JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

ADME Prediction with KNIME: Development and Validation of a Publicly Available Workflow for the Prediction of Human Oral Bioavailability

Gabriela Falcón-Cano, Christophe MOLINA, and Miguel Angel Cabrera-Pérez

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.0c00019 • Publication Date (Web): 07 May 2020

Just Accepted

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

ADME Prediction with KNIME: Development and Validation of a Publicly Available Workflow for the Prediction of Human Oral Bioavailability

Gabriela Falcón-Cano[†], *Christophe Molina*[‖], *and Miguel Ángel Cabrera-Pérez**[†,§]

[†]Unit of Modeling and Experimental Biopharmaceutics. Centro de Bioactivos Químicos. Universidad Central "Marta Abreu" de las Villas. Santa Clara 54830, Villa Clara, Cuba

‖PIKAÏROS S.A, 31650 Saint Orens de Gameville, France

§Department of Pharmacy and Pharmaceutical Technology, University of Valencia, Burjassot 46100, Valencia, Spain

Corresponding Author
*E-mail: macabreraster@gmail.com

ABSTRACT

In-silico prediction of human oral bioavailability is a relevant tool for the selection of potential drug candidates and for the rejection of those molecules with less probability of success during the early stages of drug discovery and development. However, the high variability and complexity of oral bioavailability and the limited experimental data in public domain have mainly restricted the development of reliable *in-silico* models to predict this property from the chemical structure. In this study we present a KNIME automated workflow to predict human oral bioavailability of new drug and drug-like molecules, based on five machine learning approaches combined into an ensemble model. The workflow is freely accessible and allows the quickly and easily prediction of oral bioavailability for new molecules, where users do not require any

knowledge or advanced experience in machine learning or statistical modeling to automatically obtain their predictions, increasing the potential use of the present proposal.

INTRODUCTION

The poor ADME (Absorption, Distribution, Metabolism and Excretion) properties of potential drug candidates are among the main bottlenecks in the high attrition rate in drug discovery and development.[1] One of the most studied ADME properties, considering its relevance during oral administration, is human bioavailability.[2] The identification of this property in the early stage of drug discovery and development can help researchers to better select candidates for further development by rejecting those with less chance of success.[3]

Data mining integrates all the analysis procedures required for the preparation of data, modeling, evaluation and interpretation of models in the most efficient way. The need of a systematic automated approach of each analysis demands the unification of the above processing steps under a common platform. For more than twenty years, free and public software tools such as Weka, RapidMiner, Orange and Scikit-learn have been developed to facilitate the data analysis process and offer researchers with alternatives to commercially available platforms.[4] The Konstanz Information Miner (KNIME) is a user-friendly and open source platform that provides functionality and tools for data reading, processing and analysis, through the implementation of nodes interconnected as a workflow. KNIME easily allows the integration and compatibility of different software (Weka, Keras, Scikit-learn, etc.) and can handle chemoinformatic problems with a wide variety of nodes. KNIME allows as well to create new custom nodes or integrate scripts implemented in different programming languages (R, Python, Matlab, etc.) for statistical calculation, visualization and analysis of biological data.[5]

Considering that most of the *in-silico* models available to predict oral bioavailability have been developed based on limited public databases, using statistical and artificial intelligence techniques applied individually and without the aim of making automatic the modeling process, this work is intended to develop a semi-automated QSPR modeling system, built on KNIME, to allow the automatized prediction of new drug and drug-like molecules and improve the quality of the *in-silico* bioavailability prediction during the early stages of drug discovery and development. The KNIME workflow developed is publicly available on:[21]

https://pikairos.eu/download/HOB-classification

MATERIALS AND METHODS

Computational tool. This study was performed with the open source software KNIME version 4.0.2 (available free of charge at https://www.knime.com/download-previous-versions). The "KNIME Base Chemistry Types" and the "Nodes and KNIME Chemistry Add-ons" are available at http://update.knime.com/analytics-platform/4.0. The "RDKit KNIME Integration" and the "Indigo KNIME Integration" belong to the community extensions and can be downloaded from http://update.knime.com/community-contributions/trusted/4.0 and http://update.knime.com/community-contributions/4.0, respectively. The "KNIME Weka Data Mining Integration" (3.7) and the "Enalos Nodes for KNIME" can be obtained at http://update.knime.com/community-contributions/trusted/4.0 and http://update.knime.com/community-contributions/4.0, respectively. The "AlvaDesc" extension can be downloaded from https://www.alvascience.com/knime-alvadesc/.[7] AlvaDesc 1.0.16 and academic or commercial licenses can be obtained by requesting a quote in your private area (registration required) or by contacting them directly by email (chm@kode-solutions.net).

Database of oral human bioavailability. To develop *in-silico* bioavailability models, four public data sources from the last ten years, were selected based on their reliability.[8] The first dataset was compiled by Tian et al. and includes 1013 drugs and drug-like molecules with experimental oral bioavailability values.[9] The second dataset was published by Varma et al. and consists of 309 molecules in which oral bioavailability values in humans were obtained as the product of fraction absorbed and fractions escaping from gut-wall and hepatic eliminations, respectively.[10] The third dataset was collected by Kim et al. and was composed of 995 molecules selected from public sources.[3] Finally, the last data source was the experimental bioavailability data reported in, *Lead Optimization for Medicinal Chemists. Pharmacokinetic Properties of Functional Groups and Organic Molecules.*[11] The four databases were carefully checked to compare the bioavailability values and identifiers (SMILES and CAS registry number) between them for the same molecule. In cases where the information was inconsistent, the following strategies were used: i-manual correction of the CAS registry number and SMILES using PubChem's specialized database as a reference (http://www.ncbi.nlm.nih.gov); ii-rectification of bioavailability values that differ by more than five percent from one source to another, taking as reference the values reported in, *Goodman & Gilman´s The Pharmacological Bases of Therapeutics*,[12] *The absolute oral bioavailability of selected drugs*,[13] and *Use of in vitro and in vivo data to estimate the likelihood of metabolic pharmacokinetic interactions;*[14] and iii-average of bioavailability values that differ by less than five percent.

To develop QSPR models for human oral bioavailability (F), considering that there is no global consensus to select the cut-off value to be used in a binary classification problem for predicting this property, and knowing that the strength of a classification model depends on how the upper or lower class is defined, we selected F% = 50 as the choice to achieve a better balance

of both classes (F ≥ 50% high and F < 50% low).[15] Table S1 in supporting information shows the different cut-off values that have been used for human oral bioavailability classification in previous studies.

Data curation. A first developed component "*Data Curation*" (See Figure 1) was implemented to follow the best data curation practices required for the development of QSPR model.[16] In a first stage, filters for unusual valences, salts and molecules with atoms different of H, C, N, O, S, F, Cl, Br, P, B and I were established. At the same time, considering that most of the compounds with molecular weight values higher than 1200 g/mol,[9] could have an active transport mechanism affecting the actual uptake and consequently the prediction of oral bioavailability, a filter to remove this compounds was set up. Later, unconnected structures were identified and the bigger fragment was kept. In a second stage, the molecules were standardized (this step included: remove stereo, neutralize zwitterions and standardize charges), all the molecules were aromatized and explicit hydrogens were added. Finally, in a third stage, *InChi* (IUPAC International Chemical Identifier) codes were generated and used to remove duplicate cases. The experimental bioavailability values of the duplicate cases were verified; if the difference between them was less than 5%, the average was determined; but if the difference was more than 5% then the experimental value was corrected as described above. In the case of stereoisomers, the higher value of human oral bioavailability was retained.

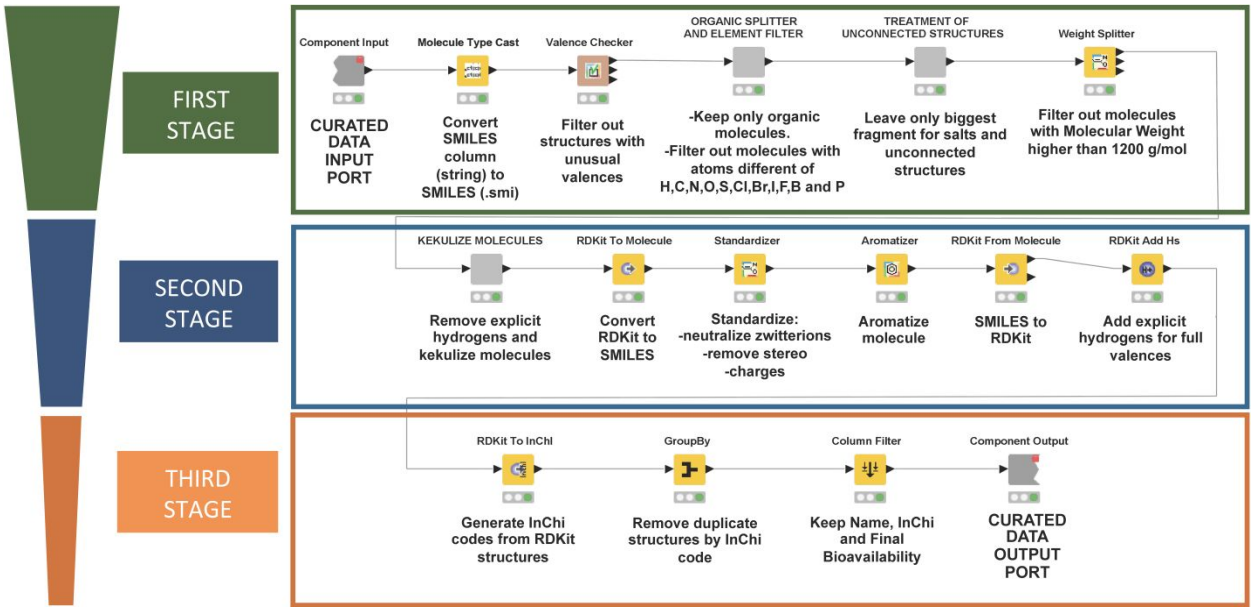


Figure 1. Workflow for the curation of the bioavailability database included in the data curation component

Descriptors Calculation. More than 3800 physicochemical descriptors (0D-2D) and molecular properties were calculated directly from molecular structure using the *"Descriptor"* node, implemented in KNIME as an extension of "*alvaDesc*" software.[7] These included the following families: constitutional indices, ring descriptors, topological indices, walk and path counts, connectivity indices, information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, P_VSA-like descriptors, ETA indices, edge adjacency indices, functional group counts, atom-centered fragments, atom-type E-state indices, pharmacophore descriptors, 2D atom pairs, charge descriptors, molecular properties, and drug-like indices.

Data partition. In order to evaluate the reliability and stability of the *in-silico* bioavailability classification models, a split of the training and validation set was carried out. From the overall data set, 20% of the molecules were randomly divided, using the "*Partitioning*" node, to create an internal test set that was never used in the construction of the models. The rest of the data set conformed the training set and a 10-fold cross validation set used for the development of the

models. Finally, to evaluate the predictive ability of the models and to demonstrate the functionality of the workflow prediction section, an external evaluation set was assembled.

Data normalization and dimensionality reduction. In order to make comparable the molecular descriptors of the training set, the Z-score normalization function was implemented. Subsequently, three methods were applied for dimensionality reduction: (i) filter the descriptors with a very low variance (< 0.001) using the "*Low Variance Filter*" node, (ii) eliminate the highly correlated molecular descriptors (r > 0.95) with the "*Correlation Filter*" node (the column with the most correlated columns is shortlisted and all its correlated columns are filtered out), and (iii) filter the descriptors with missing values based on the "*Missing Value Column Filter*" node. The normalization model was applied to internal test set and the variables excluded during dimensionality reduction were removed from this set.

Data classification. Five machine learning techniques were used for data classification: Gradient Boosted Trees (GBT), Sequential Minimal Optimization (SMO), Bagging (using Classification and Regression Trees (CART) as base classifier), Multilayer Perceptron (MLP) and Naive Bayes (NB). To avoid the risk of overfitting the last three models were trained with a forward feature selection. Gradient Boosted Trees were regularized limiting the maximum number of levels and SMO was regularized controlling the degree of the polynomial kernel. All models were optimized via 10-fold cross validation, using the error rate (*1- Accuracy*) as cost function during the optimization procedure.

In order to improve the accuracy and reliability of the individual predictions, an ensemble model was constructed based on a consensus scoring method. This method involves the application of a majority vote, and the final bioavailability class was chosen as an agreed decision of all individual classification models. In order to categorize the response of the

ensemble model, we calculated the absolute difference between the number of models classifying a molecule as 1 (high bioavailability) and 0 (low bioavailability). Based on the level of difference between the five models, the molecules were organized into three "absolute difference" categories: (i) if three models agree and two disagree, the absolute difference is 1, (ii) if four models agree and one disagree, the absolute difference is 3 and (iii) if all models agree in their classification and none disagree, the absolute difference between possible responses is 5.

Models evaluation. The quality of the individual and consensus models was evaluated by analyzing the values of specificity (SP), sensitivity (SE), precision (PR), overall accuracy (OA), correct classification rate or balanced accuracy (CCR) and Cohen's Kappa statistics for training, test and external evaluation sets. In addition, ROC (Receiver Operating Characteristic) curves were constructed for the final consensus model, and Y-randomization approach was used to evaluate the robustness. Finally, all the modeling procedure, including the optimization procedure, was repeated ten times to generate reliable statistics. In this sense, mean and standard deviation were calculated on the internal test set and the external evaluation set. For the internal test, mean and standard deviation were calculated after repeating the optimization procedure 10 times, varying the structure of the cross validation set each time. In the case of the external test set, mean and standard deviation were obtained after repeating 10 times the random split into training and internal test set, as the method of validation of the modeling process.

Applicability domain. The reliability of a QSPR model depends on its ability to achieve reliable predictions for new molecules that were not considered in the construction of the model, and it is defined by its applicability domain. In this study, the leverage approach was used to evaluating the applicability domain of the consensus model. A molecule was considered outside the

applicability domain when the leverage value is higher than the critical value of 3p/n, where p is the number of model variables plus 1 and n is the size of the training set.[17]

RESULTS

The workflow integrated two main sections (see Figure 2). The first one was related to the development of the QSPR protocol (highlighted in purple) and included the data curation and parametrization of the molecular structure, data partition, normalization and dimensionality reduction and the building, validation and applicability domain assessment of individual and ensemble models. The second section (highlighted in blue) was designed to automatize the prediction of a new external set based on the classification models obtained in the preliminary development section.

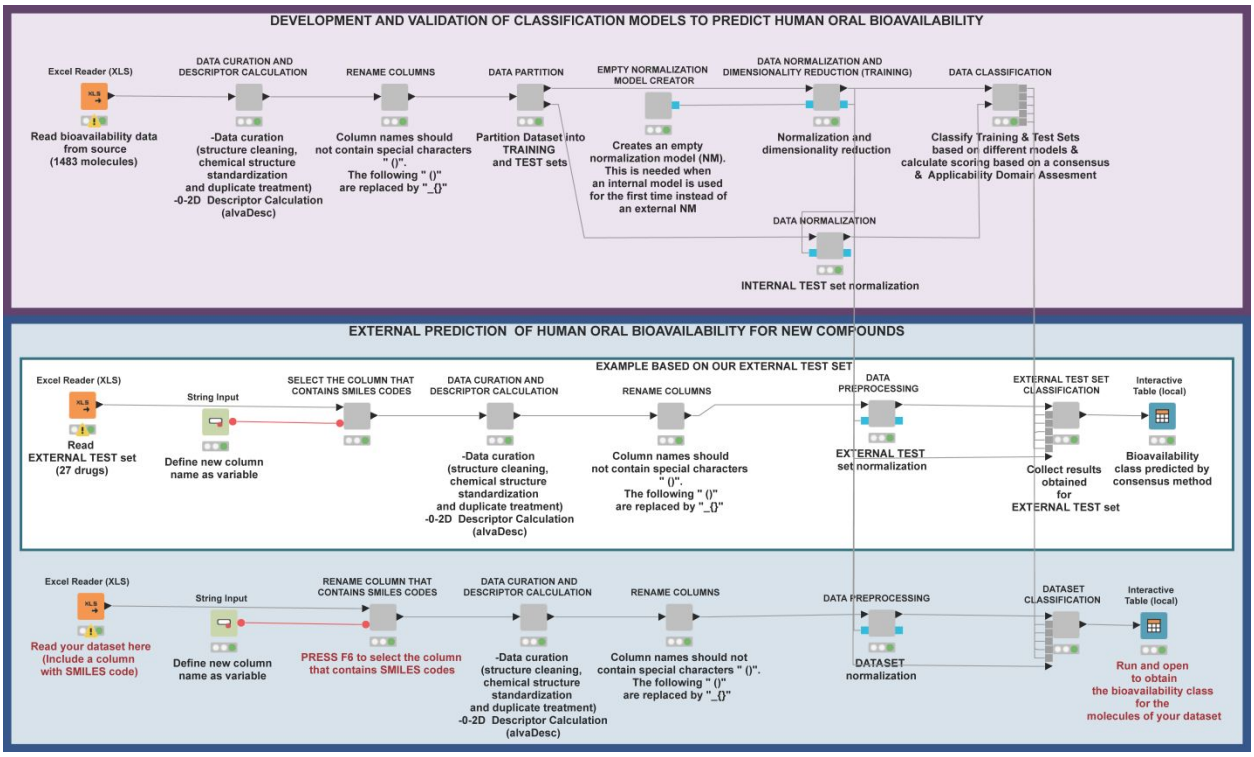


Figure 2. Workflow for modeling and predicting of human oral bioavailability. From top to bottom: First rectangle (highlighted in purple): Development and validation of QSPR protocol; second rectangle (highlighted in blue): the same sequence of steps to automatize the prediction of a new external dataset based on the classification models obtained in the development section. The text highlighted in red are instructions for users. Small section in 2nd rectangle: An example based on our external test set, is included to illustrate to users the bioavailability classification process for new databases.* This picture is also included in the supporting information (Figure S1).

Given the lack of available experimental data on human oral bioavailability, we first built a structurally diverse database, which was formed by 1483 molecules with the SMILES codes and F values (human oral bioavailability expressed as a percentage). This database was read in KNIME using the "*ExcelReader*" node. Each step during pre-processing of a database plays a critical role in the final accuracy of the *in-silico* model:[18]

Data curation. 1448 molecules were available for modeling from the original dataset of 1483 molecules after the data curation procedure. Seven molecules were excluded because they had a molecular weight greater than 1200 g/mol and one molecule was excluded because it had arsenic in its structure. The other 27 excluded molecules were duplicates in the database.

Descriptor calculation. A total of 3852 molecular descriptors (0-2D) were calculated using the "*alvaDesc*" node, and then normalized. Subsequently, with the application of KNIME "*Rule Engine*" node, the database was divided into 661 molecules with low oral bioavailability and 787 molecules with high oral bioavailability, using 50% as a cut-off value.

Data partition. The training and internal test sets had 1158 and 290 molecules, respectively. An external evaluation set of 27 new drugs gathered from the last five years, with known human bioavailability values, was used to exemplify and evaluate the classification on new databases based on the developed classification model. A detailed list of molecules belonging to each set as well as their property values are described in Table S2, S3 and S4 in the supporting information.

Dimensionality reduction. To keep only the most relevant variables in the models, a dimensionality reduction procedure was carried out. After excluding the low variance and high correlation descriptors, 1337 molecular descriptors were available for modeling.

The application of the data curation, descriptor calculation and dimensionality reduction gave us the possibility to obtain the largest and most structurally diverse database of human oral

bioavailability, with the minimum subset of molecular descriptors capable of modeling this property with good performance, lower computational/time cost and in a more robust way.

Data classification. In this component, five machine learning approaches were used to develop bioavailability classification models. Although the public version of the workflow does not include the optimization procedure to select the best parameters of each classification model, Figure 3 illustrates its application. Considering the number of descriptors after dimensionality reduction procedure and to avoid the risk of overfitting the models, Bagging - using Classification and Regression Trees (CART) as base classifier, Multilayer Perceptron (MLP) and Naive Bayes (NB) were trained with a Forward Feature Selection. Gradient Boosted Tree implicitly does variable selection (as do all the tree methods in general), and overfitting was controlled by the parameterization of the algorithm. The parameter "maximum tree depth" was optimized via cross validation, and the optimal value was 2. SVM is also auto-regularized by limiting the number of kernels used to build the SVM classifier, through the optimization of the constant of penalization and the use of a quadratic degree polynomial kernel.

Details on model fit, optimization parameters and the number of descriptors included in each model are shown in Table S5 of the supporting information. The statistics obtained for training, internal test and cross validation sets with the individual models are showed in Figure S2 of supporting information.

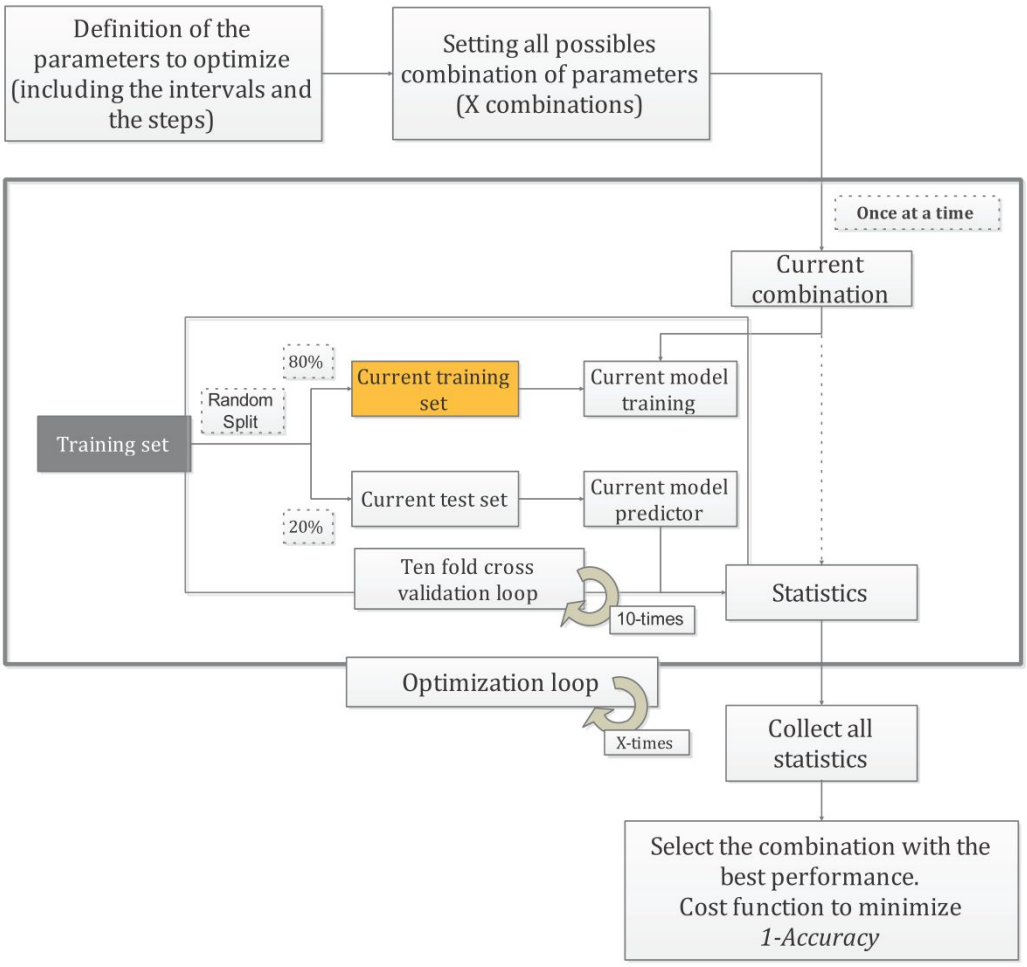


Figure 3. Optimization procedure to select the best parameters of each classification model.

Models evaluation. The performance of every individual classification model is outperformed when an ensemble model is used based on a statistical mode consensus scoring method, where at least three individual classifiers must agree. In this case, every molecule is classified active or inactive based on the majority class in the consensus. The statistics obtained with this five models ensemble method showed overall prediction accuracy for the internal test set of 78.3% ± 1. However, this scoring performance still improves for molecules for which at least four models agree, where a good rating of 82.3% for 242 molecules of the test set was obtained. Figure 4 shows the sensitivity, specificity and balanced accuracy (correct classification rate, CCR) obtained with the consensus scoring method for the test set and Figure 5 shows the ROC curves constructed for training, and internal test sets. All statistics obtained with this

method for training, test and cross validation sets are described in Table S6 of the supporting information.

Figure 4. Sensitivity, specificity and balanced accuracy (correct classification rate, CCR) values obtained with consensus scoring method for the internal test set. Results are shown as mean ± standard deviation.

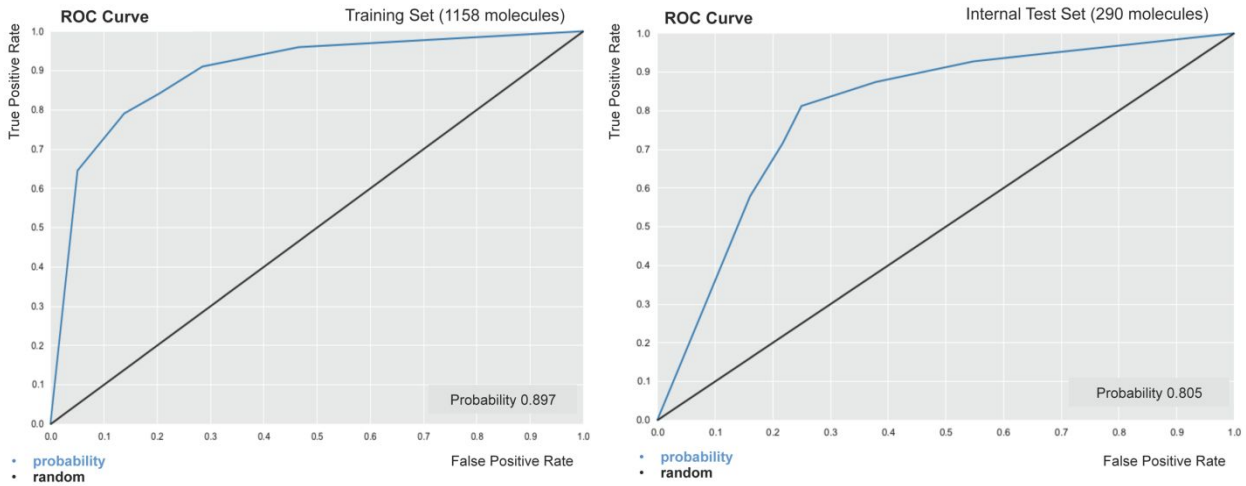


Figure 5. ROC curves, for the training (AUC=0.89 ± 0.015) and internal test set (AUC=0.80 ± 0.021), with the results obtained using the consensus model.

During the Y-randomization test, all models satisfied the acceptance criteria indicating that our models are statistically robust. The models obtained for the training set with randomized activities

had significantly lower values of CCR for the training and test sets than the models built using a training set with real activities.

The implementation of the applicability domain indicates that 21 molecules from the test set had leverage values higher than the 3p/n. In our case the leverage threshold was 0.109. When using the ensemble method and once removed the 21 outlier molecules, the global accuracy for consensus scoring was around 80%.

A clear description of the prediction error (*1-Accuracy)* of the ensemble model *versus* the range of bioavailability values was established. As can be seen in Figure 6, the most variable bioavailability predictions were in the range of 20 to 80 of bioavailability values. These results suggest that a good prediction of drug bioavailability outside these ranges should include information on the influence of efflux by P-glycoprotein (Pgp), metabolism by CYP 450, etc.

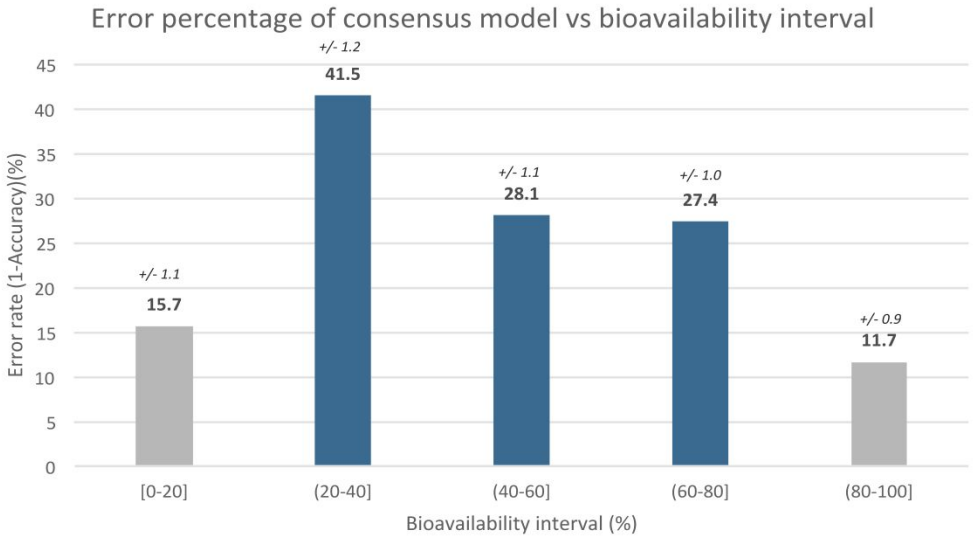


Figure 6. Distribution of prediction errors relative to experimental bioavailability. Error values lower than 25% were highlighted in gray colors. Results are shown as mean ± standard deviation.

In order to identify the limitations of our model, a sequence of nodes was designed to identify the molecules from the training and internal test sets that were incorrectly predicted by all the individual models.[16] Most of false positive structures identified are mainly drugs with high intestinal and hepatic metabolism and substrates of P-glycoprotein an most of false negative compounds are inhibitors of CYP-450 isoenzymes or P-glycoprotein, substrates of paracellular or nucleoside transport. This analysis proves the importance of including variables that describe efflux, metabolism and transport mechanisms in future bioavailability prediction models.

Prediction of human oral bioavailability for an external set. The overall accuracy obtained by the consensus scoring method on the external evaluation set was 77.8% ± 0.9. Once excluded the molecules out of the applicability domain the overall accuracy was 81% ± 0.85.

Comparison with previous studies. Several QSPR classification models have been previously published to predict oral bioavailability. Among these models, only two were derived using large datasets (over 900 molecules). Kim et al.[3] developed seven individual classification models and a consensus model with an external accuracy of 76% using a database of 995 molecules, while Ahmed et al.[19] developed a logistic algorithm with bioavailability prediction accuracy of more than 71%. To our knowledge, this study used the largest human oral bioavailability data with more than 1400 molecules. Considering the size and diversity of the dataset used, the kind of input variable (only physicochemical descriptors and molecular properties) and the predictive performances obtained either globally or by the consensus scoring method when all models agree, our results are respectively comparable or even better than those of previous published models.

CONCLUSIONS

This study reported the first automatic workflow, developed on the KNIME Analytics Platform, to predict human oral bioavailability. All steps of the QSPR were automatized, focusing on data integration and curation procedures to obtain, to our knowledge, the most extensive and structurally diverse database of human oral bioavailability. By combining five machine learning approaches into an ensemble model, a good balanced accuracy was obtained. Although the cut-off value of our model to separate compounds with low and high oral bioavailability was set to F=50%, the final consensus model was able to predict, with lower error values, compounds with very low (≤ 20%) and very high (≥ 80%) extreme bioavailability values. This result highlights the relevance of this model, during the early stage of drug discovery, to reject compounds with lower probability of success or to find candidates for further development. The freely available workflow offers researchers a tool for a rapid and automatized prediction of oral bioavailability of new molecules, where users do not require any knowledge or advanced experience in machine learning or statistical modeling to obtain their predictions, and therefore increases the potential use of this proposal. The influences of certain properties related to the final value of bioavailability will be evaluated in future automated workflows to develop more complete and useful computational models for their application in the drug discovery and development process.

Supporting Information Available: [Table S1. Different cut-off values that have been used for human oral bioavailability classification; Table S2. Training set for bioavailability prediction; Table S3. Internal test set for bioavailability prediction; Table S4. External evaluation set for bioavailability prediction; Table S5. Details on model fit, optimization parameters and the number of descriptors included in each model; Table S6. The statistics obtained for training,

internal test and cross validation sets for the consensus model; Figure S1. Workflow for modeling and prediction of human oral bioavailability; Figure S2. The statistics obtained for training, internal test and cross validation sets for each individual model].

AUTHOR INFORMATION

ORCID

Gabriela Falcón-Cano: 0000-0002-5604-6089

Christophe Molina: 0000-0001-9477-5920

Miguel Ángel Cabrera-Pérez: 0000-0001-5897-2230

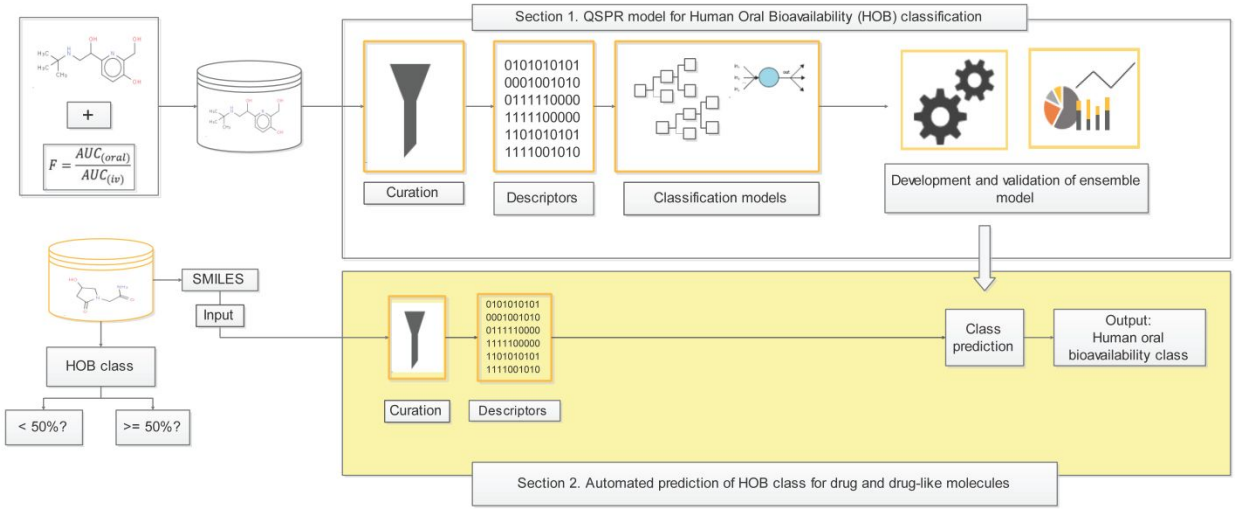
FUNDING INFORMATION

This research received no external funding.

ACKNOWLEDGEMENTS

All the authors acknowledge to Alvascience for the academic licence of alvaDesc. The author M.A.C. acknowledge to the program: Estades Temporals per a Investigadors Convidats, developed at Valencia University in 2018. All authors acknowledge to the referees for their valuable comments.

TABLE OF CONTENTS GRAPHIC



REFERENCES

(1) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S.; Wang, J.; Wallace, O. and Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat. Rev. Drug Discov.* **2015**, *14* (7), 475–486.

(2) Cabrera-Pérez, M. Á.; Pham-the, H. Computational Modeling of Human Oral Bioavailability: What Will Be Next? *Expert Opin. Drug Discov.* **2018**, *13* (6), 509–521.

(3) Kim, M. T.; Sedykh, A.; Chakravarti, S. K.; Saiakhov, R. D.; Zhu, H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. *Pharm. Res.* **2014**, *31* (4), 1002–1014.

(4) Mazanetz, M. P.; Marmon, R. J.; Reisser, C. B. T.; Morao, I. Drug Discovery Applications for KNIME : An Open Source Data Mining Platform. **2012**, 1965–1979.

(5) Trapotsi, M.-A. M.-A. Development and Evaluation of ADME Models Using Proprietary and Opensource Data, University of Hertfordshire, 2017.

(6) Varsou, D.; Nikolakopoulos, S.; Tsoumanis, A.; Melagraki, G.; Afantitis, A. Enalos+ KNIME Nodes: New Cheminformatics Tools for Drug Discovery. In *Rational Drug Design: Methods and Protocols*; 2018; Vol. 1824, pp 113–138.

(7) Alvascience srl. AlvaDesc Version 1.0.16 (Software for Molecular Descriptors Calculation). 2019.

(8) Wang, J.; Hou, T. Advances in Computationally Modeling Human Oral Bioavailability. *Adv. Drug Deliv. Rev.* **2015**, *86*, 11–16.

(9) Tian, S.; Li, Y.; Wang, J.; Zhang, J.; Hou, T. ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints. *Mol. Pharm.* **2011**, *8* (3), 841–851.

(10) Varma, M. V. S.; Obach, R. S.; Rotter, C.; Miller, H. R.; Chang, G.; Steyn, S. J.; El-kattan, A.; Troutman, M. D. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J Med Chem* **2010**, *53* (3), 1098–1108.

(11) Dorwald, F. Z. *Lead Optimization for Medicinal Chemists. Pharmacokinetic Properties of Functional Groups and Organic Compounds*; WILEY-VCH Verlag GmbH & Co. KGaA: Germany, 2012.

(12) Hardman, J. G.; Goodman, A.; Limbird, L. E. Goodman & Gilman's The Pharmacological Basis of Therapeutics. 9th ed. McGraw-Hill: New York, USA 1996.

(13) Sietsema, W. K. The Absolute Oral Bioavailability of Selected Drugs. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **1989**, *27* (4), 179–211.

(14) Bertz, R. J.; Granneman, G. R. Use of in Vitro and in Vivo Data to Estimate the Likelihood of Metabolic Pharmacokinetic Interactions. *Clin Pharmacokinet* **1997**, *32*

(3), 210–258.

(15) Olivares-Morales, A.; Hatley, O. J. D.; Turner, D.; Galetin, A.; Aarons, L.; Rostami-Hodjegan, A. The Use of ROC Analysis for the Qualitative Prediction of Human Oral Bioavailability from Animal Data. *Pharm. Res.* **2014**, *31* (3), 720–730.

(16) Fourches, D.; Muratov, E.; Tropsha, A.; Hill, C. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56* (7), 1243–1252.

(17) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space : A Review. *ATLA* **2005**, *33*, 445–459.

(18) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29* (6–7), 476–488.

(19) Ahmed, S.; Ramakrishna, V. Systems Biological Approach of Molecular Descriptors Connectivity: Optimal Descriptors for Oral Bioavailability Prediction. *PLoS One* **2012**, *7* (7), 1–10.

(20) RDKit: Open-source cheminformatics, http://www.rdkit.org (date of access: 03-13-2020)

(21) https://pikairos.eu/scientific-activity/paper-submissions/adme-prediction-with-knime-1-an-integrated-open-and-semi-automated-workflow-for-the-prediction-of-human-oral-bioavailability (date of access: 03-16-2020)