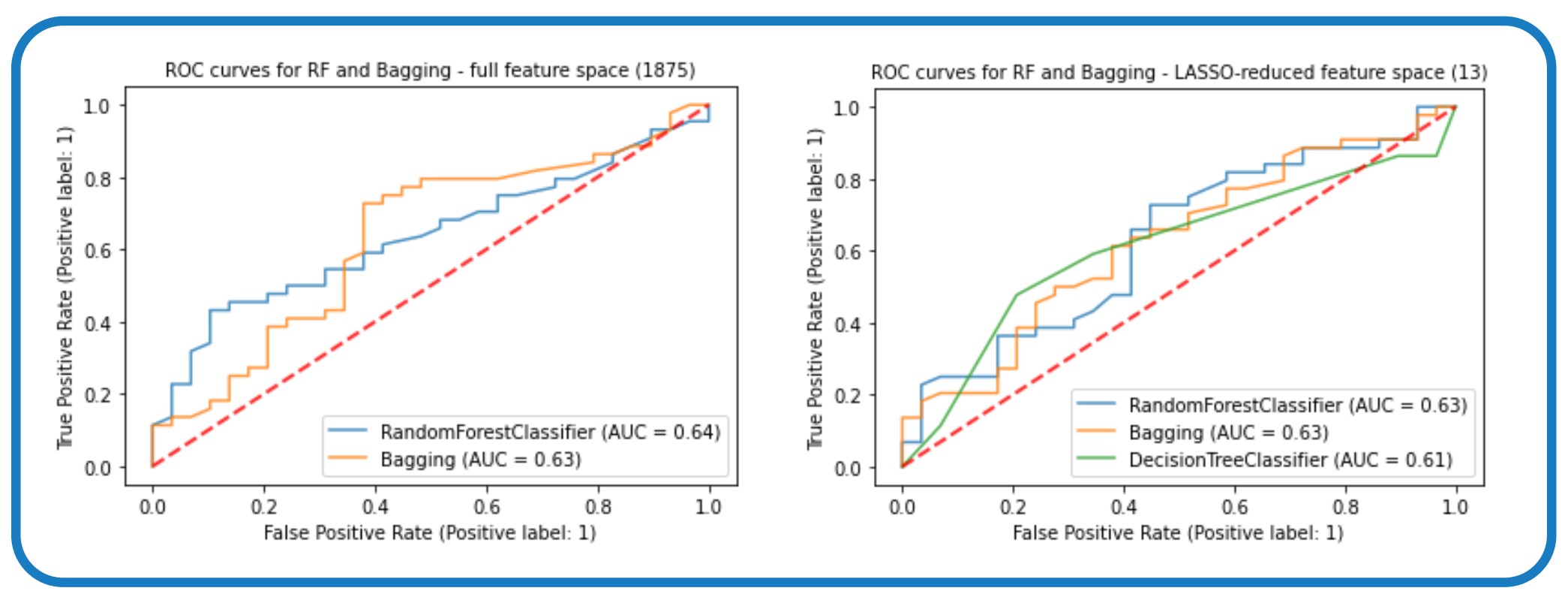

Figure 3: ROC curves for the Random Forest and the Baggin classifier using the whole feature space and using the LASSO-reduced feature space

However, with the reduced feature space from LASSO, the accuracy was slightly improved to 66.7% when using the random forest method. As the AUC using a reduced feature space is similar to the one computed using the whole feature space, it shows that there is no dramatic loss in the goodness of the model when using a reduced features set.

Artificial Neural Network

An ANN model was used to predict the bioactivity class from the molecular descriptors set. The architecture used for the model is an architecture that has proven to work with similar data in a research project from *Jürgen Drewe et al [4]*. Please refer to the related paper for further information regarding the architecture. The tuning of the parameters was performed using a brute force approach and returned the following parameters as optimal: learning rate of 0.001, dropout rate of 0.0, 3 hidden layers and batch size of 128. Such as with Random Forest, the model was assessed using the whole feature space and with the LASSO-reduced feature space. After disappointing results on the whole feature space, we decided to investigate the improvement that LASSO could bring. With those (hyper-) 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 the LASSO-reduced feature space (see Figure 4). Overall, we see that the LASSO features reduction improved the performance of the model, but not of a drastic margin. Additionally, evidences of overfitting are 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 active from inactive molecules.

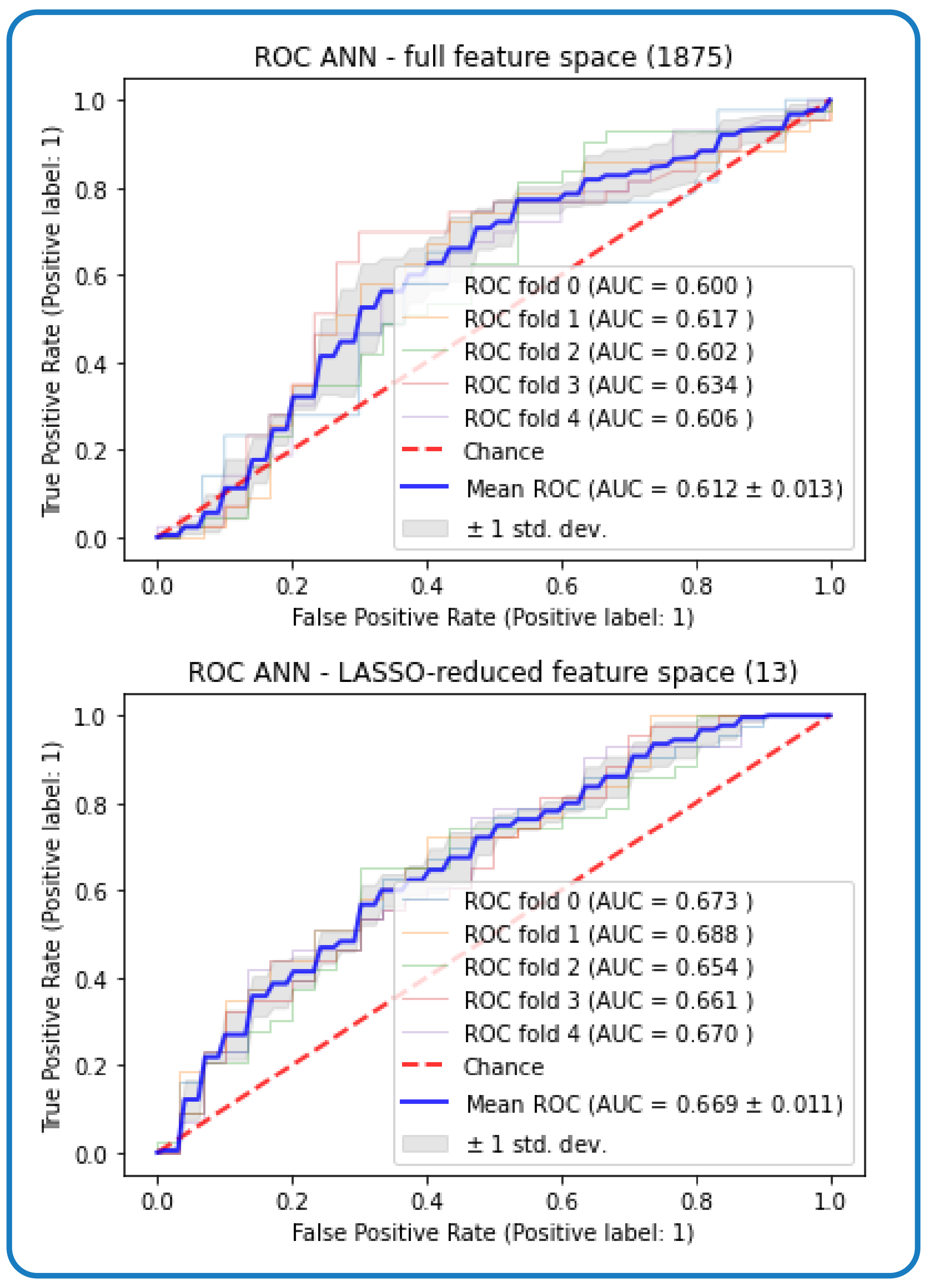

Figure 4: ROC curves for the ANN model using the whole feature space and using the LASSO-reduced feature space

Conclusion and improvements

1. ANN and Tree Based Models both yielded similar results when applied to publicly available data of AMPK-ligand activity
2. The combination of models and LASSO features selection seem to help improve the accuracy by reducing the possible overfitting
3. The models trained in the present paper have shown their efficacy in modelling the structure-activity of molecules in *Jürgen Drewe et al [4]*. 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.

Model	Train accuracy (val)	Test accuracy	Precision score	Kappa
ANN	0.982	0.603	0.640	0.1836
ANN with LASSO	0.987	0.630	0.661	0.2474
RF	0.991	0.603	0.654	0.185
Bagging	0.991	0.603	0.654	0.185
RF with LASSO	0.991	0.667	0.630	0.232
Bagging with LASSO	0.991	0.603	0.654	0.185
DecisionTree with LASSO	0.808	0.562	0.624	0.085

Table 1: Comparison of the performance metrics for the different models