BJPS

Brazilian Journal of
Pharmaceutical Sciences

http://dx.doi.org/10.1590/s2175-97902023e22373

Article

3D-QSARpy: Combining variable selection strategies and machine learning techniques to build QSAR models

Priscilla Suene de Santana Nogueira Silverio[1], Jéssika de Oliveira Viana[1], Euzébio Guimarães Barbosa[1,2*]

[1]Post-graduate Program in Bioinformatic, Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal, RN, Brazil, [2]Post-graduate Program in Pharmaceutical Sciences, Faculty of Pharmacy, Federal University of Rio Grande do Norte, UFRN, Natal, Brazil

Quantitative Structure-Activity Relationship (QSAR) is a computer-aided technology in the field of medicinal chemistry that seeks to clarify the relationships between molecular structures and their biological activities. Such technologies allow for the acceleration of the development of new compounds by reducing the costs of drug design. This work presents 3D-QSARpy, a flexible, user-friendly and robust tool, freely available without registration, to support the generation of QSAR 3D models in an automated way. The user only needs to provide aligned molecular structures and the respective dependent variable. The current version was developed using Python with packages such as scikit-learn and includes various techniques of machine learning for regression. The diverse techniques employed by the tool is a differential compared to known methodologies, such as CoMFA and CoMSIA, because it expands the search space of possible solutions, and in this way increases the chances of obtaining relevant models. Additionally, approaches for select variables (dimension reduction) were implemented in the tool. To evaluate its potentials, experiments were carried out to compare results obtained from the proposed 3D-QSARpy tool with the results from already published works. The results demonstrated that 3D-QSARpy is extremely useful in the field due to its expressive results.

Keywords*: Drug Design. 3D-QSAR. Machine learning. Variable selection.*

INTRODUCTION

For many decades the design of drugs was characterized by the search for biologically active molecules against a disease of interest (Kubinyi, Hamprecht, Mietzner, 1998). The complexity in developing new medicines is one of the most difficult processes in the pharmaceutical industry, mobilizing millions of dollars (Ooms, 2012). For a drug to reach patients, usually more than 8 years and millions of dollars in investments are needed to complete the long and tedious process of drug development (Huang *et al.*, 2010). Although these procedures were often based on rational concepts, usually they were based on trial and error.

The large amount of existing molecular information along with the presence of computational models that aid in the processing and analysis of the wide variety of data have also emerged. In this way, the use of these computational tools has demonstrated support for academia and the pharmaceutical industry in the design and development of new drugs and targets (Wilson, Lil, 2011).

Based on this, CADD (Computer-Aided Drug Design) is divided into two main classifications: Structure-Based Drug Design (SBDD) and Ligand-Based Drug Design (LBDD). SBDD is applied when 3D structural information of the molecular target is used, simulating intermolecular interactions. Extensive examples, molecular docking and dynamics have been applied in recent studies. In contrast,

***Correspondence:** E. G. Barbosa. Programa de Pós-Graduação de Ciências Farmacêuticas. Faculdade de Farmácia. Universidade Federal do Rio Grande do Norte, UFRN. Gen. Gustavo Cordeiro de Faria, s/n. 59012-570, Petrópolis, Natal, Brasil. E-mail: euzebiogb@gmail.com. ORCID: https://orcid.org/0000-0002-7685-9618

LBDD is based on investigations of small ligands with known activity, extracting information about their molecular characteristics, descriptors, and producing predictive mathematical models in pharmacophores in Quantitative Structure-Activity Relationship (QSAR) studies (Zhao *et al*., 2020).

These computational simulations and protocols have proved to be essential in early drug discovery. This scenario emerged with the classical techniques of QSAR models decades ago, such as the Hansch analysis (Hansch, 1969). Subsequently, recent advances to accelerate drug design were the automation and the improvement of these processes aided by computational chemistry methods, saving time and money (Yu, Mackerell, 2017; Rahman *et al*., 2012).

QSAR is a valuable tool for the planning of new drugs, since it decreases the number of compounds to be synthesized, facilitates the selection of more promising candidates, and saves time and financial resources. The reasons why QSAR has become a useful alternative, among other issues, is that many compounds are available due to combinatorial chemistry and HTS (High-Throughput Screening) approaches, but estimates are required for the prioritization of synthesis and screening (Verma, Khedkar, Coutinho, 2010).

Currently, there are several approaches to construct QSAR models, such as 2D-QSAR (Freitas, Brown, Martins, 2005; Casañola *et al*., 2018), 3D-QSAR (Tosco, Balle, 2011; Cramer; Patterson; Bunce, 1988; Klebe; Abraham, 1998) and 4D-QSAR (Martins *et al*., 2009), all of which differ from one another in how they represent molecular structures and define their descriptors. For example, COMFA and COMSIA take their approach from the structure of molecules correlated with their activities, representing the ligand molecules through steric, electrostatic, hydrophobic fields and hydrogen bond donor or acceptor properties. Nevertheless, all of them have the same purpose of elucidating the relationship between structure and biological response and low efficacy. The exhaustive conformational treatment of the success of such methodologies has encouraged us to develop a widely accessible QSAR method.

One of the ways to increase the existing approaches in the construction of new QSAR models is to efficiently explore the different techniques of machine learning (ML) (Devinyak, Lesyk, 2016), including the methods of variable selection (Li *et al*., 2017; Ghasemi *et al*., 2017; Jesus, Canuto, Araujo, 2018; Jesus, Canuto, Araujo, 2019) that have been researched recently. ML has been envisioned as an indispensable tool in facilitating fast, affordable, and reliable assessments by generating high-accuracy QSAR models. Studies show that ML-generated models can reduce the number of synthesized compounds, reducing time-consuming modeling and computational cost in drug discovery and development (Kausar, Falcao, 2018).

Studies show that the combination of multiple algorithms from different categories can further improve predictions and indicate the best algorithm for the evaluated dataset (Wu *et al*., 2021). The use of several ML algorithms, such as Artificial Neural Networks (ANN), Random Forest (RF) and Support Vector Machines (SVM), can show strong predictive power in QSAR modeling, being superior to other algorithms. However, in the literature there is still no availability of a 3D QSAR program with automated ML algorithms, which is free, easy to implement, and which can produce highly stable predictions.

In summary, our QSAR tool is able to create QSAR 3D models in an automated way, with a user-friendly, robust and freely available method of combining variable selection strategies and machine learning techniques. We observe that our methodology, when compared to CoMFA, had a superior performance in predictability of activity for molecules that were not used to build the model. Additionally, we offer fully accessible, modifiable and customizable hyperparameter tuning.

MATERIAL AND METHODS

Overview of the 3D-QSARpy tool

The 3D-QSARPy tool was developed using already available technologies. Python (Van Rossum, Drake, 2009) was the main programming language used to develop this tool, together with packages such as numpy, scipy library, pandas, matplotlib and sklearn.

In the current version of 3D-QSARPy, the user must upload a file containing the molecule data previously

aligned by other software. It is important to say that the quality of the alignment can be crucial to obtain accurate results and this alignment is the user's responsibility. After uploading this file, there is a preprocessing procedure to verify user's mistakes and to help identify the necessary changes to obtain a correct format before the main processing. In some cases, the user will receive an email alert to fix the problems and submit the form again.

Computing descriptors

From the input data, the molecular interaction field descriptors and the input matrix of the methods are defined (Cramer, Patterson, Bunce, 1988). A series of molecules (binders) are used which are already aligned and centered in a grid box. Subsequently, the energy interactions are calculated between the binders and a sp^3 charged +1 probe positioned at points distributed evenly along the grid.

Molecular interaction fields (MIF) were constructed to describe molecular interactions of a pharmaceutical nature (GoodFord, 1985). This describes the spatial variation of the interaction energy between a molecular target and a probe chosen within a GRID (Grisoni, Consonni, Todeschini, 2018), using the total energy of interaction between a target molecule and a probe. In this way, distinct characteristics are derived from the molecules. MIFs can be used to identify similar pharmacophore ligands, predict bioactive alignments, and derive 3D-QSAR models to predict binding affinity (Milletti *et al*., 2007; Cross, Cruciani, 2010).

Four types of descriptors were implemented to compare molecules in terms of either their similarity or diversity to the binding groups. These descriptors were chosen because they better reproduce an interaction with a biological receptor, when compared to the intrinsic information of the molecule. The descriptor count and type can be easily personalized by any user. In this process, the changes in the interaction properties through the potential energy at grid points are calculated: Lennard-Jones potential (LJ) and Coulomb's law (QQ) represent, respectively, steric (van der Waals intermolecular interactions) and electrostatic interactions (Cramer, Patterson, Bunce, 1988). Hydrogen-Bond (HB) and Hydrophobic-Bond (HF) represent, respectively, hydrogen bond and hydrophobic interacting moieties within the probe (Klebe, Abraham, 1999). Table I shows the described descriptors.

TABLE I - 3D descriptors calculated for the construction of models

Descriptor	Formula
LJ (Lennard-Jones)	$4\varepsilon\left(\frac{\sigma}{r^{12}}-\frac{\sigma}{r^{6}}\right)$
QQ (Coulomb electrostatic interactions)	$\frac{q_i q_j}{4\pi\rho r^2}$
HB (Hydrogen bond forming atoms)	$Ae^{-10(r-1)^2}$
HF (Hydrophobic atoms)	$Be^{-10(r-1)^2}$

The parameters ε and σ are the types of combined atoms derived from the MMFF94 force field, ρ middle dielectric constant (ρ = 1), *r* is the distance between the interacting atoms. A=1 for interaction with O and N. B=0 for atoms of O and N.

Building the QSAR-3D model

Once the descriptors were calculated and the matrix of descriptors is ready, prior to applying the machine learning techniques to build the QSAR-3D models, some strategies to reduce the data dimensionality are conducted. In this way we can seek a good performance from the induced model, reduce computational cost and improve the understanding of the data (Carvalho *et al*., 2011).

The approach to reduce the dimensionality of data using variable selection was specifically defined for this tool. It is a combination of different types of strategies combining selection by filter-based subsets and wrapper strategies. In the first step, the dimension reduction was performed with a variable selection strategy by filter-based subsets using three filters. The first and second filters are applied to each of the four types of descriptor matrix (QQ, LJ, HB and HF) separately. There are cutoff values by default defined in the input form for the first

and second filters, however, the user can change these values according to necessity. Before the third filter, the four matrices of descriptors are joined in a single matrix.

The first filter is a cutoff criterion based on variance. This filter aims to eliminate distant points (descriptors) from the aligned molecules because these points are not relevant for representing the interaction in the models. The second filter is a cutoff criterion based on Pearson's correlation between descriptors. This filter keeps the most correlated descriptors (independent variables) with the biological activity "**y**" (dependent variable). The variables with high correlation between them do not present different information to contribute to the models. So, when a high correlation is identified between two or more independent variables, only the most correlated to "**y**" is included in the new subset of variables selected. And finally, the third filter is another cutoff criterion based on variance applied to a single matrix of descriptors with all the descriptors remaining after the application of the first and the second filters. This third filter eliminates the descriptors (variables) which presents very small variance and that still remain in the combined matrix of descriptors. After these three filters, variable selection using wrapper is performed with strategies such as mutual info regression and the random forest algorithm. In this second step, the user must select one of the wrapper options in the form to reduce even more the number of descriptors to obtain more understandable models.

After dimensionality reduction of the data, the following different machine learning techniques were applied with variations of parameter settings to build the models, such as: linear regression methods, k-nearest neighbors, regression trees, naive bayesians, and support vector machine for regression.

The user does not have the option to choose the algorithms to be performed because the strategy here is to perform all the possibilities employed by the tool to present the best results for the users. This is one of the most relevant differentials of this tool, Partial Least Squares is the most frequent algorithm to build QSAR models. The use of several machine learning techniques from different learning paradigms, allows models to be built in different ways and, thus, increases the search for possible solutions. One of the main factors that can affect the performance of the machine learning process is diversity. This ensures that the training data can provide more discriminative information for the model (unique or complementary information) (Wu *et al*., 2021).

For each employed algorithm, models are built with different quantities of selected variables, which results in many models. Among these models, the best models will be returned to the user, as previously defined by the user in an input form. Decision trees, for example, learn through symbolic representations. Naive bayes (NB), as well as linear methods, are statistical methods. K-nearest neighbors (KNN) is a method based on examples, whereas Support Vector Regression (SVR) are techniques that are considered connectionist (Rezende, 2005). This diversity of ML techniques used is considered a key factor in increasing the chances of obtaining models with good predictive capacity (Nascimento *et al*., 2018). Besides this, the diversity in parameters of each model is another way to achieve good models (Wu *et al*., 2021).

Each experiment to build a model is performed with 10-fold-cross-validation and the R^2 score is the metric used to measure the results obtained by each model. Additionally, we calculate the Concordance Correlation Coefficient (CCC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as recently proposed by Gramatica, Sangion (2016). The best models to be applied are selected using test set results.

Datasets and comparison of model performance

Experiments were performed with data from a series of molecules that had already been published by other authors using well consolidated tools. These results are compared to the results obtained with the 3D-QSARpy tool (Patel, Ghate, 2015; Karki *et al*., 2016).

To accomplish the comparison with results already published, 3D-QSARpy was applied following the same partitioning of the molecules used for training and molecules for testing. The results were recorded after the application of the module "3D-QSARpyGrid" filters with a significant number of descriptors and after the execution of the different variable selection strategies.

The first study validated the tool by building QSAR 3D models whose series of input inhibitory molecules targets the enzyme dipeptidyl peptidase-4 involved in the treatment of type 2 diabetes mellitus. QSAR 3D models were built with a series of 36 quinolines and isoquinolines (Patel, Ghate. 2015). The second study validated the tool by building QSAR 3D models whose series of forty-five input inhibitory molecules were 2-phenol-4-aryl-6-chlorophenyl pyridine compounds, which were synthesized and evaluated for cytotoxicity against four cancer cell lines (DU145, HCT15, T47D, and HeLa), targeting topoisomerase I and II (Karki *et al*., 2016).

A summary of the information contained in these articles can be viewed in Table II.

TABLE II - Summary of data from the Patel and Ghatel (2015) and Karki *et al*. (2016) papers

Description	Patel and Ghatel (2015)	Karki *et al*. (2016)
Total molecules in series	36	45
Series of molecules	quinolines and isoquinolines	2-fenol-4-aril-6 with clorofenil piridina
Target	0	0
pIC50	4.060 - 8.744	4.318 - 6.051
Tools to build the models	0	0
Q2	0.803 and 0.826	0.820
R2	0	0.915
R2pred	0	0.985

RESULTS AND DISCUSSION

Interface

The 3D-QSARPy is a user-friendly tool, freely available without registration as an offline tool by users. The user guide is available in the supplementary material. Currently, the 3D-QSARpy tool is subdivided into three main modules, according to the scheme shown in Figure 1.

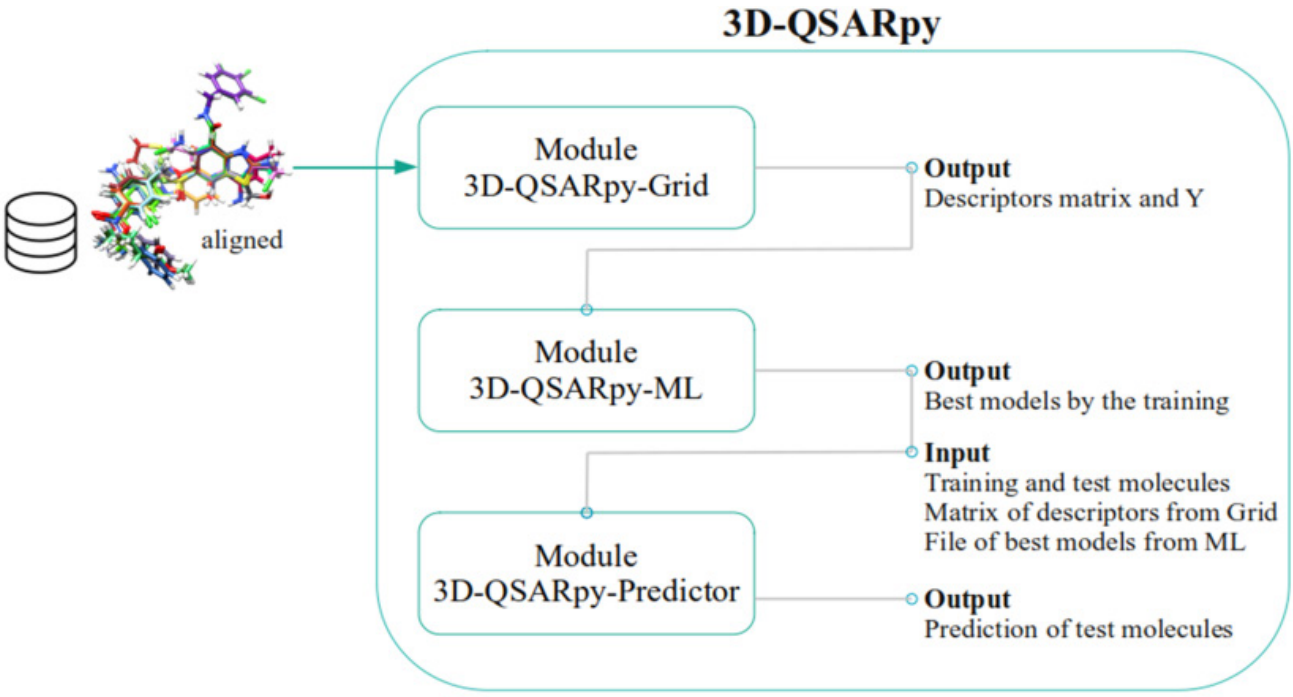

FIGURE 1 - 3D-QSARpy tool scheme.

The first module is the "3D-QSARpy-Grid" which generates the descriptor matrix and a biological activity (y) file. This module receives as input a file containing the series of input molecules and configuration information. It performs the necessary processing to achieve the data that will be used as the input for the tool, such as the coordinates and values of the charges from the series of molecules. Subsequently, when the necessary data from the file is extracted, the smallest and largest x, y and z coordinates are identified, and the values will be used to define the points that will compose the grid that encompasses the whole series of molecules. Finally, filters are applied to reduce data dimensionality.

From the minimum and maximum values of the grid coordinates, the distance from the grid edges, ranging from small, medium and large, is defined. The values of these margins correspond to, respectively, 3Å, 4Å and 5Å, and they are added to the extreme points of the boxes (minimum and maximum coordinates). The internal spacing between box points (resolution) must also be user defined and this is called the box resolution. Both measurements, edge spacing and internal spacing between points, directly influence the amount of initially calculated descriptors.

Moreover, to define which types of descriptors (LJ, QQ, HB and HF) should be calculated, the user must select True or False for each of them, except the LJ descriptor, which is always calculated. The dielectric constant value can also be changed and the calculation thresholds for variance and correlation applied to the descriptors by filters for dimensionality reduction as well. Finally, the users must provide their email to receive the result files after the processing is concluded and also receive alerts about the processing.

The second module is the training module, "3D-QSARpy-ML", which generates an output file with the information about the best models obtained from the training phase. This module receives as input the descriptors matrix and the y file obtained from the "3D-QSARpy-Grid". This matrix is reduced by variable selection strategies. Subsequently, several models are built based on different machine learning techniques. The best models built by this module are selected and the required data goes to the next module. The models are evaluated using the R^2 metric with partitioning at 10 times 10-fold-cross-validation. These data are stored for validation and testing in the subsequent module. Thus, it

is possible to use the chosen models for further validation and testing with new molecules (external set).

Variable selection is performed using three different types of strategies already implemented and imported from Scikit-Learn, these are: SelectKBest with mutual_info_regression method, SelectKBest with f_regression method and the Random Forest Learning algorithm. With all the strategies cited, models are generated with the number of attributes selected from each one until reaching the number of molecules in the input series.

In addition, all of these variable selection possibilities are used to build models based on the following algorithms: linear regression (Linear Regression, Lasso, Lasso Lars, Ridge, Bayesian Ridge), regression trees (Decision Tree Regressor, Extra Trees Regressor, Random Forest Regressor), random forest (RF), k-nearest neighbors for regression (KNN), and regression support vector machines (SVR).

Finally, the third module is the "3D-QSAR-Predictor" which generates an output file with the information about the prediction of test molecules. This module receives as input the training molecules and their predictions, the test molecules (which were not used in any preprocessing or training steps), and their predictions when biological activities are known (validation). In addition, it also receives the file containing the data regarding the best models obtained from the "3D-QSARpy-ML" module.

Thus, this module receives as input four files, these are: a file containing the series of input molecules used for training, a file containing the series of input molecules used for test, a file containing a matrix of descriptors obtained from the "3D-QSARpy-Grid", and a file with the information of the best models obtained from the "3D-QSARpy-ML" module. In the cases in which the biological activity of these molecules is already known, the input from the data of the test molecules is used to validate the models. On the other hand, in the cases of molecules that still have unknown biological activity, this module simply predicts the biological activity of these new molecules.

In the "3DQSARpy-Predictor", for each of the executed models, the descriptors for the test molecules are calculated, specifically the descriptors selected for the generation of the model in question. Several descriptor selections are tested for the different types of algorithms in the previous module. Models receive different selections, which consequently generate a specific test matrix for each model.

The test matrix is calculated similarly to the matrix generated using the input molecules in the first module, the "3D-QSARpy-Grid", but instead of calculating descriptors for all points on a grid, only specific descriptors resulting from variable selections are calculated for the test molecules. Once the test matrix is calculated for a given model, it is tested and the R^2 metric is used again to measure this prediction.

What is more, the descriptor matrix makes it possible to create a 3D representation of the generated model, highlighting the most interesting physicochemical characteristics for the compounds studied (Figure 2).

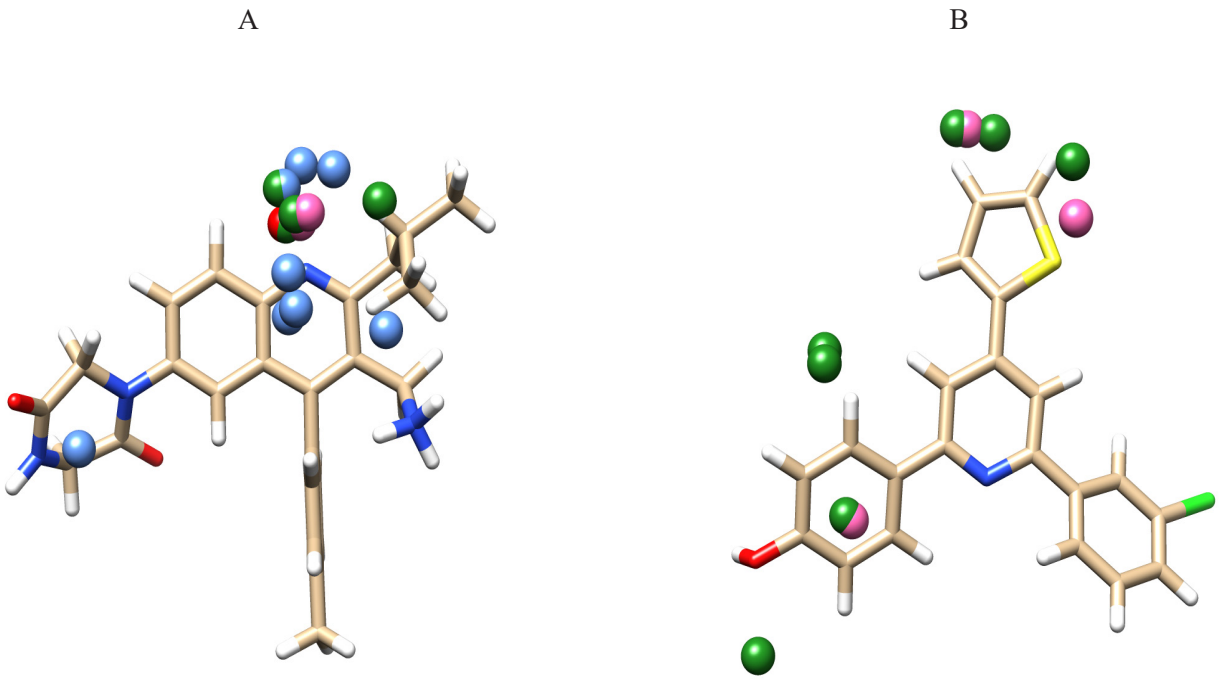

FIGURE 2 - 3D view of model representation, showing the best compound of Patel and Ghatel (2015) (compound A) and Karki *et al*. (2016) (compound B), and coordinated data of descriptors in space. The blue color represents the Hydrogen-Bond (HB), the green represents Lennard-Jones potential (LJ), the pink represents Hydrophobic-Bond (HF) and red represents Coulomb's law (QQ).

Performance

The preliminary results obtained had the objective of validating the 3D-QSARpy tool. Therefore, Table III shows some of the published results obtained using CoMFA and CoMSIA and some of the best results obtained by the present 3D-QSARpy tool.

According to information from the paper by Patel and Ghatel (2015), the final goal of treating type 2 diabetes mellitus is to control blood glucose levels for a longer time and treat its complications. Currently oral treatment options are: sulfonylurea derivatives (SU), metformin, thiazolidinedione (TZD), glycosidase inhibitors and more recently, dipeptidyl peptidase-4 (DPP-4) inhibitors and glucagon-like peptide-1 analogues (GLP-1).

In Patel and Ghatel (2015) it is possible to compare the predictive capacity of the models through the values presented in the last line. One of the models presented in Table II, which was built with the 3D-QSARpy tool, showed significantly higher prediction results compared to CoMFA and CoMSIA. This was the model using variable selection with the random forest strategy, presenting 0.918.

TABLE III - Summary table of results obtained from the publications of Patel and Ghatel (2015) and Karki *et al*. (2016) with the proposed 3D-QSARpy tool

	R2	Q2	R2pred	CCC	RMSE	MAE
Patel and Ghatel (2015)						
CoMFA	0.991	0.803	0.874	-	-	-
CoMSIA	0.983	0.826	0.847	-	-	-
3D-QSARpy (no selection)	1	0.689	0.889	-	-	-
3D-QSARpy(f_regression)	1	0.816	0.866	-	-	-
3D-QSARpy (mutual_regression)	0.967	0.698	0.853	-	-	-
3D-QSARpy (RF)	0.914	0.764	0.918	0.96	1.33	1.07
Karki *et al*. (2016)						
CoMFA	0.820	0.915	0.985	-	-	-
3D-QSARpy (no selection)	1	0.701	0.952	-	-	-
3D-QSARpy (f_regression)	1	0.683	0.979	-	-	-
3D-QSARpy (mutual_regression)	1	0.904	0.987	0.989	0.676	0.583
3D-QSARpy (RF)	1	0.682	0.979	-	-	-

It is also important to mention that the values of R^2 are generally lower when compared to the previously published models of CoMFA and CoMSIA, however, this can be easily justified in view of the divergence regarding training validation. The publication used leave-one-out, which means that only one molecule is withdrawn for validation, while experiments with 3D-QSARPy validation used the average of 10 times 10-fold-cross-validation, removing 10% of the molecules. Thus, it is expected that the results present lower values, however, the validation is considered more robust.

For CCC, RMSE and MAE metrics, we obtained representatively good values for the best model. However, this information was not included in the original paper reference and the comparison between our model and literature model was not possible.

The second study used for tool validation was a work by Karki et al. (2016), who built QSAR 3D models, and whose series of forty-five input inhibitory molecules, being 2-phenol-4-aryl-6-chlorophenyl pyridine compounds, were synthesized and evaluated for cytotoxicity against four cancer cell lines (DU145, HCT15, T47D, and HeLa), targeting topoisomerase I and II (Table III).

The main goal of this paper was to prevent the proliferation of cancer cells by inhibiting topoisomerase I and II. Human DNA topoisomerases (I and II) are expressed at different levels in different types of cancer. For example, topo I is over-expressed in colon cancer cell lines, while topo II is over-expressed in breast and ovarian cancer lines.

The authors propose a new series of 2-phenol-4-aryl-6-chlorophenyl pyridines synthesized as potential antitumor agents acting with dual inhibitory activity. For this purpose, chlorine was incorporated at the 6-phenyl position and aryl group at the 4-position of the central pyridine to investigate whether these changes cause any combined (double) effect on activity and cytotoxicity.

It was observed for Karki *et al*. (2016) that one of the models presented in Table III, which was built with

the 3D-QSARpy tool, showed higher prediction results compared to CoMFA. This model presented variable selection with the mutual regression strategy, presenting a prediction of 0.987338, while CoMFA achieved 0.985. The other models presented similar values but could not outperform the CoMFA.

CONCLUSION

The 3D-QSARpy is a friendly tool with three modules available and allows the user to build promising models with good predictive ability, surpassing other methodologies, such as CoMFA and CoMSIA. The validation of this tool achieved good results. The use of different machine learning techniques combined with variable selection strategies allowed for increased model diversity in the existing search space and the flexibility of this tool allows for possibilities for future enhancements. This tool enables the user to configure important parameters according to what is needed. What is more, it is designed to gradually include other algorithms and new configurations of parameters and the strategy of dimensional reduction to improving its performance regardless of the user's knowledge about the techniques employed. Thus, it demonstrates that this tool is robust and flexible and very useful for the drug discovery process.

ACKNOWLEDGMENT

This work was conducted during a postgraduate program at Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal and with the support of the Federal Institute of Education, Science and Technology of Rio Grande do Norte, which granted leave of absence from work activities for the completion of this postgraduate degree. This research was supported by NPAD/UFRN.

DECLARATION OF INTEREST STATEMENT

The authors report there are no conflicts of interest associated with this publication. The authors alone are responsible for the content. The manuscript has been read and approved by all named authors and the order of authors listed in the manuscript has been approved by all of us.

DECLARATIONS

Authors' contributions

Priscilla S. S. N. Silverio and Euzébio G. Barbosa contributed equally to the paper. Jéssika de O. Viana write, organized and reviewed successive versions.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 with reference number 88882.375448/2019-01.

Conflicts of interest/Competing interests

Our tool provides free services that are sold by some companies.

Availability of data and material

The 3D-QSARpy tool is available in an offline version archive that contains files with RAR compression. The program can be accessible through the Github link: https://github.com/viana1jess/3DQSARpy.git. The supplementary material contains the user guide, which is available to authorized users.

REFERENCES

Carvalho ACPLF, Faceli K, Lorena AC, Gama J. Inteligência Artificial: Uma abordagem de aprendizado de máquina. Rio de Janeiro: Editora LTC; 2011. p. 378.

Casañola-Martin GM, The HP, Garit JAC, Thu HLT. Atom based linear index descriptors in QSAR-machine learning classifiers for the prediction of ubiquitin-proteasome pathway activity. Med Chem Res. 2018;27(3):695-704.

Cramer RD, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on

Binding of Steroids to Carrier Proteins. J Amer Chem Soc. 1988;110(18):5959-67.

Cross S, Cruciani G. Molecular fields in drug discovery: getting old or reaching maturity? Drug Disc Today. 2010;15:23-32.

Devinyak O, Lesyk R. 5-Year Trends in QSAR and its Machine Learning Methods. Curr Comp Aided-Drug Design. 2016;12(4):265-71.

Freitas MP, Brown SD, Martins JA. MIA-QSAR: A simple 2D image-based approach for quantitative structure-activity relationship analysis. J Mol Struct. 2005;738(1):149-54.

Ghasemi F, Fassihi A, Pérez-Sánchez H, Dehnavi MA. The role of different sampling methods in improving biological activity prediction using deep belief network. J Comp Chem. 2017;38(4):195-203.

Goodford PJ. A computational procedure for determining energetically favorable binding sites on biological important macromolecules. J Med Chem. 1985;28:849-57.

Gramatica P, Sangion A. A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. J Chem Inf Model. 2016;56(6): 1127-31.

Grisoni F, Consonni V, Todeschini R. Impact of molecular descriptors on computational models. In: Brown J, editor. Computational Chemogenomics. Methods in Molecular Biology. New York: Humana Press; 2018; p. 171-209.

Hansch C. Quantitative approach to biochemical structure-activity relationships. Acc Chem Res. 1969;2(8):232-39.

Huang HJ, Yu HW, Chen CY, Hsu CH, Chen HY, Lee KJ, et al. Current developments of computer-aided drug design. J Taiwan Inst Chem Eng. 2010;41(6):623-35.

Jesus J, Canuto A, Araújo D. Dynamic Feature Selection Based on Pareto Front Optimization. In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018; Rio de Janeiro, p 1-1.

Jesus JKL, Canuto AMP, Araujo DSA. Investigating robustness and stability to noisy data of a dynamic feature selection method. In: Brazilian Conference on Intelligent Systems (BRACIS). 2019, Salvador. 8th Brazilian Conference on Intelligent Systems (BRACIS), p.180-5.

Karki R, Jun KY, Kadayat TM, Shin S, Magar TBT, Bist G, et al. A new series of 2-phenol-4-aryl-6-chlorophenyl pyridine derivatives as dual topoisomerase I/II inhibitors: Synthesis, biological evaluation and 3D-QSAR study. Eur J Med Chem. 2016;113:228-45.

Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1):1-23.

Klebe G, Abraham U. Comparative Molecular Similarity Index Analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Computer-Aided Mol Design. 1999;13(1):1-10.

Kubinyi H, Hamprecht FA, Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem. 1998;41(14):2553-64.

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature Selection. ACM Comp Surveys. 2017;50(6):1-45.

Martins JPA, Barbosa EG, Pasqualoto KF, Ferreira MM. LQTA-QSAR: A new 4D-QSAR methodology. J Chem Info Model. 2009;49(6):1428-36.

Milletti F, Storchi L, Sforna G, Cruciani G. New and Original pKa Prediction Method Using GRID Molecular Interaction Fields. J Chem Inf Model. 2007;47:2172-81.

Nascimento DS, Bandeira DR, Canuto AM, Ara D. Investigating the Impact of Diversity in Ensembles of Multi-label Classifiers. Proc Int Jt Conf Neural Netw. 2018;1:1-8.

Ooms F. Molecular Modeling and Computer Aided Drug Design. Examples of their Applications in Medicinal Chemistry. Curr Med Chem. 2012;7(2):141-58.

Patel BD, Ghate MD. 3D-QSAR studies of dipeptidyl peptidase-4 inhibitors using various alignment methods. Med Chem Res. 2015;24(3):1060-69.

Rahman MM, Karim MR, Ahsan MQ, Khalipha ABR, Chowdhury MR, Saifuzzaman M. Use of computer in drug design and drug discovery: A review. Inter J Pharm Life Sci. 2012;1(2):1-21.

Rezende SO. Sistemas Inteligentes: Fundamentos e Aplicações. São Paulo: Editora Manole; 2005. p. 525.

Tosco P, Balle T. Open3DQSAR: A new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. J Mol Model. 2011;17(1):201-8.

Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in Drug Design-A Review. Curr Top Med Chem. 2010;10(1):95-115.

Wilson GL, Lill MA. Integrating structure-based and ligand-based approaches for computational drug design. Future Med Chem 2011;3(6):735-50.

Wu Z, Zhu M, Kang Y, Leung ELH, Lei T, Shen C, et al. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform. 2021;22(4):1060-69.

Yu W, Mackerell AD. Computer-Aided Drug Design Methods. Meth Mol Biol. 2017;1520:85-106.

Zhao L, Ciallella HL, Aleksunes LM, Zhu H. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug Discov 2020;25(9):1624-38.

Received for publication on 10th May 2022
Accepted for publication on 07th November 2022