

Article

Modeling Structure–Activity Relationship of AMPK Activation

Jürgen Drewe ^{1,*} , Ernst Küsters ², Felix Hammann ³ , Matthias Kreuter ¹, Philipp Boss ⁴ and Verena Schöning ³ 

¹ Medical Department, Max Zeller Söhne AG, CH-8590 Romanshorn, Switzerland; matthias.kreuter@zellerag.ch

² Independent Researcher, D-79427 Eschbach, Germany; kuesterseschbach@aol.com

³ Clinical Pharmacology and Toxicology, Department of General Internal Medicine, Inselspital University Hospital, CH-3012 Bern, Switzerland; felix.hammann@insel.ch (F.H.); verena.schoening@insel.ch (V.S.)

⁴ Max Delbrück Center for Molecular Medicine in the Helmholtz Association, D-13125 Berlin, Germany; philipp.boss@posteo.de

* Correspondence: juergen.drewe@zellerag.ch

Abstract: The adenosine monophosphate activated protein kinase (AMPK) is critical in the regulation of important cellular functions such as lipid, glucose, and protein metabolism; mitochondrial biogenesis and autophagy; and cellular growth. In many diseases—such as metabolic syndrome, obesity, diabetes, and also cancer—activation of AMPK is beneficial. Therefore, there is growing interest in AMPK activators that act either by direct action on the enzyme itself or by indirect activation of upstream regulators. Many natural compounds have been described that activate AMPK indirectly. These compounds are usually contained in mixtures with a variety of structurally different other compounds, which in turn can also alter the activity of AMPK via one or more pathways. For these compounds, experiments are complicated, since the required pure substances are often not yet isolated and/or therefore not sufficiently available. Therefore, our goal was to develop a screening tool that could handle the profound heterogeneity in activation pathways of the AMPK. Since machine learning algorithms can model complex (unknown) relationships and patterns, some of these methods (random forest, support vector machines, stochastic gradient boosting, logistic regression, and deep neural network) were applied and validated using a database, comprising of 904 activating and 799 neutral or inhibiting compounds identified by extensive PubMed literature search and PubChem Bioassay database. All models showed unexpectedly high classification accuracy in training, but more importantly in predicting the unseen test data. These models are therefore suitable tools for rapid in silico screening of established substances or multicomponent mixtures and can be used to identify compounds of interest for further testing.

Keywords: AMPK activator; machine learning; random forest; support vector machine; logistic regression; deep learning; QSAR



Citation: Drewe, J.; Küsters, E.; Hammann, F.; Kreuter, M.; Boss, P.; Schöning, V. Modeling Structure–Activity Relationship of AMPK Activation. *Molecules* **2021**, *26*, 6508. <https://doi.org/10.3390/molecules26216508>

Academic Editors: Rita Guedes, Stanisław Jastrzębski and Sabina Podlowska

Received: 27 September 2021

Accepted: 26 October 2021

Published: 28 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The adenosine monophosphate (AMP) activated protein kinase (AMPK) plays a master role in regulating cellular metabolism [1]. Its regulation is critical for many cellular functions, such as lipid, glucose, and protein metabolism; cellular growth; and mitochondrial biogenesis and autophagy [2]. The cellular mode of action suggests a beneficial clinical effect in various metabolic and degenerative diseases, ageing, as well as diabetes, cancer, and viral infection [3].

AMPK adapts cellular metabolism and cell growth to the supply of energy at two levels: At the central level, its hypothalamic activity is regarded as the key negative regulator of sympathetically activated thermogenesis, integrating different peripheral hormonal signals as well as drugs such as thyroid hormone, estrogens, and metabolites with different hypothalamic networks and food signals [4–7]. In peripheral cells, AMPK senses the loss of cellular energy as an increasing AMP/ATP ratio. AMPK is activated

and in turn activates catabolic pathways, improves cellular glucose uptake, and inhibits anabolic reactions [3].

Numerous clinical studies showed the efficacy of metformin—an indirect AMPK activator—in type 2 diabetic patients, resulting in its clinical acceptance as first line initial pharmacologic management of increased blood glucose in adults with type 2 diabetes mellitus (American Diabetes Association [8]). AMPK inhibits the mTOR (mechanistic target of rapamycin) complex 1, a nutrient sensitive master regulator of cell growth, angiogenesis, and metabolism that is activated by growth factors especially in tumors [9]. Several in vitro studies with the AMPK activator metformin showed the induction of cell cycle arrest in different tumor models (among others in breast cancer [10], and prostate cancer [11]). Therefore, the indirect activator metformin has been investigated in many clinical studies in patients with different types of cancer [12].

For a variety of different compounds, direct and/or indirect AMPK activation has been described: direct activators bind to the catalytic α subunit of AMPK allowing it to be phosphorylated at Thr172 or to the regulatory γ -subunit for allosteric activation [1]. Physiological direct activators are AMP, to a smaller extent ADP; pharmacological direct activators comprise among others AICAR (5-aminoimidazole-4-carboxamide ribose), A-769662 and salicylate and a variety of newly synthesized compounds.

Indirect activation is controlled by upstream kinases (such as liver kinase B1 (LKB1), calcium/calmodulin-dependent protein kinase kinase (CaMKK β), or transforming growth factor- β (TGF β -activated kinase 1 (TAK1))), which phosphorylate Thr172 of the α -subunit [1,13]. In general, a multitude of compounds may indirectly activate AMPK by virtue of diverse mechanisms, which lower cellular levels of ATP. Examples are drugs, such as metformin, and also a variety of natural compounds [14]—in particular, constituents of *Cimicifuga racemosa* (black cohosh) [15].

These natural compounds are usually contained in mixtures of a large number of structurally different chemical compounds which can directly or indirectly activate AMPK. Therefore, our goal was to develop a screening tool that could handle the heterogeneity of different activation pathways complementing in vitro assays that provide information on direct influence on the enzyme.

Frequently, (quantitative) structure–activity analyses are performed to predict the pharmacological or toxicological activity of new compounds. Nowadays, this is often done in vitro in extensive high-throughput experiments, where the desired pharmacological effects are measured in specialized cell culture assay systems. Molecular modeling of the target structure itself has also been used to investigate the structural requirements for binding to the target. However, machine learning-based QSAR methods are also able to more quickly and with less effort linking structural information to more general questions: e.g., an effect caused by different mechanisms, such as AMPK activation, but also more complex effects such as toxic adverse drug reactions [16–20], drug–drug interactions [21,22] or to predict clinical therapeutic effect. Last but not least, these machine learning methods could be used to screen on a large-scale drugs approved for a different indication for new indications in a repurposing approach [23].

Since machine learning algorithms can model complex (unknown) relationships and patterns, we used these methods to model the structural requirements for the pharmacological effect of AMPK activation.

2. Results

The activators of AMPK and controls (either no activation or inhibition of AMPK) were compared by the *t*-distributed stochastic neighbor embedding (tSNE) analysis to graphically assess the applicability domain of the database compounds (Figure 1).

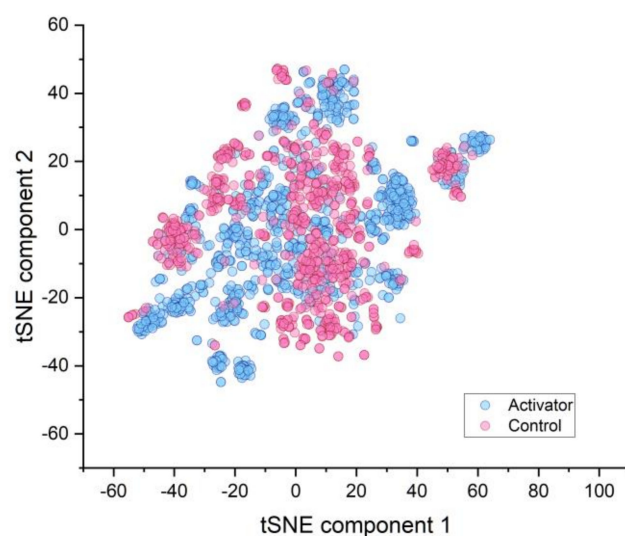


Figure 1. *t*-distributed stochastic neighbor embedding (tSNE) analysis: AMPK activators and controls.

In this comparison, the controls appeared to be mainly distributed in the center of the plot, while activators were preferentially distributed in the periphery, indicating different clusters (Figure 1). The chemical characterization of activators and inhibitors was done according to suggestions by Sharma and Kumar [14], who described groups of chemical compounds that were frequently associated with AMPK activation in the literature (Figure 2). The visualization of these proposed chemical groups showed clear qualitative differences between the activators and the controls.

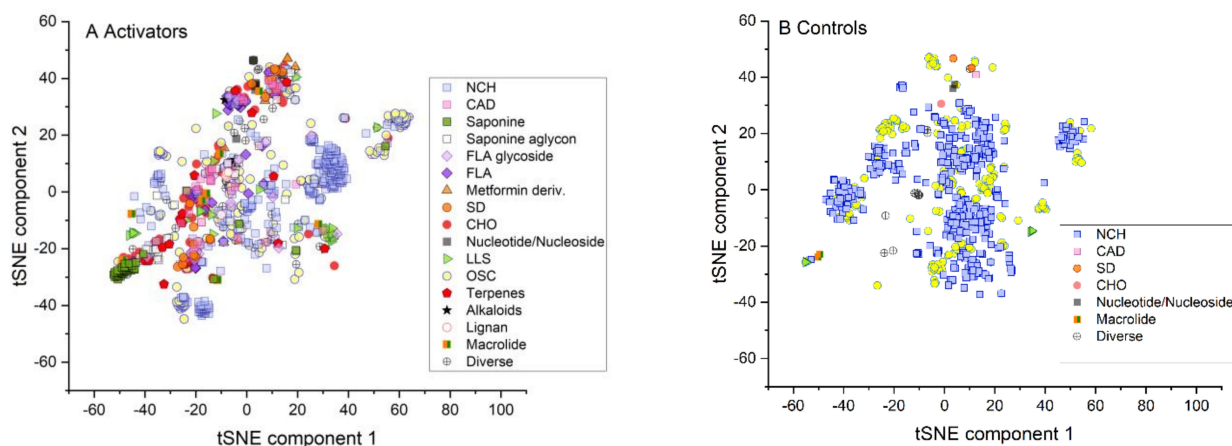


Figure 2. tSNE analysis of chemically classified activators and controls separated by chemical structure. (A) AMPK activators (N = 904); (B) AMPK control (N = 799).

2.1. Similarity of Groups

The mean standardized distances (mean; standard deviation) was higher (45.8; 18.8) within the activators than within the controls (40.2; 22.2), indicating a higher heterogeneity in the former group. This reflects the most diverse activation pathways of the AMPK.

2.2. Statistical Comparison of Datasets

As described in the method section, for each database instance (compound), 1445 chemical descriptors (PaDEL) have been calculated. All descriptors (features) have been compared between controls and activators by unpaired *t*-tests assuming unequal variances. From all 1445 features, 835 were significantly different between the two groups. Among others,

selection of intuitive features is reported in Table 1. Compared to the controls, the activators had on average more acidic groups ($p < 0.001$), more aromatic atoms and aromatic bonds (both $p < 0.001$) and a higher number of nitrogen atoms ($p < 0.02$). In addition, they had a significantly ($p < 0.001$) smaller number of 5- to 7-membered rings containing heteroatoms (N, O, P, S, or halogens).

Finally, activators had a significantly smaller molecular weight ($p < 0.001$) and a higher lipophilicity ($p < 0.001$). Both, controls and activators had on average less than one failure of the Lipinski's rule of five, which qualify them as drug-like compounds.

There were also large differences between features in classification, as demonstrated, for example, in the random forest classification (Figure 3).

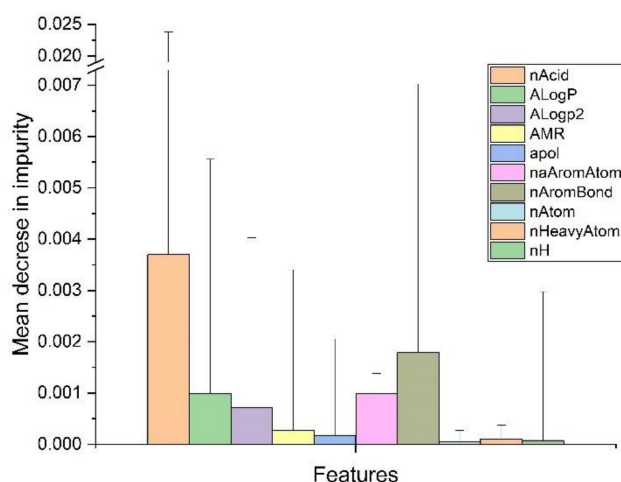


Figure 3. Feature importance (standard deviation) of the first 10 features for random forest classification; nAcid = number of acidic groups; ALogP = Ghose-Crippen LogKow; ALogP2 = square of ALogP; AMR = molar refractivity; apol = sum of the atomic polarizabilities (including implicit hydrogens); naAromAtom = number of aromatic atoms; nAromBond = number of aromatic bonds; nAtom = number of atoms; nHeavyAtom = number of heavy atoms; nH = number of hydrogen atoms.

2.3. Random Forest Classification (RFC)

The following hyperparameter settings were found to be optimal by grid-search analysis with a mean accuracy of 91.6% on the training set: using the Gini impurity criterion; determined minimal samples split = 6; minimal samples leaf = 2; the number of estimators (created trees) was set to 100. The bootstrap option was set to "False". Finally, the maximum number of features was determined by the square root of the number of features. For all other settings, default values were used, including automatic calculation of weights to account for different class sizes.

2.4. Support Vector Machine Classification (SVM-C)

The following hyperparameter settings were found to be optimal by grid-search analysis with a mean accuracy of 91.0% on the trainings set: the regularization parameter $C = 10$. The strength of the regularization is inversely proportional to C . This value is strictly positive. The penalty is a squared l2 penalty. The radial basis function (RBF) kernel was chosen with kernel coefficient $\gamma = \text{"scale"}$. That means that the program uses $1/(\text{n_features} \times X.\text{var}())$ as value of γ . Finally, the class weight was set to "balanced" to include automatic calculation of weights to account for different class sizes.

2.5. Stochastic Gradient Boosting (SGB) Analysis

The following hyperparameter settings were found to be optimal by grid-search analysis with a mean accuracy of 99.3%: the optimal loss function was "exponential" that recovers the AdaBoost algorithm. The optimal learning rate was 0.1, maximum dept was

3, the number of estimators 1000 and the fraction of samples to be used for fitting the individual base learners was 0.7. Furthermore, the class weight was set to “balanced” to include automatic calculation of weights to account for different class sizes.

2.6. Logistic Regression Classification (LRC)

Different optimizers (“Newton-cg” (Newton conjugate gradient optimizer), “lbfgs” (limited memory Broyden-Fletcher-Goldfarb-Shanno optimizer)—a quasi-Newton method and “liblinear”—a linear classifier for smaller datasets) have been tested. The following hyperparameter settings were found to be optimal by grid-search analysis: the regularization parameter C was set to 0.1; for regularization, the penalty by the l_2 -norm and as optimizer the Newton-cg was applied. The class weight was “balanced” to include automatic calculation of weights to account for different class sizes.

2.7. Deep Neural Network (DNN) Analysis

As the last approach, a deep neural network analysis was performed using tensorflow and its API keras. The following sequential model approach was applied: The optimum hyperparameter were: Adams optimizer with learning rate = 0.001, batch size = 128, and 50 learning epochs. The input layer used 1000 units (neurons), three hidden layers with neurons, and the activation function Elu. The code of the model is given in Best_DNN_model.pdf in https://github.com/cptbern/QSAR_AMPK, accessed on 27 October 2021.

2.8. Test Performance

The performance of these models is given in Table 1. The ROC-curves for all models are given in Figure 4.

Table 1. Summary of results of classification of different machine learning methods.

Method	Training Accuracy (%)	Test Accuracy (%)	Y-Randomization ** (%)	Test Precision (%)	Sensitivity (%)	Specificity (%)	AUC *
RFC	91.6	92.6	52.7 ± 2.3	90.3	91.2	94.0	0.968 ± 0.013
SVM-C	91.0	93.0	53.2 ± 2.2	90.1	93.5	92.4	0.962 ± 0.009
SGB	91.3	93.0	52.8 ± 2.2	90.7	92.0	94.0	0.968 ± 0.012
LRC	90.8	91.0	52.6 ± 2.1	89.2	97.4	94.8	0.948 ± 0.014
DNN	91.6	90.6	53.0 ± 1.8	87.6	90.2	91.1	0.970 ± 0.002

Test set (number): activator (262), control (249); * AUC = Area under the receiver operating characteristics curve. ** N = 100 permutations.

All investigated models showed an unexpectedly high accuracy in the training but more importantly in the unseen test data set. The accuracy of all models was comparable between 91.0% and 93.0%. No clear advantage of one method over the others could be identified.

To rule out overfitting of the data, Y-randomization of the response variables was applied, achieving only about 52% accuracy for all methods (an indication that final models have picked up on actual patterns rather than statistical noise). In addition, a 5-fold cross-validation was performed, which confirmed the good predictivity. ROC curves confirmed the high accuracy of the models.

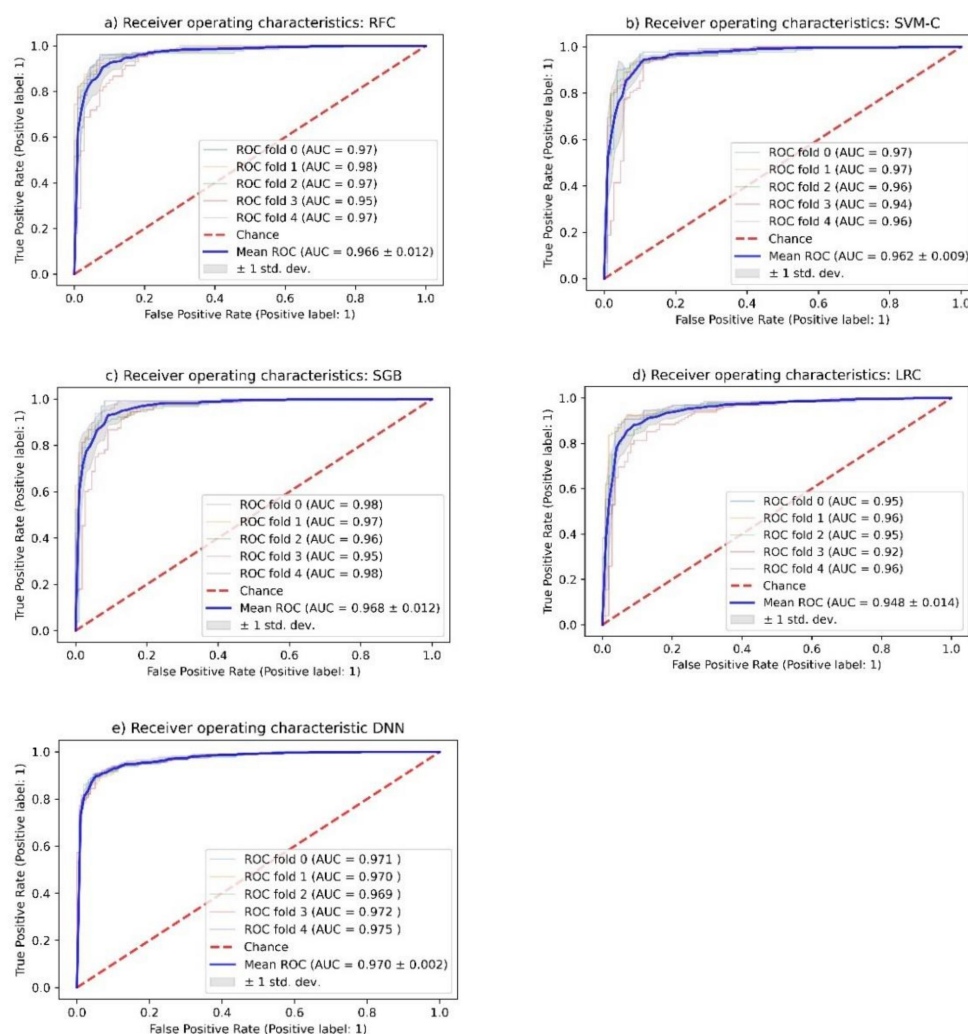


Figure 4. Receiver operating characteristic (ROC) of the investigated methods. (a) Random Forest classifier, (b) Support Vector Machine classifier, (c) Stochastic Gradient Boosting classifier, (d) Logistic Regression classifier, and (e) Deep Neural Network classifier.

3. Discussion

All models showed very good performance in discriminating AMPK activators from controls. The 5-fold cross-validation and the Y-randomization of the response variable makes it unlikely that the high accuracy of all methods used is due to overfitting of the data or to chance.

However, the DNN model appeared to have lower variability compared to the others (Figure 4). It can be speculated that deep learning networks cope better with complex, heterogeneous datasets than the other machine learning methods.

It is obvious that activation of AMPK is obtained via different (direct and indirect) mechanisms (Figure 4) explaining the structural heterogeneity (Figure 2). In order to achieve comprehensive prediction, it is therefore also necessary to have a sufficient number of different activators in the database acting through any of these mechanisms.

Although we had an excellent prediction of our unseen test data set, we could not exclude missing mechanisms. Furthermore, we have to consider that machine learning modeling always carries the limitation that classification provides only probabilities which require verification by direct in vitro or in vivo experiments. However, these experiments are time- and resource-consuming. Therefore, one goal of machine learning prediction is to facilitate the selection of suitable candidates from a range of possible candidates.

In the literature (PubMed search with terms “AMPK” and “QSAR”), several QSAR models for prediction of AMPK activation have been published so far [24–30]. However, all of these models used pharmacophore docking analyses or homology models or structure-, ligand-, or fragment-based design and focused exclusively on direct AMPK activators. In contrast to these approaches, as far as we know, we were the first to explicitly extend our models by including all compounds that showed evidence of AMPK activation, regardless of the activation pathway, whether direct or indirect. Therefore, our approach is more general and better accounts for the well-described heterogeneity of AMPK activation (Figure 4). Finally, it is easily adaptable to unseen data. However, there is also a limitation of our approach, because it does not provide information about the activation pathway and the type of activation (direct or indirect).

Only a smaller fraction of activators interacts directly with AMPK by binding to specific activating or allosteric binding sites of the enzyme. The majority of activators act indirectly. They bind to upstream regulatory sites that, when activated, in turn phosphorylate and activate AMPK. This places structural requirements on these activators to bind to these sites, contributing to the structural diversity of these compounds (Figure 2).

AMPK is an important enzyme sensing and controlling energy supply and different cellular functions—e.g., carbohydrate cellular entry and metabolism, reactive oxygen species (ROS) generation, apoptosis, cellular growth, and mitochondrial biogenesis and autophagy [1]. Since the enzyme is at the intersection of several important cellular pathways, it is obvious to assume the existence of different modes of activation (Figure 5).

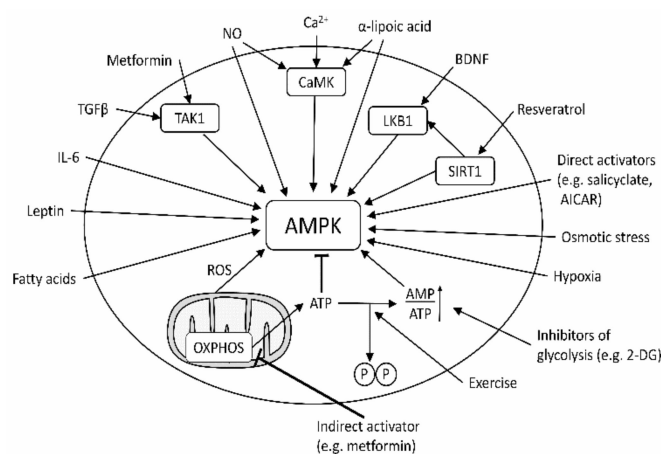


Figure 5. Activation pathways of AMPK.

Machine learning methods for AMPK activation are important: the relevance of AMPK in the pathogenesis of different diseases and their treatment was discussed above and new and old treatment modalities will be assessed with regard to their potential to modulate AMPK activity in the sense of repurposing of established (herbal) drugs.

To investigate herbal drugs, experiments are complicated by the fact that extracts are multi-substance mixtures and the required pure substances are often not sufficiently isolated and/or therefore not sufficiently available. Therefore, our extended method offers attractive new applications in the extended screening of these multi-substance mixtures to identify lead substances of interest for further intensified testing.

4. Materials and Methods

4.1. Data

The AMPK data set was based on extensive literature search of AMPK activators and inhibitors that was performed in PubMed (<https://pubmed.ncbi.nlm.nih.gov/>, accessed on 27 October 2021) using search terms:

- “AMPK AND activation”
- “AMPK AND inhibition”

In addition, the Bioassay database of PubChem Substance and Compound databases (<https://pubchem.ncbi.nlm.nih.gov/>, accessed on 27 October 2021) was used when EC_{50} was $\leq 0.1 \mu\text{M}$ to identify proven activators. In addition, compounds were included that were confirmed activators by at least one PubMed-listed publication. On the other hand, tested compounds shown to be inactive for AMPK activation or showing inhibitory function or compounds described in the literature as inhibitors of AMPK formed the control group for this analysis.

4.2. Data Preprocessing

Chemical structures were coded by the Simplified Molecular-Input Line-Entry System (SMILES; isomeric if available, canonical otherwise) that were taken directly from PubChem or when no PubChem ID was available were determined by MarvinSketch version 19.22 (ChemAxon). SDF Files were generated by Open Babel (version 2.21) <https://openbabel.org>, accessed on 27 October 2021. These files were used to calculate physicochemical descriptors with PaDEL descriptor software <http://www.yapcsoft.com/dd/padeldescriptor/>, accessed on 27 October 2021. We computed the entire range of available 1D and 2D descriptors [31] for all compounds.

Data preprocessing (curation) involved removing any double entries, salts and mixtures, and proteins from SMILES structures and focused on small molecular weight drug-like compounds. Tautomers were not standardized. Compounds and/or descriptors with empty descriptor values were excluded from analysis yielding $N = 904$ and $N = 799$ valid cases for activators and controls, respectively and $N = 1445$ features (descriptors). A complete list of used activators and controls is given:

- with their names, smiles codes, PubChem IDs, and PubMed IDs (Compounds.csv); and
- with all calculated PaDel descriptors (Data.csv) in https://github.com/cptbern/QSAR_AMPK, accessed on 27 October 2021.

4.3. Validation

Validation of models were based on OECD Principles for (Q)SAR Validation [32]. Validation was based on the random split of data into training and test data. From the full data set ($N = 1703$), 70% were randomly chosen for training ($N = 1192$) and 30% for the final test data set ($N = 511$), respectively. From the training data, 80% ($N = 953$) of the data were randomly chosen as validation training dataset and 20% ($N = 239$) as the validation test dataset. Following internal validation, the final model was used on the full training data set ($N = 1192$) to predict the previously unseen instances of the test data set (30% of activators ($N = 262$) and controls ($N = 249$), each) as external validation.

Training data were used to optimize model hyperparameters and train the models. After training hyperparameters were optimized on the training data set, final model learning was performed using 5-fold internal cross-validation. Furthermore, the training was repeated after randomization ($N = 100$) of the response variable (*Y-randomization*) as additional validation [33].

4.4. Similarity

Similarity or heterogeneity within the activators and controls was calculated in each case by mean Euclidean distance between each element to all other elements of the group based on the standardized values of the descriptors

$$\text{Mean standardised distance} = \frac{\sqrt{\sum_{i,j} (x_i - x_j)^2}}{S} \quad (1)$$

where, x_i and x_j denote the i -th and j -th instance and each descriptor values of the respective group, respectively ($i, j = 0, \dots, N$, the number of elements of the group, $i \neq j$), S denotes the number of pairwise differences.

For classification of data, different machine learning algorithms from the open scikit-learn python-based machine learning framework (<https://scikit-learn.org/stable/>, accessed on 27 October 2021) were applied: tuning of hyperparameters was done by grid search estimation (sklearn *GridSearchCV*) where relevant.

4.5. Machine Learning Models

All calculations were performed using Python 3.9.1 (<https://www.python.org/>, accessed on 27 October 2021). Graphical analysis was done with OriginPro, Version 2021. OriginLab Corporation, Northampton, MA, USA and Matplotlib, version 3.3.3 (<https://matplotlib.org/>, accessed on 27 October 2021).

The structural relationship of the high dimensional data of activators, controls and random samples were visualized by non-linear embedding into a two-dimensional space by the *t*-distributed stochastic neighbor embedding technique (*tSNE*) to visualize the applicability domain of the database [34]. In order to better visualize the structural diversity, activators and controls were chemically classified similar to the suggestions of Sharma and Kumar [14] as alkaloids; cinnamic acid derivatives (CAD); carbohydrates (CHO); flavonoid derivatives (FLA); lignans, lipid-like structures (LLS: \geq C8-chain); macrolides; metformin derivatives; nitrogen-containing heterocycles (NCH); nucleotide/nucleoside derivatives; organic sulfur-containing structures (OSC); saponins and their aglycons; sugar derivatives (SD); stilbenes and terpenes.

4.6. Random Forest Classification (RFC)

Random forest is an ensemble method [35]. These methods combine several base estimators in order to improve generalizability and robustness compared to single estimators (decision trees). A sequence of base estimators is built and each of these estimators tries to reduce the bias of the combined estimator. Random forests are a powerful decision tree algorithm for classification. Hyperparameters were tuned by grid search analysis on number of estimators, maximum features used, maximum depth of trees, minimum samples split, minimum samples leaf, and impurity criterion. No bootstrap sampling was performed.

4.7. Stochastic Gradient Boosting Classification (SGB)

Stochastic gradient boosting (SGB) classification [36] also belongs to the ensemble methods. Hyperparameters were tuned by grid search analysis on number of estimators, maximum depth of trees, the loss function (deviance (=logistic regression) or exponential (=AdaBoost algorithm)), the learning rate, and the fraction of samples to be used for fitting the individual base learners.

4.8. Support Vector Machine Classification (SVM-C)

Support Vector Machine Classification [37] tries to separate the two classes in the *n*-dimensional feature space by constructing a (*n*-1)-dimensional hyperplane that maximizes the margin between the two classes. Most important is the choice of the best kernel, a function to transform the data to a higher-dimensional space: Here, grid search evaluated the radial basis function (RBF, Gaussian kernel), the sigmoid and the polynomial kernels. Furthermore, the regularization parameter *C* and the best the kernel coefficient gamma were estimated.

4.9. Logistic Regression Classification (LRC)

Logistic Regression Classification [38] is used to estimate the probability \hat{p} that an instance belongs to a class

$$\hat{p} = h_{\theta}(x) = \sigma(\theta^T \cdot x), \quad (2)$$

using the logistic function

$$\sigma(t) = \frac{1}{1 + e^{-t}}. \quad (3)$$

Classification for two classes denotes with 0 and 1 will be obtained

$$\hat{y} = \sigma(t) = \begin{cases} 0, & \hat{p} < 0.5 \\ 1, & \hat{p} \geq 0.5 \end{cases} \quad (4)$$

Different solvers (Newton-cg, lbfgs and liblinear), the regularization parameter C and the penalty (l_1 -norm, l_2 -norm, and elastic net) were evaluated.

4.10. Deep Learning Neural Network (DNN)

As last approach, a deep neural network analysis was performed using tensorflow, version 2.3.1 (<https://www.tensorflow.org/>, accessed on 27 October 2021) and its associated neural networking API tensorflow.keras (https://www.tensorflow.org/api_docs/python/tf/keras, accessed on 27 October 2021). The following sequential model approach was applied: an input layer with 1000 units and “elu” activation and a variable number of hidden dense layers with “elu” activation and finally a dense output layer with sigmoid activation was used: The input layer was connected to one to three hidden dense layer and these to an output layer giving the prediction of the model. An extensive grid search was performed to estimate the unknown parameters (n_hidden: number of hidden dense layers: 0, 1, 2, 3 for the number of hidden dense and dropout layers; number of neurons: 750, 500, and 250 units for the hidden dense layers; the Adams or RMSprop optimizers; 0.0001 to 5 for the learning rates of the optimizer and 16 to 1024 for the batch sizes). As kernel initializer the HeNormal, and as bias initializer Constant(value = 0) were used. Number of features (descriptors) was 1445 and the number of targets was 1. This search was done using the tensorflow scikit learn wrapper and the GridSearchCV module from scikit learn with 5-fold cross-validation. As activation functions the exponential linear unit [ELU(z) = z , for $z > 0$ and $\alpha \times (\exp(x) - 1)$ if $x < 0$] and sigmoid [$\sigma(z) = 1/(1 + \exp(-z))$] functions were used for hidden and output layers, respectively and as loss function the binary cross-entropy [39] was used.

4.11. Model Evaluation

To account for differences in class sizes between activators and controls the ‘balanced’ option was chosen, if available, to perform automatic calculation of class weights. Data were divided into disjunct parts using the *sklearn.model_selection.preprocessing* procedure *train_test_split* using a test_size of 30% and 70% for training. A 5-fold cross-validation (CV) was applied.

Activators and controls populations were compared with 100,000 random samples drawn from PubChem database using PubChemPy tools that provide a way to interact with PubChem in Python (<https://pubchempy.readthedocs.io/en/latest/#>, accessed on 27 October 2021). The data distributions were compared by *t*-distributed stochastic neighbor embedding analysis using the *sklearn.manifold.TSNE* procedure. This method visualizes high-dimensional data by a two-dimensional representation to graphically assess the applicability domains. Goodness of machine learning modeling is reported as

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (8)$$

where TP = true positive (activator correctly predicted); FP = false positive (activator incorrectly predicted); TN = true negative (control correctly predicted); FN = (false negative control incorrectly predicted).

Validation was performed as 5-fold cross-validation using *sklearn.KFold* procedure. To evaluate binary classifier output independently of thresholds the receiver operating characteristic (ROC) and its area under the curve (AUC) scores [40] were calculated that assesses the tradeoff between sensitivity by plotting sensitivity versus (1 – specificity).

Author Contributions: Conceptualization, J.D., V.S., F.H. and P.B.; Methodology, J.D., F.H. and P.B.; Software, J.D. and P.B.; Validation, J.D., V.S. and P.B.; Formal analysis, J.D.; Investigation, J.D., E.K. and M.K.; Data curation, J.D., E.K. and M.K.; Writing—original draft preparation, J.D. and V.S.; Writing—review and editing, J.D.; Visualization, J.D. and V.S.; Supervision, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Complete list of used activators and controls are given: with their names, smiles codes, PubChem IDs, and PubMed IDs (Compounds.csv) and with all calculated PaDel descriptors (Data.csv) and source codes of all models in https://github.com/cptbern/QSAR_AMPK, accessed on 27 October 2021.

Conflicts of Interest: J.D. and M.K. both work at Max Zeller Söhne AG, a phytopharmaceutical company. The authors declare no conflict of interest.

Sample Availability: Samples of the compounds not available from the authors, however, names and descriptors are given at https://github.com/cptbern/QSAR_AMPK, accessed on 27 October 2021.

References

1. Hardie, D.G.; Carling, D. The AMP-activated protein kinase—Fuel gauge of the mammalian cell? *Eur. J. Biochem. FEBS* **1997**, *246*, 259–273. [[CrossRef](#)]
2. Hardie, D.G.; Ross, F.A.; Hawley, S.A. AMPK: A nutrient and energy sensor that maintains energy homeostasis. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 251–262. [[CrossRef](#)]
3. Hardie, D.G. Adenosine Monophosphate-Activated Protein Kinase: A Central Regulator of Metabolism with Roles in Diabetes, Cancer, and Viral Infection. *Cold Spring Harb. Symp. Quant. Biol.* **2011**, *76*, 155–164. [[CrossRef](#)]
4. López, M.; Seoane, L.; Tovar, S.; Senaris, R.M.; Diéguez, C. Thyroid status regulates CART but not AgRP mRNA levels in the rat hypothalamus. *Neuroreport* **2002**, *13*, 1775–1779. [[CrossRef](#)]
5. Blanco Martínez de Morentin, P.; González, C.R.; Saha, A.K.; Martins, L.; Diéguez, C.; Vidal-Puig, A.; Tena-Sempere, M.; López, M. Hypothalamic AMP-activated protein kinase as a mediator of whole body energy balance. *Rev. Endocr. Metab. Disord.* **2011**, *12*, 127–140. [[CrossRef](#)]
6. López, M.; Nogueiras, R.; Tena-Sempere, M.; Dieguez, C. Hypothalamic AMPK: A canonical regulator of whole-body energy balance. *Nat. Rev.* **2016**, *12*, 421–432. [[CrossRef](#)]
7. Merchenthaler, I.; Lane, M.V.; Numan, S.; Dellovade, T.L. Distribution of estrogen receptor alpha and beta in the mouse central nervous system: In vivo autoradiographic and immunocytochemical analyses. *J. Comp. Neurol.* **2004**, *473*, 270–291. [[CrossRef](#)] [[PubMed](#)]
8. American Diabetes Association. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2019. *Diabetes Care* **2019**, *42*, S90–S102. [[CrossRef](#)]
9. Shackelford, D.B.; Shaw, R.J. The LKB1-AMPK pathway: Metabolism and growth control in tumour suppression. *Nat. Rev. Cancer* **2009**, *9*, 563–575. [[CrossRef](#)]
10. Zhuang, Y.; Miskimins, W.K. Cell cycle arrest in Metformin treated breast cancer cells involves activation of AMPK, downregulation of cyclin D1, and requires p27Kip1 or p21Cip1. *J. Mol. Signal* **2008**, *3*, 18. [[CrossRef](#)] [[PubMed](#)]
11. Ben Sahra, I.; Laurent, K.; Loubat, A.; Giorgetti-Peraldi, S.; Colosetti, P.; Auberger, P.; Tanti, J.F.; Le Marchand-Brustel, Y.; Bost, F. The antidiabetic drug metformin exerts an antitumoral effect in vitro and in vivo through a decrease of cyclin D1 level. *Oncogene* **2008**, *27*, 3576–3586. [[CrossRef](#)]
12. Ben Sahra, I.; Le Marchand-Brustel, Y.; Tanti, J.F.; Bost, F. Metformin in cancer therapy: A new perspective for an old antidiabetic drug? *Mol. Cancer Ther.* **2010**, *9*, 1092–1099. [[CrossRef](#)] [[PubMed](#)]
13. Li, J.; Li, S.; Wang, F.; Xin, F. Structural and biochemical insights into the allosteric activation mechanism of AMP-activated protein kinase. *Chem. Biol. Drug Des.* **2017**, *89*, 663–669. [[CrossRef](#)] [[PubMed](#)]
14. Sharma, H.; Kumar, S. Natural AMPK Activators: An Alternative Approach for the Treatment and Management of Metabolic Syndrome. *Curr. Med. Chem.* **2017**, *24*, 1007–1047. [[CrossRef](#)] [[PubMed](#)]

15. Moser, C.; Vickers, S.P.; Brammer, R.; Cheetham, S.C.; Drewe, J. Antidiabetic effects of the *Cimicifuga racemosa* extract Ze 450 in vitro and in vivo in ob/ob mice. *Phytomedicine* **2014**, *21*, 1382–1389. [CrossRef]
16. Hammann, F.; Gutmann, H.; Vogt, N.; Helma, C.; Drewe, J. Prediction of adverse drug reactions using decision tree modeling. *Clin. Pharm.* **2010**, *88*, 52–59. [CrossRef]
17. Schöning, V.; Hammann, F.; Peinl, M.; Drewe, J. Identification of any structure-specific hepatotoxic potential of different pyrrolizidine alkaloids using Random Forests and artificial Neural Networks. *Toxicol. Sci.* **2017**, *160*, 361–370. [CrossRef]
18. Schöning, V.; Krähenbühl, S.; Drewe, J. The hepatotoxic potential of protein kinase inhibitors predicted with Random Forest and Artificial Neural Networks. *Toxicol. Lett.* **2018**, *299*, 145–148. [CrossRef]
19. Hammann, F.; Schöning, V.; Drewe, J. Prediction of clinically relevant drug-induced liver injury from structure using machine learning. *J. Appl. Toxicol.* **2019**, *39*, 412–419. [CrossRef]
20. Helma, C.; Schöning, V.; Drewe, J.; Boss, P. A comparison of nine machine learning mutagenicity models and their application for predicting pyrrolizidine alkaloids. *Front. Pharmacol.* **2021**, *12*, 1–15. [CrossRef]
21. Hammann, F.; Gutmann, H.; Jecklin, U.; Maunz, A.; Helma, C.; Drewe, J. Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein. *Curr. Drug Metab.* **2009**, *10*, 339–346. [CrossRef]
22. Zhou, L.; Li, Z.; Yang, J.; Tian, G.; Liu, F.; Wen, H.; Peng, L.; Chen, M.; Xiang, J.; Peng, L. Revealing Drug-Target Interactions with Computational Models and Algorithms. *Molecules* **2019**, *24*, 1714. [CrossRef] [PubMed]
23. Tejera, E.; Munteanu, C.R.; Lopez-Cortes, A.; Cabrera-Andrade, A.; Perez-Castillo, Y. Drugs Repurposing Using QSAR, Docking and Molecular Dynamics for Possible Inhibitors of the SARS-CoV-2 M(pro) Protease. *Molecules* **2020**, *25*, 5172. [CrossRef]
24. Balaramnavar, V.M.; Srivastava, R.; Rahuja, N.; Gupta, S.; Rawat, A.K.; Varshney, S.; Chandasana, H.; Chhonker, Y.S.; Doharey, P.K.; Kumar, S.; et al. Identification of novel PTP1B inhibitors by pharmacophore based virtual screening, scaffold hopping and docking. *Eur. J. Med. Chem.* **2014**, *87*, 578–594. [CrossRef] [PubMed]
25. Hao, J.; Yang, Z.; Li, J.; Han, L.; Zhang, Y.; Wang, T. Discovery of natural adenosine monophosphate-activated protein kinase activators through virtual screening and activity verification studies. *Mol. Med. Rep.* **2021**, *23*, 203. [CrossRef] [PubMed]
26. Li, Y.; Peng, J.; Li, P.; Du, H.; Li, Y.; Liu, X.; Zhang, L.; Wang, L.L.; Zuo, Z. Identification of potential AMPK activator by pharmacophore modeling, molecular docking and QSAR study. *Comput. Biol. Chem.* **2019**, *79*, 165–176. [CrossRef]
27. Nanduri, R.; Kalra, R.; Bhagyaraj, E.; Chacko, A.P.; Ahuja, N.; Tiwari, D.; Kumar, S.; Jain, M.; Parkesh, R.; Gupta, P. AutophagyS-MDB: A curated database of small molecules that modulate protein targets regulating autophagy. *Autophagy* **2019**, *15*, 1280–1295. [CrossRef]
28. Ramesh, M.; Vepuri, S.B.; Oosthuizen, F.; Soliman, M.E. Adenosine Monophosphate-Activated Protein Kinase (AMPK) as a Diverse Therapeutic Target: A Computational Perspective. *Appl. Biochem. Biotechnol.* **2016**, *178*, 810–830. [CrossRef]
29. Shi, Q.; Pei, F.; Silverman, G.A.; Pak, S.C.; Perlmutter, D.H.; Liu, B.; Bahar, I. Mechanisms of Action of Autophagy Modulators Dissected by Quantitative Systems Pharmacology Analysis. *Int. J. Mol. Sci.* **2020**, *21*, 2855. [CrossRef]
30. Wang, Z.; Huo, J.; Sun, L.; Wang, Y.; Jin, H.; Yu, H.; Zhang, L.; Zhou, L. Computer-aided drug design for AMP-activated protein kinase activators. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 214–227. [CrossRef]
31. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
32. OECD. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 49. ENV/JM/MONO(2004)24. Available online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.7793&rep=rep1&type=pdf> (accessed on 22 October 2004).
33. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488. [CrossRef]
34. Van der Maaten, L.J.P.; Hinton, G.E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
35. Ho, T.K. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
36. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Math. Intell.* **2005**, *27*, 83–85. [CrossRef]
37. Fan, R.-E.; Chang, K.-W.; Hsieh, C.J.; Wang, X.-R.; Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
38. Tolles, J.; Meurer, W.J. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA* **2016**, *316*, 533–534. [CrossRef] [PubMed]
39. Goodfellow, I.; Bengio, Y.; Courville, A. Universal Approximation Properties and Depth. In *Deep Learning*; MIT Press: Cambridge, UK; London, UK, 2016.
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]