https://dx.doi.org/10.21577/0103-5053.20200160

J. Braz. Chem. Soc., Vol. 32, No. 1, 110-122, 2021
Printed in Brazil - ©2021 Sociedade Brasileira de Química

Article

Automated Framework for Developing Predictive Machine Learning Models for Data-Driven Drug Discovery

Bruno J. Neves,[a,b] José T. Moreira-Filho,[b] Arthur C. Silva,[b] Joyce V. V. B. Borba,[b] Melina Mottin,[b] Vinicius M. Alves,[c] Rodolpho C. Braga,[d] Eugene N. Muratov[b,c,e] and Carolina H. Andrade *,[b]

[a]Laboratório de Quimioinformática (LabChem), Centro Universitário de Anápolis (UniEVANGÉLICA), 75083-515 Anápolis-GO, Brazil

[b]Laboratório de Planejamento de Fármacos e Modelagem Molecular (LabMol), Faculdade de Farmácia, Universidade Federal de Goiás, 74605-170 Goiânia-GO, Brazil

[c]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, 27955-7568 Chapel Hill-NC, USA

[d]InsilicAll Ltda., 04363-090 São Paulo-SP, Brazil

[e]Departamento de Ciências Farmacêuticas, Universidade Federal da Paraíba, 58059-900 João Pessoa-PB, Brazil

The increasing availability of extensive collections of chemical compounds associated with experimental data provides an opportunity to build predictive quantitative structure-activity relationship (QSAR) models using machine learning (ML) algorithms. These models can promote data-driven decisions and have the potential to speed up the drug discovery process and reduce their failure rates. However, many essential aspects of data preparation and modeling are not available in any standalone program. Here, we developed an automated framework for the curation of chemogenomics data and to develop QSAR models for virtual screening using the open-source KoNstanz Information MinEr (KNIME) program. The workflow includes four modules: (*i*) dataset preparation and curation; (*ii*) chemical space analysis and structure-activity relationships (SAR) rules; (*iii*) modeling; and (*iv*) virtual screening (VS). As case studies, we applied these workflows to four datasets associated with different endpoints. The implemented protocol can efficiently curate chemical and biological data in public databases and generates robust QSAR models. We provide scientists a simple and guided cheminformatics workbench following the best practices widely accepted by the community, in which scientists can adapt to solve their research problems. The workflows are freely available for download at GitHub and LabMol web portals.

Keywords: drug discovery, KNIME, predictive modeling, machine learning, virtual screening

Introduction

Quantitative structure-activity relationship (QSAR) modeling is a major cheminformatics approach in computer-aided drug discovery.[1,2] Nowadays, machine learning (ML) methods can be used to generate QSAR models that accurately predict chemicals and how chemical modifications might influence biological properties.[2] In contrast to classical QSAR models that used simple multivariate regression approaches to correlate biological activity with structure and chemical properties, advanced cheminformatics and ML techniques are able to model more complex and nonlinear data. ML uses pattern recognition algorithms to discern mathematical relationships between experimental observations of small molecules and extrapolate them to predict the biological properties of novel compounds.[3,4]

One of the primary application areas for ML in drug discovery is to predict chemicals lacking of biological data.[5-8] Over the past decade, there has been a remarkable increase in the amount of available bioassay data in repositories such as ChEMBL[9] and PubChem[10] owing to

*e-mail: carolina@ufg.br

the emergence of new experimental techniques such as high-throughput screening (HTS).[11,12] This rapid increase in publicly available data has allowed the training of ML algorithms to guide the development of lead compounds in drug discovery as well as in the assessment of chemical safety of untested compounds.[2]

A variety of ML methods such as random forest (RF),[13] support vector machines (SVM),[14] and deep neural networks (DNN)[15] have been utilized for drug discovery.[16] In particular, ML models can benefit the drug discovery process, due to its low cost and ability to screen a large number of compounds in a short period of time.[4,16,17] However, the success of ML in cheminformatics requires a series of pre-processing steps, such as chemical and biological data curation,[18-20] dataset balancing, descriptors calculation, modeling, validation, statistical analysis, etc.,[2] that can be performed using a wide variety of programming languages and computational tools. Additionally, the development and implementation of high-quality models require that users have a thorough understanding of the modeled bioassay data, expert comprehension of best practices for model development, validation and application,[21] and computational skills.

Until this date, many essential aspects of data preparation and modeling are not available in a single standalone program. Here, we developed an automated computational framework to curate and prepare datasets, to generate and validate predictive ML models, and to perform virtual screening (VS) of chemical libraries using the KoNstanz Information MinEr (KNIME) program.[22,23] KNIME is an open-source platform that provides a customizable framework for data management and modeling through a user-friendly graphical interface.[24,25]

Results and Discussion

Automated framework

In the current work, we illustrate and describe the development of an automated framework to curate, model, analyze, and screen chemicals using the KNIME platform. In addition, we tested the workflow by developing four case studies for the prediction of antiplasmodial and antischistosomal activity, as well as cardiotoxicity and mutagenicity. Previously, other groups have proposed automated[26] and semi-automated[27] KNIME workflows for the development of ML models and cheminformatics analysis. Similar workflows[28] using the commercial Pipeline Pilot[29] program have also been published. More recently, Java and JavaFX based program was developed for the removal of duplicates, activity cliffs, and modeling focusing specifically on small datasets.[30] However, our workflow is the first to integrate all the aspects of the best practices for the development and validation of QSAR modeling.

The main framework was subdivided into four tasks: (*i*) dataset preparation and curation;[18-20] (*ii*) chemical space analysis and structure-activity relationships (SAR) rules (see Supplementary Information section); (*iii*) ML modeling;[2] and (*iv*) VS.[31] Each subtask is performed within the workflows (see sections below) that are encapsulated inside metanodes for improving the organization and the layout of the routines, increasing their flexibility for future adaptations.

Dataset preparation and curation

Each task in the data preparation and curation workflow is critical to the development of predictive QSAR models.[18-20] Usually, data stored in public databases contains a fraction of erroneous records resulted from measurement variations and insufficient quality assessment. The first step is the curation of the chemical data, which allows for the identification and correction of errors in chemical structures.[20] Mixtures of components, inorganic compounds and organometallic compounds are removed (if these are underrepresented in the dataset) and standardization of specific chemotypes such as aromatic rings, nitro groups, and tautomeric forms is required. Counterions are removed and any duplicate compounds identified should be analyzed and removed. Duplicate analysis, the next step of our workflow, is critical because it allows for the evaluation of the quality of experimental data and for the removal of (*i*) chemicals associated with duplicate records with contradictory experimental results and (*ii*) records repeating the same experimental outcome for the same compound. The presence of duplicates directly affects the quality of models, i.e., duplicates with identical activity present in both training and test sets lead to an overestimation of the quality of the models. Manual inspection is required at the end of the process to ensure that all structures are correct. Unreliable sources must be identified and removed. The amount of time and effort spent at this phase will depend on the dataset used.

Input data

The data were downloaded in comma-separated values (CSV) format (bioactivities) and standard structure-data file (SDF) format (chemical structures) from PubChem Bioassay[32] or in CSV format (bioactivities + chemical structures) from the ChEMBL database.[9,33] We used four reference datasets: *Plasmodium falciparum* 3D7 strain (*Pf*3D7),[34] *Schistosoma mansoni* thioredoxin glutathione reductase (*Sm*TGR),[35] human ether-a-go-go-related gene (hERG),[36] and Ames mutagenicity,[37] as case studies to develop an integrated pipeline for data preparation and curation for drug discovery and toxicity. An overview of the datasets is shown in Table 1.

Data files gathered from ChEMBL and PubChem databases have their particularities about column names and exported files' extension when exported from the database. Therefore, considering these differences among the raw data that will serve as input for the user, two separate workflows were developed to treat bioassay data from PubChem bioassay (Figure 1a) and ChEMBL (Figure 1b). In the PubChem bioassay workflow, bioactivities (loaded in "CSV Reader" node) and chemical structures (loaded

Table 1. Summary of the curated datasets

Dataset	Source	Activity type		Activity threshold	No. of entries before curation	No. of entries after curation	Concordance / %	Balanced dataset (1:1)
*Pf*3D7	CHEMBL2366922[34]	categorical	pEC_{50}	6.0	1,855	1,337	86.2	1,134
		continuous	pEC_{50}	–	1,855	1,173	18.3	–
*Sm*TGR	PubChem 485364[35]	categorical	pIC_{50}	5.0	359,841	316,663	99.2	5,024
Ames	literature[37]	categorical	phenotype	–	7,546	6,931	94.2	6,114
hERG	ChEMBL 240[36]	categorical	pIC_{50}	5.0	9,859	4,673	88.7	4,656

*Pf*3D7: *Plasmodium falciparum* 3D7 strain; pEC_{50}: negative logarithm of the half maximal effective concentration; *Sm*TGR: *Schistosoma mansoni* thioredoxin glutathione reductase; Ames: mutagenicity; hERG: human ether-a-go-go-related gene; pIC_{50}: negative logarithm of the half maximal inhibitory concentration.

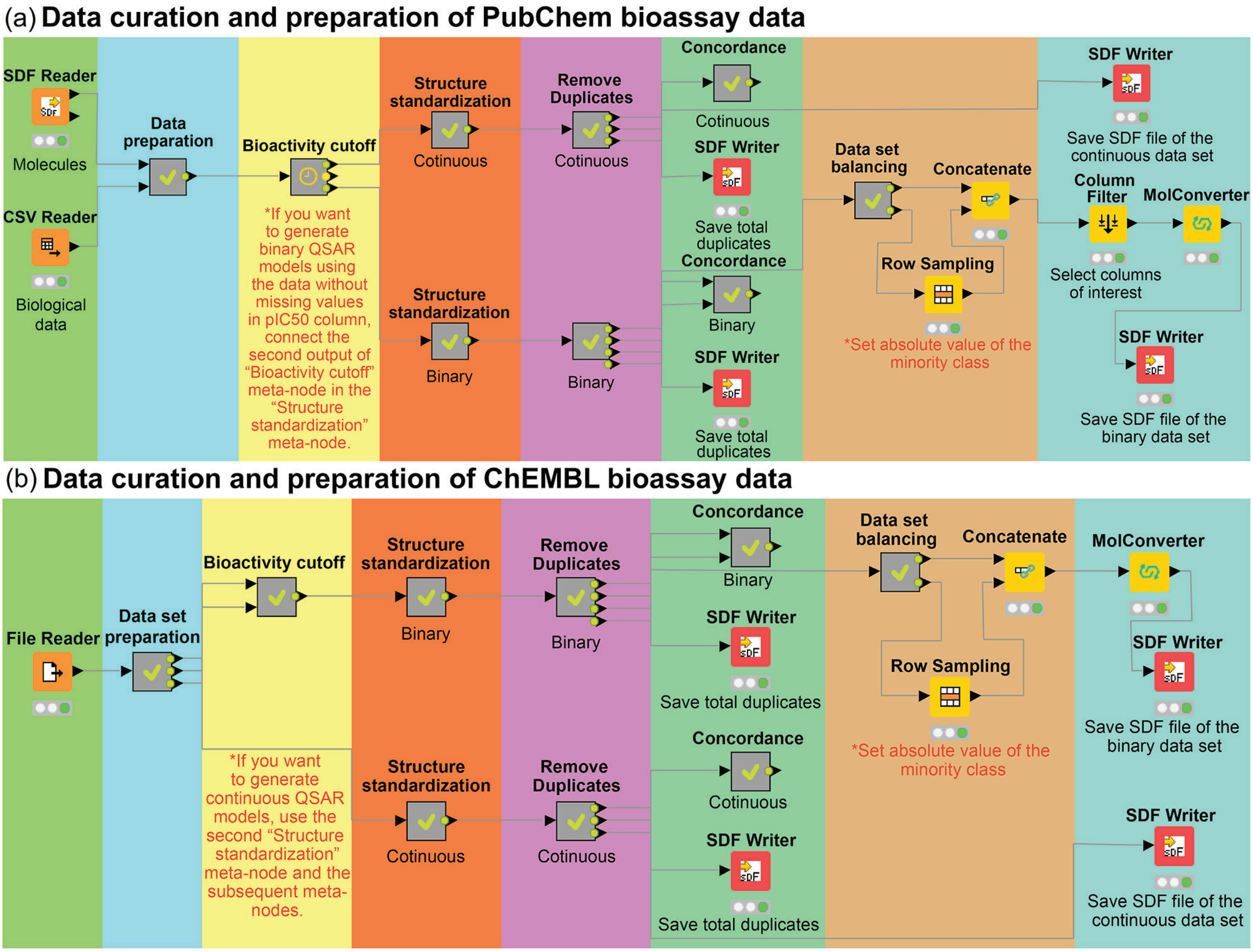

(a) Data curation and preparation of PubChem bioassay data
SDF Reader
Molecules
CSV Reader
Biological data
Data preparation
Bioactivity cutoff
*If you want to generate binary QSAR models using the data without missing values in pIC50 column, connect the second output of "Bioactivity cutoff" meta-node in the "Structure standardization" meta-node.
Structure standardization
Cotinuous
Structure standardization
Binary
Remove Duplicates
Cotinuous
Binary
Concordance
Cotinuous
SDF Writer
Save total duplicates
Concordance
Binary
SDF Writer
Save total duplicates
Data set balancing
Concatenate
Row Sampling
*Set absolute value of the minority class
SDF Writer
Save SDF file of the continuous data set
Column Filter
Select columns of interest
MolConverter
SDF Writer
Save SDF file of the binary data set
(b) Data curation and preparation of ChEMBL bioassay data
File Reader
Data set preparation
Bioactivity cutoff
*If you want to generate continuous QSAR models, use the second "Structure standardization" meta-node and the subsequent meta-nodes.
Structure standardization
Binary
Structure standardization
Cotinuous
Remove Duplicates
Binary
Remove Duplicates
Cotinuous
Concordance
Binary
SDF Writer
Save total duplicates
Concordance
Cotinuous
SDF Writer
Save total duplicates
Data set balancing
Concatenate
Row Sampling
*Set absolute value of the minority class
MolConverter
SDF Writer
Save SDF file of the binary data set
SDF Writer
Save SDF file of the continuous data set

Figure 1. General overview of data preparation and curation workflow scheme developed for PubChem Bioassay (a) and ChEMBL (b) data. These two workflows were prepared to deal with the particularities of the gathered data from PubChem Bioassay and ChEMBL databases.

in "SDF Reader" node) are merged using compound IDs. In the ChEMBL workflow, a CSV file containing the simplified molecular input line entry specification (SMILES) notations of molecules is read in using the "CSV Reader" node. The input parameters of each workflow are configured separately. The user must double-click on them to open the configuration window, load the curated dataset saved as SDF and CSV formats, then click the "OK" button.

Data preparation

Next, the input data is passed through the "Data preparation" metanode to normalize or transform different measures of binding affinity. For example, bioactivities (half maximal inhibitory concentration (IC_{50}) or half maximal effective concentration (EC_{50})) on the mass scale (e.g., µg mL^{-1}) were transformed to the molar scale (µM mL^{-1}), then normalized to negative logarithm (–log) units (i.e., pEC_{50} and pIC_{50}). Fundamentally, dose-dependent inhibition is a logarithmic phenomenon, so it makes sense to work in this manner. Subsequently, the "Bioactivity cutoff" metanode is used to set a threshold value that differentiates active/toxic compounds from inactive/non-toxic compounds. The bioactivity cutoffs were selected according to hit and lead criteria in drug design.[38] Details on the threshold values used in each dataset are shown in Table 1. The selected thresholds are based on data distribution and on literature for that particular endpoint. It is worth noting that this step is data-dependent, and the user must perform all the transformations according to their own data.

Structure standardization

Very often, public datasets contains chemicals represented in different formats due to the experimental protocols they were evaluated or due to different protocols for drawing/storing chemicals. To solve these problems, the "Structure standardization" metanode is employed to standardize and clean all chemical structures according to protocols developed by Fourches *et al.*[18-20] Explicit hydrogens are added, whereas polymers, salts, metals, organometallic compounds, and mixtures are removed to follow the best practices in data curation, since most of descriptor-generating program do not properly process these structures, generating errors in descriptors and fingerprints calculation. In addition, specific chemotypes such as aromatic rings and nitro groups are normalized, and valences are validated or corrected. This allows for the detection of the most common errors in chemical structures, such as abbreviations of functional groups (e.g., Phe as phenyl) and incorrectly assigned valences or aromaticity. Finally, International Chemical Identifier Keys (InChIKey) are generated for each entry in the dataset. InChIKey is an efficient method for detecting duplicates, once molecules have been adequately standardized.

Analysis of duplicates

Datasets ready for modeling must have unique compounds that are structurally different from all other compounds in the dataset. However, the same compound may be present many times in the same dataset. If modelers build models using datasets containing these structural duplicates in both modeling and external sets, the predictivity of these models will be overestimated.[18-20] Therefore, duplicates must be identified and removed prior to any modeling study. Here, InChIKey notations are used to automatically identify duplicate entries in the "Analysis of duplicates" metanode. Once duplicates are identified, an analysis of their bioactivities is performed using the "Concordance" metanode. In this step, the intra- and inter-laboratory assay concordance between duplicate records is investigated to ensure consistency and quality of the datasets. Lastly, duplicates are removed as follows (Figure 2):

Binary models

(*i*) If duplicates have discordant outcomes, both entries are excluded; and (*ii*) if the reported outcomes of the duplicates were the same, one entry is retained in the dataset and the other excluded.[18]

Continuous models

(*i*) If duplicates presented difference > 0.2 logarithmic units (as proposed by Fourches *et al.*),[18] both entries are excluded; and (*ii*) if the reported potencies are ≤ 0.2, an average of the values was calculated, and one entry is retained in the dataset.[18] The number of duplicates identified in each dataset and overall concordance are shown in Table 1.

Dataset balancing

Data balancing will lead to loss of important data and might reduce chemical coverage. One should always try to develop models with balancing the data. However, this is not always possible. Usually, in HTS data, the number of active compounds is much smaller than the number of inactive compounds.[39] In this case, binary ML models built from imbalanced datasets may be biased toward the prediction of the majority class and may be poorly predictive for the minority class.[39] Other approaches for dealing with imbalanced datasets are discussed elsewhere.[39] Considering the unbiased characteristics of studied datasets (see details in Table 1), an under-sampling approach (i.e., reducing the size of the majority class) is applied through the "Dataset balancing" metanode.[3] Datasets used in

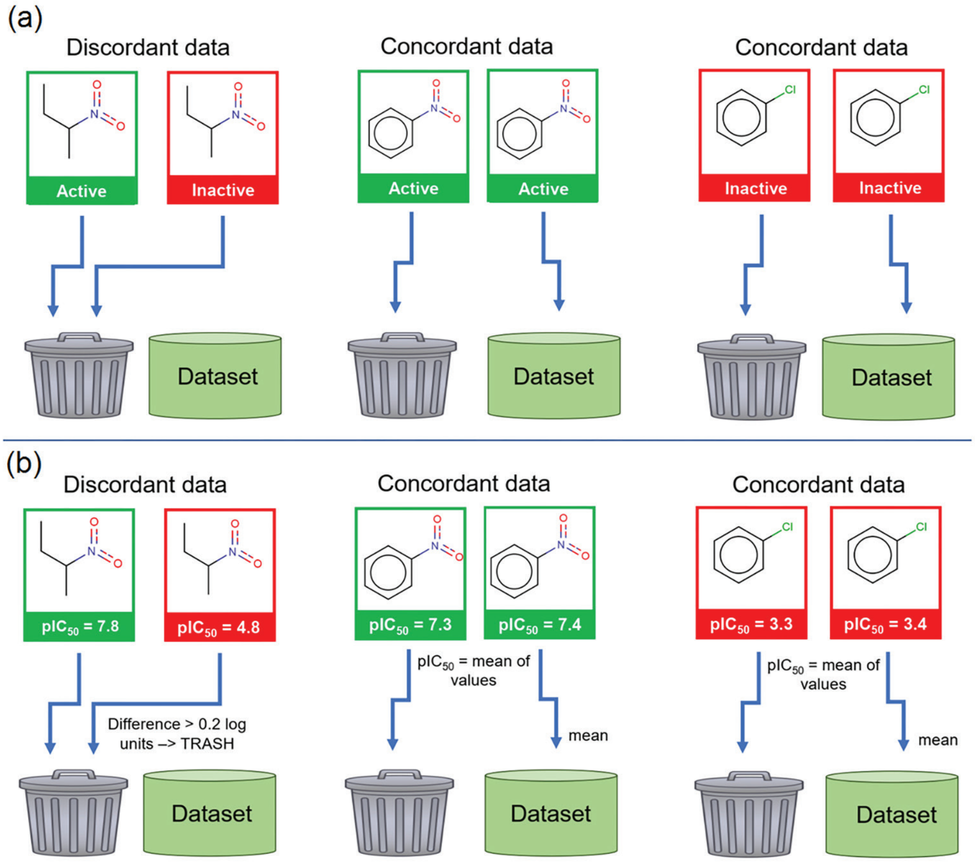


Figure 2. Criteria for duplicate data analysis in categorical (a) and continuous (b) bioassay data.

continuous ML models must be saved without performing the balancing step.

The under-sampling strategy used here retains most of the representative structures of the majority class, thus ensuring a structural diversity that is most representative of the original chemical space.[3] Initially, the Euclidean distances between each compound in the majority class and those present in the minority class are measured using k-nearest neighbor (k-NN) algorithm. Then, representative molecules with high, medium or low chemical similarity in the majority class were selected using k-distances and extracted to generate balanced datasets.[40] The user must check out the number of compounds in minority class to see how many compounds will be necessary to linearly select from the majority class. This number must be inserted in the configuration of the node "Row sampling," in the "Absolute" field. Finally, each balanced dataset may be saved by "SDF Writer" node in the user-defined directory. Details of the number of compounds in each balanced dataset are shown in Table 1.

Automated model building

The developed workflow (Figure 3) aims to simplify and automate the model-building protocol according to the best practices for building predictive models.[2,21] The details of each step of modeling procedure are covered in the following sections.

Input data

To run the workflow, the "SDF Reader" node must be loaded with a curated dataset in SDF format. This node was configured to allow for a column containing chemical structures and associated biological data (binary or continuous). Datasets with binary data must have the experiment result column labeled as "outcome" and datasets with continuous data must have the pIC_{50} data labeled as "pIC_{50}" (Figure 3).

Molecular fingerprints

Molecular descriptors are the result of mathematical procedures that transform chemical structure of a molecule into relevant numerical data. These descriptors can be used to establish relationships between the chemical structure and biological property of interest.[41] Fingerprints are a type of descriptors encoded as bit strings which encode the absence (0) or presence (1) of a fragment or atom in a chemical structure. The current workflow

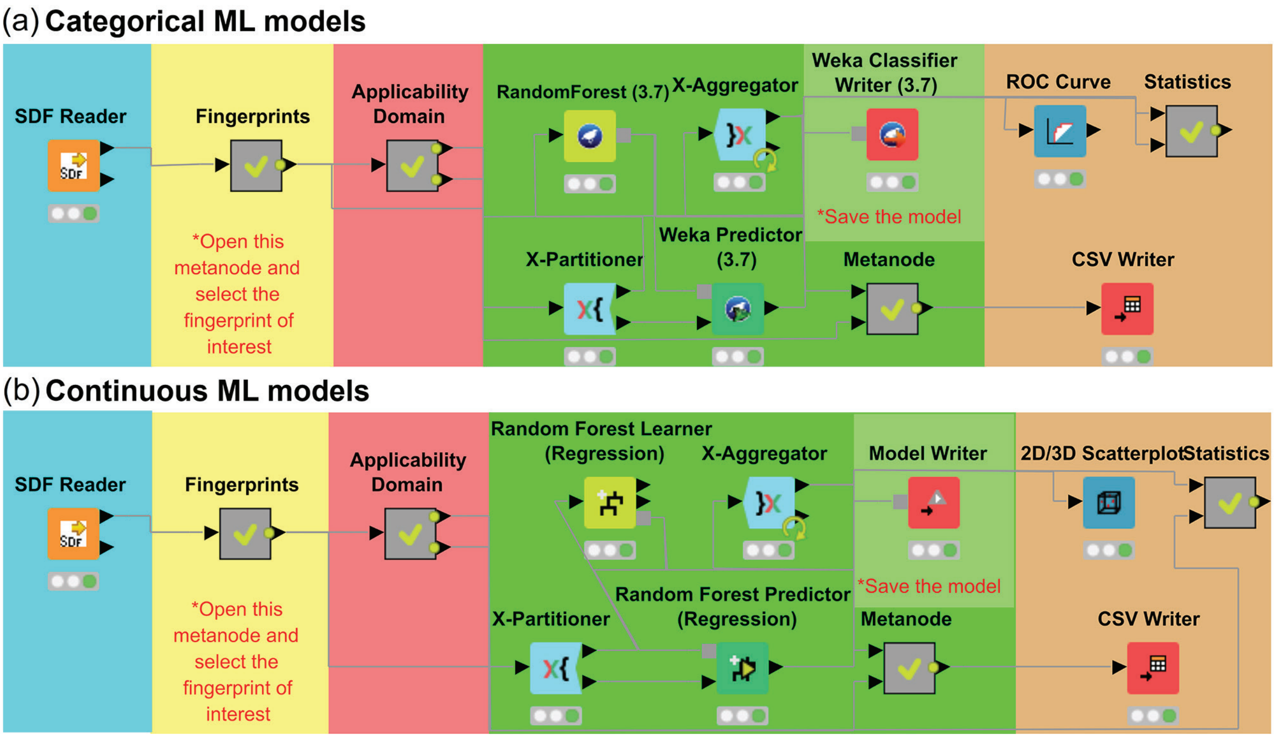


Figure 3. General workflow for automated QSAR modeling of (a) categorical and (b) continuous data.

automatically calculates different fingerprint types, including substructure-based fingerprints (Avalon),[42] and circular fingerprints (Morgan[43] and FeatMorgan).[43] Within the "Fingerprints" metanode, the user must click on "RDKit fingerprints" node to select the desired fingerprint type. In addition, circular fingerprints may be adjusted according to bond radius and number of bits, whereas path-based fingerprints may be adjusted by number of bits and path length. In this work, all fingerprints were generated for the chosen datasets using radius 2 (for the circular fingerprints) and bit vector of 2,048 bits (for the circular and path-based fingerprints).

Highly correlated descriptors are linearly dependent and have similar effect on the dependent variable.[2] Some algorithms are more prone to bias if correlated variables are used. Although ensemble trees are less prone to bias, removing correlated variables can make model building faster and facilitate interpretation. After calculating fingerprint descriptors, the workflow calculates the correlation between all descriptors and removes one of the highly correlated.

5-Fold external cross-validation (5FECV)

After the molecular fingerprint calculations, the dataset is split into five subsets of equal size using the "X-Partitioner" node. Four of these subsets form the training set (80% of the full set), while the remaining subset (20% of all compounds) serves as the test set. This procedure is repeated five times, allowing each of the five subsets to be used as a test set. Models are built using the training set while the compounds in the test set (fold) are employed to evaluate the predictive performance. At the end of each iteration, the "X-Aggregator" node collects the prediction results. All nodes in between these two nodes are executed as many times as repetitions should be performed.[21]

Applicability domain

One of the most important problems in any QSAR modeling is establishing the applicability domain (AD). The AD must be determined for the given chemical space of predictive models in order to localize "reliable" and "unreliable" regions for prediction.[44] Users should be able to trust the model's predictions if they have evidence that the chemical space used for training matches the chemical space of the compounds not previously seen by the model. We used the "Applicability Domain" metanode to estimate the AD of the developed models (Figure 3). Within this metanode, the "Domain-Similarity" node utilizes Euclidean distances to define chemical similarity among all training compounds and each compound in the test set. This prediction may be unreliable if the distance of a compound not present in the test set to its nearest neighbor in the training set is higher than an arbitrary parameter ($Z = 0.5$) that controls the significance level.[45]

Table 2. Statistical characteristics of developed QSAR models using RF and assessed by 5-fold external cross validation

				Categorical model			
Endpoint	Fingerprint	CCR	SE	SP	PPV	NPV	Coverage
	Avalon	0.85	0.82	0.88	0.87	0.83	0.59
*Pf*3D7	Morgan	0.87	0.86	0.88	0.87	0.86	0.83
	FeatMorgan	0.84	0.83	0.85	0.85	0.83	0.80
	Avalon	0.84	0.82	0.87	0.87	0.83	0.72
*Sm*TGR	Morgan	0.85	0.84	0.86	0.86	0.84	0.51
	FeatMorgan	0.84	0.85	0.84	0.84	0.85	0.54
	Avalon	0.80	0.81	0.79	0.79	0.80	0.95
Ames	Morgan	0.80	0.81	0.79	0.79	0.80	0.67
	FeatMorgan	0.79	0.80	0.77	0.78	0.79	0.56
	Avalon	0.77	0.79	0.75	0.76	0.78	0.74
hERG	Morgan	0.77	0.78	0.76	0.76	0.77	0.92
	FeatMorgan	0.78	0.79	0.77	0.78	0.79	0.89
				Continuous model			
Endpoint	Fingerprint	R^2	Q^2_{ext}	RMSECV	k	k'	Coverage
	Avalon	0.73	0.73	0.61	1.00	0.99	0.28
*Pf*3D7	Morgan	0.73	0.72	0.62	1.00	0.99	0.83
	FeatMorgan	0.73	0.72	0.62	1.00	0.99	0.78

CCR: correct classification rate; SE: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value; Coverage: percentage of test set compounds within the applicability domain; *Pf*3D7: *Plasmodium falciparum* 3D7 strain; *Sm*TGR: *Schistosoma mansoni* thioredoxin glutathione reductase; Ames: mutagenicity; hERG: human ether-a-go-go-related gene; R^2: correlation coefficient; Q^2_{ext}: predictive squared correlation coefficient for the test set; RMSECV: root mean square error of cross validation; k and k': slopes of regression lines through the origin.

Machine learning

To validate the workflow, we built categorical models using datasets previously studied by our group (i.e., *Sm*TGR,[3] *Pf*3D7,[6] Ames mutagenicity and hERG),[46] as well as continuous models using *Pf*3D7 dataset. For this purpose, KNIME contains nodes for most popular ML algorithms, such as RF[13] and SVM[14] and the user can decide which algorithm to use. As an example, we only used RF algorithm to build models for our case-studies. Categorical (Figure 3a) and continuous (Figure 3b) ML models can be generated with any of the mentioned algorithms using the "Weka" or "Analytics" nodes, respectively. Subsequently, optimization parameters of each algorithm may be adjusted by the user in their corresponding node. For example, in the "RandomForest (3.7)", the number of trees and maximum depth parameters may be adjusted to increase model predictivity and to avoid overfitting. After completion of the model building step, the output of the developed model is saved in the user's defined working directory using the "Weka Classifier Writer (3.7)" node (categorical models) or "Model Writer" node (continuous models).

Performance of ML models

The external predictivity resulting from a 5FECV procedure can be adequately assessed with the "Statistics" metanode. During this step, the categorical models are evaluated using correct classification rate (CCR), sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV), whereas continuous models are evaluated using correlation coefficient (R^2), root mean square error of cross validation (RMSECV), and predictive squared correlation coefficient for the test set (Q^2_{ext}). After this step, users can easily access statistical results of built models by right-clicking on the metanode and selecting the option "Connected to: Filtered table". The statistical characteristics of the models developed in this study are summarized in Table 2. As one can see, all the models were robust and predictive, with CCR for external sets in the range of 0.77-0.87. The binary *Sm*TGR models showed predictive power similar to that obtained in our previous ML study.[3] For the Ames mutagenicity data, all individual models presented a CCR 2% and SP 13%, higher than the consensus model developed by Alves *et al.*[47] For the continuous *Pf*3D7 models, the combination of fingerprints with RF led to predictive models (Table 2) with Q^2_{ext} values ranging between 0.72-0.73 and R^2 of 0.73. The remaining models were not compared with public models since they were developed using unexplored or old datasets. Hence, the implemented protocol efficiently generates robust models with reliable predictive performance.

Virtual screening

The VS is a computational procedure to filter down large chemical libraries (i.e., 10^5 to 10^8 compounds) to prioritize a smaller number of compounds that will then be tested experimentally (i.e., 10^1 to 10^3 compounds).[48] Although VS was introduced 30 years ago, many molecular databases still contain errors and molecules with undesired physicochemical properties. In order to address this issue, we developed a comprehensive workflow (Figure 4) for the VS of potentially active and non-toxic compounds by a practical application on the ChemBridge EXPRESS-Pick Collection.[49]

Input data

To execute the workflow, the "SDF Reader" node must be loaded with the EXPRESS-Pick Collection (504,599 compounds) or any other library in SDF format.[49] This node was configured to provide a column containing chemical structures and associated physicochemical properties.

Data curation

The first step of the VS module contains a set of procedures to guarantee that structures are well represented and standardized using the same protocol for those employed in generating the models. In this step, structures are standardized, problematic molecules and duplicates are removed from the library. Subsequently, 2D chemical structures are stored within a KNIME table, and tagged with a ChemBridge identifier and experimental physicochemical properties. Several rules based on molecular property distribution were developed to characterize specific subsets of chemical libraries such as lead-like molecules (molecular weight ≤ 460, –4 ≤ logP (logarithm of octanol/water partition coefficient) ≤ 4.5, logSw (logarithm of aqueous solubility) ≥ –5, ≤ 5 H-bond donors, ≤ 9 H-bond acceptors, ≤ 9 rotatable bonds, and ≤ 4 number of rings).[50] Finally, we adopted the "Aggregator advisor" metanode as a filter to identify molecules that are known to aggregate or may aggregate in prospective experimental assays. The criteria used to predict high probability for aggregation are logP > 3.0 and Tanimoto coefficient ≥ 85% to the closest

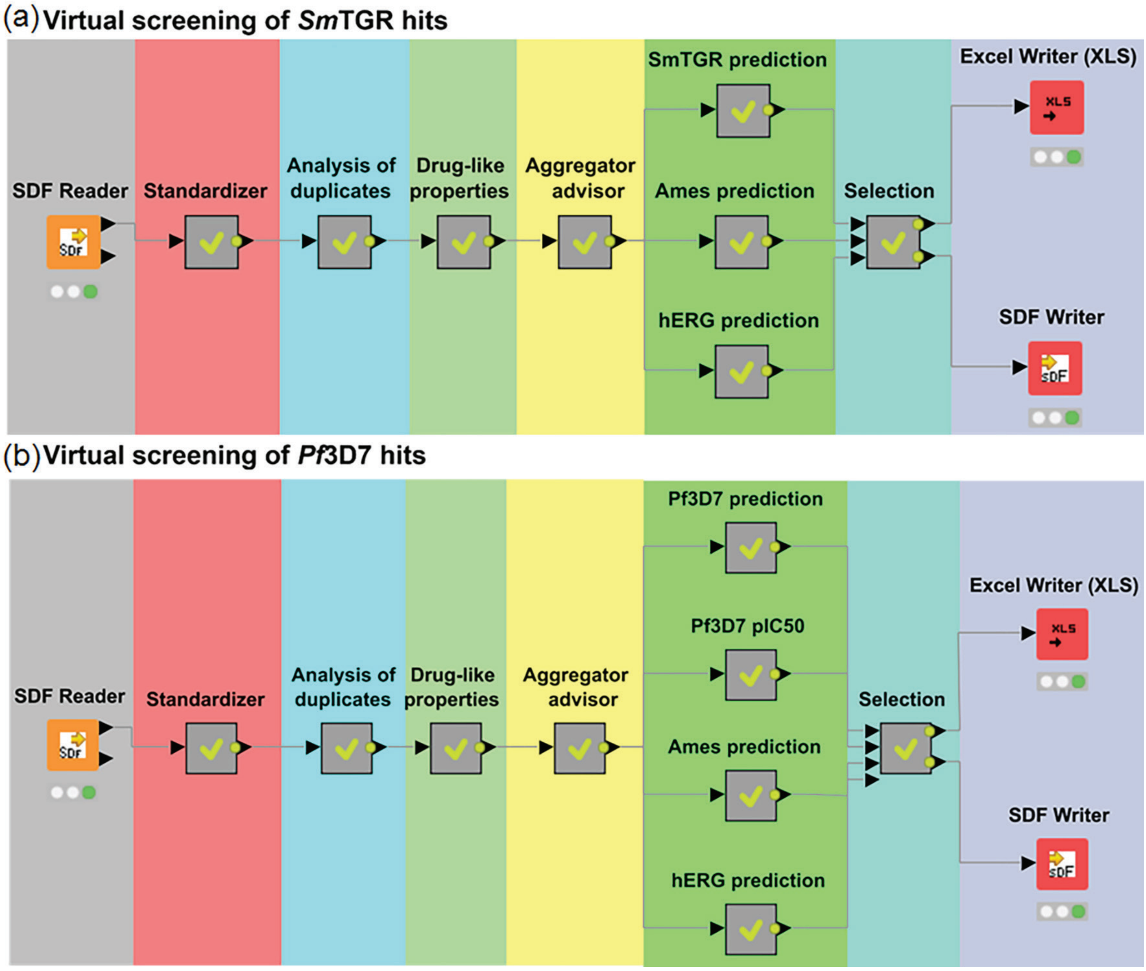


Figure 4. General workflows for the VS of new *Sm*TGR (a) and *Pf*3D7 (b) hits.

known experimental aggregators.[51] After these steps, 244,194 compounds were excluded.

ML filtering

To perform ML predictions, the "Model Reader" nodes must be loaded with output files from the most predictive models, while the fingerprints must be defined in the "RDKit fingerprints" nodes using the same modeling parameters. In parallel, the "SDF Reader" nodes must be loaded with corresponding curated datasets in SDF format to estimate applicability domains. Finally, the most promising hit compounds appearing at the top of the VS list can be exported through the "SDF Writer" and "Excel Writer (XLS)" nodes. The user can set their own hit criteria configuring the meta-node "Selection" after the ML predictions.

After data curation, we performed two independent VS on the compounds from the ChemBridge EXPRESS-Pick Collection using developed models. The first one (Figure 4a) applied the categorical models constructed with Avalon fingerprint for *Sm*TGR activity (probability to be active ≥ 0.7), Avalon fingerprint for Ames mutagenicity, and FeatMorgan fingerprint for hERG blocking. The second one (Figure 4b) applied the categorical model developed with Morgan fingerprint (probability to be active > 0.7) and the continuous model obtained from Morgan fingerprint (predicted $pIC_{50} \geq 6$) for *Pf*3D7 activity and the same categorical models for Ames mutagenicity and hERG blocking for *Sm*TGR endpoint. The Figure 5 shows the number of compounds prioritized in each step of the VS campaigns. The top three virtual hits for each target and their respective predictions are found in Table 3. The complete list of *Sm*TGR and *Pf*3D7 virtual hits and their respective predictions are listed in Supplementary Information, Files S2 and S3, respectively.

Conclusions

In this work, we developed an automated computational workflow for building robust and predictive QSAR models employing ML algorithms following the best practices for model development and validation. The worfklow

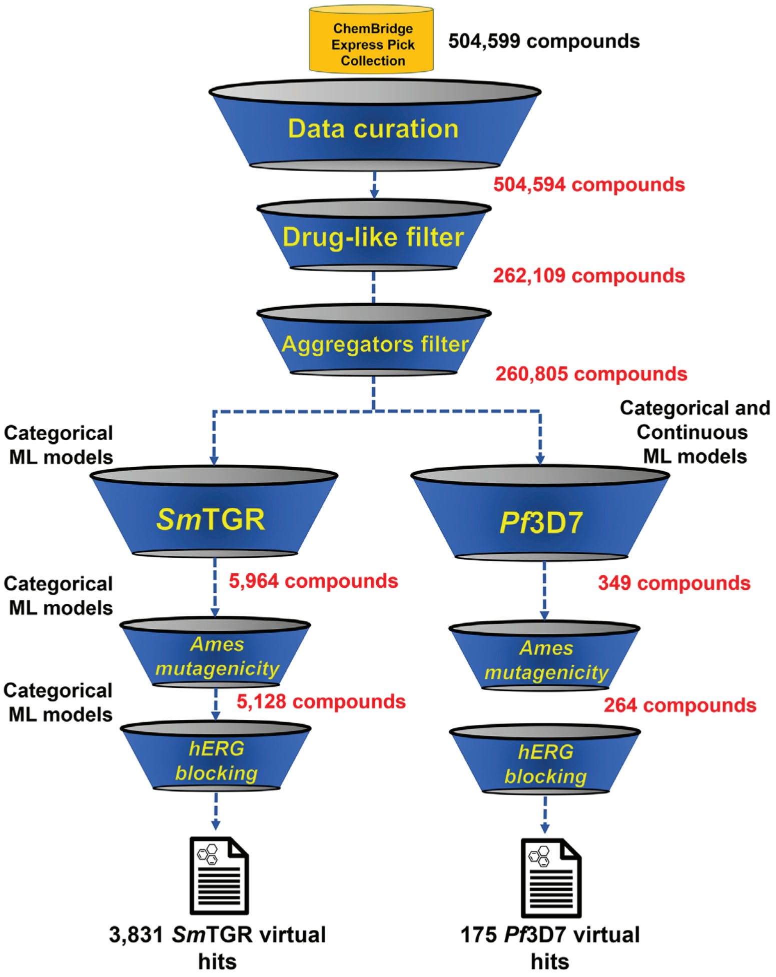
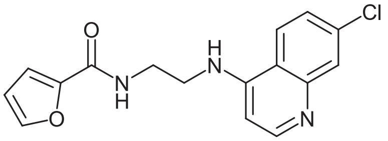
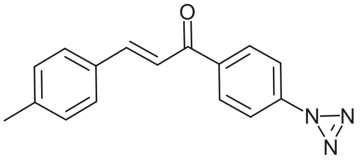
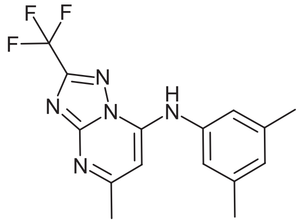
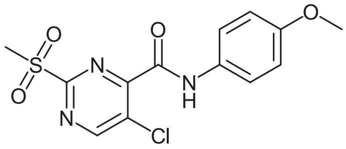
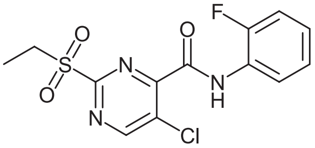
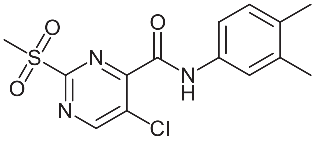


Figure 5. General scheme of case study VS filters and results of each step.

Table 3. Case study: top three virtual hits for each biological property and their predictions by ML model

Target	Compound	Predicted pIC_{50}	Confidence	Applicability domain	Ames mutagenicity	hERG inhibition
*Pf*3D7		7.13	high probability (96%)	reliable	nonmutagenic (–)	noncardiotoxic (–)
		6.66	good probability (88%)	reliable	nonmutagenic (–)	noncardiotoxic (–)
		7.99	good probability (88%)	reliable	nonmutagenic (–)	noncardiotoxic (–)
*Sm*TGR		–	high probability (99%)	reliable	nonmutagenic (–)	noncardiotoxic (–)
		–	high probability (99%)	reliable	nonmutagenic (–)	noncardiotoxic (–)
		–	high probability (99%)	reliable	nonmutagenic (–)	noncardiotoxic (–)

pIC_{50}: negative logarithm of the half maximal inhibitory concentration; Ames: mutagenicity; hERG: human ether-a-go-go-related gene; *Pf*3D7: *Plasmodium falciparum* 3D7 strain; *Sm*TGR: *Schistosoma mansoni* thioredoxin glutathione reductase.

has four modules, containing data curation, modeling, chemical space analysis (Supplementary Information), and VS. Moreover, we employed these workflows to curate, analyze, and model four datasets previously used by our group, as a benchmark. This workflow provides scientists a simple and guided open-source cheminformatics platform for development of QSAR models. This framework could be used as a validated starting point for beginner cheminformatics scientists to implement modeling in their laboratory routines following the best practices for building predictive models approved by the community. The workflows are freely available for download at GitHub[52] and in Supplementary Information (File S1).

Methodology

Architecture

The automated framework described herein was developed within KNIME 3.6.0[22,23] containing the RDKit,[53] Weka,[54] Enalos,[55] Chemistry Development Kit (CDK),[56] and the Indigo[57] nodes distributed as part of the 'community contributions'.[58]

General protocol

Initially, compounds with bioactivity data are retrieved from ChEMBL[9] and PubChem[10] database and imported separately into KNIME 3.6.0.[22,23] Subsequently, the potency values are converted to –log units, and an activity threshold is defined for discrimination between active and inactive compounds. All chemical structures and correspondent biological properties are carefully standardized and curated using Indigo nodes[57] and according to the protocols proposed by Fourches *et al*.[18-20] Then, curated datasets were balanced using a linear under-sampling approach[3] implemented in Enalos nodes.[55] At the end of dataset balancing, molecular fingerprints may be calculated for all chemical structures using RDKit nodes.[53] The binary and continuous ML models are developed using RF algorithm[13] implemented in Weka nodes.[54] Models will be validated using 5FECV procedure using "X-Partitioner" nodes. The AD was calculated using "Domain-Similarity" node implemented by Enalos.[55] ML models are fully compliant to best practices for predictive modeling,[16,20] and Organization for Economic Co-operation and Development (OECD) recommendations,[59] such as (*i*) a defined end point, (*ii*) an unambiguous algorithm, (*iii*) a defined domain of applicability, (*iv*) appropriate measures of goodness-of-fit, robustness, and predictivity, and (*v*) mechanistic interpretation, if possible. Finally, ML models were saved using "Weka Classifier Writer (3.7)" node or "Model Writer" node and implemented as filters in VS workflow to prioritize new compounds for further testing in experimental assay platforms. A previous version of this article has been published as preprint.[60]

Supplementary Information

Supplementary information (detailed description of structure-activity relationship workflow) is available free of charge at http://jbcs.sbq.org.br as PDF file. Source code of KNIME workflows (File S1) at .knar format, and complete list of *Sm*TGR and *Pf*3D7 virtual hits (Files S2 and S3, respectively) in PDF file are also available at http://jbcs.sbq.org.br.

Acknowledgments

The authors would like to thank Brazilian funding agencies, CNPq, FAPEG, FAPESP, and CAPES for financial support and fellowships. M. M. appreciates financial support of CNPq/FAPEG DCR (grant 300508/2017-4); E. N. M. thanks NIH (grants 1U01CA207160 and GM5105946). C. H. A. thanks CNPq (grant 400760/2014-2). C. H. A. is CNPq research productivity fellow. C. H. A. also thanks the "L'Oréal-UNESCO-ABC Para Mulheres na Ciência" and "L'Oréal-UNESCO International Rising Talents" for the awards and fellowships received, which partially funded this work. We are thankful to Tesia Bobrowski for her kind help with editing the language of the manuscript.

Author Contributions

Bruno J. Neves, Melina Mottin, Rodolpho C. Braga, Vinicius M. Alves, Carolina H. Andrade, and Eugene N. Muratov designed and supervised the study. José T. Moreira-Filho, Melina Mottin, Joyce V. V. B. Borba, and Arthur C. Silva developed the automated workflows under the guidance and support of Bruno J. Neves. The manuscript was written through contribution of all authors, under the guidance of Carolina H. Andrade. All authors read and approved the final manuscript.

References

1. Romanha, A. J.; de Castro, S. L.; Soeiro, M. N. C.; Lannes-Vieira, J.; Ribeiro, I.; Talvani, A.; Bourdin, B.; Blum, B.; Olivieri, B.; Zani, C.; Spadafora, C.; Chiari, E.; Chatelain, E.; Chaves, G.; Calzada, J. E.; Bustamante, J. M.; Freitas-Junior, L. H.; Romero, L. I.; Bahia, M. T.; Lotrowska, M.; Soares, M.; Andrade, S. G.; Armstrong, T.; Degrave, W.; Andrade, Z. A.; *Mem. Inst. Oswaldo Cruz* **2010**, *105*, 233.
2. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A.; *J. Med. Chem.* **2014**, *57*, 4977.
3. Neves, B. J.; Dantas, R. F.; Senger, M. R.; Melo-Filho, C. C.; Valente, W. C. G.; de Almeida, A. C. M.; Rezende-Neto, J. M.; Lima, E. F. C.; Paveley, R.; Furnham, N.; Muratov, E.; Kamentsky, L.; Carpenter, A. E.; Braga, R. C.; Silva-Junior, F. P.; Andrade, C. H.; *J. Med. Chem.* **2016**, *59*, 7075.
4. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B.; *Drug Discovery Today* **2018**, *23*, 1538.
5. Riniker, S.; Landrum, G. A.; *J. Cheminf.* **2013**, *5*, 26.
6. Neves, B. J.; Braga, R. C.; Alves, V. M.; Lima, M. N. N.; Cassiano, G. C.; Muratov, E. N.; Costa, F. T. M.; Andrade, C. H.; *PLoS Comput. Biol.* **2020**, *16*, e1007025.
7. Alves, V. M.; Braga, R. C.; Muratov, E.; Andrade, C. H.; *J. Braz. Chem. Soc.* **2018**, *29*, 982.
8. Neves, B. J.; Agnes, J. P.; Gomes, M. N.; Donza, M. R. H.; Gonçalves, R. M.; Delgobo, M.; de Souza Neto, L. R.; Senger, M. R.; Silva-Junior, F. P.; Ferreira, S. B.; Zanotto-Filho, A.; Andrade, C. H.; *Eur. J. Med. Chem.* **2020**, *189*, 111981.
9. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P.; *Nucleic Acids Res.* **2012**, *40*, D1100.

10. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annual Reports in Computational Chemistry*, vol. 4; American Chemical Society: Washington, D.C., 2008, p. 217-241.
11. Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K.; *Chem. Res. Toxicol.* **2014**, *27*, 1643.
12. Singh, S.; Carpenter, A. E.; Genovesio, A.; *J. Biomol. Screening* **2014**, *19*, 640.
13. Breiman, L. E. O.; *Mach. Learn.* **2001**, *45*, 5.
14. Vapnik, V. V.; *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
15. Schmidhuber, J.; *Neural Networks* **2015**, *61*, 85.
16. Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M.; *Nat. Mater.* **2019**, *18*, 435.
17. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T.; *Drug Discovery Today* **2018**, *23*, 1241.
18. Fourches, D.; Muratov, E.; Tropsha, A.; *J. Chem. Inf. Model.* **2016**, *56*, 1243.
19. Fourches, D.; Muratov, E.; Tropsha, A.; *Nat. Chem. Biol.* **2015**, *11*, 535.
20. Fourches, D.; Muratov, E.; Tropsha, A.; *J. Chem. Inf. Model.* **2010**, *50*, 1189.
21. Tropsha, A.; *Mol. Inf.* **2010**, *29*, 476.
22. Aguiar-Pulido, V.; Gestal, M.; Cruz-Monteagudo, M.; Rabuñal, J. R.; Dorado, J.; Munteanu, C. R.; *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 206.
23. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B.; *SIGKDD Explor.* **2009**, *11*, 26.
24. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *Data Analysis, Machine Learning and Applications*; Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., eds.; Springer: Berlin, Heidelberg, 2008, p. 319-326.
25. Fillbrunn, A.; Dietz, C.; Pfeuffer, J.; Rahn, R.; Landrum, G. A.; Berthold, M. R.; *J. Biotechnol.* **2017**, *261*, 149.
26. Kausar, S.; Falcao, A. O.; *J. Cheminf.* **2018**, *10*, 1.
27. Gadaleta, D.; Lombardo, A.; Toma, C.; Benfenati, E.; *J. Cheminf.* **2018**, *10*, 60.
28. Cox, R.; Green, D. V. S.; Luscombe, C. N.; Malcolm, N.; Pickett, S. D.; *J. Comput.-Aided Mol. Des.* **2013**, *27*, 321.
29. Warr, W. A.; *J. Comput.-Aided Mol. Des.* **2012**, *26*, 801.
30. Ambure, P.; Gajewicz, A.; Cordeiro, M. N. D. S.; Roy, K.; *J. Chem. Inf. Model.* **2019**, *59*, 4070.
31. Braga, R. C.; Alves, V. M.; Silva, A. C.; Nascimento, M. N.; Silva, F. C.; Liao, L. M.; Andrade, C. H.; *Curr. Top. Med. Chem.* **2014**, *14*, 1899.
32. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H.; *Nucleic Acids Res.* **2012**, *40*, D400.
33. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R.; *Nucleic Acids Res.* **2017**, *45*, D945.
34. https://www.ebi.ac.uk/chembl/target/inspect/CHEMBL2366922, accessed in July 2020.
35. https://pubchem.ncbi.nlm.nih.gov/bioassay/485364, accessed in July 2020.
36. https://www.ebi.ac.uk/chembl/target/inspect/CHEMBL240, accessed in July 2020.
37. Sushko, I.; Novotarskyi, S.; Ko, R.; Pandey, A. K.; Cherkasov, A.; Liu, H.; Yao, X.; Tomas, O.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V.; *J. Chem. Inf. Model.* **2010**, *50*, 2094.
38. Katsuno, K.; Burrows, J. N.; Duncan, K.; van Huijsduijnen, R. H.; Kaneko, T.; Kita, K.; Mowbray, C. E.; Schmatz, D.; Warner, P.; Slingsby, B. T.; *Nat. Rev. Drug Discovery* **2015**, *14*, 751.
39. Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C.; *J. Chem. Inf. Model.* **2014**, *54*, 705.
40. Altman, N.; *Am. Stat.* **1992**, *46*, 175.
41. Consonni, V.; Todeschini, R. In *Recent Advances in QSAR Studies - Methods and Applications*; Puzyn, T.; Leszczynski, J.; Cronin, M. T. D., eds.; Springer Netherlands: Dordrecht, 2010, p. 29-93.
42. Gedeck, P.; Rohde, B.; Bartels, C.; *J. Chem. Inf. Model.* **2006**, *46*, 1924.
43. Rogers, D.; Hahn, M.; *J. Chem. Inf. Model.* **2010**, *50*, 742.
44. Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O.; *Int. J. Quant. Struct.-Prop. Relat.* **2016**, *1*, 45.
45. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A.; *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241.
46. Braga, R. C.; Andrade, C. H.; *Curr. Top. Med. Chem.* **2013**, *13*, 1127.
47. Alves, V. M.; Golbraikh, A.; Capuzzi, S. J.; Liu, K.; Lam, W. I.; Korn, D. R.; Pozefsky, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A.; *J. Chem. Inf. Model.* **2018**, *58*, 1214.
48. Tanrikulu, Y.; Krüger, B.; Proschak, E.; *Drug Discovery Today* **2013**, *18*, 358.
49. https://www.chembridge.com/screening_libraries/, accessed in July 2020.
50. Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D.; *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308.
51. Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K.; *J. Med. Chem.* **2015**, *58*, 7076.
52. https://github.com/LabMolUFG/automated-qsar-framework, accessed in July 2020.

53. Riniker, S.; Landrum, G. A.; *J. Cheminf.* **2013**, *5*, 26.

54. Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H.; *Bioinformatics* **2004**, *20*, 2479.

55. Varsou, D.-D.; Nikolakopoulos, S.; Tsoumanis, A.; Melagraki, G.; Afantitis, A. In *Methods in Molecular Biology*, vol. 1824; Mavromoustakos, T.; Kellici, T., eds.; Humana Press: New York, NY, USA, 2018, p. 113-138.

56. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E.; *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493.

57. Pavlov, D.; Rybalkin, M.; *Indigo Toolkit*, 1.4.0; EPAM SolutionsHub, USA, 2017.

58. http://tech.knime.org/community/, accessed in July 2020.

59. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf, accessed in July 2020.

60. Neves, B. J.; Moreira-Filho, J. T.; Silva, A. C.; Borba, J. V. V. B.; Mottin, M.; Alves, V. M.; Braga, R. C.; Muratov, E. N.; Andrade, C. H.; *ChemRxiv*, 2020, DOI: 10.26434/chemrxiv.12250046.v1, available at https://chemrxiv.org/articles/preprint/Automated_Framework_for_Developing_Predictive_Machine_Learning_Models_for_Data-Driven_Drug_Discovery/12250046/1, accessed in July 2020.

Submitted: March 11, 2020
Published online: August 7, 2020