UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

A new method for ligand-based virtual screening using linear algebra

ORIENTADO: **Carmelina Figueiredo Vieira Leite**
ORIENTADOR: **Prof. Dr. Marcos Augusto dos Santos**
CO-ORIENTADORA: **Dra. Lucianna Helene Santos**

BELO HORIZONTE - MG
Novembro de 2020

A new method for ligand-based virtual screening using linear algebra

Tese apresentada ao programa
de Pós-Graduação em Bioinformatica,
do Instituto de Ciências Biológicas,
da Universidade Federal de Minas Gerais
como requisito parcial para a obtenção do grau
de Doutor em Bioinformática

ORIENTANDO: **Carmelina Figueiredo Vieira Leite**
ORIENTADOR: **Prof. Dr. Marcos Augusto dos Santos**
CO-ORIENTADORA: **Dra. Lucianna Helene Santos**

043 Leite, Carmelina Figueiredo Vieira.
A new method for ligand-based virtual screening using linear algebra [manuscrito] / Carmelina Figueiredo Vieira Leite. – 2020.
152 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Marcos Augusto dos Santos. Coorientadora: Dra. Lucianna Helene Santos.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. Algoritmos. 3. Álgebra linear. 4. Modelos Logísticos. 5. Aprendizado de Máquina. 6. Medicamentos de Referência. 7. Reposicionamento de Medicamentos. I. Santos, Marcos Augusto dos. II. Santos, Lucianna Helene. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Rosilene Moreira Coelho de Sá – CRB 6 - 2726

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-graduação em Bioinformática

ATA DA DEFESA DE TESE

CARMELINA FIGUEIREDO VIEIRA LEITE

Às dez horas do dia **27 de novembro de 2020**, reuniu-se, no aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho de **Carmelina Figueiredo Vieira Leite,** intitulado: **"A new method for ligand-based virtual screening using linear algebra"**, requisito para obtenção do grau de Doutora em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Marcos Augusto dos Santos**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**Indicação**
Dr. Marcos Augusto dos Santos	UFMG	Aprovada
Dra. Lucianna Helene Silva dos Santos	UFMG	Aprovada
Dr. José Miguel Ortega	UFMG	Aprovada
Dr. Anderson Rodrigues dos Santos	UFU	Aprovada
Dr. Bráulio Roberto Gonçalves Marinho Couto	UniBH	Aprovada
Dr. Carlos Ernesto Ferreira Starling	Hospital Life Center	Aprovada

Pelas indicações, a candidata foi considerada: **Aprovada**

O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 27 de novembro de 2020.

Documento assinado eletronicamente por **Lucianna Helene Silva dos Santos**, **Usuário Externo**, em 27/11/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bráulio Roberto Gonçalves Marinho Couto**, **Usuário Externo**, em 27/11/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 27/11/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Anderson Rodrigues dos Santos**, **Usuário Externo**, em 27/11/2020, às 13:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carlos Ernesto Ferreira Starling**, **Usuário Externo**, em 27/11/2020, às 13:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcos Augusto dos Santos**, **Membro de comissão**, em 27/11/2020, às 13:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0441047** e o código CRC **F34FEFB3**.

Referência: Processo nº 23072.241537/2020-31 SEI nº 0441047

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

PARECER Nº **21/2020**
PROCESSO Nº 23072.241537/2020-31

FOLHA DE APROVAÇÃO

Carmelina Figueiredo Vieira Leite

"A new method for ligand-based virtual screening using linear algebra"

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Marcos Augusto dos Santos - Orientador
UFMG

Profa. Lucianna Helene Silva dos Santos - Coorientadora
UFMG

Prof. José Miguel Ortega
UFMG

Prof. Anderson Rodrigues dos Santos
UFU

Prof. Bráulio Roberto Gonçalves Marinho Couto
UniBH

Prof. Carlos Ernesto Ferreira Starling
Hospital Life Center

Belo Horizonte, 27 de novembro de 2020.

Documento assinado eletronicamente por **Lucianna Helene Silva dos Santos**, **Usuário Externo**, em 27/11/2020, às 13:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bráulio Roberto Gonçalves Marinho Couto**, **Usuário**

Externo, em 27/11/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 27/11/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Anderson Rodrigues dos Santos**, **Usuário Externo**, em 27/11/2020, às 13:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carlos Ernesto Ferreira Starling**, **Usuário Externo**, em 27/11/2020, às 13:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcos Augusto dos Santos**, **Membro de comissão**, em 27/11/2020, às 13:29, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0441113** e o código CRC **49FDA8C9**.

Referência: Processo nº 23072.241537/2020-31 SEI nº 0441113

To my parents, always.

ACKNOWLEDGMENT

To the Financial Assistance FUNDEP e CNPQ;

To my advisor, Professor Marcos Augusto dos Santos, for the support and friendship;

To Ph.D. Lucianna Helene Santos, for the availability;

To my colleagues of LBS, for the discussions and laughs;

To Manoel Felipe Santiago and his workers, for the confidence;

To my relatives and friends, for friendship and support;

To God, for the opportunity.

“and to knowledge, self-control; and to self-control, patience; and to patience, godliness.”

(2 Peter 1:6)

ABSTRACT

Ligand-based virtual screening of large molecular databases can help reduce costs with experiments by filtering and ranking promising compounds in an initial stage of the drug developing process. However, some ligand-based methods can be ineffective when presented with a high-dimensional number of attributes extracted from an extensive dataset of compounds. Herein, we propose a drug-mining algorithm that can screen ligands and repurpose known drugs from any dataset for any target. The Milk-Way algorithm combines mathematical and regression methods to select promising compounds from a high-dimensional and imbalanced dataset without massive computational power. The significant advantages of Milk-Way algorithm are non-recursive, and the utilization of more features than individuals in the same model. To validate the algorithm, we used literature data of known ligands and compared Milk-Way performance with the methods of Support Vector Machine (SVM) and Random Forest (RF). The chosen datasets of HIV-1 reverse transcriptase receptors showed that our algorithm had better AUC (Area under curve of Receiver Operating Characteristics Curve) than SVM and RF. We also worked with 17 targets from a different database to evaluate the new algorithm, which were consistent with previous, reaching the AUC=1.00. The feature selection done through the Milk-Way algorithm has been improved the values of AUC of itself but, also, the AUC of SVM, and Logistic Regression (LR). Moreover, a prospective screening targeting cyclin-dependent kinase type two (CDK-2) was carried out. The combined use of the algorithm metrics and molecular docking (DOCK6.8) suggested five promising drugs to be repositioned. Three were already mentioned as possible inhibitors of related diseases in the literature. In order to complementary my thesis with a structure-based virtual screening technique, I explored the vector space of protein targets of approved drugs. This strategy results in a suggestion of treatment to COVID-19, the tetrachlorodecaoxide. The product of this dissertation is the Milk-Way algorithm, and two others sub-products: a feature selection procedure and, a mathematical model of protein targets of approved drugs. These products resulted in two deposit patents, one paper published, and a draft of another.

KEYWORDS: ligand-based virtual screening; logistic regression; algorithm; machine learning; drug discovery; drug repurposing; vector space.

RESUMO

A triagem virtual baseada em ligantes de grandes bancos de dados moleculares pode auxiliar a reduzir custos com experimentos, ao filtrar e classificar compostos promissores numa etapa inicial do processo de desenvolvimento de novas drogas. No entanto, alguns métodos baseados em ligantes demonstram ser ineficazes quando aplicados em um grande número de características extraído de um diverso conjunto de dados. Propomos, nesta tese, um algoritmo de mineração de drogas que pode ser usado para selecionar ligantes e reposicionar fármacos já comercializados, a partir de qualquer conjunto de dados para qualquer alvo. O algoritmo Milk-Way combina métodos matemáticos e de regressão a fim de selecionar compostos promissores a partir de um elevado conjunto de dados e desbalanceado, sem o utilizar um grande infra-estrutura computacional. As principais vantagens do algoritmo Milk-Way são: não-recursivo e a utilização de mais características do que indivíduos no mesmo modelo. Para validar o algoritmo, utilizamos dados da literatura de ligantes conhecidos e comparamos o desempenho do Milk-Way com os métodos Máquina de Vetores de Suporte (SVM) e Floresta Aleatória (RF). Os conjuntos de dados escolhidos dos inibidores dos receptores de transcriptase reversa do HIV-1 mostraram que nosso algoritmo teve uma AUC (Área sob a Curva de Característica de Operação do Receptor) mais elevada, em relação ao SVM e RF. Também trabalhamos com 17 alvos de um banco de dados diferente para avaliar o novo algoritmo, que foram consistentes com o anterior, atingindo AUC=1.00. A seleção de características feita através do algoritmo Milk Way melhorou os valores de AUC do próprio, mas, também, a AUC de SVM e a Regressão Logística (LR). Além disso, foi realizada uma triagem virtual prospectiva utilizando a quinase dependente de ciclina tipo dois (CDK-2). O uso combinado das métricas do algoritmo e do atracamento molecular (DOCK6.8) sugeriu cinco fármacos promissores a serem reposicionados, das quais três já foram citados na literatura como possíveis inibidores da CDK-2. A fim de complementar a tese com uma técnica de triagem virtual baseada em estrutura, exploramos o espaço vetorial de alvos protéicos de fármacos aprovados. Esta última estratégia resultou em uma sugestão de tratamento para COVID-19, o tetraclorodecaoxido. O produto desta tese é o algoritmo Milk-Way e dois outros sub-produtos: um procedimento para a seleção de características e um modelo matemático de alvos protéicos de fármacos aprovados. Estes produtos resultaram em dois depósitos de patentes, uma publicação de artigo e um esboço de outro.

PALAVRAS-CHAVE: Triagem virtual baseada no ligante; regressão logística; algoritmo; aprendizado de máquina; desenvolvimento de novos medicamentos; reposicionamento; espaço vectorial.

LIST OF ABBREVIATIONS

AUC	Area under curve of Receiver Operating Characteristics Curve
CDK-2	cyclin-dependent kinase type 2
HTS	High-Throughput Screening
LR	Logistic Regression
LBVS	Ligand-Based Virtual Screening
MUV	Maximum Unbiased Validation
NB	Naïve Bayes
RF	Random Forest
SBVS	Structure-Based Virtual Screening
SVM	Support Vector Machine

LIST OF FIGURES

SUMMARY

1. Introduction

Computational approaches, such as virtual screening (VS), have emerged as alternatives to screen large libraries of small molecules in a cost-efficient manner. Although VS procedures do not substitute experimental assays, they can speed up and rationalize the process of drug discovery, enriching the number of hits since it can downsize the number of candidates to be tested (NEVES; BRAGA; MELO-FILHO; MOREIRA-FILHO *et al.*, 2018; SEIFERT; WOLF; VITT, 2003). In structure-based virtual screening (SBVS), the three-dimensional structure of the target is known from x-ray crystallographic, NMR, or computational modeling (LI; SHAH, 2017; STOCKWELL, 2004). SBVS usually involves the molecular docking methodology, which places and rank the compounds in a target binding site according to an algorithm that predicts their possible binding affinity. In the absence of three-dimensional structures of the targets, the molecular and chemical properties of known actives and tested compounds are gathered to create models of their binding using ligand-based virtual screening (LBVS) (GEPPERT; VOGT; BAJORATH, 2010). Some LBVS methodologies can assume that one or more actives share a binding mode. Thus, the screening will be done as a similarity or matching search to select potential new binders with similar chemical features to the known ones (CARPENTER; HUANG, 2018). Chemical compounds to compose screening libraries are available as a free resource of bioactivity data for small molecules in various databases such as ChEBI (DEGTYARENKO; DE MATOS; ENNIS; HASTINGS *et al.*, 2008), ZINC (IRWIN; SHOICHET, 2005), PubChem (WANG; BOLTON; DRACHEVA; KARAPETYAN *et al.*, 2010), DrugBank (KNOX; LAW; JEWISON; LIU *et al.*, 2011), IUPHAR-DB (SHARMAN; MPAMHANGA; SPEDDING; GERMAIN *et al.*, 2011), and KEGG (KANEHISA; GOTO; FURUMICHI; TANABE *et al.*, 2010). Alternatively, in-house compound datasets can also be created from previously tested compounds and analogs.

For a specific target, known actives can be employed as training data in classification methods. These methods use this information to separate a database of compounds with unknown activity into predicted actives and inactives (PLEWCZYNSKI; SPIESER; KOCH, 2006). Classification methods are usually machine learning approaches that build models, such as decision trees (HAN; WANG; BRYANT, 2008; RINIKER; WANG; JENKINS; LANDRUM, 2014), neural networks (CHEN; WILD; GUHA, 2009; PAOLINI; SHAPLAND; VAN HOORN; MASON *et*

al., 2006), and support vector machines (HAO; WANG; BRYANT, 2014), and can perform exceptionally well in enriching actives (PLEWCZYNSKI; SPIESER; KOCH, 2006). Established algorithms for data mining (Naive Bayes, SVM, Random Forrest, J48) are also used to classify chemical compounds (SCHIERZ, 2009). However, classification methods can demand extensive knowledge over the many methodologies and need high computing power. Furthermore, these methods might not perform well when subjected to imbalanced or high-dimensional data, *i.e.*, it can lead to an inadequate exploration of the ligands, and lack of accurate results, essential to a screening (DAI; GUO, 2019; TRUNK, 1979; YIN; GE; XIAO; WANG *et al.*, 2013). Therefore, the proposition of an improved *in silico* approach to classified chemical compounds is a relevant issue.

All those methods normally use molecular descriptors, which are a result of a logical and mathematical procedure where chemical information of a given molecule is transformed into numeric values represented by vectors (GRISONI; CONSONNI; TODESCHINI, 2018).

With the High-Throughput Screening (HTS), it is possible to screen millions of ligands against a target. However, it is an expensive process. Therefore, alternatives with lower cost for sorting several ligands as computational processes are sought out—this helps decrease the number of ligands for an experimental phase. In 1988, the Nobel Laureate James Black stated that "the most fruitful basis of the discovery of a new drug is to start with an old drug" (RAJU, 2000), indicating both the reality in the chemical design field and a prelude to the integrated vision of drug targets and their respective drugs. Drug repurposing is the intended strategy of discovering new uses or conditions for approved drugs and to not only assess the effects of the drug in a new target but also to reduce the cost of developing a new drug

Inspired by the diversity of methods, the present dissertation proposes combining the modified logistic regression method with linear algebra for classification compounds according to their possible activity to a specific target. In this dissertation, we offer a new approach that contributes to the way of classifying compounds to a chosen target, with metrics higher than 90%, named "Milk-Way algorithm (Mathematical Interpretation of the Logistic ranK, a WAY). It will help selecting possible hit drugs to be validated. The algorithm enables us to calculate and project the probability of each ligand (P(x), where x is the ligand), which is used to distinguish possible high performing ligands. The Milk-Way algorithm was validated using

literature data by performing studies using HIV-1 reverse transcriptase inhibitors. We also applied the Milk-Way algorithm to classify a challenge database, Maximum Unbiased Validation (MUV). Using the same literature data, we compared Milk-Way performance with other known algorithms: Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Logistic Regression (LR). Moreover, an application of repurposing commercially available drugs with the Milk-Way algorithm was conducted with the cyclin-dependent kinase type 2 (CDK-2) as a target to exemplify the algorithm's predictive power. To overcome the limitation of the method, which requires ligands to build it, we also proposed a technique for exploring vector space with protein projections. It allowed us to explore the neighbourhood of the chosen target, described through their signatures.

I would like to comment on the architecture of the work presented. I chose to show the results published and deposited of the method created during these four years. There are more results, but our partner, Coordenadoria de Transferência e Inovação Tecnológica, suggested not mention the products that are being negotiated. This department led the process of depositing the patents, disclosed the work (https://youtu.be/TRu3OQOiwkE), and is arranging possible purchasers. The first article and the first patent summarize the Milk-Way details, and it also had case studies. The draft of the case study with MUV is the validation of the ability to work with large and imbalanced databases, choosing the best features for a classification model. And, finally, I explored the vector space of therapeutic targets for a viable drug candidate to integrate the pharmacological treatment drug for COVID-19, resulting in another patent. It served to eliminate the lag of the Milk-Way algorithm, but above all, it is the duty of every scientist to fight the pandemic we are living in.

2. Hypothesis

Compounds can be classified about their activity to a specific target using data-mining techniques, as logistic regression, vector space representation, and linear algebra.

3. Objectives

3.1 General Objective

Build and validate an algorithm classified compounds, using modified logistic regression and targets with a therapeutic interest, to propose a possible new drug or a possibility repurposed.

3.2 Specific Objectives

1) Validate the Milk-Way algorithm with different types of case studies;

2) Use a benchmark study with Maximum Unbiased Validation (MUV) database to compare the performance of the Milk-Way algorithm with other known algorithms;

3) Use the Milk-Way algorithm to perform the stratified feature selection.

4) Obtain active and inactive compounds using the database PubChem and DrugBank;

5) Describe compounds through different fingerprints;

6) Propose a strategy to explore the vector space, using only protein therapeutic targets.

4. Contributions: Patents and Papers

The Milk-Way algorithm had the three requirements for a patent - new, inventive activity, and industrial application - it was possible to register it (BR 10 2019 027703 3). As this text had to be a self-explain to a specialist in the area and, I put it before the paper published (DOI is 10.31300/TDB.13.2020.1-20). The first document also had a case study of reverse transcriptase. While in the second, CDK-2.

Another paper, which is finished, and under the last considerations of the co-authors, is "Stratified feature selection with Milk-Way algorithm applied in the MUV database: a study case". The aim of this paper was the consolidation of the performance of the Milk-Way algorithm and extended exploration of the stratified feature selection done by it. Riniker and Landrum (2013) used the three databases: the directory of useful decoys (DUD) (IRWIN, 2008), ChEMBL (GAULTON; BELLIS; BENTO; CHAMBERS *et al.*, 2011), and Maximum Unbiased Validation (MUV) (ROHRER; BAUMANN, 2009). Riniker *et al*.(2013) proposed a classifier to improve performance. However, the MUV database had the worst metrics comparing with the two others.

We chose three fingerprints of the different types, of the earlier work (RINIKER; LANDRUM, 2013), to evaluate the compounds: Circular fingerprint (ECFP4), path-based (Atom Pair)(CARHART; SMITH; VENKATARAGHAVAN, 1985), RDK5 (RDKit: Cheminformatics and Machine Learning Software, 2019). It was 17 targets described, being 30 actives and 15000 inactives compounds, and all the fingerprints have 2048 of length. We applied the Milk-Way algorithm, using all the features and with the stratified feature selection. We validate the method using the *k*-cross-validation, a popular strategy for algorithm selection, to calculate the metrics of new models. The main idea was to divide the data, *k* times, for estimating the quality of each algorithm: part of the data was used for training each algorithm, and the remaining part was used for estimating the risk of the algorithm. Then, we calculated the recall, specificity, precision, f1, fa, Area under curve of Receiver Operating Characteristics Curve (AUC), and Enrichment Factor 5% (EF5%) to evaluate the performance (ARLOT; CELISSE, 2010; GIMENO; OJEDA-MONTES; TOMÁS-HERNÁNDEZ; CERETO-MASSAGUÉ *et al.*, 2019). In other words, for each of the *k* experiments, we used *k*-1 folds for training and the remaining one for testing. We reproduced all this procedure to different known algorithms: SVM, RF, NB, and LR. The stratified feature

selection had improved all metrics, not with only the Milk-Way algorithm but also the SVM and LR.

To validate the algorithms and to compare them, we used well-known metrics. The sensitivity (or recall) measures the proportion of correct predictions of active ligands compared to the whole number of active ligands. It is calculated as follows:

$$\text{Sensitivity} = (TP/(TP + FN)) \times 100, \quad (1)$$

where TP is the number of correct predictions (true positive), while FN is the number of incorrect predictions (false negative). The specificity measures the proportion of the number of inactive ligands compared to the whole number of inactive ligands and is calculated as follows:

$$\text{Specificity} = (TN /(TN + FP)) \times 100, \quad (2)$$

where TN is the number of correctly inactive (true negative) samples, while FP is the number of incorrectly accepted ligands (false positive).

Another metric is the precision, which relates the true actives with all the individuals classified positively (true positive and false positive).

$$\text{Precision} = (TP /(TP + FP)) \times 100 \quad (3)$$

The two metrics, precision, and sensitivity, are often combined as their harmonic mean, known as the F-measure, which can be formulated as follows (HRIPCSAK; ROTHSCHILD, 2005):

$$f1 = (2* \text{Precision}* \text{Sensitivity}/(\text{Precision}+ \text{Sensitivity})); \quad (4)$$

It is also possible to calculate the harmonic mean between sensitivity and specificity, with the following formula:

$$fa = (2* \text{Sensitivity}*\text{Specificity}/(\text{Sensitivity}+\text{Specificity})); \quad (5)$$

The area under such a multipoint curve is thus of some value, but the optimum in practice is the area under the simple trapezoid defined by the model:

$$\text{AUC} = (1\text{-Specificity vs Sensitivity}) \quad (6)$$

Maximizing AUC is thus equivalent to maximizing sensibility and specificity or minimizing a sum of normalized error FP+FN(POWERS, 2011).

Lastly, the accuracy relates to the true positives (TP) and negatives (TN) with the total individuals (N):

$$\text{Accuracy} = (TP+TN\ /N)\times 100 \quad (7)$$

Another metric, which is combinatory probably is enrichment factor metric is how many actives we find within a defined "early recognition" fraction of the ordered list relative to a random distribution. It is calculated as follows (MISHRA; BASU, 2013):

$$\text{Enrichment Factor} = (TP_{x\%}/N_{x\%})/(TP/N) \quad (8)$$

Being the $TP_{x\%}$ is the true positive of the x% early fraction and $N_{x\%}$, the number of compounds of assay x% first fraction.

The Milk-Way algorithm is used as LBVS. To complete the search for new drugs, we develop an SBVS strategy using the latent semantic retrieval technique (ÉLDEN, 2006). To do this, we used all approved targets of all species, until June 2018, and described by InterPro (HUNTER; APWEILER; ATTWOOD; BAIROCH *et al.*, 2009). This is an annotation that provides information about individual protein families, domains, and important sites. The proteins were the entities, and the InterPro, the attributes. Here, as we can interpret the proteins as vectors here, they can be projected into the vector space.

Now, it is possible to calculate the distance between then, identifying latent targets seeing that there has not an InterPro in common – indirectly relation between them. The beauty of SVD, however, is that it allows a simple strategy for optimal approximate fit using smaller matrices. Everitt and Dunn (1991) proposed an alternative approach where singular values whose relative variance is less than $0.7/n$, where n is the number of individuals in the matrix, must be ignored. The 3D plot (Figure 1) reveals that all targets are interrelated in the vector space (rank= 334; Sum (Sr=0.7252)).

We applied this technique to integrate the pharmacological therapy of SARS-CoV-2 and generates one proposal: Tetrachlorodecaoxide (BR 10 2020 007050 9).

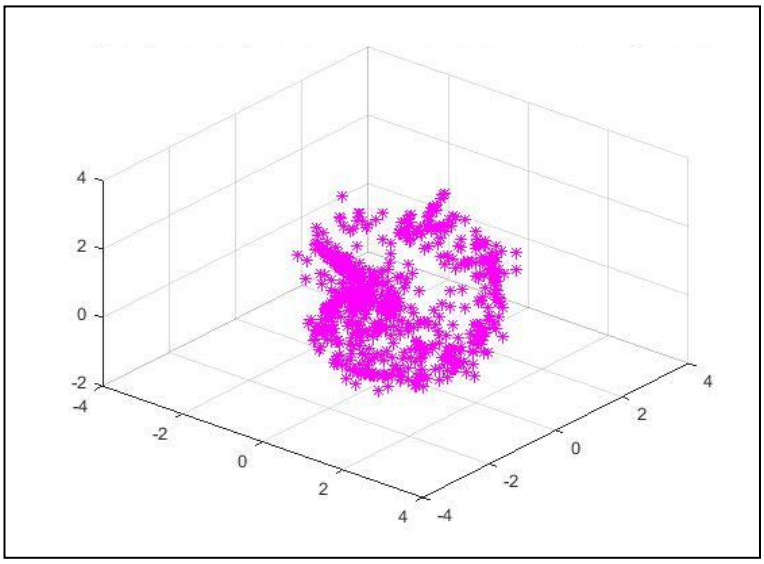


Figure 1 – An idea of a 3D plot of targets vector space, described through InterPro

4.1. Patent BR 10 2019 027703 3 - Método de Triagem de Compostos Baseado em Regressão Logística Modificada

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

23/12/2019 870190138531 14:44

29409161911878203

Pedido nacional de Invenção, Modelo de Utilidade, Certificado de Adição de Invenção e entrada na fase nacional do PCT

Número do Processo: BR 10 2019 027703 3

Dados do Depositante (71)

Depositante 1 de 1

Nome ou Razão Social: UNIVERSIDADE FEDERAL DE MINAS GERAIS

Tipo de Pessoa: Pessoa Jurídica

CPF/CNPJ: 17217985000104

Nacionalidade: Brasileira

Qualificação Jurídica: Instituição de Ensino e Pesquisa

Endereço: Av. Antônio Carlos, 6627 - Unidade Administrativa II - 2º andar- sala 2011

Cidade: Belo Horizonte

Estado: MG

CEP: 31270-901

País: Brasil

Telefone: (31) 3409-6430

Fax:

Email: patentes@ctit.ufmg.br

Dados do Pedido

Natureza Patente: 10 - Patente de Invenção (PI)

Título da Invenção ou Modelo de Utilidade (54): MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA

Resumo: A presente invenção consiste em um método de triagem de compostos que utiliza um modelo de regressão logística modificada e que permite a utilização de um número superior de atributos (descritores moleculares e/ou físico-químicos e/ou topológicos e/ou estruturais e/ou farmacológicos) em relação ao número de entidades (fármacos ou ligantes), podendo ser utilizado, por exemplo, no processo de reposicionamento de fármacos e no descobrimento de novos compostos. O referido método de triagem permite construir um modelo específico para cada alvo, através da seleção das entidades mais próximas (ad hoc), de modo a torná-lo mais exclusivo e capaz de diferenciar pequenas sutilezas. Além disso, é possível analisar os coeficientes de regressão logística, fazendo uma seleção estratificada de melhores atributos. Com esta combinação de potencialidades, a invenção permite a utilização de bases de dados contendo falsos positivos e, ainda assim, obter um desempenho satisfatório. Quando implementado através de um software, o método proposto não necessita de uma infraestrutura computacional complexa e robusta para a sua execução.

Dados do Inventor (72)

Inventor 1 de 6

Nome: CARMELINA FIGUEIREDO VIEIRA LEITE
CPF: 01818638630
Nacionalidade: Brasileira
Qualificação Física: Pesquisador
Endereço: Rua Cláudio Manoel, 210, apt 101. Funcionários
Cidade: Belo Horizonte
Estado: MG
CEP: 30140-100
País: BRASIL
Telefone: (31) 340 93932
Fax:
Email: patentes@ctit.ufmg.br

Inventor 2 de 6

Nome: MARCOS AUGUSTO DOS SANTOS
CPF: 27456510644
Nacionalidade: Brasileira
Qualificação Física: Professor do ensino superior
Endereço: Alameda Serra da Canastra, 197. Vila Del Rei
Cidade: Nova Lima
Estado: MG
CEP: 34000-000
País: BRASIL
Telefone:
Fax:
Email: patentes@ctit.ufmg.br

Inventor 3 de 6

Nome: LUCIANNA HELENE SILVA DOS SANTOS

CPF: 11112916750

Nacionalidade: Brasileira

Qualificação Física: Pesquisador

Endereço: Rua Itajubá, 2105 – Sagrada Família

Cidade: Belo Horizonte

Estado: MG

CEP: 31035-540

País: BRASIL

Telefone:

Fax:

Email: patentes@ctit.ufmg.br

Inventor 4 de 6

Nome: LARISSA FERNANDO LEIJÔTO

CPF: 10150370628

Nacionalidade: Brasileira

Qualificação Física: Pesquisador

Endereço: Rua Amélia Procópio Correia, n.45. Bom Retiro

Cidade: Sabará

Estado: MG

CEP: 34710-270

País: BRASIL

Telefone:

Fax:

Email: patentes@ctit.ufmg.br

Inventor 5 de 6

Nome: DIEGO CÉSAR BATISTA MARIANO

CPF: 08518676690

Nacionalidade: Brasileira

Qualificação Física: Pesquisador

Endereço: Rua Conde Dolabella, 1179. Várzea.

Cidade: Lagoa Santa

Estado: MG

CEP: 33400-000

País: BRASIL

Telefone:

Fax:

Email: patentes@ctit.ufmg.br

Inventor 6 de 6

Nome: RAFAEL EDUARDO OLIVEIRA ROCHA

CPF: 37155509884

Nacionalidade: Brasileira

Qualificação Física: Pesquisador

Endereço: Rua Guarda Custódio, 322, apto 208. Ouro Preto

Cidade: Belo Horizonte

Estado: MG

CEP: 31310-140

País: BRASIL

Telefone:

Fax:

Email: patentes@ctit.ufmg.br

Documentos anexados

Tipo Anexo	Nome
Comprovante de pagamento de GRU 200	1 - Comprovante de pagamento GRU 29409161911878203.pdf
Portaria	2 - Portaria 010-2019 - Prof. Gilberto UFMG.pdf
Relatório Descritivo	3 - Relatório descritivo.pdf
Reivindicação	4 - Reivindicações.pdf
Resumo	5 - Resumo.pdf

Acesso ao Patrimônio Genético

☑ Declaração Negativa de Acesso - Declaro que o objeto do presente pedido de patente de invenção não foi obtido em decorrência de acesso à amostra de componente do Patrimônio Genético Brasileiro, o acesso foi realizado antes de 30 de junho de 2000, ou não se aplica.

Declaração de veracidade

☑ Declaro, sob as penas da lei, que todas as informações acima prestadas são completas e verdadeiras.

INSTRUÇÕES:

A data de vencimento não prevalece sobre o prazo legal. O pagamento deve ser efetuado antes do protocolo. Órgãos públicos que utilizam o sistema SIAFI devem utilizar o número da GRU no campo Número de Referência na emissão do pagamento. Serviço: 200-Pedido nacional de Invenção, Modelo de Utilidade, Certificado de Adição de Invenção e entrada na fase nacional do PCT

Clique aqui e pague este boleto através do Auto Atendimento Pessoa Física.

Clique aqui e pague este boleto através do Auto Atendimento Pessoa Jurídica.

Recibo do Pagador

BANCO DO BRASIL | **001-9** | **00190.00009 02940.916196 11878.203170 6 80820000007000**

Nome do Pagador/CPF/CNPJ/Endereço
UNIVERSIDADE FEDERAL DE MINAS GERAIS CPF/CNPJ: 17217985000104
AV ANTONIO CARLOS 6627 UNIDADE ADMINISTRATIVA II 2 ANDAR SALA 2011, BELO HORIZONTE -MG CEP:31270901
Sacador/Avalista

Nosso-Número	Nr. Documento	Data de Vencimento	Valor do Documento	(=) Valor Pago
29409161911878203	29409161911878203	23/11/2019	70,00	

Nome do Beneficiário/CPF/CNPJ/Endereço
INSTITUTO NACIONAL DA PROPRIEDADE INDUST CPF/CNPJ: 42.521.088/0001-37
RUA MAYRINK VEIGA 9 24 ANDAR ED WHITE MARTINS , RIO DE JANEIRO - RJ CEP: 20090910

Agência/Código do Beneficiário
2234-9 / 333028-1

Autenticação Mecânica

BANCO DO BRASIL | **001-9** | **00190.00009 02940.916196 11878.203170 6 80820000007000**

Local de Pagamento
PAGÁVEL EM QUALQUER BANCO ATÉ O VENCIMENTO

Data de Vencimento
23/11/2019

Nome do Beneficiário/CPF/CNPJ
INSTITUTO NACIONAL DA PROPRIEDADE INDUST CPF/CNPJ: 42.521.088/0001-37

Agência/Código do Beneficiário
2234-9 / 333028-1

Data do Documento	Nr. Documento	Espécie DOC	Aceite	Data do Processamento
25/10/2019	29409161911878203	DS	N	25/10/2019

Nosso-Número
29409161911878203

Uso do Banco	Carteira	Espécie	Quantidade	xValor
29409161911878203	17	R$		

(=) Valor do Documento
70,00

Informações de Responsabilidade do Beneficiário

A data de vencimento não prevalece sobre o prazo legal.
O pagamento deve ser efetuado antes do protocolo.
Órgãos públicos que utilizam o sistema SIAFI devem utilizar o número da GRU n
o campo Número de Referência na emissão do pagamento.
Serviço: 200-Pedido nacional de Invenção, Modelo de Utilidade, Certificado de
Adição de Invenção e entrada na fase nacional do PCT

(-) Desconto/Abatimento

(+) Juros/Multa

(=) Valor Cobrado

Nome do Pagador/CPF/CNPJ/Endereço
UNIVERSIDADE FEDERAL DE MINAS GERAIS CPF/CNPJ: 17217985000104
AV ANTONIO CARLOS 6627 UNIDADE ADMINISTRATIVA II 2 ANDAR SALA 2011,
BELO HORIZONTE-MG CEP:31270901

Código de Baixa

Sacador/Avalista

Autenticação Mecânica - **Ficha de Compensação**

CTIT

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Gabinete da Reitora

PORTARIA Nº 010, DE 24 DE JANEIRO DE 2019

A REITORA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, no uso de suas atribuições legais e estatutárias, considerando o disposto nos artigos 11 e 12 do Decreto-Lei n° 200, de 25 de fevereiro de 1967, bem como no Decreto nº 7.689, de março de 2012; na Portaria nº 249, de 13 de junho de 2012, do Ministério do Planejamento, Orçamento e Gestão (MPOG), no Decreto nº 9.189, de 1º de novembro de 2017, e no inciso II do art. 2º, combinado com o art. 3º, ambos da Portaria MEC nº 36, de 18 de janeiro de 2018, do Ministério da Educação (MEC),

RESOLVE:

Art. 1º. Tornar sem efeito a portaria nº 174, de 8 de agosto de 2018;

Art. 2º Delegar competência ao Diretor da Coordenadoria de Transferência e Inovação Tecnológica (CTIT), professor GILBERTO MEDEIROS RIBEIRO, inscrição UFMG nº 247405, matrícula SIAPE nº 1964486, e a seu substituto eventual, para, no âmbito da CTIT:

a) assinar, por meio eletrônico ou físico, documentos ou instrumentos jurídicos concernentes ao exercício das atividades de competência da CTIT, no âmbito da Lei 10.973/04 – Lei de Inovação Tecnológica, da Política de Inovação da UFMG e suas resoluções específicas, tais como Contrato de Transferência de *Know-How*, Contrato de Licenciamento de Tecnologia, Contrato de Partilhamento de Titularidade de Tecnologia, Acordos de Confidencialidade e Termos de Sigilo, Termos de Autorização de Teste e documentos afins;

b) assinar, por meio eletrônico ou físico, documentação necessária para depósito, processamento, adição, retificação, substituição, modificação, ampliação e resposta de relatórios referentes a objeto de proteção de propriedade intelectual junto aos órgãos competentes, em âmbito nacional e internacional;

c) autorizar a realização de despesas dentro dos limites orçamentários da CTIT;

b) autorizar a concessão de suprimento de fundos a servidores da Unidade, bem como determinar a baixa de responsabilidade;

c) requisitar passagens e transportes em geral, por quaisquer vias, nos limites da dotação orçamentária da CTIT;

(...)

Avenida Antônio Carlos, 6627 · 31270-901 · Belo Horizonte · Minas Gerais
+55 31 3409-4127 · Fax: 31 3409-4130 · chefia@gabinete.ufmg.br · www.ufmg.br

ORDM nº 3221/2019

PORTARIA Nº 010, DE 24 DE JANEIRO DE 2019

d) autorizar viagens de servidores, a serviço da Unidade, arbitrando-lhes as respectivas diárias, obedecidas as disposições legais pertinentes;

e) assinar contratos, decorrentes de licitação, de sua dispensa ou inexigibilidade, no âmbito da CTIT;

f) prover arrecadação de receitas em geral no âmbito da CTIT;

g) apurar dívidas de terceiros para com a Universidade, oriundas de contratos de cotitularidade, licenciamento, transferência, dentre outros, adotando as medidas necessárias à regularização delas, no âmbito da CTIT.

Art. 3º Subdelegar competência ao Diretor da CTIT para:

a) celebrar novos contratos administrativos decorrentes de licitação, de sua dispensa ou de inexigibilidade ou prorrogar contratos em vigor relativos a atividades de custeio cujos valores sejam inferiores a R$500.000,00 (quinhentos mil reais); e

b) autorizar a realização de despesas relativas a atividades de custeio cujos valores sejam inferiores a R$500.000,00 (quinhentos mil reais).

Art. 4º A presente Portaria entra em vigor nesta data.

Belo Horizonte, 24 de janeiro de 2019.

Profa. Sandra Regina Goulart Almeida
Reitora

CONFERE COM O ORIGINAL
EM 28 / 01 / 19
Juliana Campideli Neves Garcia
Secretária Executiva
Inscrição nº 22547-9

SCG/jcng

L:Document/Portaria/p19-010

“MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA”

[01] A presente invenção consiste em um método de triagem de compostos que utiliza um modelo de regressão logística modificada e que permite a utilização de um número superior de atributos (descritores moleculares e/ou físico-químicos e/ou topológicos e/ou estruturais e/ou farmacológicos) em relação ao número de entidades (fármacos ou ligantes), podendo ser utilizado, por exemplo, no processo de reposicionamento de fármacos e no descobrimento de novos compostos. O referido método de triagem permite construir um modelo específico para cada alvo, através da seleção das entidades mais próximas (*ad hoc*), de modo a torná-lo mais exclusivo e capaz de diferenciar pequenas sutilezas. Além disso, é possível analisar os coeficientes de regressão logística, fazendo uma seleção estratificada de melhores atributos. Com esta combinação de potencialidades, a invenção permite a utilização de bases de dados contendo falsos positivos e, ainda assim, obter um desempenho satisfatório. Quando implementado através de um software, o método proposto não necessita de uma infraestrutura computacional complexa e robusta para a sua execução.

[02] A triagem de alto rendimento é um método automático para ensaios de triagem de potenciais compostos. Em geral, a triagem virtual utiliza técnicas computacionais para analisar os resultados da triagem de alto rendimento, construindo modelos treinados e indicando compostos promissores. Capaz de trabalhar com todos os atributos de cada composto, a presente invenção classifica e/ou ordena os compostos segundo a sua atividade para determinado alvo terapêutico.

[03] O documento EP1111533A2, intitulado *Logistic regression trees for drug analysis*, apresenta um método para filtrar compostos com propriedades farmacocinéticas potencialmente ruins, como metabolismo desfavorável ou baixa solubilidade no processo de busca de compostos para o reposicionamento de fármacos. Porém, diferente do método antecipado, na presente invenção, as variáveis escolhidas para a descrição dos compostos podem ser de qualquer natureza; pode-se ou não englobar as características de farmacocinética; e a

regressão logística utilizada é modificada, sendo capaz de utilizar um número superior de características do que de compostos, o que possibilita a realização do método de forma não recursiva. Ademais, a presente invenção é útil tanto para o descobrimento de novos compostos como para o reposicionamento de fármacos, visto que ela classifica e/ou ordena os compostos segundo a sua atividade para determinado alvo terapêutico.

[04] O artigo intitulado *Heterogeneous Classifier Fusion for Ligand-Based Virtual Screening: Or, How Decision Making by Committee Can Be a Good Thing*, elaborado por Sereina Riniker et.al (2013), transparece a dificuldade em se analisar diferentes bases de dados, caso estejam mais ou menos curadas. Embora os autores tenham melhorado o desempenho do classificador, combinando diferentes algoritmos com diferentes descritores dos compostos, essa melhoria não foi efetiva para todas as bases de dados, o que sugere que, na referida metodologia, o algoritmo utilizado depende de uma base de dados curada. A presente invenção, além de usar uma infraestrutura mais simples, conseguiu superar os resultados do trabalho citado e pode ser aplicada a diferentes bases de dados.

[05] A presente invenção torna mais ágil o processo de desenvolvimento de medicamentos e de reposicionamento de fármacos e, por consequência, reduz os custos relacionados ao processo. Sua implementação e aplicação podem ser realizadas através de um software e não requerem uma infraestrutura computacional complexa e robusta. Por fim, da forma como foi proposta, a invenção avança no sentido de permitir a utilização de um número superior de atributos (descritores moleculares e/ou físico-químicos e/ou topológicos e/ou estruturais e/ou farmacológicos) em relação ao número de entidades (fármacos ou ligantes), por exemplo, no processo de triagem e escolha de compostos para a então realização de testes *in vitro* quando do desenvolvimento de medicamentos e do reposicionamento de fármacos pela indústria farmacêutica.

DESCRIÇÃO DETALHADA DA TECNOLOGIA

[06] A presente invenção consiste em um método de triagem de compostos que utiliza um modelo de regressão logística modificada e que permite a utilização de um número superior de atributos (descritores moleculares e/ou físico-químicos e/ou topológicos e/ou estruturais e/ou farmacológicos) em relação ao número de entidades (fármacos ou ligantes), podendo ser utilizado, por exemplo, no processo de reposicionamento de fármacos e no descobrimento de novos compostos. O referido método de triagem permite construir um modelo específico para cada alvo, através da seleção das entidades mais próximas (*ad hoc*), de modo a torná-lo mais exclusivo e capaz de diferenciar pequenas sutilezas. Além disso, é possível analisar os coeficientes de regressão logística, fazendo uma seleção estratificada de melhores atributos. Com esta combinação de potencialidades, a invenção permite a utilização de bases de dados contendo falsos positivos e, ainda assim, obter um desempenho satisfatório. Quando implementado através de um software, o método proposto não necessita de uma infraestrutura computacional complexa e robusta para a sua execução.

[07] A presente invenção pode ser realizada através de programas de computador e aplicada a diferentes bases de compostos e diferentes descritores. Deve-se, preferencialmente, utilizar como atributos dos ligantes ou fármacos os seus descritores binários, porém, pode-se utilizar também como atributos, por exemplo, os descritores moleculares, físico-químicos, topológicos, estruturais ou farmacológicos dos compostos.

[08] Uma vez escolhido o alvo de interesse e pelo menos um composto-teste, faz-se necessário armazenar os descritores de uma série de compostos (ligantes ou fármacos) de interesse em uma matriz *A*, a partir da qual serão identificados os compostos ativos e os inativos. Opcionalmente, pode-se reduzir o ruído das informações da matriz *A*, utilizando o método de decomposição de valor singular (SVD).

[09] Cada composto, com seus respectivos descritores, pode ser projetado no espaço vetorial. A partir disso, opcionalmente, pode-se realizar uma seleção *ad hoc* dos compostos ativos e inativos na construção do modelo, os quais são escolhidos de acordo com a sua proximidade espacial em relação ao composto-

teste. Essa avaliação pode ser realizada, por exemplo, utilizando a distância de *Hamming*, similaridade de cossenos ou a distância euclidiana. Caso se utilize descritores binários como atributos, preferencialmente, deve-se utilizar a distância de *Hamming*. Embora essa etapa de seleção *ad hoc* não seja obrigatória, sua aplicação possibilita a criação de modelos mais homogêneos entre os compostos ativos e os compostos inativos, permitindo que características sutis sejam detectadas.

[010] Em seguida, aplica-se um modelo de Regressão Logística modificada. O resultado será uma probabilidade *P(x)* associada a cada composto, que por sua vez, compreende um conjunto de descritores *x*. O cálculo da probabilidade *P(x)* do composto de ser ativo ou inativo para o alvo de interesse se dá pela equação:

[011] $$P(x) = \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_i x_i + \alpha_{n+1}}} \qquad (1)$$

[012] Onde *α* representa o peso associado a cada atributo para o qual *P(x)* seja zero ou um. O índice *i* indica cada descritor e o índice *n* indica o número total destes. Quanto maior for o valor de *α*, mais próximo de um será o *P(x)* e, quanto menor for o valor de *α*, mais próximo de zero será o *P(x)*.

[013] A resolução da Equação (1) é um problema combinatório e uma das possíveis soluções se dá através da soma da minimização dos quadrados dos resíduos e da minimização, ao quadrado, da norma dos valores *α*. A atribuição dos parâmetros *α* aplicados à Equação (1) se dá a partir da aplicação da equação a seguir:

$$f(\alpha) = \|\alpha\|^2 + \|A\alpha - b\|^2 \qquad (2)$$

[014] A função *f(α)* é convexa e o valor do argumento α é escolhido igualando-se a derivada de *f(α)* em relação à α à zero, o que corresponde ao ponto de mínimo que é obtido com a solução do seguinte sistema de equações lineares:

$$\alpha^T\alpha + A^T A\,\alpha - A^T b = 0 \Leftrightarrow (I + A^T A)\,\alpha = A^T b \quad (3)$$

[015] Onde b é a aproximação dos logaritmos das chances de *P(x)* ser um ou zero, associada a cada composto; e *I* é uma matriz identidade com a mesma ordem da maior dimensão da matriz *A*.

[016] Uma vez obtidos os valores de *α,* obtêm-se os valores de *P(x)*, os quais indicam a probabilidade do composto ser ativo ou inativo para o alvo em estudo.

Ao adicionar a parcela de minimização do quadrado da norma dos valores de *α* na Equação (2), cujo efeito recai sobre a Equação (3), é possível utilizar um número superior de atributos em relação ao número de entidades.

[017] Uma vez estabelecida uma dada linha de corte para o *P(x)*, é possível determinar se o composto é ativo ou inativo para determinado alvo. A linha de corte se dá, preferencialmente, a partir de métodos de validação cruzada.

[018] Opcionalmente, pode-se aplicar no método proposto uma seleção estratificada de características. Primeiramente, faz-se a escolha de um percentual dos atributos cujos valores de *α* estão mais distantes de zero, tanto positivos quanto negativos. Embora esta não seja uma etapa obrigatória, sua aplicação traz como vantagens a melhoria do desempenho da classificação em bases com elevado número de compostos falsos positivos. Caso se realize esta etapa, deve-se calcular novamente o valor da probabilidade *P(x)*, considerando o referido subconjunto de características escolhidas.

[019] Opcionalmente, é possível aplicar outros algoritmos em substituição à Regressão Logística modificada proposta, tais como *Support Vector Machine* (SVM), *Random Forest* (RF) e *Naive Bayes* (NB), para a classificação dos compostos, utilizando apenas a seleção estratificada de características feita pela presente invenção.

[020] O número de indivíduos que compõem o modelo pode ser o mesmo número de valores singulares, conforme aplicação de SVD, ou outro, de acordo com um estudo de aferição. Este estudo pode ser realizado variando o número de compostos ativos e inativos para a composição do modelo, até se atingir uma validação satisfatória, preferencialmente, equivalente à área sob a curva maior ou igual a 0,70.

[021] O método de triagem de compostos baseado em regressão logística modificada compreende as seguintes etapas:

a. Escolher o alvo de interesse e pelo menos um composto-teste;
b. Armazenar uma série de compostos e seus descritores em uma matriz *A*, sendo uma das dimensões da matriz *A* correspondente aos compostos e a outra dimensão, aos respectivos descritores;

c. Obter, através da Equação (4), os valores de α para cada atributo da matriz *A*, onde *α* representa o peso associado a cada atributo para o qual a probabilidade de o composto ser ativo ou inativo para o alvo de interesse *P(x)* é igual a zero ou igual a um; *b* é a aproximação dos logaritmos das chances de *P(x)* ser igual a zero ou igual a um associada a cada composto; e *I* é a matriz identidade de *A*;

$$(I + A^T A)\,\alpha = A^T b \qquad (4)$$

d. Obter, através da Equação (1), os valores de *P(x)* para cada composto compreendendo um conjunto de descritores *x*, onde o índice *i* indica cada descritor e o índice *n* indica o número total destes.

$$P(x) = \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_i x_i + \alpha_{n+1}}} \qquad (1)$$

e. Estabelecer uma linha de corte sobre o valor de *P(x)* e selecionar os compostos ativos e inativos para o alvo de interesse.

[022] Na etapa (a), o(s) tipo(s) de descritores pode(m) ser selecionado(s) do grupo que compreende descritores binários, descritores moleculares, descritores físico-químicos, descritores topológicos, descritores estruturais e descritores farmacológicos dos compostos.

[023] Opcionalmente, pode-se reduzir o ruído das informações da matriz *A*, utilizando o método de decomposição de valor singular (SVD), logo após a realização da etapa (a).

[024] Opcionalmente, pode-se projetar espacialmente cada composto, com seus respectivos descritores, e realizar uma seleção *ad hoc* dos compostos ativos e inativos de acordo com a sua proximidade espacial em relação ao composto-teste, imediatamente antes à realização da etapa (c).

[025] Quando se optar por realizar a seleção ad hoc dos compostos ativos e inativos, esta poderá se basear em um método a ser selecionado do grupo que compreende distância de *Hamming*, similaridade de cossenos e distância euclidiana. Caso na etapa (a) tenha se utilizado descritores binários, é recomendável que seja utilizada a distância de *Hamming*.

[026] Opcionalmente, pode-se utilizar o método de validação cruzada para estabelecer a linha de corte sobre o valor de *P(x)* na etapa (e).

[027] Opcionalmente, após a etapa (e), pode-se escolher um percentual dos atributos cujos valores de *α* estão mais distantes de zero, tanto os valores positivos quanto os valores negativos, e então repetir a etapa de cálculo do valor da probabilidade *P(x)*, considerando apenas com o referido subconjunto de características escolhidas.

[028] Em alternativa às etapas (c) e (d), pode-se utilizar um algoritmo de aprendizado de máquina a ser selecionado do grupo que compreende *Support Vector Machine* (SVM), *Random Forest* (RF) e *Naive Bayes* (NB) para o cálculo da probabilidade *P(x)*.

[029] A presente invenção pode ser mais bem compreendida através do exemplo a seguir, não limitante da tecnologia.

[030] **Exemplo 1 – Aplicação do método de triagem de compostos baseado em regressão logística modificada tendo como alvo a transcriptase reversa**

[031] O alvo escolhido foi a transcriptase reversa (*Proteína NCBI: Q72547*). A base de treino para este alvo foi obtido da base de dados pública *PubChem*.

[032] Foram encontrados 1892 compostos ativos testados experimentalmente e 51 inibidores inativos para a transcriptase reversa, compondo assim os dados de entrada. Foram removidos sete fármacos já comercializados quando da realização do teste (*abacavir, didanosina, emtricitabina, lamivudina, estavudina, zalcitabina e zidovudina*) para compor o teste cego, ou seja, os compostos que serão submetidos e classificados pelo modelo. Todos os compostos coletados foram descritos de forma binária, pelos pares de fragmentos e impressões digitais de farmacóforo, pelo programa *PowerMV* (882 características no total).

[033] Em seguida, os descritores foram agrupados em uma matriz de entrada, onde as colunas corresponderam aos ligantes e as linhas, aos seus atributos. Foram aplicadas a regressão logística modificada e a seleção *ad hoc*. Depois de calculado o valor de *α*, foi realizada a seleção estratificada de características com diferentes quantidades, a fim de avaliar o melhor caso (5%, 25%, 50%, 75% e 100%). No treinamento, foi realizada a validação cruzada (*k=5*) e, em seguida, foi

realizado o teste cego. Para avaliar a eficácia da metodologia proposta, foi calculada a sensibilidade; especificidade; precisão; f1; fa; AUC; e precisão.

[034] O algoritmo proposto teve desempenho máximo na validação cruzada apenas com apenas 50% ou 75% dos atributos usados no conjunto de dados dos inibidores análogos nucleosídeos da transcriptase reversa (Tabela 1). No teste cego, o desempenho foi melhor para 50% das características (Tabela 2), o que indicou a necessidade de se ter especial atenção tanto aos verdadeiros positivos quanto aos verdadeiros negativos. Todas as métricas utilizaram o valor da probabilidade associada a cada composto.

[035] **Tabela 1** – Métricas do modelo de treinamento dos inibidores análogos nucleosídeos da transcriptase reversa, usando a validação cruzada (*k*=*5*).

CARACTERÍSITCAS	5%(44*)	25%(220*)	50%(441*)	75%(661*)	100%(882*)
Sensibilidade	0.89233	0.997879	1	1	0.925242
Especificidade	0.901818	1	1	1	1
Precisão	0.997161	1	1	1	1
f1	0.940677	0.998938	1	1	0.961037
fa	0.886212	0.998938	1	1	0.961037
AUC	0.924012	0.99894	1	1	0.963224
Acurácia	0.89263	0.997934	1	1	0.927229

* Número de caraterísticas utilizadas pela seleção estratificada.

[036] **Tabela 2** – Métricas do teste cego do dos inibidores análogos nucleosídeos da transcriptase reversa, utilizando diferente número de caraterísticas da seleção estratificada.

CARACTERÍSITICAS	5%(44*)	25%(220*)	50%(441*)	75%(661*)	100%(882*)
Sensibilidade	1	1	1	1	1
Especifidade	0.894286	0.945714	0.965714	0.902857	0.448571
Precisão	0.159091	0.269231	0.368421	0.170732	0.035
f1	0.27451	0.424242	0.538462	0.291667	0.067633
Fa	0.944193	0.9721	0.982558	0.948949	0.619329
AUC	0.947143	0.972857	0.982857	0.951429	0.724286
Acurácia	0.896359	0.946779	0.966387	0.904762	0.459384

* Número de caraterísticas utilizadas pela seleção estratificada.

[037] Em seguida, a presente invenção foi comparada com outros algoritmos já conhecidos, o *Random Forest* (RF) e o *Support Vector Machine* (SVM). Na Tabela 3, é possível observar que a presente invenção supera os resultados obtidos com a utilização dos referidos algoritmos, através da aplicação da Regressão Logística modificada, uma vez que a sua aplicação ajuda a sanar a influência dos falsos positivos.

[038] **Tabela 3** – Métricas do teste cego do dos inibidores análogos nucleosídeos da transcriptase reversa, utilizando diferentes algoritmos com a validação cruzada *k=5*.

	Sensibilidade	Especificidade	Precisão	f1	fa	AUC	Acurácia
Presente Invenção	1.00	0.97	0.37	0.54	0.98	0.98	0.97
RF	1.00	0	0.02	0.04	0	0.50	0.02
SVM	1.00	0	0.02	0.04	0	0.50	0.02

[039] O método em questão pode ser aplicado, por exemplo, para seleção de compostos no início do processo de desenvolvimento de medicamentos, assim como no processo de reposicionamento de fármacos já comercializados.

REIVINDICAÇÕES

1. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, caracterizado por compreender as seguintes etapas:

a. Escolher o alvo de interesse e pelo menos um composto-teste;

b. Armazenar uma série de compostos e seus descritores em uma matriz *A*, sendo uma das dimensões da matriz *A* correspondente aos compostos e a outra dimensão, aos respectivos descritores;

c. Obter, através da Equação (4), os valores de *α* para cada atributo da matriz *A*, onde *α* representa o peso associado a cada atributo para o qual a probabilidade de o composto ser ativo ou inativo para o alvo de interesse *P(x)* é igual a zero ou igual a um; *b* é a aproximação dos logaritmos das chances de *P(x)* ser igual a zero ou igual a um associada a cada composto; e *I* é a matriz identidade de *A*;

$$(I + A^T A)\,\alpha = A^T b \qquad (4)$$

d. Obter, através da Equação (1), os valores de *P(x)* para cada composto compreendendo um conjunto de descritores *x*, onde o índice *i* indica cada descritor e o índice *n* indica o número total destes;

$$P(x) = \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_i x_i + \alpha_{n+1}}} \qquad (1)$$

e. Estabelecer uma linha de corte sobre o valor de *P(x)* para selecionar os compostos ativos e inativos para o alvo de interesse.

2. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por utilizar na etapa (a) pelo menos um tipo de descritores a ser selecionado(s) do grupo que compreende descritores binários, descritores moleculares, descritores físico-químicos, descritores topológicos, descritores estruturais e descritores farmacológicos dos compostos.

3. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por reduzir o ruído das informações da matriz *A*, utilizando o

método de decomposição de valor singular (SVD), logo após a realização da etapa (a).

4. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por projetar espacialmente cada composto, com seus respectivos descritores, e realizar uma seleção *ad hoc* dos compostos ativos e inativos de acordo com a sua proximidade espacial em relação ao composto-teste, imediatamente antes à realização da etapa (c).

5. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 4, caracterizado pela seleção ad hoc dos compostos ativos e inativos ser realizada através de um método selecionado do grupo que compreende distância de *Hamming*, similaridade de cossenos e distância euclidiana.

6. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por utilizar o método de validação cruzada para estabelecer a linha de corte sobre o valor de *P(x)* na etapa (e).

7. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por, após a etapa (e), escolher um percentual dos atributos cujos valores de α estão mais distantes de zero, tanto os valores positivos quanto os valores negativos, e então repetir a etapa de cálculo do valor da probabilidade *P(x)*, considerando apenas com o referido subconjunto de características escolhidas.

8. MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA, de acordo com a reivindicação 1, caracterizado por aplicar, em alternativa às etapas (c) e (d), um algoritmo de aprendizado de máquina a ser selecionado do grupo que compreende *Support Vector Machine* (SVM), *Random Forest* (RF) e *Naive Bayes* (NB) para o cálculo da probabilidade *P(x)*.

RESUMO

“MÉTODO DE TRIAGEM DE COMPOSTOS BASEADO EM REGRESSÃO LOGÍSTICA MODIFICADA”

A presente invenção consiste em um método de triagem de compostos que utiliza um modelo de regressão logística modificada e que permite a utilização de um número superior de atributos (descritores moleculares e/ou físico-químicos e/ou topológicos e/ou estruturais e/ou farmacológicos) em relação ao número de entidades (fármacos ou ligantes), podendo ser utilizado, por exemplo, no processo de reposicionamento de fármacos e no descobrimento de novos compostos. O referido método de triagem permite construir um modelo específico para cada alvo, através da seleção das entidades mais próximas (*ad hoc*), de modo a torná-lo mais exclusivo e capaz de diferenciar pequenas sutilezas. Além disso, é possível analisar os coeficientes de regressão logística, fazendo uma seleção estratificada de melhores atributos. Com esta combinação de potencialidades, a invenção permite a utilização de bases de dados contendo falsos positivos e, ainda assim, obter um desempenho satisfatório. Quando implementado através de um software, o método proposto não necessita de uma infraestrutura computacional complexa e robusta para a sua execução.

4.2. Published Paper: Milk-Way algorithm for ligand-based virtual screening: a CDK-2 case study

Trends in Developmental Biology
Vol. 13, 2020

Original Article

Milk-Way algorithm for ligand-based virtual screening: CDK2 case study

Carmelina Figueiredo Vieira Leite[1,*], Lucianna Helene Silva Santos[2], Larissa Fernandes Leijôto[1], Diego César Batista Mariano[1], Rafael Eduardo Oliveira Rocha[2] and Marcos Augusto dos Santos[3]

[1]Laboratory of Bioinformatics and Systems; [2]Department of Biochemistry and Immunology; [3]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil.

ABSTRACT

Ligand-based screening of large molecular databases can help reduce costs with experiments by filtering and ranking promising compounds in an initial stage of the drug developing process. However, some ligand-based methods can be ineffective when presented with a high-dimensional number of attributes extracted from an extensive dataset of compounds. Herein, we propose a drug-mining algorithm that can be used to screen ligands and repurpose known drugs, from any dataset for any target. The Milk-Way algorithm combines mathematical and regression methods to select promising compounds from a high-dimensional dataset without the use of massive computational power. We carried out a prospective screening targeting cyclin-dependent kinase two (CDK2), an attractive target for therapeutics designed to arrest or recover control of the cell cycle. The combined use of the algorithm metrics and molecular docking suggested five promising drugs to be repositioned (Pramocaine, Prochlorperazine, Trifluoperazine, Methionine, and Pergolide), in which three were already mentioned as possible inhibitors of related diseases in the literature.

KEYWORDS: algorithm, drug discovery, drug repurposing, ligand-based virtual screening, logistic regression, machine learning, development.

*Corresponding author: cleite@ufmg.br

1. INTRODUCTION

We present a new algorithm to screen novel compounds using CDK2 as the target. This is an enzyme that phosphorylates many proteins involved in cell cycle progression, DNA replication, histone synthesis, centrosome duplication, among other processes [1, 2]. Because of these functions, CDK2 represents an attractive target for therapeutics designed to arrest or recover control of the cell cycle in dividing cells [3], and since the enzyme is not essential for the cell cycle, its toxicity is not severe [4]. Despite the importance of the CDK2 protein, not many commercial drugs act against it. Thus, we investigated the use of drug repurposing as an aid to CDK2 drug development. Drug repurposing is the strategy of discovering new uses or conditions for approved drugs to not only assess the effects of the drug on a new target but also to reduce the cost of developing a new drug.

Computational approaches, such as virtual screening (VS), have emerged as alternatives to screen large libraries of small molecules in a cost-efficient manner. Although VS approaches do not substitute experimental assays, they can speed up and rationalize the process of drug discovery, enriching the number of hits since it can downsize the number of candidates to be tested [5, 6]. In structure-based virtual screening (SBVS), the three-dimensional structure of the target is known, from x-ray crystallographic, NMR, or computational

modeling [7, 8]. Recently, cryo-EM (cryo-electron microscopy) was introduced as a powerful three-dimensional source and is starting to make an impact in drug discovery [9]. SBVS usually involves the molecular docking methodology, which places and ranks the compounds in the binding site according to an algorithm that predicts their possible binding affinity [10].

In the absence of three-dimensional structures of the targets, the molecular and chemical properties of known actives and tested compounds are gathered to create models of their binding using ligand-based virtual screening (LBVS) [11]. Some LBVS methodologies can assume that one or more actives share a binding mode. Thus, the screening will be done as a similarity or matching search to select potential new binders with similar chemical features to the known ones [12]. Chemical compounds to compose screening libraries are available as a free resource of bioactivity data for small molecules in various databases such as ChEBI [13], ZINC [14], PubChem [15], DrugBank [16], IUPHAR-DB [17], and KEGG [18]. Alternatively, in-house compound datasets can also be created from previously tested compounds and analogs. A powerful LBVS method is the Quantitative structure-activity relationship (QSAR). QSAR models starts by calculating chemical descriptors collected from compounds found in databases or the literature. These descriptors are correlated with biological properties using a variety of machine learning techniques [5, 19]. After created and validated, QSAR models are applied to predict novel compounds in virtual screening campaigns. Although QSAR is effective and with the use of SBVS methods also target orientated, it is still a time and computationally demanding method [5].

For many LBVS methods, in the model's generation step, known actives can be employed as training data in classification methods. These methods use this information to separate a database of compounds with unknown activity into predicted actives and inactives [20]. Classification methods are usually machine learning approaches that build models, such as decision trees [21, 22], neural networks [23, 24], and support vector machines [25], and can perform particularly well in enriching actives [20]. Established algorithms for data mining (Naive Bayes, SMO, Random Forrest, J48) are also used to classify chemical compounds [26].

However, classification methods can demand extensive knowledge over the many methodologies and need high computing power. Furthermore, these methods might not perform well when subjected to imbalanced or high-dimensional data, *i.e.*, when all features describing the chemical properties of the compounds are used. These problems can lead to an inadequate exploration of the ligands, and lack of accurate results, essential to screening [27-29]. Therefore, the proposition of an improved *in silico* approach to classified chemical compounds is a relevant issue.

In this paper, we suggest the Milk-Way algorithm (a WAY of Mathematical Interpretation of the Logistic ranK), a robust combination of well-known techniques of data-mining, logistic regression, vector space representation, and linear algebra to contribute to the ligand-based approaches to rational drug design. Our algorithm does not demand a complex computational infrastructure to select potential hits in a screening campaign. A first step in the algorithm is to collect a library of actives and inactives compound structures to be designated through descriptors (Fragment Pair/ Pharmacophore). Then, a model is created, validated, and trained to classify potential hits. The model enables us to calculate and project the probability of each ligand ($P(x)$, where x is the ligand) outcome, which is used to distinguish possible high performing ligands. An application of repurposing commercially available drugs with the Milk-Way algorithm was conducted using the cyclin-dependent kinase 2 (CDK2) as a target to exemplify the algorithm's predictive power. CDK is a family of serine/threonine protein kinases which act as critical regulatory element in cell cycle progression and development. As the name reveals, those are enzymes that catalyze the transfer of a phosphate from ATP to a protein substrate, more precisely on a serine or threonine amino acid residue [30].

2. MATERIALS AND METHODS

All the data were processed in MATLAB R2017a, using a laptop with 4 GB RAM, 320 GB hard drive, and a processor Intel Core i5 2.53 GHz.

The Milk-Way methodology runs in the Windows operating systems.

2.1. Data collection

The starting point, in our algorithm, is the construction of a matrix, whose entities are the ligands (columns) and their attributes (lines). All the attributes were generated through the PowerMV program [31]. The attributes are binary descriptors that define both active and inactive ligands. We have chosen fragment pair (735 features) and pharmacophore fingerprints (147 features) as molecular features [32-34]. For fragment-based descriptors, PowerMV replaces atom types with groups of atoms and counts the shortest path between them. For example, two phenyl rings, which are separated by two bonds, are expressed as AR_02_AR. In total 14 groups of atoms are considered. Whereas pharmacophore fingerprints are built based on bioisosteric principles [35], *i.e.*, two atoms (or groups), predicted to have similar biological effect, are classified as the same type. For example, the disulfide (-S-) is often used to replace ester group (-O-); hence we assign these two groups to the same type. Therefore, only six classes are considered in the pharmacophore-based descriptors. However, it is possible to use any database and attributes to build the input matrix, such as molecular, physicochemical, topological, structural, pharmacological descriptors, or any property of the ligands. A binary representation is also not obligatory. The only imperative premise is a representation of ligands. To the algorithm, this is the most critical step since the projection of the ligand into the vector space depends on that, to calculate the probability.

Literature data of known inhibitors were extracted from two databases, PubChem [36] and DrugBank [16], depending on the case study. In the CDK2 study case, we used in the training set the only approved drug to CDK2 to compose the actives and 152 compounds as decoys. The test set included all 2389 commercial drugs according to the DrugBank [16], at the time the search was performed.

2.2. The screening algorithm

The algorithm consists of several consecutive steps: (i) SVD; (ii) Ad-hoc choice; (iii) Modified Logistic Regression and, (iv) Stratified feature selection through the alpha values.

2.2.1. Singular value decomposition

This step in our screening algorithm helps reducing the noise and retrieve latent patterns of the input matrix. A technique for information retrieval, using a linear algebra approach, is the singular value decomposition (SVD). This rank reduction procedure is closely related to matrix factorization, data compression, dimension reduction, and feature selection/extraction [37]. When SVD is utilized, it allows the matrix to be represented as a set of derived matrices, which can have different depictions of data without loss in their semantic meaning [37, 38]. A matrix submitted to the SVD method can be represented as:

$$A = U\Sigma V^T, \quad (1)$$

where A is a matrix of real numbers or complex numbers composed of m rows by n columns. However, now, m represents the attributes and n, the entities. The U is an orthonormal m x m matrix and the eigenvectors of AA^T; the Σ is a m x n matrix, known as the diagonal matrix, with real and non-negative numbers and contain the singular values of A. The matrix V^T is known as a conjugate transpose, a n x n unit matrix. As the diagonal values of Σ are ordered in descending order, Σ is a direct function of matrix A and distinguishes the singular values of this matrix. This sorting is from the most meaningful to the least significant values. Whereas from a subset of singular values of size $k < m$, we can obtain A_k, the approximate matrix of A, with k-dimensional:

$$A_k = U_k\Sigma_k V_k^T \quad (2)$$

The approximation will be related to how many singular eigenvalues are used [37, 39, 40]. This strategy enables an information extraction based on less data, and the data analysis execution time does not increase exponentially when the matrix size is increased. A data set represented by a smaller number of singular values than the original full-size dataset tends to cluster data that would not be clustered together if the original one was used. Therefore, the derived representation, which captures associations, is used for retrieval [38-41]. The representation in the reduced space depiction is economical, in the sense that N original index features have been replaced by the $k < M$ best-approximated surrogates. It is essential for the method that the derived k-dimensional

factor space does not reconstruct the original term space correctly. The beauty of SVD, however, is that it allows a simple strategy for optimal approximate fit using smaller matrices. Everitt and Dunn [42] proposed an alternative approach where singular values whose relative variance is less than 0.7/*n*, where *n* is the number of individuals in the matrix, must be ignored. If the singular values in Σ, are ordered by size, the first and largest *k* may be kept and the remaining smaller ones set to zero [43].

2.2.2. Ad-hoc choice

After determining the number of singular values of our input matrix of molecular features of active and inactive compounds, we define the number of singular values as a criterion of the number of individuals to collect. Otherwise, we would have an infinitive of possibilities as well as combinations. The model was constructed for each query (ligand to predict the P(*x*)) using the closest active and inactive ligands. The main objective of this particular strategy is to create a homogeneous and specific system through the choice of closely spaced individuals. We chose Hamming distance, on account of better adjusting the matrix that is composed of zeros and ones. This distance can be defined as the number of positions in which a codeword differs between two code words [44], or, in other words, it is the minimum number of *errors* that could have transformed one string into the other.

2.2.3. Modified logistic regression

The regression method is helpful to any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The logistic regression analysis proposes the classification of individuals in different categories, with an accurate estimation for that possibility [45]. The logistic regression equation consists of assigning values to $\alpha = (\alpha_0, \alpha_1,\ldots,\alpha_n)$ to fit the logit function (3)

$$P(x) = \frac{e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}} \quad (3)$$

such that $P(x) = 0/1$ associated with the activity of each ligand (inactive or active, respectively). The data is represented by matrix *A*, with *m* rows and *n* columns; so the value of each position $a_{m,n}$ represents the attribute *n* of the ligand *m*. Here, the matrix is the transpose of the previous SVD-treated matrix. We will omit the indication of row *m* in the elements of vector *x*, that is $x = (x_1, x_2,\ldots,x_n)$. Associated with each row *m*, there is $P_i(x) = 0/1$ that informs the activity of the ligand. We observed that when $e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}$ drops to zero, $P_i(x)$ also goes to zero. On the other hand, if $e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}$ tends to infinity, $P_i(x)$ approximates one. Viewing $P_i(x)$ as the probability, the odds $C_i(x)$ is given by:

$$C_i(x) = \frac{P(x)}{1-P(x)} = e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}} \quad (4)$$

To implement the method, we use $\hat{C}i(x) \approx Ci(x) = (0.99999 / (1\text{-}0.99999))$ instead of $C_i(x)$ when the odds are related to $P_i(x) = 1$. When $P_i(x) = 0$, we consider $\hat{C}_i(x) \approx C_i(x) = (0.00001 / (1\text{-}0.00001))$.

Taking the logarithm on both sides of (equation 4) we have:

$$\mathrm{In}\left[\frac{P(x)}{1-P(x)}\right] = \mathrm{In}\left[e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}\right] = \sum_i \alpha_1 x_1 + \alpha_{n+1} \quad (5)$$

The system (equation 5) is a linear algebraic model created to determine α:

$$bi = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \quad (6)$$

where $i = 1, 2, \ldots, m$.

Let $\bar{e} = (1, \ldots, 1)^T$ be a vector of *m* ones and $b = (b_1, b_2,\ldots, b_m)^T$. The system of linear equations (equation 6) may be represented by:

$$B\alpha = b, \text{ with } B = [\bar{e}\ A] \quad (7)$$

The solution of equation 7 is given by the solution of a least square problem but, in our case (equation 7), it has an infinite number of solutions, since $n + 1 >> m$. It is usual to circumvent this difficulty by suppressing the model and keeping only a small subset of the *n* attributes. This procedure resembles the feature selection in data mining, an open problem research area [35].

We approximate a solution of the combinatorial problem related to the feature selection solving the following:

$$\alpha = \textit{argument that minimize } f(\alpha) \quad (8)$$

$$\textit{where } f(\alpha) = \alpha^T\alpha + (B\alpha - b)^T * (B\alpha - b)$$

The solution of equation (8) is a convex unconstrained optimization problem [46]; after applying the optimality conditions, it can be shown that the optimal solution α* is such that it verifies the following system of linear equations

$$(I + B^TB)\,\alpha = B^Tb, \quad (9)$$

where I is an identity matrix of dimension n. We point out that in Golub [47], an identity matrix to solve the rank deficiency in systems of linear equations was used. One should note that the identity matrix in equation 9 does not allow the rank to become deficient.

It was observed that the optimal solution α^* of (equation 8) is unique. So, given a query $q = (q_1, q_2,\ldots, q_n)$, the probability of q be an active ligand, is given by

$$P(q) = g(q) / (1 + g(q)), \quad (10)$$

where, $g(q) = \exp([1\ q]\ \alpha)$.

On this step, the regression gives us a probability associated with each ligand. We would like to highlight the fact that modification in the logit function allows us to use more features than ligands.

2.3. Molecular docking

A docking protocol with DOCK6 [48] was performed to establish the use of the generated docking score as the binding energy. DOCK6 provides multiple scoring functions, from force-field-based to pharmacophore-based, and the possibility of combining them. Therefore, in our case, it was useful to incorporate chemical features of known inhibitors and the binding sites to the docking of the selected compounds. Bosutinib was also subject to docking simulations since no crystallographic structure of the complex CDK2-Boustinib is available. The docking poses of the selected drugs were analyzed with the help of the Discovery Studio Visualizer [49] that showed possible interactions with CDK2 active site.

3. RESULTS

3.1. Data collection, model building, and testing of the Milk-Way algorithm

Milk-Way method is divided into five steps: (A) selection of active and inactive ligands; (B) fingerprint construction, where ligand properties are collected and stored in a matrix; (C) noise reduction using singular value decomposition (SVD); (D) model construction based on ad hoc selection; and (E) prediction using a modified logistic regression, which selects ligands based on high values of P(x) (Figure 1).

The Milk-Way algorithm requires an initial input matrix of individuals to be capable of building a classification model. In this matrix, two sets of compounds can be found — compounds with experimentally tested activity towards the desired target and compounds without any evident activity. The latter set can be formed by confirmed inactives or artificially created compounds, the so-called decoys.

Interestingly, when the weights of the attributes (α_i values in equation 3) were analyzed, we could understand the impact that each feature had on the classification into active or inactive. Since we have a significant number of features, we expected that the highest α values corresponded to active compounds. Moreover, the inverse is also true – the lowest α values correspond to inactive compounds.

3.2. Application of the Milk-Way algorithm

For the construction of the model, we selected the only commercial drug that acts against CDK2 (Bosutinib [15]) as the active entity. Decoys were generated from DUD-E [50] based on Bosutinib and two commercialized drugs for CDK4 and CDK6 [16], Ribociclib, and Palbociclib. CDK4 and CDK6 are homologous proteins to CDK2 [1] (Supplementary material Table S1). We put them with the inactives to distinguish the effect of the homology between the enzymes. The same molecular features (Fragment Pair/Pharmacophore fingerprints) previously described were also used as attributes (Supplementary material Table S2). The screening was performed using commercialized drugs retrieved from the DrugBank [15]. The drugs which obtain a probability (P(x)) of 0.98 or

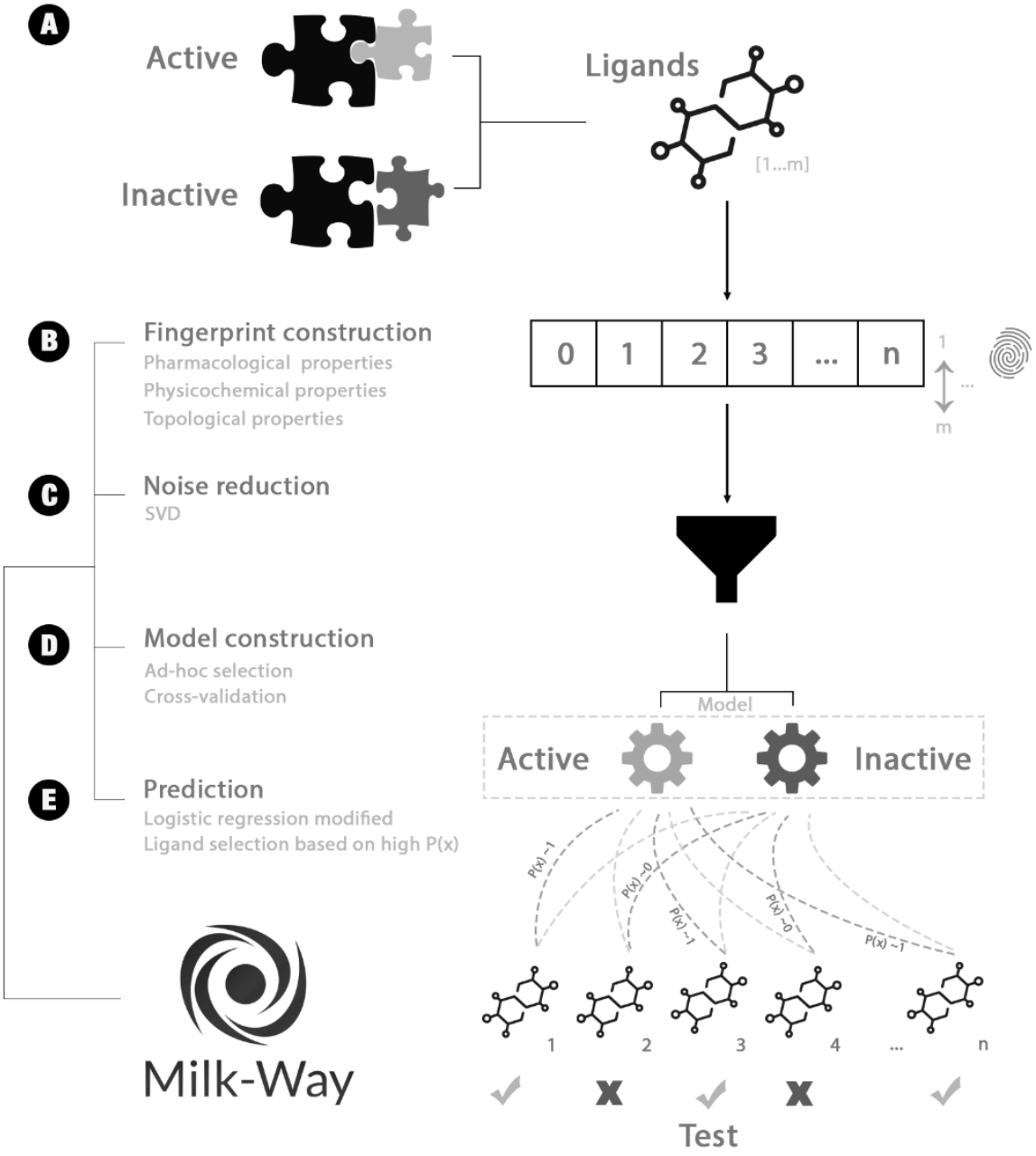


Figure 1. Workflow of the Milk-Way algorithm. It is divided into five steps: **(A)** selection of active and inactive ligands; **(B)** fingerprint construction; **(C)** noise reduction; **(D)** model construction; and **(E)** prediction. The input of the algorithm is a matrix of ligands to a specific target describe through descriptors. After the Singular Value Decomposition, the training model is ready either to calculate the probability of new binders or for repositioning marketed drugs by ad hoc selection. The output is the suggestion of new compounds to be used as ligands to the target.

higher were selected. In total, five drugs were selected: pramocaine, prochlorperazine, trifluoperazine, methionine and pergolide (Supplementary material Table S3).

We analyzed these drugs through molecular docking to probe a possible complementarity between the CDK2 structure and them. Molecular docking provided a rank through the chosen

scoring function, prioritizing three compounds with predicted high affinity besides bosutinib (-26.17 kcal/mol): pramocaine (-30.83 kcal/mol), trifluoperazine (-23.67 kcal/mol), and prochlorperazine (-17.87 kcal/mol) (Supplementary material Table S4). Interestingly, pramocaine [51], prochlorperazine [52], and trifluoperazine [52, 53] are described in the literature to act on CDK2-related diseases such as Glioblastoma, breast cancer, and other tumor effects (Supplementary material Table S3).

The binding modes from docking showed that all compounds fitted very well in the CDK2 active site (Supplementary material Table S5) and interacted with key residues in the active site (Supplementary material Table S6). For instance, bosutinib binding mode was complementary to the active site (Figure 2a), and it achieved interactions with the same residues as the inhibitor present in the crystal structure, such as hydrogen bond interaction with LYS 33, hydrophobic interactions with VAL 18 and GLN 120 (Figure 2b). Similar behavior was observed with molecular docking highest-ranked compound, pramocaine (Figure 3a). Although pramocaine showed less interaction than bosutinib in the CDK2 binding site (Figure 3b), the molecular docking binding mode of this compound presented the same interacted residues as the crystal inhibitor (ILE 10, VAL 18, LYS 33, LEU 72, GLN 120, and LEU 123). The presence of these interactions from known inhibitors might indicate a possible binding of the predicted compounds.

4. DISCUSSION

Machine learning approaches are robust methodologies capable of screening drug leads from a dataset of many compounds with reasonable accuracy in a faster and cheaper manner than experimental testing. Most methods act as a classifier, separating the compounds into actives and inactives. To accomplish this classification, a model is created and validated through training sets using known actives to a specific target [12]. Milk-Way provides a novel and alternative approach to machine learning using a combination of data-mining techniques.

An interesting characteristic of Milk-Way is the usefulness of alpha values (equation 3) to perform the stratified feature selection. The positive values of the components α_i of α contribute to approximate $P(\alpha)$ to 1, and the negative values approximate $P(\alpha)$ to 0. Absolute values of α_i close to zero don't have significant impact in the

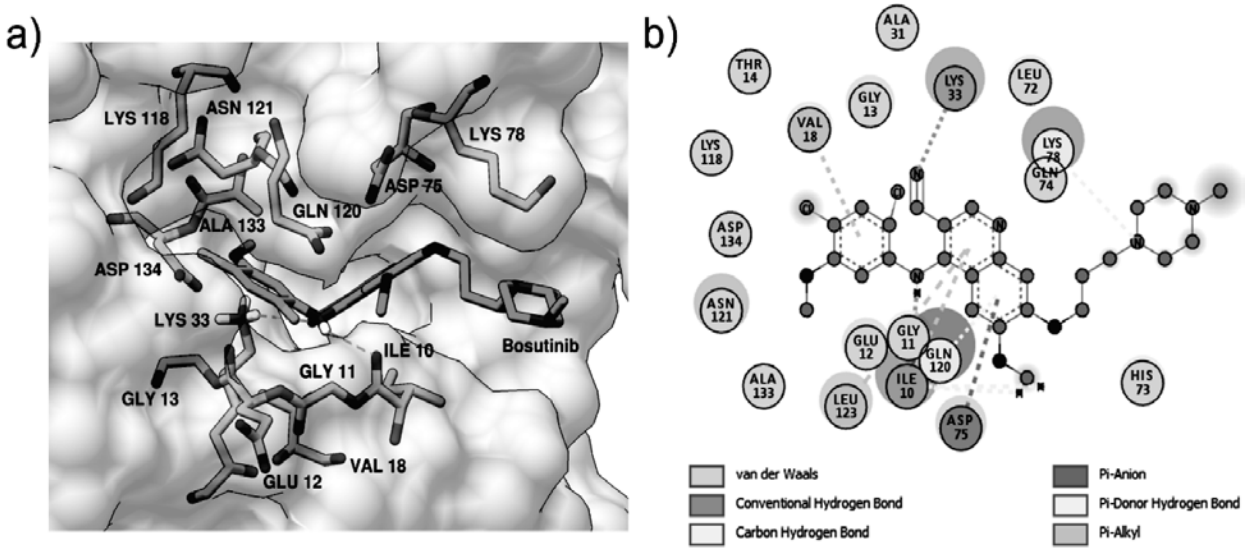


Figure 2. Bosutinib molecular docking binding mode. **a)** Bosutinib in the active site of CDK2 with the interacting residues and hydrogen bond interactions. **b)** 2D representation of the interactions between Bosutinib and the active site of CDK2.

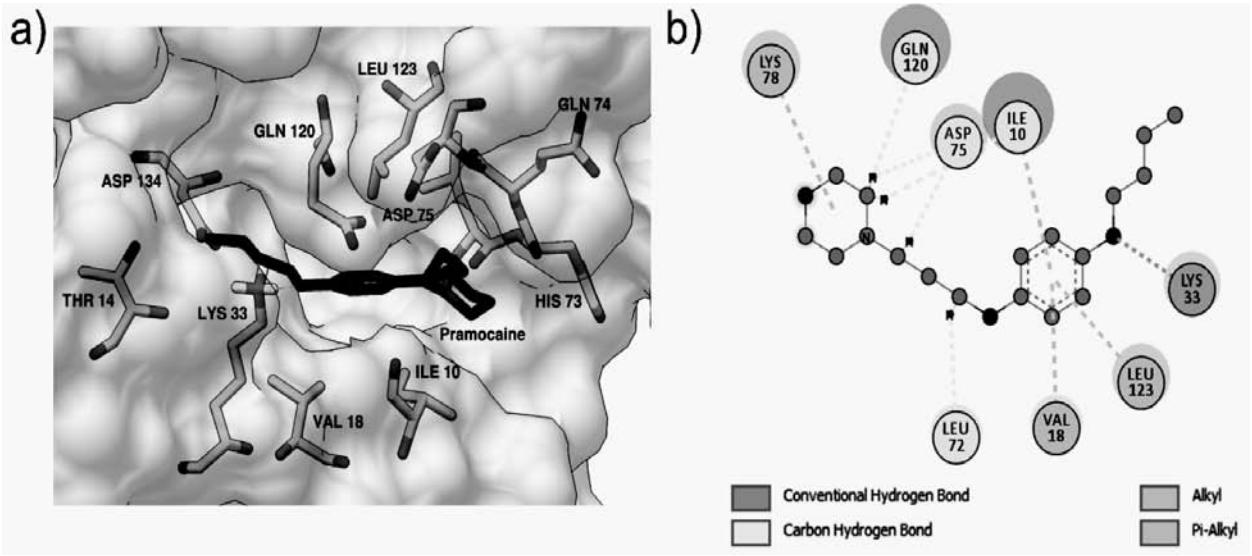


Figure 3. Pramocaine molecular docking binding mode. **a)** Pramocaine in the active site of CDK2 with the interacting residues and hydrogen bond interactions. **b)** 2D representation of the interactions between Pramocaine and the active site of CDK2.

computation of P(α). The modified logistic regression can identify the highest alpha values, which carried critical molecular features present on active compounds (Supplementary material Table S2). Low alpha values were usually observed in inactive compounds. Therefore, one advantage of the Milk-Way algorithm is that more features than compounds can be processed since the rank of the matrix is the one used in the calculations, not the entire matrix.

We performed a simulation study that showed the combination of our method to other in silico techniques. Since our algorithm is ligand-based, the ligands selected can benefit from molecular docking, a structure-based approach, when a structure of the target is known. The docking scores (Supplementary material Table S4) and interaction analysis can provide a better atomistic understanding of a possible inhibition of the CDK2 enzyme by the selected compounds (Supplementary material Tables S5 and S6). For instance, the highest scored compound pramocaine displayed all the interactions presented in another known CDK2 inhibitor (ILE 10, VAL 18, LYS 33, LEU 72, GLN 120, and LEU 123), indicating a possible strong binding of this compound. Furthermore, three of the five selected ligands (Pramocaine, Prochlorperazine, and Trifluoperazine) by the Milk-Way to repurpose, have already been described in the literature to act on CDK2-related diseases (Supplementary material Table S3) [51-53].

5. CONCLUSIONS

The development of new drugs takes about fourteen years, and the cost ranges are estimated at around 1.0 billion USD [54]. With the HTS method, it is possible to screen millions of compounds against a chosen target experimentally. However, it is an expensive process, and therefore lower-cost alternatives capable of sorting promising compounds from several other ones are sought out. These help to decrease the number of ligands to be tested in a possible experimental phase.

Our holistic approach can classify ligands with the support of the selected case studies. The use of literature data and datasets were appropriate for testing the algorithm and measuring the results. It is essential to notice, the cases investigated throughout the paper are unrelated to each other, demonstrating a practical way to prove the efficiency of the proposed algorithm for LBVS. Nevertheless, it is essential to highlight the acceptance of a higher number of attributes (ligands' features) than entities (ligands), without

a problem of rank deficiency. This added factor is opposed to the classical logistic regression in which it is obligatorily to have more entities than attributes.

The proposed mathematical modulation does not require a massive infra-structure apparatus to be performed and constitutes a good strategy for the selection of promising compounds. The Milk-Way algorithm demonstrated excellent performance, and with less computational infrastructure. For a more in-depth and broader study of this algorithm, we are already applying it to other targets and other data-sets. Since there is an attempt to continually improve the efficiency of computational processes for the development of new and repositioning drugs, we proposed a robust approach that provides a general classifier to separate actives from inactives present in a dataset of ligands for any data-driven LBVS.

The Milk-Way algorithm already has an associated patent BR 10 2019 027703 3 [55].

AUTHOR CONTRIBUTIONS

Methodology, CFVL, MAS; Data curation, CFVL; Investigation, CFVL; Writing original draft, CFVL, LHS; Formal analysis, CFVL, LHS, LFL; Software, CFVL, LHS, LFL; Validation, CFVL, LHS, LFL; Project Administration, MAS; Resources, MAS; Supervision, MAS; Funding Acquisition, MAS; Writing - review & editing: CFVL, LHS, LFL, DCBM, REOR, and MAS.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

ACKNOWLEDGMENTS

The authors acknowledge Pedro Magalhães Martins for technical support.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ABBREVIATIONS

CDK2 : cyclin-dependent kinase 2
HTS : High-throughput screening
LBVS : ligand-based virtual screening
RTI : reverse transcriptase inhibitors
SBVS : structure-based virtual screening

SUPPLEMENTARY MATERIAL

CDK2 model

Data collection

The matrix was composed of drugs described through zeros and ones, done by the Power MV [31]. The descriptors generated were: 735 fragment pair and 147 Pharmacophore fingerprints [32-34]. They are a topological representation of a chemical structure of ligands, some of which have already been used in data mining [26]. This model was based on the only drug commercialized to CDK2, bosutinib (DB06616), used for chronic myelogenous leukemia [56]. The inactives were made by generating 50 decoys, from DUD-E [50] of this drug and using two commercialized drugs for CDK4/6 and their respective decoys. Ribociclib (DB11730) and palbociclib (DB09073) inhibit tumor growth across a diversity of retinoblastoma cancers (Rb+) [57, 58]. We assign drugs for CDK4/6 in the group of inactive ligands for the sake of guaranteeing the specificity for CDK2, since the fact that they belonged to the same family could generate false positives, due to their homology. It has already been proven that CDK2 is structurally and functionally related to CDK1. Also, CDK2 has a considerably broader substrate profile than CDK4 and CDK6 [1].

As we are working with drug repositioning, the matrix of queries consisted of drugs already marketed according to the DrugBank database [16]. Each ligand was described through the same 882 descriptors of the model.

Singular value decomposition

The singular value decomposition (SVD) helps to reduce the dimension of the matrix. All the eigenvalues were chosen based on the Everitt *et al.* criteria [42].

Ad-hoc choice

It is also essential to analyze the P(x) of the drug which acts in CDK4/6, inferring about the specificity of the model, in the family of CDK, due to the fact of homology between then.

Table S1. Number of ligands and features of the training matrix CDK2 and the result of singular value decomposition.

		882*
Original matrix	Active	1
	Inactive	152
SVD	Sum Sr**	0.7005
	Eigenvalue	61
Each training model	Active	1
	Inactive	152

*number of features (fragment pair and Pharmacophore fingerprints).
**Sr = (diagonal(S)/sum(S)) [42].

Table S2. The 1% features with the highest alpha value.

Alpha features
“ARC_06_-O-”
“HY1_03_HY2”
“POS_04_POS”
“POS_04_POS”
“HY1_04_HY1”
“ARC_04_-O-”
“POS_06_-O-”
“HAL_05_HY2”

Alpha values analysis

With the system of equations represented in equation 3, the alpha value translates the impact that each feature has on the categorization of the active or inactive ligand. To get an analysis of the alpha values and how the validation model is categorizing, we compared the 1% of the highest alphas of the model. By the logic and rigor of the equation, it is expected that the higher alpha values belong to the active reference active bosutinib.

Molecular docking protocol

Molecular docking can provide a better understanding of the interactions between a target macromolecule when the target structure is known. Docking begins by sampling different orientations and conformations of the ligand within the target binding site [59, 60]. Afterward, the best positions, the so-called pose, for each ligand are determined by ranking them according to a scoring function [61]. With this strategy, we can predict a possible affinity between ligand and target. We chose DOCK6.8 [48] since it provides multiple scoring functions, from force-field based to pharmacophore-based, and the possibility of combining them.

Therefore, in our case it was useful to incorporate chemical features of known inhibitors and the binding sites to the docking of the selected compounds. We chose a protocol that calculates

grid parameters to every residue interacting with the reference, the so-called Multigrid energy score (MGE). The sum of the interactions in each grid equals the interaction of a single grid representing the entire target.

Since DOCK6.8 allows the combinations of score functions, we also included the Pharmacophore Matching Similarity (PHS) score combined with the MGE. PHS is a scoring function that calculates the level of pharmacophore overlap between a reference molecule and a candidate molecule in three-dimensional space. Since MGE already uses a reference molecule, we found that including its pharmacophore overlap in the score component would be useful.

To perform the docking of the compounds chosen by the Milk-Way algorithm, we chose the 2R3Q [62] structures, since it achieved good results in pose reproduction and cross-docking experiments (data not showed). We used its native ligand as reference for the MGE and PHS scoring. OpenBabel [63] was used to convert the ligands from SMILES to the mol2 format. Geometry optimization was performed for all ligands using GAMESS [64], while AM1-BCC partial charges were assigned using AMBER's antechamber [65].

Table S3. The references of the selected ligands (P(x) > = 0.98).

CID	Name	P(x)	Citation	Reference
DB09345	Pramocaine	0.9970	Pramocaine induced expression changes in 'Signaling Pathways in Glioblastoma'	[51]
DB00433	Prochlorperazine	0.9919	Drugs with potential antitumor effects	[52]
DB00831	Trifluoperazine	0.9875	Trifluoperazine might be a potential available drug for treating triple-negative breast cancer with brain metastasis, which urgently needs novel treatment options	[52, 53]
DB00134	Methionine	0.9830	-	-
DB01186	Pergolide	0.9813	-	-

Table S4. Values of energy using DOCK scoring MGE + PHS.

Name:	Bosutinib	Pramocaine	Prochlorperazine	Trifluoperazine	Methionine	Pergolide
Reference PDB:	**PDB 2R3Q**					
Descriptor_Score:	-26.16698	-30.83058	-17.86901	-23.67142	-8.72689	-14.40792
MGE_Score:	-30.64930	-35.99844	-21.49072	-28.90000	-13.62098	-19.95671
PHS_Score:	4.954826	5.610463	4.244604	5.756922	5.129757	5.86296

Table S5. Binding mode of the selected ligands with P(x) > = 0.98. The system used was DOCK scoring – MGE + PHS SCORE – using PDB 2R3Q.

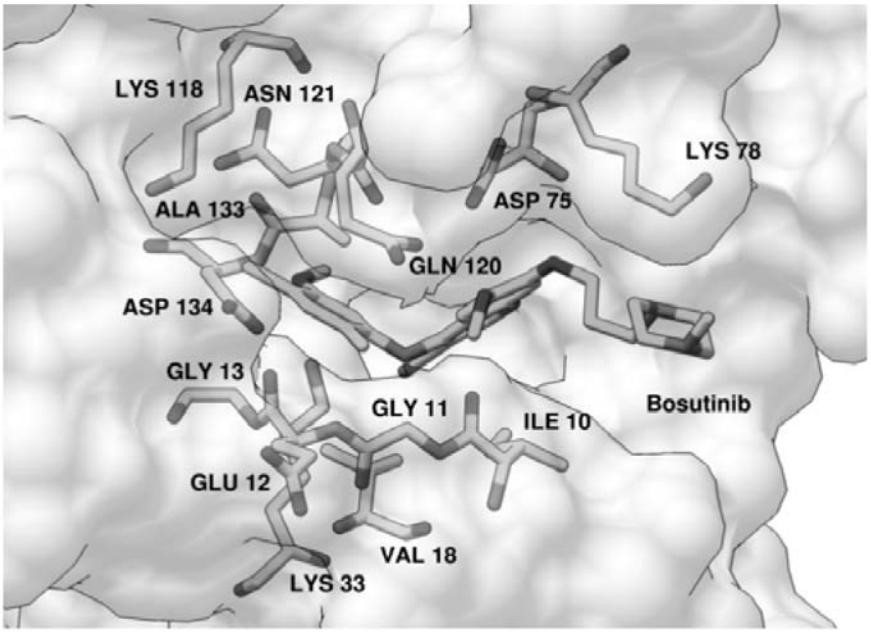
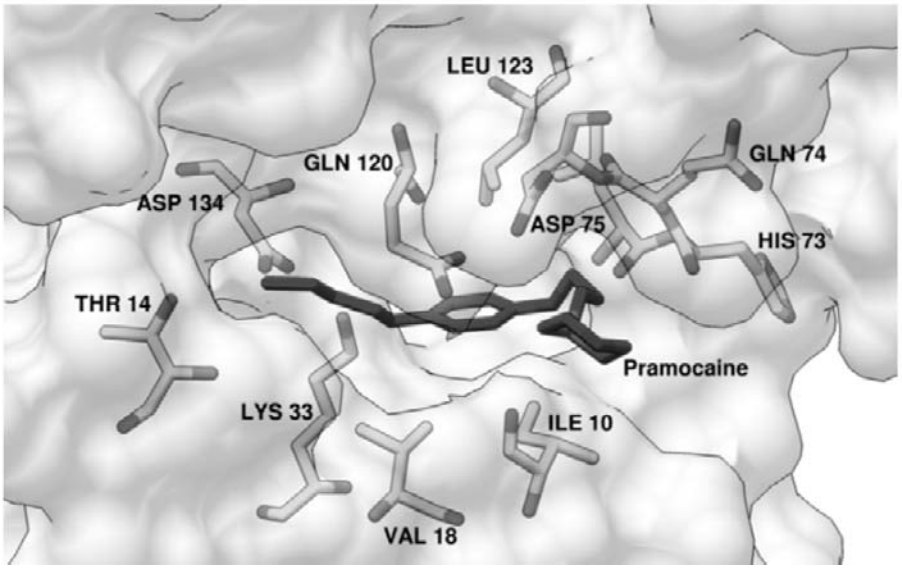
CID	Name	P(x)
DB06616	Bosutinib	-
INTERACTIONS*		
DB09345	Pramocaine	0.9970
INTERACTIONS*		

Table S5 continued..

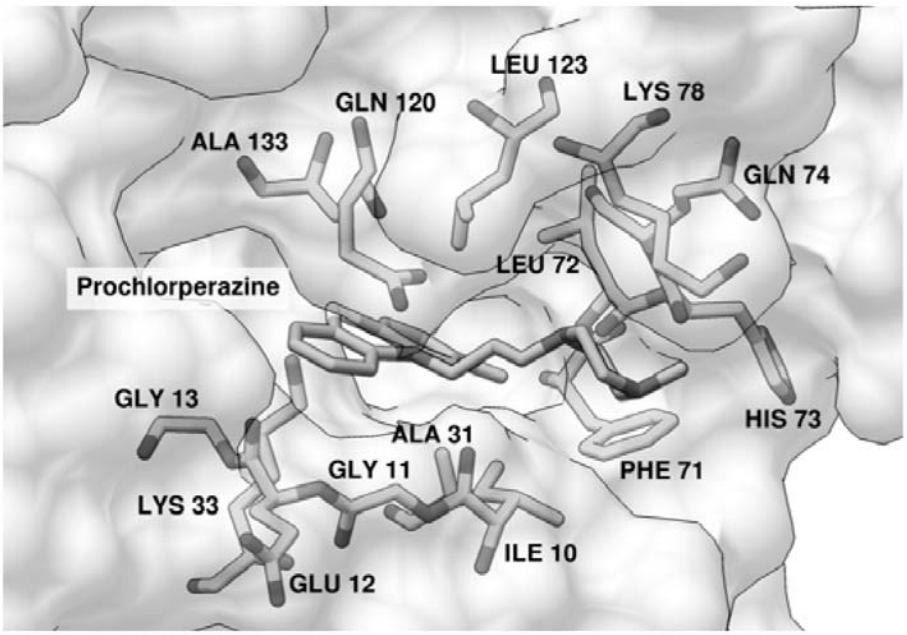
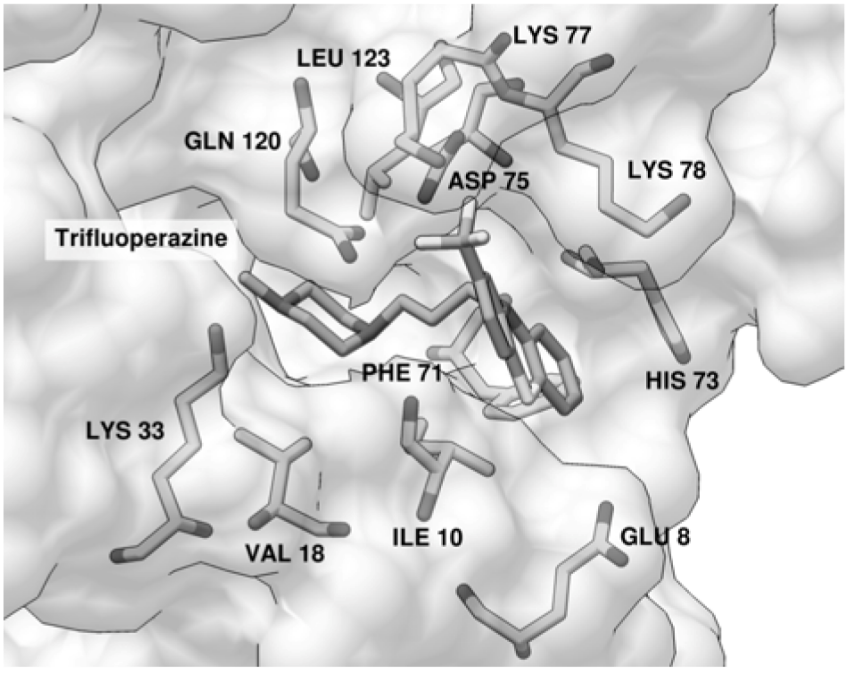
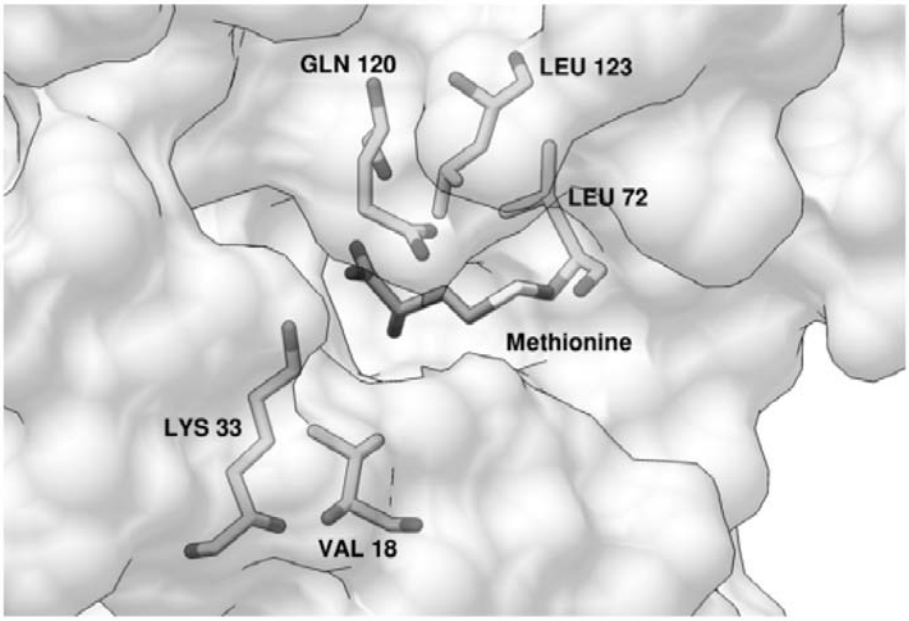
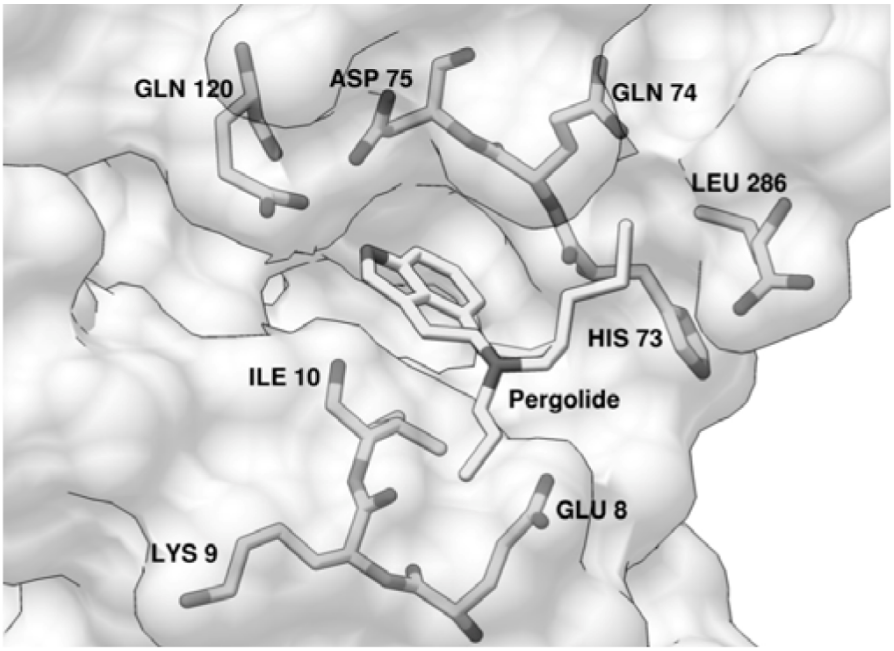
CID	Name	P(x)
DB00433	Prochlorperazine	0.9919
INTERACTIONS*		
DB00831	Trifluoperazine	0.9875
INTERACTIONS*		

Table S5 continued..

CID	Name	P(x)
DB00134	Methionine	0.9830
INTERACTIONS*		
DB01186	Pergolide	0.9813
INTERACTIONS*		

*All images were generated with UCSF Chimera [66].

Table S6. Interactions of the selected ligands with P(x) > = 0.98. The system used was DOCK scoring – MGE + PHS SCORE – using PDB 2R3Q.

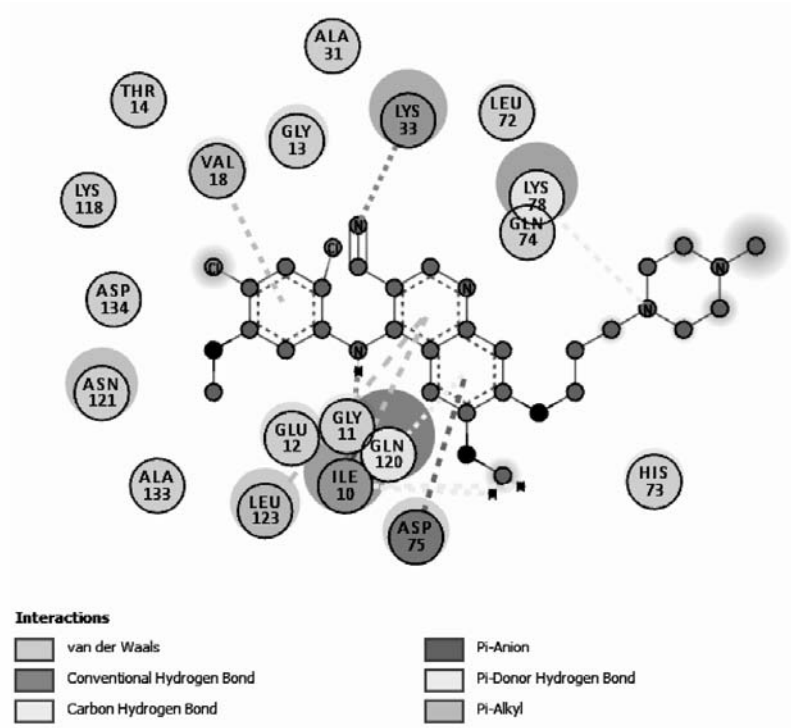
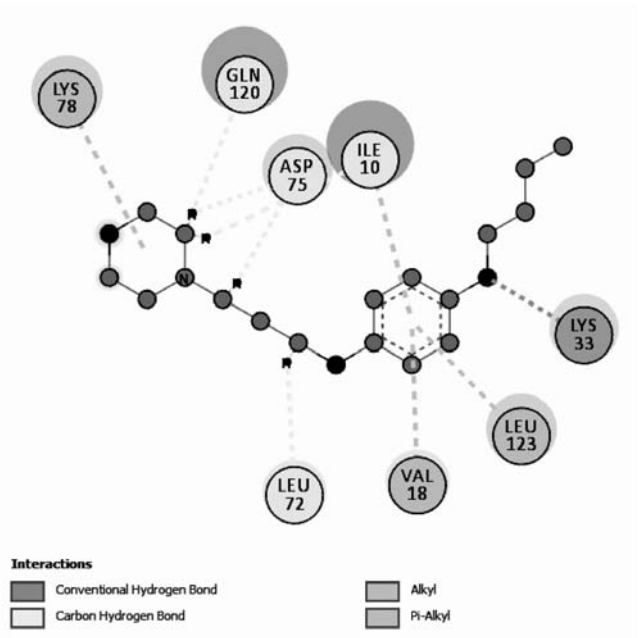
CID	Name	P(x)
DB06616	Bosutinib	-
INTERACTIONS*		
DB09345	Pramocaine	0.9970
INTERACTIONS*		

Table S6 continued..

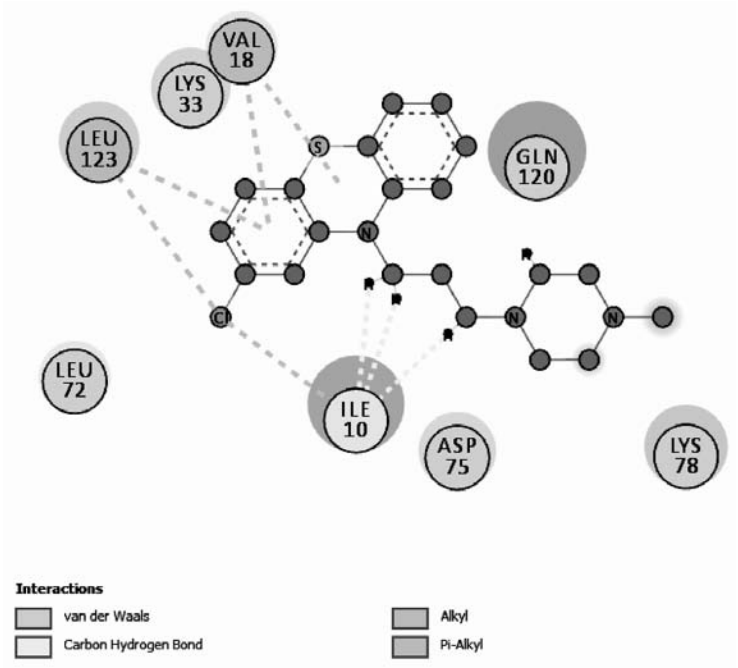
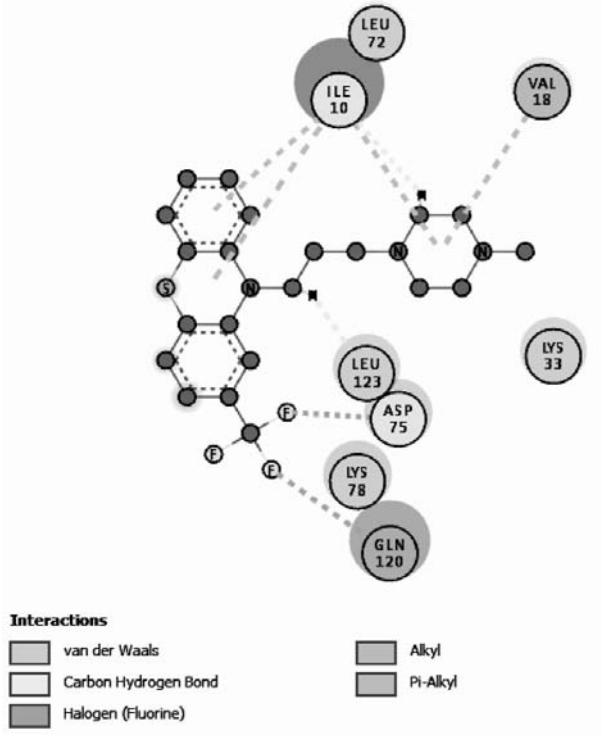
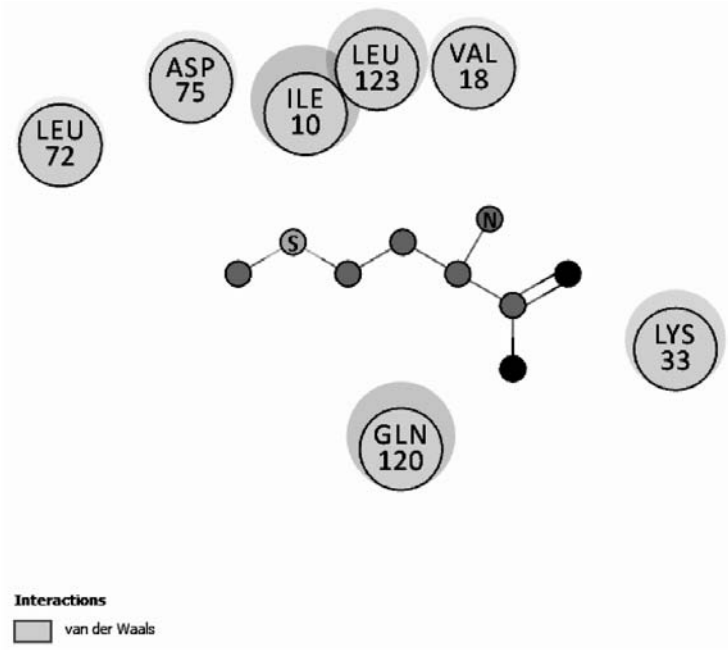
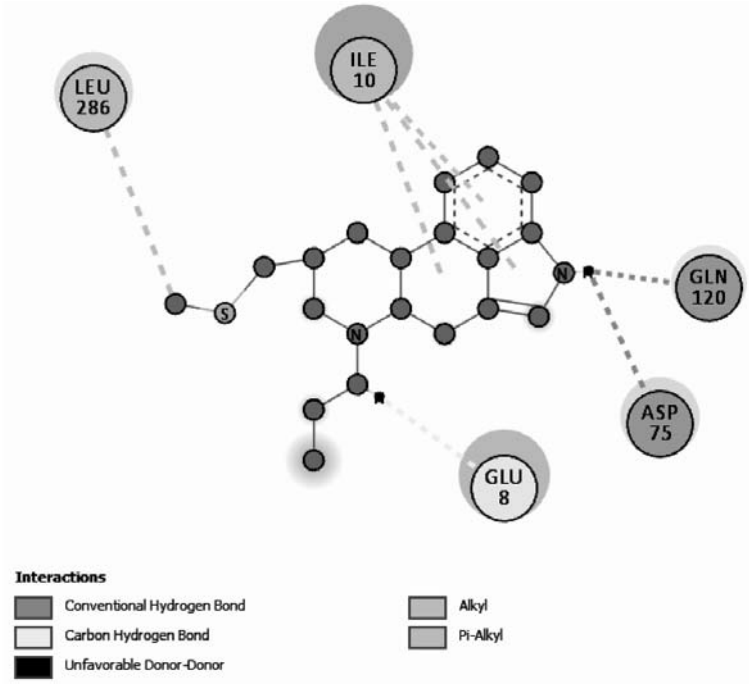
CID	Name	P(x)
DB00433	Prochlorperazine	0.9919
INTERACTIONS*		
DB00831	Trifluoperazine	0.9875
INTERACTIONS*		

Table S6 continued..

CID	Name	P(x)
DB00134	Methionine	0.9830
INTERACTIONS*		
DB01186	Pergolide	0.9813
INTERACTIONS*		

*All images were generated in Discovery Studio [49].

REFERENCES

1. Asghar, U., Witkiewicz, A. K., Turner, N. C. and Knudsen, E. S. 2015, Nature Reviews. Drug Discovery, 14(2), 130.
2. Kaldis, P. and Richardson, H. E. 2012, Development, 139(2), 225-230.
3. Betzi, S., Alam, R., Martin, M., Lubbers, D. J., Han, H., Jakkaraj, S. R., Georg, G. I. and Schönbrunn, E. 2011, ACS Chemical Biology, 6(5), 492-501.
4. Johnson, L. 2007, Biochemical Society Transactions, 35(1), 7.
5. Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N. and Andrade, C. H. 2018, Frontiers in Pharmacology, 9, 1275.
6. Seifert, M. H., Wolf, K. and Vitt, D. 2003, Biosilico, 1(4), 143-149.
7. Li, Q. and Shah, S. 2017, Methods Mol. Biol., 1558, 111-124.
8. Stockwell, B. R. 2004, Nature, 432(7019), 846-854.
9. Ceska, T., Chung, C.-W., Cooke, R., Phillips, C. and Williams, P. A. 2019, Biochemical Society Transactions, 47(1), 281-293.
10. Gimeno, A., Ojeda-Montes, M. J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G. and Garcia-Vallvé, S. 2019, International Journal of Molecular Sciences, 20(6), 1375.
11. Geppert, H., Vogt, M. and Bajorath, J. 2010, Journal of Chemical Information and Modeling, 50(2), 205-216.
12. Carpenter, K. A. and Huang, X. 2018, Curr. Pharm. Des., 24, 3347-3358.
13. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. 2008, Nucleic Acids Research, 36(Suppl. 1), D344-D350.
14. Irwin, J. J. and Shoichet, B. K. 2005, Journal of Chemical Information and Modeling, 45(1), 177-182.
15. Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J., Xiao, J., Zhang, J. and Bryant, S. H. 2010, Nucleic Acids Research, 38(Suppl. 1), D255-D266.
16. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., and Neveu, V. 2011, Nucleic Acids Research, 39(Suppl. 1), D1035-D1041.
17. Sharman, J. L., Mpamhanga, C. P., Spedding, M., Germain, P., Staels, B., Dacquet, C., Laudet, V. and Harmar, A. J. 2011, Nucleic Acids Research, 39(Suppl. 1), D534-D538.
18. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. 2010, Nucleic Acids Research, 38(Suppl. 1), D355-D360.
19. Soufan, O., Ba-alawi, W., Magana-Mora, A., Essack, M. and Bajic, V. B. 2018, Scientific Reports, 8(1), 9110.
20. Plewczynski, D., Spieser, S. A. H. and Koch, U. 2006, Journal of Chemical Information and Modeling, 46(3), 1098-1106.
21. Han, L., Wang, Y. and Bryant, S. H. 2008, BMC Bioinformatics, 9(1), 1.
22. Riniker, S., Wang, Y., Jenkins, J. L. and Landrum, G. A. 2014, Journal of Chemical Information and Modeling, 54(7), 1880-1891.
23. Chen, B., Wild, D. and Guha, R. 2009, Journal of Chemical Information and Modeling, 49(9), 2044-2055.
24. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. and Hopkins, A. L. 2006, Nature Biotechnology, 24(7), 805-815.
25. Hao, M., Wang, Y. and Bryant, S. H. 2014, Analytica Chimica Acta, 806, 117-127.
26. Schierz, A. C. 2009, Journal of Cheminformatics, 1, 21.
27. Trunk, G. V. 1979, IEEE Transactions on Pattern Analysis & Machine Intelligence, 1(3), 306-307.
28. Dai, W. and Guo, D. 2019, Molecules, 24(13), 2414.
29. Yin, L., Ge, Y., Xiao, K., Wang, X. and Quan, X. 2013, Neurocomputing, 105, 3-11.
30. Diallo, A. and Prigent, C. 2011, Bull Cancer, 98(11), 1335-1345.
31. Liu, K., Feng, J. and Young, S. S. 2005, Journal of Chemical Information and Modeling, 45(2), 515-522.
32. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S. and Pujadas, G. 2015, Methods, 71, 58-63.

33. Ahmed, H. E., Vogt, M. and Bajorath, J. R., 2010, Journal of Chemical Information and Modeling, 50(4), 487-499.
34. Awale, M. and Reymond, J. L. 2014, Journal of Chemical Information and Modeling, 54(7), 1892-1907.
35. Liu, H., Motoda, H., Setiono, R. and Zhao, Z. 2010, Feature Selection in Data Mining, 4-13.
36. Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., Shoemaker, B. A., Gindulyte, A. and Bryant, S. H. 2014, Nucleic Acids Research, 42(D1), D1075-D1082.
37. Élden, L. 2006, Acta Numerica, 15.
38. Berry, M. W., Dumais, S. T. and OBrien, G. W. 1995, SIAM Review, 37.
39. Santos, A. R., Santos, M. A., Baumbach, J., McCulloch, J. A., Oliveira, G. C., Silva, A., Miyoshi, A. and Azevedo, V. 2011, BMC Genomics, 12(4), 1-15.
40. Kumar, N., Nasser, M. and Sarker, S. C. 2011, Journal of Geography and Geology, 3(1), 227.
41. Silverio-Machado, R., Couto, B. R. and Dos Santos, M. A. 2014, Bioinformatics, 31(8), 1267-1273.
42. Everitt, B. S. D., Everitt, G. B. S. and Dunn, G. 1991, Applied Multivariate Data Analysis.
43. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990, Journal of the American Society for Information Science, 41(6), 391.
44. Morgan, S. P. 1998, Notices Amer. Math. Soc., 45(8), 972-977.
45. Cokluk, O. 2010, Educational Sciences: Theory and Practice, 10(3), 1397-1407.
46. Luenberger, D. G. and Ye, Y. 2016, Linear and Nonlinear Programming, Springer.
47. Golub, G. H. 1965, Numerical Mathematics, 7(3), 206-216.
48. Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D. and Rizzo, R. C. 2015, Journal of Computational Chemistry, 36(15), 1132-1156.
49. Biovia, D. S. 2017, Discovery Studio Modeling Environment, Dassault Systèmes: San Diego.
50. Mysinger, M. M., Carchia, M., Irwin, J. J. and Shoichet, B. K. 2012, Journal of Medicinal Chemistry, 55(14), 6582-6594.
51. Hardt, C., Beber, M. E., Rasche, A., Kamburov, A. and Herwig, R. 2019, ToxDB. Vertebrate Genomics Department at the Max Planck Institute for Molecular Genetics in Berlin, Germany.
52. Qi, L. and Ding, Y. 2013, Science China Life Sciences, 56(11), 1020-1027.
53. Feng, Z., Xia, Y., Gao, T., Xu, F., Lei, Q., Peng, C., Yang, Y., Xue, Q., Hu, X., Wang, Q., Wang, R., Ran, Z., Zeng, Z., Yang, N., Xie, Z. and Yu, L. 2018, Cell Death & Disease, 9(10), 1006.
54. Ou-Yang, S.-S., Lu, J.-Y., Kong, X.-Q., Liang, Z.-J., Luo, C. and Jiang, H. 2012, Acta Pharmacologica Sinica, 33(9), 1131-1140.
55. Figueiredo Vieira Leite, C., Dos Santos, M. A., Silva Dos Santos, L. H., Fernando Leijôto, L., Batista Mariano, D. C. and Oliveira Rocha, R. E. 2019, Método de Triagem de compostos baseados em Regressão Logística Modificada.
56. Amsberg, G. K. and Schafhausen, P. 2013, Biologics, 7, 115-122.
57. Infante, J. R., Cassier, P. A., Gerecitano, J. F., Witteveen, P. O., Chugh, R., Ribrag, V., Chakraborty, A., Matano, A., Dobson, J. R., Crystal, A. S., Parasuraman, S. and Shapiro, G. I. 2016, Clinical Cancer Research, 28(23), 5696-705.
58. DeMichele, A., Clark, A. S., Tan, K. S., Heitjan, D. F., Gramlich, K., Gallagher, M., Lal, P., Feldman, M., Zhang, P., Colameco, C., Lewis, D., Langer, M., Goodman, N., Domchek, S., Gogineni, K., Rosen, M., Fox, K. and O'Dwyer, P. 2015, Clinical Cancer Research, 21(5), 995-1001.
59. Meng, X. Y., Zhang, H. X., Mezei, M. and Cui, M. 2011, Current Computational Aided Drug Design, 7(2), 146-157.
60. Yuriev, E. and Ramsland, P. A. 2013, Journal Molecular Recognition, 26(5), 215-239.
61. Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T. and Altman, R. B. 2012, Journal of the Royal Society Interface, 9(72), 1409-1437.

62. Fischmann, T. O., Hruza, A., Duca, J. S., Ramanathan, L., Mayhood, T., Windsor, W. T., Le, H. V., Guzi, T. J., Dwyer, M. P., Paruch, K., Doll, R. J., Lees, E., Parry, D., Seghezzi, W. and Madison, V. 2008, Biopolymers, 89(5), 372-379.
63. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R. 2011, Journal of Cheminformatics, 3(1), 33.
64. Gordon, M. S. and Schmidt, M. W. 2005, Advances in electronic structure theory: GAMESS a decade later. Theory and Applications of Computational Chemistry: the first forty years, C. E. Dykstra, G. Frenking, K. S. Kim and G. E. Scuseria (Eds), Elsevier, 1167-1189.
65. Wang, J., Wang, W., Kollman, P. A. and Case, D. A. 2006, Journal of Molecular Graphics and Modelling, 25(2), 247-260.
66. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. 2004, Journal of Computational Chemistry, 25(13), 1605-1612.

4.3. Draft of a paper: Stratified feature selection with Milk-Way algorithm applied in the MUV database: a study case

Stratified feature selection with Milk-Way algorithm applied in the MUV database: a study case

Carmelina Figueiredo Vieira Leite[1§], Luan Carvalho Martins[2§], Larissa Fernandes Leijôto[1], Pedro Magalhães Martins[1], Lucianna Helene Silva Santos[2], Marcos Augusto dos Santos[3]* and Rafaela Salgado Ferreira[2]

[1] *Laboratory of Bioinformatics and Systems, Institute of Biological Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

[2] *Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil*

[3] *Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

§ Equal contribution

*Address correspondence to:

Marcos Augusto dos Santos, PhD

Departamento de Ciências da Computação

Instituto de Ciências Exatas

Universidade Federal de Minas Gerais

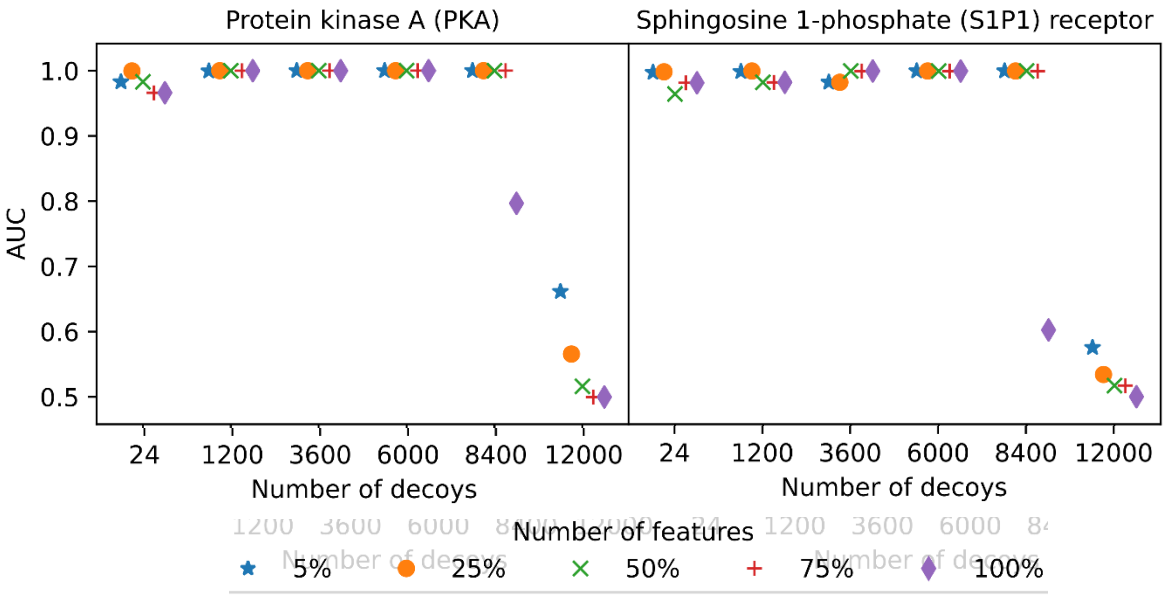
Belo Horizonte, MG, Brazil

Tel: +55 31 3409 5896

E-mail: marcos@dcc.ufmg.br

Abstract

Virtual screening (VS) is widely applied in drug development to score, rank, and filter a set of molecules of biological interest, using computational methods. The aim of a VS is prioritizing compounds from an extensive library so that the selected molecules should have an enhanced probability of binding to a target. In ligand-based VS, descriptors represent molecular features, which are used to calculate the molecular similarity between compounds with known and unknown activity. Some features are essential and can enhance the discrimination between binders and non-binders, while other features are not relevant, increasing computational time while adding noise to the model. Herein, we applied the stratified feature selection method to the Maximum Unbiased Validation (MUV), a challenging dataset designed to test VS methods. The stratified feature selection ranks the most and least relevant features according to their weight in the model. We tested the ECFP4, Atom Pairs, and rdkit's hashed topological fingerprints to represent molecular features. The MUV dataset was investigated using five machine learning-based classifiers: Classic Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and the Milk-Way method, a modified logistic regression developed by our group. Stratified feature selection combined with Milk-Way, SVM, and LR improved the sensitivity in these methods on most of the 17 MUV targets, outperforming previously reported results. The stratified feature selection showed a useful non-recursive way to improve the metrics, having low computational cost.



Abstract plot – Two examples of used targets, classified by Milk-Way algorithm, showing that performance of algorithm reach better AUC with the stratified feature selection than using more compounds.

KEYWORDS

ligand-based virtual screening; logistic regression; classifier; algorithm; bioinformatics; machine learning; virtual screening; drug discovery; drug mining; data mining

LIST OF ABBREVIATIONS

AUC	Area under curve of Receiver Operating Characteristics Curve
FP	Fingerprint
HTVS	High-Throughput Virtual Screening
LR	Logistic Regression
LBVS	Ligand-Based Virtual Screening
MUV	Maximum Unbiased Validation
NB	Naïve Bayes
RF	Random Forest
SVM	Support Vector Machine

1. Introduction

In ligand-based-virtual screening (LBVS), knowledge of binders – and, if available, non-binders – of a target is used to classify a broad set of molecules. LBVS requires no information about the target, so it is useful when the target is unknown or when experimental structures of it are absent. Several methodologies have been applied in LBVS, which can be roughly classified in (a) methods that construct a mathematical model to predict biological activity from a set of molecular structures; and (b) methods that use 2D or 3D similarities to known binders (Sliwoski *et al.*, 2014). e-Pharmacophores (Seidel *et al.*, 2017) and LigandScout (Wolber and Langer, 2005), are examples of the former, using pharmacophores, i.e., three-dimensional representations of chemical features that are possibly relevant for activity (Roy *et al.*, 2015). Examples of the latter are ROCS (Hawkins *et al.*, 2007) and OptiPharm (Puertas-Martín *et al.*, 2019), that apply shape-based similarity methods to score molecules. In shape similarity, the global shape of a queried molecule is compared to a database of potential drug candidates. Other similarity-based methods calculate the similarities of each query molecule to be classified to known ligands and use that information to construct a classifier. In general, the similarity and dissimilarity between molecules are assessed using fingerprints (FPs), bit-reduced representation of the molecule that can be compared using extremely fast bit-wise comparison methods (Cereto-Massagué *et al.*, 2015).

Different machine learning (ML) methods have been explored to construct classifiers for LBVS, as Naïve Bayes classifiers (Bender *et al.*, 2005), k-nearest neighbors (Luo *et al.*, 2016), artificial neural networks (Carpenter and Huang, 2018), supporting vector machines (SVM) (Subramaniam *et al.*, 2011), and other methods (Ma *et al.*, 2009; Mballo and Makarenkov, 2010; Dai and Guo, 2019). The development of a classifier method characterized by its high speed, straightforward methodology for model construction and yielding good results, would be of great interest to the LBVS community. Finally, the possibility to recover desirable and undesirable characteristics and correlating them to the chemical structure can be very useful in drug discovery. ML methods are usually assessed by data sets with known actives and known or artificial inactives. Their performance depends on the chosen validation data sets and the choice of molecular descriptors to calculate the similarity (Riniker and Landrum, 2013).

For instance, Riniker *et al.**(2013)*** investigated the performance of heterogeneous classifier fusion in LBVS using three machine learning methods: Classical Logistic Regression (LR), Naïve Bayes (NB) and Random Forest (RF), and four standard 2D FPs. The methods were compared using benchmark data sets, with challenging targets filtered to provide difficulty, from three different public sources: the directory of useful decoys (DUD)(Huang *et al.*, 2006; Irwin, 2008; Jahn *et al.*, 2009), the maximum unbiased validation data sets (MUV) (Rohrer and Baumann, 2009), and a selection of targets from ChEMBL(Gaulton *et al.*, 2011). All the methods performed poorly in targets of the MUV dataset, with an average area under the ROC curve (AUC) value of 0.67, although the same methods had good performances on DUD (Irwin, 2008), ChEMBL (Gaulton et al., 2011) targets with average AUC values of 0.8 (Riniker *et al.*, 2013). MUV is a challenging dataset for ML methods for its structurally diverse actives integrated in sets of structurally diverse decoys (Rohrer and Baumann, 2009) (Riniker and Landrum, 2013). MUV consists of assay data from 17 targets, each with 30 actives and 15000 decoys based on PubChem (Wang *et al.*, 2013) bioactivity data. These actives can be structurally diverse and contain compounds that bind to the target in a different manners, so that the SAR maybe complex. Previous studies of this dataset showed that similarity search methods were susceptible to the different binding modes of the actives and the overwhelming presence of false negatives, causing them to perform poorly (Tiikkainen *et al.*, 2009) (Riniker *et al.*, 2013). Another example, Schierz (2009) worked with a selection of Weka cost-sensitive classifiers using two types of FP.

The low performance of similarity-based methods might also be related to the use of 2D FPs that hold redundant and, sometimes, irrelevant features (Nasser *et al.*, 2018). Since unrelated features can mislead VS results, removing them can improve the recall of similarity measures (Liu and Motoda, 2007; Vogt *et al.*, 2010). The approach to maintain only relevant features contributing to the model construction is known as feature selection (Xue *et al.*, 2015). Although feature selection can improve performance and minimize training time, discriminating relevant features from noise ones is not trivial because of the vast search space and complex interactions between features (Nasser *et al.*, 2018). For example, a feature classified as irrelevant could be beneficial for model building when it interacts with other features, while a relevant one could become redundant when combined with other features (Ahmed *et al.*, 2013). Previous studies have

applied the concept of using only a subset of selected molecular features, removing unimportant fragments and, therefore, reducing the number of features, but keeping the percentage of compounds with biological activity the same (Xue *et al.*, 2001; Xue *et al.*, 2003; Ahmed *et al.*, 2013; Nasser *et al.*, 2018). A method to perform feature selection is the stratified feature selection, which emphasizes the features that most contributed to classifying the compound as a binder or a non-binder. First, a weight is attributed to all features available and. Then, these weights are used in the algorithm to select the features with the higher and the lower values.

In this work, we applied a stratified feature selection to the MUV dataset using Milk-Way, a classifier developed by our group (Figueiredo Vieira Leite, 2019(Leite *et al.*, 2020)), and four other classifiers: RF, NB, LR, and Support Vector Machine (SVM). We also investigated Milk-Way's performance on MUV (Rohrer and Baumann, 2009) and compared it to RF, NB, LR, and Support Vector Machine (SVM). The effects of the ligands number to construct the model, FP choice, and possible VS enrichment were also investigated.

2. Methods

2.1.Data collection

In this work, we used the Maximum Unbiased Validation (MUV) dataset. We evaluated the performance of three different types of FPs: ECFP4, Atom Pair (Carhart et al., 1985), and a hashed FP with maximum path 5 (RDK5) (RDKit: Cheminformatics and Machine Learning Software, 2019). All the FPs were 2048 bits long. Virtual screening was performed for 17 targets, considering 30 active and 15000 inactive compounds for each target (Table 1).

Table 1 – Description of the data set used

MUV code	Name of the Target	Number of active compounds	Number of inactive compounds
466	Sphingosine 1-phosphate (S1P1) receptor	30	15000
548	Protein kinase A (PKA)	30	15000
600	Steroidogenic factor 1 (SF1): inhibitors	30	15000
644	Rho-kinase 2	30	15000
652	HIV-1 RT-Rnase H	30	15000
689	Ephrin receptor A4	30	15000
692	Steroidogenic factor 1 (SF1): agonists	30	15000
712	Heat shock protein 90 (HSP90)	30	15000
713	Estrogen receptor (ER)α: inhibitors	30	15000
733	Estrogen receptor (ER) β	30	15000
737	Estrogen receptor (ER) α: potentiators	30	15000
810	Focal adhesion kinase (FAK)	30	15000
832	Cathepsin G	30	15000
846	Factor XIa (FXIa)	30	15000
852	Factor XIIa (FXIIa)	30	15000
858	Dopamine receptor D1	30	15000
859	Muscarinic receptor M1	30	15000

2.2.The Milk Way algorithm

The Milk-Way algorithm was applied as previous described by Figueiredo Vieira Leite *et al. (2019, 2020)*. Briefly, this method consists of a classifier that uses a modified logistic regression model, allowing the use of a higher number of attributes (descriptors) than the number of entities (compounds). The algorithm enables the building of a specific model for each target by selecting the closest entities (*ad hoc*). Building a model for each target can improve the performance on severely unbalanced datasets. As the algorithm fest, is nonrecursive, and can be implementated straightforwardly.

The Milk-Way algorithm calculates of P(x), the estimated probability, ranging from 0 to 1, of a compound to be active. Delineating a cut-off, we can define if the ligand is predicted to be active (near to one) or inactive (near to zero). The variables x are the descriptors, which, combined with an alpha value (α), a weight value associated to each feature, is used to calculate P(x). The term α_{n+1} is a stabilizer term:

$$P(x) = \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}, \quad (1)$$

Since α is a combinatorial problem, and one polynomial way of solving is:

$$Minimize f(\alpha) = \|\alpha\|^2 + \|B\alpha - b\|^2 \quad (2)$$

As $f(\alpha)$ is a convex function, the argument α^* that minimizes (2) is given by the derivative of $f(\alpha)$, equal to zero (optimal point). This results in the following system of linear equations:

$$\alpha^T + B^T B\,\alpha - B^T b = 0 \Leftrightarrow (I + B^T B)\,\alpha = B^T b, \quad (3)$$

Solving equation (3) allows us to use more features (descriptors, characteristics) than compounds (entities, individuals).

2.3. Model Validation

Genuine Milk-Way with all features

Using the first version of the Milk-Way algorithm, all features (2048) were used. For the compounds, we picked the number of eigenvalues of each model, which translates the number of patterns. For this, we used the Hamming distance to the *ad hoc* choice (Figueiredo Vieira Leite *et al.*, 2019). The Hamming distance was used because it offered better performance on binary matrices (data not shown). This distance can be defined as the number of positions in which two codewords differ (Morgan, 1998), or, in other words, it is the minimum number of errors that could transform one string into the other. This distance was compared to cosine and Euclidean distances, and showed a better performance (data not shown). To evaluate the models, we used the *k*-cross-validation (*k*=10). After analyzing the best AUC value of each target, we picked the corresponding FP to continue the analysis process – submit to the stratified feature selection.

Stratified feature selection

Using Equation 1, it is possible to sort the highest to lowest alpha value, which contributes to a P(*x*) close to 1 (active) or close to 0 (inactive). This step occurs when all

features and compounds are still in the model. The ranked alpha values will, then, be used to build alternative models with the stratified feature selection. By analyzing the alpha value, it is possible to choose a subgroup of features which maximizes the contributions, thus reducing noise. We screened five different cutoffs for the number features selected to build the alternative model - 2048 (100%), 1536 (75%), 1024 (50%), 512 (25%) and 102 (5%) - 2048 (100%),

Number of inactive compounds used in the model Stratified feature selection

We tested in a data sample the effects of varying the number of compounds and different number of features simultaneously over the metrics (EF5%, sensitivity, specificity, precision, f1, fa, AUC, and accuracy).

We evaluated the performance of the method when using different ratios of inactive to active compounds to build the model. As we used *k*-cross-validation (k=5), the training set had 24 actives and 12000 inactive compounds and test set had 3000 inactive and 6 active compounds. We built six models to each target: 24 (24A:24I), 1200 (24A:1200I), 3600 (24A:3600I), 6000 (24A:6000I), 8400 (24A:8400I), 12000 (24A:12000I) - with the same strategy of *ad hoc* choice. The Hamming distance was used, and the validation was *k*-cross-validation (k=5).

2.4.Validation of Milk-Way Algorithm and comparison with other algorithms

RF, NB, LR, and SVM were applied to the same dataset, using the *k*-cross-validation (*k*=5) was used to all targets, and the same metrics to evaluate them. Each fold had 3000 inactive and 6 active compounds. See for **Supplementary Material A** further details regarding the parameters for each classifier. All these algorithms were optimized.

2.5. Machine details

The data was processed in MATLAB R2019b, using a laptop with 8 GB RAM, 551 GB hard drive, and a processor Intel Core i7 2.70GHz. The Milk-Way methodology runs

in the Windows operating systems. For the building and validation of SVM, NB, and RF models, a server with 32 cores with 96 GB RAM was used.

2.6. Statistical test

Welch's t-test was used to test the equality of the means. A confidence level of 95% was employed to reject the null hypothesis, and the p-value is reported.

3. **Results and Discussion**

3.1.Choosing the best FP to each target

Riniker and Landrun (2013) proposed an open-source platform to benchmark FPs for ligand-based virtual screening. The reported performance of all FPs is generally similar (except for the baseline FPs) and MUV data sets were the most difficult of those studied. Nevertheless, posteriorly, Riniker et al. (2013) used only the four top FPs of the same platform (AP, topological torsions, RDKit FP, and circular FP) to train ML classifiers for the same targets. Although most of the methods/FPs combinations were statistically indistinguishable, the top performers were LR(ECFP4), NB(ECFP4), and RF(AP). The three databases used were: the directory of useful decoys (DUD) (Irwin, 2008), ChEMBL (Gaulton et al., 2011), and MUV (Rohrer and Baumann, 2009), the latter being the most challenging dataset, in which no method was able to reach an AUC above 0.90. Even a robust combination of known algorithms and FPs did not yield good results for this dataset. Here, we tested ECFP4, AP, and RDKIT5 FPs for the MUV. Using the Milk-Way algorithm (Figueiredo Vieira Leite *et al.*, 2019; Leite *et al.*, 2020), we calculated the sensitivity, specificity, precision, f1, fa, AUC, and accuracy to evaluate the model (Gimeno *et al.*, , 2019) using *k*-cross-validation (*k*=10) (Figure 1 and Supplementary B). The relatively high number of folds were used to assess subtitle differences between different FPs. Small differences between FPs were, for instance, reported by Riniker and Landrum (2013). We observed that the performance of AUC is highly target-dependent, and there is not a FP clearly better. For all targets, sensitivity was lower than specificity, translating the challenges of an imbalanced dataset, since the

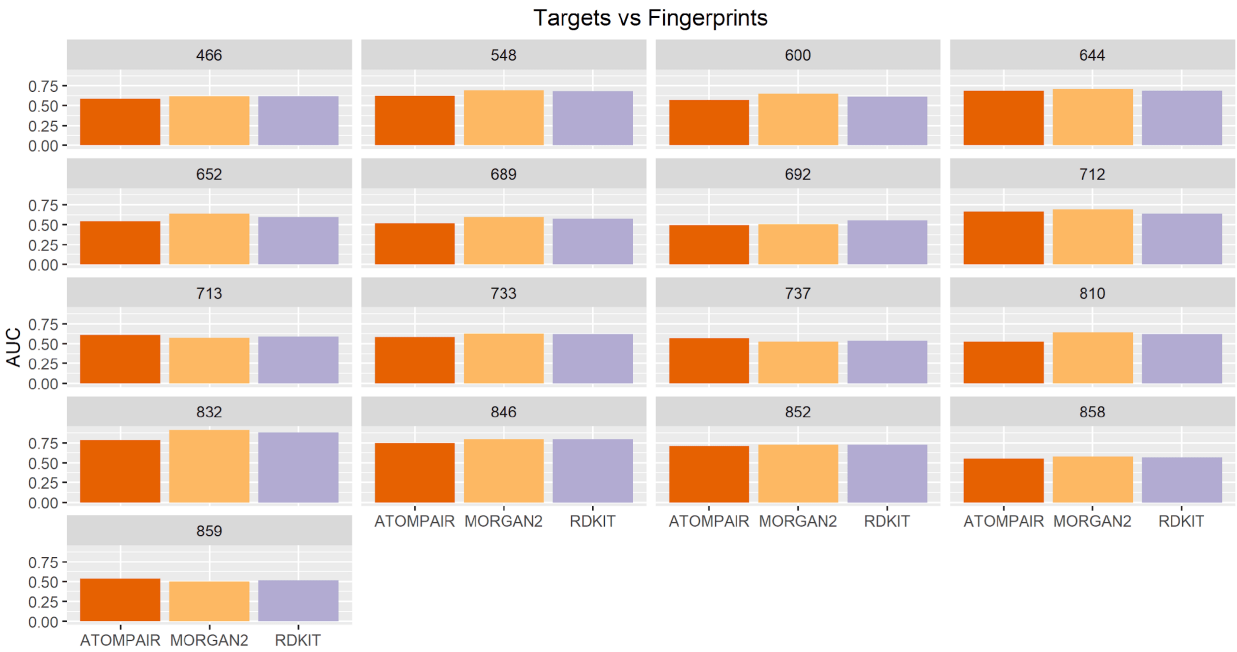


Figure 1 – Results of *k*-cross-validation (*k*=10) of AUC, using Milk Way algorithm dark orange =Atom Pair; light orange= ECFP4; purple= RDKit)

model does not have an enough active ligands to represent an efficient training. Because of that, accuracy was all high.

Knowing the best performing FP for each target, we sought to study the RF, NB, LR, and SVM algorithms and the effects of feature selection. Therefore, we selected the best FP for each target as ranked by highest AUC The ECFP4 FP yielded the highest AUC among the three FPs in 70% of the targets (12 targets); RDkit yielded the best AUC only for 12% of the targets with high AUC values (2 targets); the Atom Pair had 18% of the targets with high AUC values (3 targets) (Supplementary B). Using the unmodified Milk-Way algorithm, ECFP4 FP offered the best representation of ligands for this dataset. This result is in line with previous reported data Riniker and Landrum (2013).

We, then, applied the FP yielding the best AUC using Milk-Way for each target to build new models using SVM, RF, NB, and classical LR. We evaluated the same metrics: sensitivity, specificity, precision, f1, accuracy, AUC, and EF5% (**Supplementary Material D**). We observed that the AUC using other algorithms were significantly different from the Milk-Way algorithm (*p-values* < 0.05, **Table 2**). RF had the worst performance, in terms of AUC, of all five algorithms, yielding AUC no better than random selection on all cases. Milk-Way outperformed SVM, RF, NB, and classical

LR for all targets, albeit by small margins for many of them. When analyzing EF5% , we observed a different pattern. The SVM and RF were statistically different from the Milk-Way algorithm, within a 5% significance level (*p-values* < 0.05, **Table 3**). SVM had the best performance for 13 out of 17 targets. NB and classical LR EF5% were not statistically different from the Milk-Way algorithm (*p-value*> 0.05).

Table 2 - Mean of AUC of *k*-cross-validation (*k*=10) without Stratified Feature Selection.

Target	Type of Feature	MILK-WAY	SVM	RF	NB	LR
466	ECFP4	0.621433333	0.5159	0.5	0.5164	0.516333
548	ECFP4	0.692288889	0.5664	0.5	0.516533	0.566267
600	ECFP4	0.654978	0.5328	0.5	0.516467	0.516533
644	ECFP4	0.710622222	0.5658	0.5	0.499933	0.533
652	ECFP4	0.643366667	0.549467	0.5	0.5166	0.516433
689	ECFP4	0.593188889	0.533067	0.5	0.4999	0.516667
692	RDKit	0.551822222	0.5	0.5	0.4793	0.5
712	ECFP4	0.694411	0.533	0.5	0.5	0.549667
713	Atom Pair	0.609089	0.5165	0.5	0.529433	0.516567
733	ECFP4	0.623456	0.532667	0.5	0.499733	0.516533
737	Atom Pair	0.568978	0.516567	0.5	0.502733	0.516567
810	ECFP4	0.641511	0.533033	0.5	0.4999	0.516467
832	ECFP4	0.912667	0.682767	0.5	0.5166	0.682867
846	ECFP4	0.7968	0.6326	0.5	0.6165	0.6328
852	RDKit	0.730567	0.666567	0.533333	0.7478	0.633233
858	ECFP4	0.574756	0.5	0.5	0.4999	0.5
859	Atom Pair	0.539967	0.5	0.5	0.476933	0.5
***p-value* (95%)**	-	-	0.000582	4.87677E-06	6.26E-05	0.000236

Table 3 - Mean of EF5% of *k*-cross-validation (k=10) without Stratified Feature Selection.

Target	Type of Feature	MILK-WAY	SVM	RF	NB	LR
466	ECFP4	7.251315789	15.92583	0	15.92583	11.94437
548	ECFP4	10.54737	19.90728	0	7.962914	15.92583
600	ECFP4	6.592105	15.92583	0	15.92583	11.94437
644	ECFP4	13.18421	19.90728	0	3.981457	19.90728
652	ECFP4	6.592105	19.90728	0	7.962914	19.90728
689	ECFP4	6.592105	15.92583	0	7.962914	3.981457
692	RDKit	2.636842	0	0	8.286675	0
712	ECFP4	8.569737	11.94437	0	0	7.962914
713	Atom Pair	4.614474	15.92583	0	18.13515	11.94437
733	ECFP4	5.932895	11.94437	0	11.94437	11.94437
737	Atom Pair	2.636842	11.94437	0	18.72211	7.962914
810	ECFP4	6.592105	19.90728	0	7.962914	19.90728
832	ECFP4	16.48026	19.90728	0	7.962914	19.90728
846	ECFP4	13.18421	19.90728	0	15.92583	19.90728
852	RDKit	11.86579	15.92583	7.962914	11.01702	15.92583
858	ECFP4	3.955263	0	0	7.962914	0
859	Atom Pair	2.636842	0	0	18.58828	0
***p-value* (95%)**	-	-	0.005002	1.34651E-06	0.054575	0.05806

3.2. Number of inactive compounds used in the model

The Milk Way algorithm and all other algorithms yielded low AUC values for most targets (**Table 2** and **Figure 1**). We, then, sought to to evaluate the effect of the active to inactive compounds rate and the *ad hoc choice*. While the original Milk-Way algorithm used the same number of eigenvalues to select the number of inactive compounds, using the *ad hoc choice* (Figueiredo Vieira Leite et al., 2019, Leite *et al.*, 2020), we evaluated the effect of several actives to inactives ratios (**Table 4**).

We made a sampling of 4 targets (24%) – 466, 548, 600, 659 – to evaluate effects of increasing the number of inactive compounds in the model (**Table 4** for average AUC). The 1:1 proportion (24 active compounds *vs.* 24 inactive compounds) yielded 0.6155 to 1.000 average AUC while minimizing the required computational resources. The precision and f1 varied coherently with AUC while the specificity was almost always high (0.99) (See **Supplementary Material C** for other metrics). Comparing the average AUC *per* target, we observed that no clear trend was present. However, adding more inactive compounds to the model did not, in general, improve the performance of the classifier.

On the other hand, the variation of the number of features had clear impact on the average AUC. Using all features yielded the worst performance in all models, while applying stratified feature selection with 50%, and 75% of features yielded an average AUC=1.0 on three of the four targets. We observed that aggregating more features and more individuals did not improve the models. According to these results, we chose the models which used fewer resources - with 50% of the characteristics and with fewer inactives. In other words, more features and more individuals did not translate to a better representation of the compounds. We also observed that the performance for all four targets was similar to each other (**Table 4**, and **Supplementary Material C**).

Table 4 – Mean of AUC (cross-validation *k*=5) varying the number of inactive compounds and features, when training the model (Targets: 466 Sphingosine 1-phosphate receptor, 548 - Protein kinase A, 600 - Steroidogenic factor 1: inhibitors, 692 - Steroidogenic factor 1: agonists)

TARGET	FEATURES	NUMBER ACTIVES vs. NUMBER INACTIVES					
		24A:24I	24A:1200I	24A:3600I	24A:6000 I	24A:8400I	24A:12000I
466	5%	0,9979	0,9988	0,9658	0,9825	0,9824	0,9991
	25%	1,0000	0,9832	0,9832	0,9832	0,9833	0,9832
	50%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	75%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	100%	0,6155	0,5894	0,5494	0,5330	0,5331	0,5165
548	5%	0,9829	0,9996	0,9828	0,9662	0,9662	0,9994
	25%	1,0000	1,0000	0,9999	1,0000	1,0000	0,9999
	50%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	75%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	100%	0,7966	0,6614	0,5656	0,5162	0,4997	0,4998
600	5%	0,9647	0,9485	0,9324	0,9489	0,9490	0,9658
	25%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	50%	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	75%	1,0000	1,0000	1,0000	0,9833	0,9833	0,9833
	100%	0,7128	0,6251	0,5319	0,4997	0,4998	0,4998
692	5%	0,9449	0,8498	0,7499	0,6666	0,6500	0,6166
	25%	0,9972	0,9666	0,8833	0,7833	0,7500	0,7333
	50%	0,9935	0,9661	0,9000	0,8833	0,8500	0,8000
	75%	0,9969	0,9832	0,9333	0,9500	0,9500	0,9333
	100%	0,6232	0,5812	0,5155	0,4999	0,5000	0,5000

3.2.Stratified feature selection

After selecting the best proportion of individuals (24 active compounds *vs.* 24 inactive compounds), we expanded the calculations to all the targets (**Figure 2**). The results were similar between all of them: using all features (2048 features, 100%) produced the worst results, followed by 102 features (5%). With 50% of features, 16 targets reached an average AUC=1.0 (see **Supplementary Material C** for other metrics to evaluate the models). The precision and f1 variated in a similar manner as AUC.

Our results suggest that using 102 features (5%) failed to represent the ligands, because the sensibility was high, but the precision, low – a evidence of the FP failing to treat the imbalanced dataset, thus and not being able to distinguish active and inactive compounds. All targets have similar behavior with different models, corroborating the previous findings that more features do not translate into a better representation, and, therefore, into better metrics.

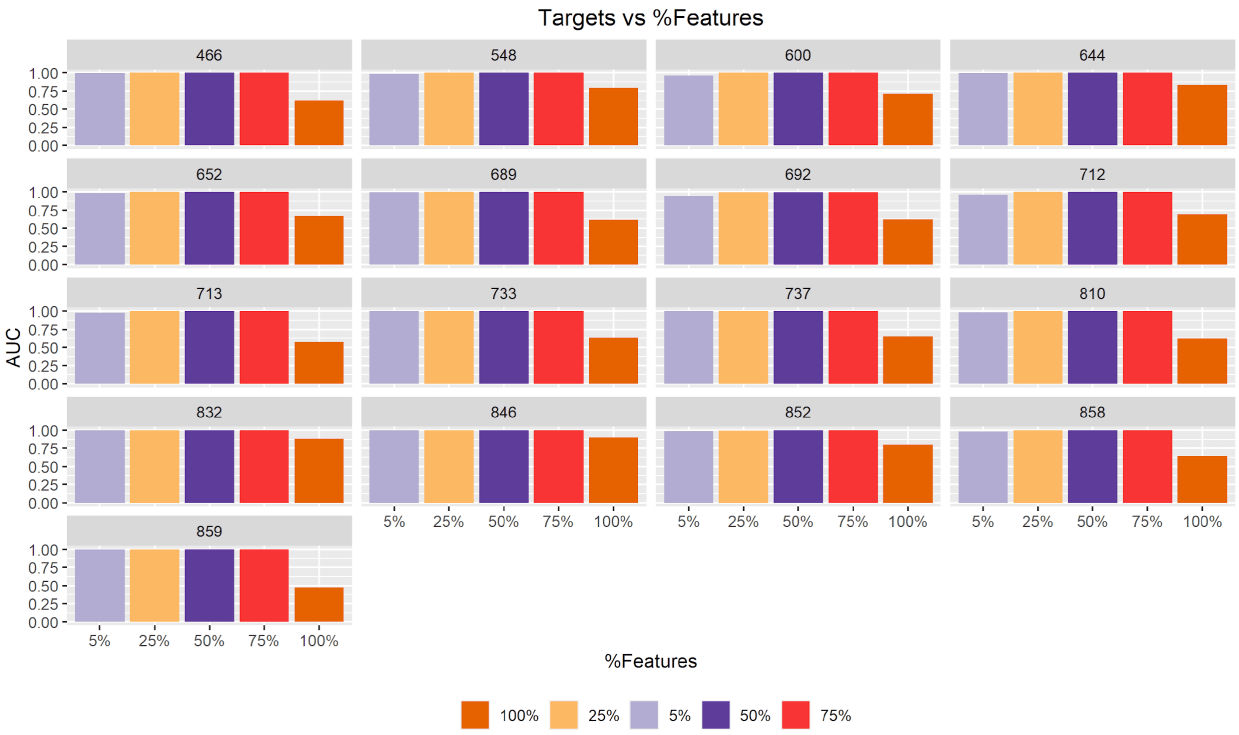


Figure 2 – Milk-Way algorithm using the best proportion of individuals (24 inactives compounds vs. 24 actives compounds) describe with different % of features (purple - 5%; light orange – 25%; dark purple – 50%; red – 75%; dark orange – 100%).

3.4. Comparison between Milk-Way Algorithm to other algorithms

To validate the stratified feature selection and the results of the Milk-Way algorithm with the MUV dataset, we applied the same feature selection strategy for other algorithms (ie, the same 1024 features) (**Figure 2 and Supplementary C**). We analyzed all 17 targets using RF, SVM, RF, and LR with *k*-cross-validation (*k*=5).

We observed that stratified feature selection was efficient in SVM and LR (**Table 4**). Statistically, the results with SVM and LR are not significantly different from Milk-Way, within a 95% confidence interval, with *p-values* of 0.06 and 0.13, respectively. On the other hand, AUC obtained with RF and NB were statistically different from the Milk-Way AUC, with a *p-value* < 0.05 (8.09E-9 and 3.2E-9, respectively). We achieved better average AUC in all targets, with the Milk-Way, SVM, and LR algorithms (see **Supplementary Material E** for other metrics). The precision and the f1, which were the worst metrics without stratified feature selection, improved from mean 0.10±0.08 to mean 0.91±0.25, and 0.14±0.11 to mean 0.93±0.21, respectively. The mean of EF5% reached up to 19.90728, suggesting a good earlier enrichment when using Milk-Way, SVM, and LR. This good early enrichment was observed for all targets and there was no significant difference between them (*p-values*=1). These results suggest that stratified feature selection is an efficient approach to improve different classifier methods.

Table 4 - Average of AUC of *k*-cross-validation (k=5) with 1024 features (50%).

Target	Type of Feature	MILK-WAY	SVM	RF	NB	LR
466	ECFP4	1	1	0.55	0.516667	1
548	ECFP4	1	1	0.583333	0.533333	1
600	ECFP4	1	1	0.55	0.516667	1
644	ECFP4	1	1	0.566667	0.516667	1
652	ECFP4	1	0.95	0.533333	0.516667	0.983333
689	ECFP4	1	1	0.533333	0.5	1
692	RDKit	0.9935278	0.816667	0.5	0.958433	0.95
712	ECFP4	1	1	0.5	0.5	1
713	Atom Pair	1	0.983333	0.883333	0.6	1
733	ECFP4	1	0.983333	0.516667	0.5	1
737	Atom Pair	1	0.966667	0.95	0.566667	1
810	ECFP4	1	0.983333	0.533333	0.516667	0.983333
832	ECFP4	1	1	0.533333	0.516667	1
846	ECFP4	1	1	0.666667	0.65	1
852	RDKit	0.9978333	0.866667	0.583333	0.9635	0.9
858	ECFP4	1	1	0.5	0.516667	1
859	Atom Pair	1	1	0.883333	0.516667	1
p-value (95%)	-	-	0.05795	8.0953E-9	3.2090E-9	0.12743

Table 5 – Mean of EF5% of *k*-cross-validation (*k*=5) with 1024 features (50%).

Target	Type of Feature	MILK-WAY	SVM	RF	NB	LR
466	ECFP4	19.90728	19.90728	11.94437	3.981457	19.90728
548	ECFP4	19.90728	19.90728	15.92583	7.962914	19.90728
600	ECFP4	19.90728	19.90728	7.962914	3.981457	19.90728
644	ECFP4	19.90728	19.90728	3.981457	3.981457	19.90728
652	ECFP4	19.90728	19.90728	3.981457	3.981457	19.90728
689	ECFP4	19.90728	19.90728	7.962914	0	19.90728
692	RDKit	19.90728	19.90728	0	19.52583	19.90728
712	ECFP4	19.90728	19.90728	0	0	19.90728
713	Atom Pair	19.90728	19.90728	19.90728	15.92583	19.90728
733	ECFP4	19.90728	19.90728	3.981457	0	19.90728
737	Atom Pair	19.90728	19.90728	19.90728	11.94437	19.90728
810	ECFP4	19.90728	19.90728	7.962914	3.981457	19.90728
832	ECFP4	19.90728	19.90728	0	3.981457	19.90728
846	ECFP4	19.90728	19.90728	15.92583	19.90728	19.90728
852	RDKit	19.90728	19.90728	15.92583	19.90728	19.90728
858	ECFP4	19.90728	19.90728	0	3.981457	19.90728
859	Atom Pair	19.90728	19.90728	19.90728	3.981457	19.90728
***p-value* (95%)**	-	-	1	2.30979E-05	2.11E-06	1

4. Conclusion

In this work, we applied a novel method to tackle a challenging dataset for LBVS. We observed that the original Milky-Way algorithm yielded poor results with the three FPs studied. Consistently with previous work with this dataset, with the original Milk-Way algorithm, we did not see a clear best FP, even though ECFP4 yielded the best metrics more often. Using the FP with the best AUC for each target, we applied stratified feature selection. Applying stratified feature selection improved the results significantly, obtaining an average AUC=1.0 for most targets and the EF5%=19.90728. This strategy also produced high average AUC values with SVM and LR, while RF and NB had statistically different worse results. Our results show that an efficient selection of the most representative features to describe the compounds in the dataset is essential to the performance of a LBVS classifier. Increasing the number of inactive compounds during the model training, and the number of features did not improve the metrics. In other words, the addition of lines and columns to the matrix can have no significance and impair the training of the model. It is remarkable that equation 1 attributes a weight (alpha value) for each feature, considering the whole model. Overall, stratified feature selection was able to consistently improve the results of different classifiers, allowing them to score a near-perfect recovery of binders in a hard LBVS dataset .

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

BENDER, A.; MUSSA, H. Y.; GLEN, R. C. Screening for dihydrofolate reductase inhibitors using MOLPRINT 2D, a fast fragment-based method employing the naïve Bayesian classifier: limitations of the descriptor and the importance of balanced chemistry in training and test sets. **J Biomol Screen,** v. 10, n. 7, p. 658-66, Oct 2005. ISSN 1087-0571. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/16170051>.

CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom pairs as molecular features in structure-activity studies: definition and applications. **Journal of Chemical Information and Computer Sciences,** v. 25, n. 2, p. 64-73, 1985. ISSN 0095-2338.

CARPENTER, K. A.; HUANG, X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. In: (Ed.). **Curr Pharm Des**, v.24, 2018. p.3347-58. ISBN 1381-6128 (Print)1873-4286 (Electronic).

CERETO-MASSAGUÉ, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods,** v. 71, p. 58-63, 2015/01/01/ 2015. ISSN 1046-2023. Available at: <http://www.sciencedirect.com/science/article/pii/S1046202314002631>.

DAI, W.; GUO, D. A Ligand-Based Virtual Screening Method Using Direct Quantification of Generalization Ability. **Molecules,** v. 24, n. 13, Jun 2019. ISSN 1420-3049. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/31262005>.

GAULTON, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. **Nucleic Acids Research,** v. 40, n. D1, p. D1100-D1107, 2011. ISSN 0305-1048. Available at: <https://doi.org/10.1093/nar/gkr777>. Accessed on: 10/1/2019.

GIMENO, A. et al. The Light and Dark Sides of Virtual Screening: What Is There to Know? **International journal of molecular sciences,** v. 20, n. 6, p. 1375, 2019.

HAWKINS, P. C.; SKILLMAN, A. G.; NICHOLLS, A. Comparison of shape-matching and docking as virtual screening tools. **J Med Chem,** v. 50, n. 1, p. 74-82, Jan 2007. ISSN 0022-2623. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/17201411>.

IRWIN, J. J. Community benchmarks for virtual screening. **Journal of Computer-Aided Molecular Design,** v. 22, n. 3, p. 193-199, March 01 2008. ISSN 1573-4951. Available at: <https://doi.org/10.1007/s10822-008-9189-4>.

LOVING, K.; SALAM, N. K.; SHERMAN, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. **J Comput Aided Mol Des,** v. 23, n. 8, p. 541-54, Aug 2009. ISSN 1573-4951. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/19421721>.

LUO, M.; WANG, X. S.; TROPSHA, A. Comparative Analysis of QSAR-based vs. Chemical Similarity Based Predictors of GPCRs Binding Affinity. **Mol Inform,** v. 35, n. 1, p. 36-41, 01 2016. ISSN 1868-1751. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/27491652>.

MA, X. H. et al. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. **Comb Chem High Throughput Screen,** v. 12, n. 4, p.

344-57, May 2009. ISSN 1875-5402. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/19442064>.

MBALLO, C.; MAKARENKOV, V. Using machine learning methods to predict experimental high-throughput screening data. **Comb Chem High Throughput Screen,** v. 13, n. 5, p. 430-41, Jun 2010. ISSN 1875-5402. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/20236062>.

PUERTAS-MARTÍN, S. et al. OptiPharm: An evolutionary algorithm to compare shape similarity. **Sci Rep,** v. 9, n. 1, p. 1398, Feb 2019. ISSN 2045-2322. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/30718737>.

RDKit: Cheminformatics and Machine Learning Software. 2019.

RINIKER, S.; FECHNER, N.; LANDRUM, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. **Journal of chemical information and modeling,** v. 53, n. 11, p. 2829-2836, 2013. ISSN 1549-9596.

ROHRER, S. G.; BAUMANN, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. **Journal of Chemical Information and Modeling,** v. 49, n. 2, p. 169-184, 2009/02/23 2009. ISSN 1549-9596. Available at: <https://doi.org/10.1021/ci8002649>.

SEIDEL, T. et al. 3D pharmacophore modeling techniques in computer-aided molecular design using LigandScout. **Tutorials Chemoinform,** v. 281, p. 279-309, 2017.

SLIWOSKI, G. et al. Computational methods in drug discovery. **Pharmacological reviews,** v. 66, n. 1, p. 334-395, 2014. ISSN 1521-0081.

SUBRAMANIAM, S.; MEHROTRA, M.; GUPTA, D. Support vector machine based classification model for screening Plasmodium falciparum proliferation inhibitors and non-inhibitors. **Biomedical Engineering and Computational Biology,** v. 3, p. BECB. S7503, 2011. ISSN 1179-5972.

AHMED, A.; SALIM, N.; ABDO, A. Fragment reweighting in ligand-based virtual screening. **Advanced Science Letters,** v. 19, n. 9, p. 2782-2786, 2013. ISSN 1936-6612.

GAULTON, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. **Nucleic Acids Research,** v. 40, n. D1, p. D1100-D1107, 2011. ISSN 0305-1048. Available at: < https://doi.org/10.1093/nar/gkr777 >. Accessed on: 10/1/2019.

HUANG, N.; SHOICHET, B. K.; IRWIN, J. J. Benchmarking sets for molecular docking. **Journal of medicinal chemistry,** v. 49, n. 23, p. 6789-6801, 2006. ISSN 0022-2623.

IRWIN, J. J. Community benchmarks for virtual screening. **Journal of Computer-Aided Molecular Design,** v. 22, n. 3, p. 193-199, March 01 2008. ISSN 1573-4951. Available at: < https://doi.org/10.1007/s10822-008-9189-4 >.

JAHN, A. et al. Optimal assignment methods for ligand-based virtual screening. **Journal of cheminformatics,** v. 1, n. 1, p. 14, 2009. ISSN 1758-2946.

LEITE, C. F. V. et al. Milk-Way algorithm for ligand-based virtual screening: CDK2 case study. **Trends in Developmental Biology,** v. 13, 2020.

LIU, H.; MOTODA, H. **Computational methods of feature selection**. CRC Press, 2007. ISBN 1584888792.

MORGAN, S. P. Richard Wesley Hamming. **Notices of the AMS,** v. 45, n. 8, p. 972-977, 1998.

NASSER, M. et al. Deep Belief Network for Molecular Feature Selection in Ligand-Based Virtual Screening. International Conference of Reliable Information and Communication Technology, 2018, Springer. p.3-14.

RINIKER, S.; FECHNER, N.; LANDRUM, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. **Journal of chemical information and modeling,** v. 53, n. 11, p. 2829-2836, 2013. ISSN 1549-9596.

ROY, K.; KAR, S.; DAS, R. N. **Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment**. Academic press, 2015. ISBN 0128016337.

TIIKKAINEN, P. et al. Critical comparison of virtual screening methods against the MUV data set. **Journal of chemical information and modeling,** v. 49, n. 10, p. 2168-2178, 2009. ISSN 1549-9596.

VOGT, M.; WASSERMANN, A. M.; BAJORATH, J. Application of information—Theoretic concepts in chemoinformatics. **Information,** v. 1, n. 2, p. 60-73, 2010.

WANG, Y. et al. PubChem bioassay: 2014 update. **Nucleic acids research**, p. gkt978, 2013. ISSN 0305-1048.

WOLBER, G.; LANGER, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. **J Chem Inf Model,** v. 45, n. 1, p. 160-9, 2005 Jan-Feb 2005. ISSN 1549-9596. Available at: < https://www.ncbi.nlm.nih.gov/pubmed/15667141 >.

XUE, B. et al. A survey on evolutionary computation approaches to feature selection. **IEEE Transactions on Evolutionary Computation,** v. 20, n. 4, p. 606-626, 2015. ISSN 1089-778X.

XUE, L. et al. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. **Journal of chemical information and computer sciences,** v. 43, n. 4, p. 1218-1225, 2003. ISSN 0095-2338.

_____. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. **Journal of chemical information and computer sciences,** v. 41, n. 2, p. 394-401, 2001. ISSN 0095-2338.

FIGUEIREDO VIEIRA LEITE, C. et al. **Método de Triagem de compostos baseados em Regressão Logística Modificada**. INDUSTRIAL, I. N. D. P. Brazil: Universidade Federal de Minas Gerais. BR 10 2019 027703 3**:** 22 p. 2019.

LEITE, C. F. V. et al. Milk-Way algorithm for ligand-based virtual screening: CDK2 case study. **Trends in Developmental Biology,** v. 13, 2020.

MORGAN, S. P. Richard Wesley Hamming. **Notices of the AMS,** v. 45, n. 8, p. 972-977, 1998.

RINIKER, S.; LANDRUM, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. **J Cheminform,** v. 5, n. 1, p. 26, May 2013. ISSN 1758-2946. Available at: < https://www.ncbi.nlm.nih.gov/pubmed/23721588 >.

ROHRER, S. G.; BAUMANN, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. **Journal of Chemical Information and Modeling,** v. 49, n. 2, p. 169-184, 2009/02/23 2009. ISSN 1549-9596. Available at: < https://doi.org/10.1021/ci8002649 >.

SCHIERZ, A. C. Virtual screening of bioassay data. **Journal of cheminformatics,** v. 1, p. 21, 2009.

SUPPLEMENTARY A

Best Parameters

Without Stratified Feature Selection

Table A.1 - Best Parameters of SVM Without Stratified Feature Selection

Target_466_ECFP4	C: 1000	gamma: 0.001	kernel: rbf
Target_548_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_600_ECFP4	C: 1000	gamma: 0.001	kernel: rbf
Target_644_ECFP4	C: 1	kernel: linear	
Target_652_ECFP4	C: 1	kernel: linear	
Target_689_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_692_RDKIT	C: 1	gamma: 0.001	kernel: rbf
Target_712_ECFP4	C: 1000	gamma: 0.001	kernel: rbf
Target_713_Atom Pair	C: 100	gamma: 0.001	kernel: rbf
Target_733_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_737_Atom Pair	C: 100	gamma: 0.001	kernel: rbf
Target_810_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_832_ECFP4	C: 1000	gamma: 0.001	kernel: rbf
Target_846_ECFP4	C: 1000	gamma: 0.001	kernel: rbf
Target_852_RDKIT	C: 100	gamma: 0.0001	kernel: rbf
Target_858_ECFP4	C: 1	gamma: 0.001	kernel: rbf
Target_859_Atom Pair	C: 1	gamma: 0.001	kernel: rbf

Table A.2 - Best Parameters of LR without Stratified Feature Selection

Target_466_ECFP4	C: 10.0	penalty : l2
Target_548_ECFP4	C: 0.001	penalty : none
Target_600_ECFP4	C: 10.0	penalty : l2
Target_644_ECFP4	C: 1000.0	penalty : l2
Target_652_ECFP4	C: 10.0	penalty : l2
Target_689_ECFP4	C: 1.0	penalty : l2
Target_692_RDKIT	C: 0.001	penalty : l2
Target_712_ECFP4	C: 0.001	penalty : none
Target_713_Atom Pair	C: 100.0	penalty : l2
Target_733_ECFP4	C: 1.0	penalty : l2
Target_737_Atom Pair	C: 10.0	penalty : l2
Target_810_ECFP4	C: 100.0	penalty : l2
Target_832_ECFP4	C: 1000.0	penalty : l2
Target_846_ECFP4	C: 1000.0	penalty : l2
Target_852_RDKIT	C: 100.0	penalty : l2
Target_858_ECFP4	C: 0.001	penalty : l2
Target_859_Atom Pair	C: 0.001	penalty : l2

Table A.3 - Best Parameters of RF without Stratified Feature Selection

Target_466_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_548_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_600_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_644_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_652_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_689_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_692_RDKIT	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_712_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_713_Atom Pair	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_733_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_737_Atom Pair	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_810_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_832_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_846_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_852_RDKIT	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 200
Target_858_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_859_Atom Pair	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100

With Stratified Feature Selection

Table A.4 - Best Parameters of SVM with Stratified Feature Selection

Target_466_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_548_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_600_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_644_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_652_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_689_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_692_RDKIT	C: 100	gamma: 0.0001	kernel: rbf
Target_712_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_713_Atom Pair	C: 10	gamma: 0.001	kernel: rbf
Target_733_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_737_Atom Pair	C: 1	gamma: 0.001	kernel: rbf
Target_810_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_832_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_846_ECFP4	C: 100	gamma: 0.001	kernel: rbf
Target_852_RDKIT	C: 100	gamma: 0.0001	kernel: rbf
Target_858_ECFP4	C: 10	gamma: 0.001	kernel: rbf
Target_859_Atom Pair	C: 10	gamma: 0.001	kernel: rbf

Table A.5 - Best Parameters of LR with Stratified Feature Selection

Target_466_ECFP4	C: 0.001	penalty: none
Target_548_ECFP4	C: 0.001	penalty: none
Target_600_ECFP4	C: 0.001	penalty: none
Target_644_ECFP4	C: 0.001	penalty: none
Target_652_ECFP4	C: 0.001	penalty: none
Target_689_ECFP4	C: 0.001	penalty: none
Target_692_RDKIT	C: 0.001	penalty: none
Target_712_ECFP4	C: 0.001	penalty: none
Target_713_Atom Pair	C: 0.001	penalty: none
Target_733_ECFP4	C: 0.001	penalty: none
Target_737_Atom Pair	C: 0.001	penalty: none
Target_810_ECFP4	C: 0.001	penalty: none
Target_832_ECFP4	C: 0.001	penalty: none
Target_846_ECFP4	C: 0.001	penalty: none
Target_852_RDKIT	C: 0.001	penalty: none
Target_858_ECFP4	C: 0.001	penalty: none
Target_859_Atom Pair	C: 0.001	penalty: none

Table A.6 - Best Parameters of RF with Stratified Feature Selection				
Target_846_ECFP4	max_depth: 90	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 300
Target_689_ECFP4	max_depth: 100	min_samples_leaf: 3	min_samples_split: 10	n_estimators: 100
Target_852_RDKIT	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_810_ECFP4	max_depth: 110	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 200
Target_737_Atom Pair	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_832_ECFP4	max_depth: 110	min_samples_leaf: 3	min_samples_split: 12	n_estimators: 100
Target_713_Atom Pair	max_depth: 110	min_samples_leaf: 3	min_samples_split: 14	n_estimators: 100
Target_859_Atom Pair	max_depth: 80	min_samples_leaf: 3	min_samples_split: 10	n_estimators: 100
Target_652_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_733_ECFP4	max_depth: 90	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_600_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 200
Target_712_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_858_ECFP4	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_644_ECFP4	max_depth: 100	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_466_ECFP4	max_depth: 90	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 200
Target_692_RDKIT	max_depth: 80	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100
Target_548_ECFP4	max_depth: 100	min_samples_leaf: 3	min_samples_split: 8	n_estimators: 100

SUPPLEMENTARY B

Metrics of all 17 targets, described through three fingerprints

	10 folds/466 RDKIT		**10 folds/466 ATOM PAIR**		**10 folds/466 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.266667	0.249444	0.233333333	0.213437475	0.266666667	0.290593263
Specificity	0.9818	0.008417	0.973866667	0.011348226	0.984	0.00412041
Precision	0.045131	0.050264	0.021630543	0.01845109	0.040019359	0.046871041
f1	0.063518	0.08222	0.039423632	0.03361054	0.069137807	0.079769485
fa	0.358339	0.311959	0.327853357	0.280767186	0.338419987	0.353782852
AUC	0.620544	0.12688	0.590566667	0.108253628	**0.621433333**	0.146817502
Accuracy	0.980373	0.008666	0.972388556	0.011505697	0.982568197	0.004520369
EF5%	5.932894737	4.614473684	4.614473684	4.221006899	***7.251315789***	6.218979667

	10 folds/548 RDKIT		**10 folds/548 ATOM PAIR**		**10 folds/548 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.4	0.326598632	0.3	0.314466038	0.4	0.359010987
Specificity	0.981533333	0.015640546	0.970866667	0.007429969	0.989466667	0.006358197
Precision	0.054505173	0.049599141	0.024578909	0.026678762	0.106994048	0.090248744
f1	0.092644942	0.083629081	0.045307081	0.049093955	0.160838509	0.134032765
fa	0.485600577	0.352382658	0.376798357	0.342477236	0.478262905	0.363554172
AUC	0.685511111	0.161339372	0.6239	0.162219912	**0.692288889**	0.181249015
Accuracy	0.980372588	0.01540436	0.969527611	0.007831771	0.988290086	0.006554151
EF5%	8.569737	6.624984	5.932895	6.21898	***10.54737***	6.041764

	10 folds/600 RDKIT		**10 folds/600 ATOM PAIR**		**10 folds/600 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.266667	0.249444	0.2	0.221108	0.333333	0.258199
Specificity	0.977733	0.009611	0.972267	0.01266	0.9864	0.008215
Precision	0.027049	0.02543	0.020852	0.022721	0.050808	0.038311
f1	0.04823	0.044668	0.03758	0.040866	0.08521	0.062277
fa	0.357068	0.310389	0.278383	0.290673	0.437322	0.311008
AUC	0.614078	0.125695	0.574511	0.113651	**0.654978**	0.128954
Accuracy	0.976314	0.009495	0.970725	0.012905	0.985096	0.007918
EF5%	5.273684	4.93308	3.955263	4.372708	***6.592105***	5.106223

	10 folds/644 RDKIT		**10 folds/644 ATOM PAIR**		**10 folds/644 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.4	0.2	0.433333	0.213437	0.433333333	0.260341656
Specificity	0.9872	0.010003555	0.9734	0.010248	0.993066667	0.004772141
Precision	0.124105682	0.128929452	0.035372	0.018619	0.12855042	0.065587308
f1	0.177454386	0.160899437	0.065016	0.033868	0.190036075	0.093807306
fa	0.538528119	0.224220545	0.564779	0.233644	0.558529787	0.252573278
AUC	0.688211111	0.10352667	0.691056	0.11024	**0.710622222**	0.130929377
Accuracy	0.986027944	0.01028154	0.972322	0.010244	0.991949434	0.004671114
EF5%	9.228947	3.229459	7.910526	3.955263	***13.18421***	5.896158

	10 folds/652 RDKIT		**10 folds/652 ATOM PAIR**		**10 folds/652 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.2	0.221108319	0.133333333	0.221108319	0.3	0.276887462
Specificity	0.992533333	0.003649049	0.973333333	0.011873406	0.992266667	0.004239497
Precision	0.0725	0.087245662	0.019192982	0.034987242	0.114457014	0.132760653
f1	0.101509972	0.118277876	0.033066514	0.059911252	0.156130717	0.168309978
fa	0.279475915	0.292005438	0.179098772	0.284264904	0.38920019	0.338254959
AUC	0.591788889	0.111859374	0.543722222	0.114556787	**0.643366667**	0.139845559
Accuracy	0.99095143	0.003777796	0.971656687	0.012090823	0.990884897	0.004493358
EF5%	5.273684	4.93308	2.636842	4.372708	***6.592105***	5.106223

	10 folds/689 RDKIT		**10 folds/689 ATOM PAIR**		**10 folds/689 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.166666667	0.166666667	0.1	0.152752523	0.2	0.221108319
Specificity	0.988133333	0.003804091	0.952133333	0.013405803	0.992533333	0.003636848
Precision	0.031259023	0.034434271	0.005167297	0.00958807	0.059074074	0.062942262
f1	0.051890035	0.055865358	0.009676808	0.017658611	0.087303807	0.089847252
fa	0.249276041	0.249276405	0.148028776	0.226123068	0.279227093	0.291443195
AUC	0.572222222	0.083518801	0.513166667	0.075798372	**0.593188889**	0.110440812
Accuracy	0.986493679	0.003845734	0.950432468	0.01334741	0.99095143	0.003523141
EF5%	5.932895	3.549957	0.659211	1.977632	***6.592105***	4.169213

	10 folds/692 RDKIT		**10 folds/692 ATOM PAIR**		**10 folds/692 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.133333333	0.163299316	0.066666667	0.133333333	0.033333333	0.1
Specificity	0.9852	0.003873557	0.955333333	0.00818671	0.982266667	0.004818483
Precision	0.019210741	0.023706418	0.003389831	0.006779661	0.005	0.015
f1	0.033549451	0.041323411	0.006451613	0.012903226	0.008695652	0.026086957
fa	0.199326539	0.244124177	0.099004463	0.198008926	0.049840148	0.149520444
AUC	**0.551822222**	0.081859594	0.494622222	0.066383421	0.502811111	0.050928879
Accuracy	0.983499667	0.003978687	0.953559548	0.008271601	0.980372588	0.004882406
EF5%	***2.636842***	3.229459	1.318421	2.636842	1.977632	3.020882

	10 folds/712 RDKIT		**10 folds/712 ATOM PAIR**		**10 folds/712 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.3	0.276887	0.4	0.249444	0.4	0.249444
Specificity	0.9878	0.004382	0.963267	0.016241	0.9936	0.002969
Precision	0.057176	0.055145	0.025318	0.018517	0.130206	0.107612
f1	0.095167	0.090538	0.047235	0.034041	0.188195	0.136795
fa	0.388581	0.337558	0.513298	0.288119	0.518806	0.291734
AUC	0.639656	0.139809	0.669867	0.127449	**0.694411**	0.12532
Accuracy	0.986427	0.00464	0.962142	0.016225	0.992415	0.002949
EF5%	5.932895	5.475814	6.592105	5.106223	***8.569737***	4.221007

	10 folds/713 RDKIT		**10 folds/713 ATOM PAIR**		**10 folds/713 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.2	0.163299	0.266667	0.249444	0.166667	0.166667
Specificity	0.985533	0.006346	0.974133	0.012795	0.988467	0.00525
Precision	0.037155	0.034277	0.021517	0.018307	0.039663	0.045417
f1	0.061732	0.055493	0.039152	0.032925	0.062456	0.069409
fa	0.299098	0.244213	0.356062	0.308968	0.24936	0.24936
AUC	0.585311	0.082791	**0.609089**	0.122234	0.573233	0.083867
Accuracy	0.983965	0.006502	0.972721	0.012532	0.986826	0.005338
EF5%	3.955263	3.229459	4.614474	4.221007	***5.273684***	3.955263

	10 folds/733 RDKIT		**10 folds/733 ATOM PAIR**		**10 folds/733 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.266667	0.249444	0.233333	0.152753	0.266667	0.133333
Specificity	0.9838	0.008465	0.9654	0.009691	0.9906	0.003705
Precision	0.031323	0.028751	0.015432	0.011313	0.073692	0.067215
f1	0.054347	0.048201	0.02882	0.0209	0.108114	0.077885
fa	0.357324	0.310547	0.347088	0.227226	0.39911	0.199555
AUC	0.619456	0.122979	0.583378	0.07526	**0.623456**	0.067008
Accuracy	0.982369	0.008096	0.963939	0.009782	0.989155	0.003788
EF5%	5.273684	4.93308	4.614474	3.020882	***5.932895***	3.549957

	10 folds/737 RDKIT		**10 folds/737 ATOM PAIR**		**10 folds/737 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.1	0.152753	0.2	0.221108	0.066667	0.133333
Specificity	0.986133	0.00451	0.9646	0.012541	0.988933	0.005122
Precision	0.02193	0.034424	0.011569	0.011922	0.042424	0.100686
f1	0.035758	0.055708	0.021694	0.022247	0.047619	0.104328
fa	0.149664	0.228616	0.276286	0.287462	0.0999	0.199799
AUC	0.537011	0.078018	**0.568978**	0.105197	0.5224	0.06799
Accuracy	0.984365	0.004726	0.963074	0.01233	0.987092	0.005299
EF5%	***3.296053***	3.296053	2.636842	3.229459	2.636842	3.229459

	10 folds/810 RDKIT		**10 folds/810 ATOM PAIR**		**10 folds/810 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.266667	0.249444	0.133333	0.163299	0.3	0.233333
Specificity	0.986	0.009951	0.950933	0.014701	0.991733	0.005418
Precision	0.052083	0.056435	0.007089	0.008724	0.100607	0.089752
f1	0.085865	0.091606	0.013459	0.016556	0.144468	0.125267
fa	0.358643	0.312592	0.198088	0.242607	0.40911	0.290749
AUC	0.619078	0.126862	0.525544	0.083665	**0.641511**	0.118214
Accuracy	0.984564	0.010037	0.949301	0.014889	0.990353	0.005545
EF5%	5.273684	4.93308	2.636842	3.229459	***6.592105***	4.169213

	10 folds/832 RDKIT		**10 folds/832 ATOM PAIR**		**10 folds/832 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.766667	0.260342	0.6	0.290593	0.833333	0.223607
Specificity	0.994467	0.002109	0.986733	0.007708	0.993133	0.003724
Precision	0.235882	0.096824	0.091022	0.044167	0.223847	0.098729
f1	0.354495	0.131062	0.155066	0.071341	0.348025	0.139742
fa	0.83804	0.190145	0.705668	0.217521	0.887319	0.157398
AUC	0.879722	0.130765	0.787133	0.148631	**0.912667**	0.112852
Accuracy	0.994012	0.002324	0.985961	0.007423	0.992814	0.003877
EF5%	***16.48026***	4.422119	12.525	6.21898	***16.48026***	4.422119

	10 folds/846 RDKIT		**10 folds/846 ATOM PAIR**		**10 folds/846 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.6	0.249444	0.533333	0.266667	0.6	0.290593
Specificity	0.9936	0.003617	0.976667	0.016414	0.9962	0.002291
Precision	0.179838	0.068893	0.05667	0.036971	0.262381	0.108279
f1	0.259161	0.07202	0.099614	0.061554	0.355519	0.154381
fa	0.717854	0.192259	0.643039	0.264525	0.708965	0.220578
AUC	0.794333	0.125003	0.747356	0.132905	**0.7968**	0.146274
Accuracy	0.992814	0.003284	0.975782	0.016164	0.995409	0.002389
EF5%	***13.18421***	5.106223	9.888158	5.314725	***13.18421***	5.106223

	10 folds/852 RDKIT		**10 folds/852 ATOM PAIR**		**10 folds/852 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.466667	0.266667	0.466667	0.305505	0.466667	0.266667
Specificity	0.997467	0.000777	0.971133	0.015894	0.996467	0.001815
Precision	0.274524	0.144255	0.034717	0.01756	0.234177	0.130117
f1	0.340159	0.17829	0.063576	0.032118	0.286642	0.125736
fa	0.589507	0.262183	0.572699	0.271829	0.589111	0.261746
AUC	**0.730567**	0.134136	0.7082	0.15488	0.729856	0.133423
Accuracy	0.996407	0.001081	0.970126	0.015613	0.995409	0.001587
EF5%	*11.86579*	3.955263	7.910526	5.746864	10.54737	5.273684

	10 folds/858 RDKIT		**10 folds/858 ATOM PAIR**		**10 folds/858 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.166667	0.223607	0.166667	0.223607	0.166667	0.223607
Specificity	0.981467	0.008583	0.969133	0.011948	0.990333	0.00294
Precision	0.0265	0.035827	0.012951	0.016011	0.042273	0.059027
f1	0.045626	0.061588	0.023792	0.029317	0.067302	0.093204
fa	0.229172	0.292293	0.227736	0.289786	0.229514	0.292784
AUC	0.566489	0.114926	0.552289	0.115218	**0.574756**	0.112861
Accuracy	0.97984	0.008832	0.967532	0.011953	0.988689	0.003246
EF5%	3.296053	4.422119	3.296053	4.422119	*3.955263*	4.372708

	10 folds/859 RDKIT		**10 folds/859 ATOM PAIR**		**10 folds/859 ECFP4**	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.066667	0.133333	0.133333	0.221108	0.033333	0.1
Specificity	0.979667	0.007878	0.966533	0.006138	0.9806	0.004526
Precision	0.009973	0.024937	0.008248	0.014065	0.006667	0.02
f1	0.016458	0.040057	0.015508	0.026409	0.011111	0.033333
fa	0.099392	0.198787	0.178122	0.282528	0.049883	0.149648
AUC	0.516078	0.066003	**0.539967**	0.1101	0.499956	0.051653
Accuracy	0.977844	0.007811	0.96487	0.006183	0.978709	0.004667
EF5%	1.318421	2.636842	***2.636842***	4.372708	1.977632	3.020882

SUPPLEMENTARY C

Metrics of all 17 targets, described with different % of fingerprints using Stratified feature Selection (models:24Actives*vs*24Inactives):

	5 folds/466 – 102 (5%)		5 folds /466 – 512 (25%)		5 folds /466 – 1024 (50%)		5 folds /466 – 1536 (75%)		5 folds /466 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.666667	0.105409
Specificity	0.996533	0.001454	0.999933	0.000133	1	0	1	0	0.620133	0.073575
Precision	0.386022	0.076454	0.971429	0.057143	1	0	1	0	0.003678	0.001036
f1	0.552307	0.085539	0.984615	0.030769	1	0	1	0	0.007314	0.002054
fa	0.998263	0.000731	0.999967	6.67E-05	1	0	1	0	0.639346	0.076632
AUC	0.997889	0.000972	0.999989	2.22E-05	1	0	1	0	0.615489	0.095551
Accuracy	0.99654	0.001452	0.999933	0.000133	1	0	1	0	0.666667	0.105409
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.317881	2.967603

	5 folds/548 – 102 (5%)		5 folds /548 – 512 (25%)		5 folds /548 – 1024 (50%)		5 folds /548 – 1536 (75%)		5 folds /548 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.966667	0.066667	1	0	1	0	1	0	0.833333	0
Specificity	0.999133	0.000499	0.999933	0.000133	1	0	1	0	0.775133	0.059859
Precision	0.71619	0.148327	0.971429	0.057143	1	0	1	0	0.007942	0.002271
f1	0.813846	0.094856	0.984615	0.030769	1	0	1	0	0.015722	0.004451
fa	0.981396	0.036291	0.999967	6.67E-05	1	0	1	0	0.801981	0.032235
AUC	0.98285	0.033412	0.999967	6.67E-05	1	0	1	0	0.796633	0.018481
Accuracy	0.999069	0.000489	0.999933	0.000133	1	0	1	0	0.77525	0.05974
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	10.61722	3.869281

	5 folds/600 – 102 (5%)		5 folds /600 – 512 (25%)		5 folds /600 – 1024 (50%)		5 folds /600 – 1536 (75%)		5 folds /600 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.933333	0.08165	1	0	1	0	1	0	0.8	0.066667
Specificity	0.995733	0.002015	0.999933	0.000133	1	0	1	0	0.649933	0.022504
Precision	0.33913	0.098747	0.971429	0.057143	1	0	1	0	0.004552	0.000301
f1	0.483713	0.092754	0.984615	0.030769	1	0	1	0	0.009052	0.000597
fa	0.961567	0.043524	0.999967	6.67E-05	1	0	1	0	0.714961	0.023532
AUC	0.964672	0.040411	0.999967	6.67E-05	1	0	1	0	0.712817	0.050377
Accuracy	0.995609	0.001863	0.999933	0.000133	1	0	1	0	0.650233	0.022371
EF5%	19.24371	1.327152	19.90728	0	19.90728	0	19.90728	0	4.645033	1.625423

	5 folds/644 – 102 (5%)		5 folds /644 – 512 (25%)		5 folds /644 – 1024 (50%)		5 folds /644 – 1536 (75%)		5 folds /644 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.866667	0.124722
Specificity	0.9974	0.001083	1	0	1	0	1	0	0.815067	0.038478
Precision	0.461107	0.11585	1	0	1	0	1	0	0.009776	0.002756
f1	0.622917	0.104484	1	0	1	0	1	0	0.019327	0.005418
fa	0.998698	0.000543	1	0	1	0	1	0	0.83642	0.072527
AUC	0.9989	0.000327	1	0	1	0	1	0	0.835028	0.074756
Accuracy	0.997405	0.001081	1	0	1	0	1	0	0.81517	0.038518
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	11.28079	3.383588

	5 folds/652 – 102 (5%)		5 folds /652 – 512 (25%)		5 folds /652 – 1024 (50%)		5 folds /652 – 1536 (75%)		5 folds /652 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.966667	0.066667	1	0	1	0	1	0	0.766667	0.08165
Specificity	0.998667	0.000471	1	0	1	0	1	0	0.605667	0.064507
Precision	0.603333	0.101016	1	0	1	0	1	0	0.00398	0.000793
f1	0.740252	0.091235	1	0	1	0	1	0	0.007918	0.001573
fa	0.981186	0.036461	1	0	1	0	1	0	0.673164	0.055113
AUC	0.982578	0.033623	1	0	1	0	1	0	0.670322	0.085021
Accuracy	0.998603	0.000572	1	0	1	0	1	0	0.605988	0.064392
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.317881	3.634556

	5 folds/689 – 102 (5%)		5 folds /689 – 512 (25%)		5 folds /689 – 1024 (50%)		5 folds /689 – 1536 (75%)		5 folds /689 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.666667	0.182574
Specificity	0.993867	0.002491	0.9998	0.000267	1	0	1	0	0.593933	0.071374
Precision	0.269744	0.08132	0.921429	0.10202	1	0	1	0	0.003484	0.001396
f1	0.418519	0.099487	0.956044	0.057732	1	0	1	0	0.006931	0.002773
fa	0.996922	0.001254	0.9999	0.000133	1	0	1	0	0.622981	0.109059
AUC	0.997006	0.001119	0.9999	0.000133	1	0	1	0	0.613689	0.132494
Accuracy	0.993879	0.002486	0.9998	0.000266	1	0	1	0	0.594079	0.071506
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.981457	3.250846

	5 folds/692 – 102 (5%)		5 folds /692 – 512 (25%)		5 folds /692 – 1024 (50%)		5 folds /692 – 1536 (75%)		5 folds /692 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.933333	0.08165	1	0	1	0	1	0	0.8	0.163299
Specificity	0.953933	0.019194	0.994333	0.00244	0.9838	0.011876	0.9938	0.004118	0.502667	0.061618
Precision	0.044709	0.014664	0.290453	0.095403	0.163751	0.102852	0.346803	0.224291	0.003307	0.000915
f1	0.08488	0.026653	0.441818	0.112903	0.268857	0.142576	0.477967	0.22371	0.006586	0.001821
fa	0.941824	0.045908	0.997157	0.001227	0.991797	0.006073	0.996886	0.002072	0.613605	0.088805
AUC	0.944883	0.042074	0.997167	0.00122	0.993528	0.003136	0.9969	0.002059	0.623161	0.120285
Accuracy	0.953892	0.019202	0.994345	0.002436	0.983832	0.011853	0.993812	0.00411	0.50326	0.061711
EF5%	17.91656	2.654305	19.90728	0	19.90728	0	19.90728	0	2.654305	2.482875

	5 folds/712 – 102 (5%)		5 folds /712 – 512 (25%)		5 folds /712 – 1024 (50%)		5 folds /712 – 1536 (75%)		5 folds /712 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.933333	0.08165	1	0	1	0	1	0	0.733333	0.08165
Specificity	0.993667	0.00152	0.999933	0.000133	1	0	1	0	0.680533	0.056553
Precision	0.233111	0.034118	0.971429	0.057143	1	0	1	0	0.004792	0.001427
f1	0.371512	0.046628	0.984615	0.030769	1	0	1	0	0.009519	0.002817
fa	0.960656	0.044026	0.999967	6.67E-05	1	0	1	0	0.703346	0.054191
AUC	0.962983	0.040763	0.999967	6.67E-05	1	0	1	0	0.695406	0.050818
Accuracy	0.993546	0.001467	0.999933	0.000133	1	0	1	0	0.680639	0.05648
EF5%	19.24371	1.327152	19.90728	0	19.90728	0	19.90728	0	7.962914	3.383588

	5 folds/713 – 102 (5%)		5 folds /713 – 512 (25%)		5 folds /713 – 1024 (50%)		5 folds /713 – 1536 (75%)		5 folds /713 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.966667	0.066666667	1	0	1	0	1	0	0.7	0.124722
Specificity	0.994467	0.001203698	1	0	1	0	1	0	0.4722	0.022065
Precision	0.267073	0.055810145	1	0	1	0	1	0	0.002658	0.000518
f1	0.416005	0.068154917	1	0	1	0	1	0	0.005297	0.001032
fa	0.979142	0.036205002	1	0	1	0	1	0	0.560232	0.05116
AUC	0.979233	0.033324168	1	0	1	0	1	0	0.576522	0.083938
Accuracy	0.994411	0.001215946	1	0	1	0	1	0	0.472655	0.022134
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.981457	3.869281

	5 folds/733 – 102 (5%)		5 folds /733 – 512 (25%)		5 folds /733 – 1024 (50%)		5 folds /733 – 1536 (75%)		5 folds /733 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.766667	0.08165
Specificity	0.995933	0.000327	1	0	1	0	1	0	0.577933	0.089426
Precision	0.33065	0.018201	1	0	1	0	1	0	0.00372	0.000562
f1	0.496696	0.020448	1	0	1	0	1	0	0.007403	0.001113
fa	0.997962	0.000164	1	0	1	0	1	0	0.649674	0.050944
AUC	0.998089	0.000281	1	0	1	0	1	0	0.634028	0.047385
Accuracy	0.995941	0.000326	1	0	1	0	1	0	0.57831	0.089144
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	2.654305	3.869281

	5 folds/737 – 102 (5%)		5 folds /737 – 512 (25%)		5 folds /737 – 1024 (50%)		5 folds /737 – 1536 (75%)		5 folds /737 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.766667	0.169967
Specificity	0.9986	0.000573	1	0	1	0	1	0	0.590333	0.015887
Precision	0.604732	0.098974	1	0	1	0	1	0	0.003715	0.000773
f1	0.748921	0.077387	1	0	1	0	1	0	0.007394	0.001539
fa	0.999299	0.000287	1	0	1	0	1	0	0.657114	0.065072
AUC	0.999283	0.000472	1	0	1	0	1	0	0.647433	0.093144
Accuracy	0.998603	0.000572	1	0	1	0	1	0	0.590685	0.015668
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	4.645033	1.625423

	5 folds/810 – 102 (5%)		5 folds /810 – 512 (25%)		5 folds /810 – 1024 (50%)		5 folds /810 – 1536 (75%)		5 folds /810 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.966667	0.066667	1	0	1	0	1	0	0.666667	0.105409
Specificity	0.995667	0.001563	0.999933	0.000133333	1	0	1	0	0.650533	0.089105
Precision	0.32327	0.061323	0.971429	0.057142857	1	0	1	0	0.004069	0.001115
f1	0.48046	0.072469	0.984615	0.030769231	1	0	1	0	0.008086	0.002211
fa	0.979704	0.036004	0.999967	6.66778E-05	1	0	1	0	0.656207	0.08654
AUC	0.981278	0.033432	0.999967	6.66667E-05	1	0	1	0	0.619578	0.108889
Accuracy	0.995609	0.001523	0.999933	0.000133067	1	0	1	0	0.650566	0.089065
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.981457	3.250846

	5 folds/832 – 102 (5%)		5 folds /832 – 512 (25%)		5 folds /832 – 1024 (50%)		5 folds /832 – 1536 (75%)		5 folds /832 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.933333	0.133333
Specificity	0.9996	0.000327	1	0	1	0	1	0	0.840067	0.065222
Precision	0.847619	0.106053	1	0	1	0	1	0	0.012811	0.003199
f1	0.913846	0.064248	1	0	1	0	1	0	0.025228	0.006199
fa	0.9998	0.000163	1	0	1	0	1	0	0.874655	0.060284
AUC	0.9998	0.000163	1	0	1	0	1	0	0.881433	0.068296
Accuracy	0.999601	0.000326	1	0	1	0	1	0	0.840253	0.064931
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	12.60795	4.876269

	5 folds/846 – 102 (5%)		5 folds /846 – 512 (25%)		5 folds /846 – 1024 (50%)		5 folds /846 – 1536 (75%)		5 folds /846 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.933333	0.08165
Specificity	0.999333	0.000869	0.999867	0.000267	1	0	1	0	0.847467	0.017964
Precision	0.813736	0.199689	0.95	0.1	1	0	1	0	0.012159	0.000943
f1	0.88236	0.136282	0.971429	0.057143	1	0	1	0	0.024001	0.001842
fa	0.999666	0.000435	0.999933	0.000133	1	0	1	0	0.885893	0.031656
AUC	0.9998	0.00024	0.999933	0.000133	1	0	1	0	0.899167	0.041964
Accuracy	0.999335	0.000867	0.999867	0.000266	1	0	1	0	0.847638	0.017809
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	15.26225	1.625423

	5 folds/852 – 102 (5%)		5 folds /852 – 512 (25%)		5 folds /852 – 1024 (50%)		5 folds /852 – 1536 (75%)		5 folds /852 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.833333	0.105409
Specificity	0.972867	0.014528	0.9906	0.002284	0.995667	0.002087	0.994067	0.002453	0.805067	0.033597
Precision	0.087817	0.039291	0.183665	0.042011	0.354329	0.120813	0.281139	0.099138	0.008573	0.000749
f1	0.159046	0.066717	0.308267	0.05822	0.511765	0.129091	0.430044	0.114576	0.016966	0.001458
fa	0.986192	0.007484	0.995276	0.001153	0.997828	0.001048	0.997023	0.001234	0.813402	0.037888
AUC	0.9884	0.006031	0.9953	0.001142	0.997833	0.001043	0.997033	0.001227	0.801489	0.041381
Accuracy	0.972921	0.014499	0.990619	0.00228	0.995675	0.002083	0.994079	0.002448	0.805123	0.033346
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	9.953642	2.098412

	5 folds/858 – 102 (5%)		5 folds /858 – 512 (25%)		5 folds /858 – 1024 (50%)		5 folds /858 – 1536 (75%)		5 folds /858 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	0.966667	0.066667	1	0	1	0	1	0	0.733333	0.133333
Specificity	0.993533	0.001784	0.999667	0.000298	1	0	1	0	0.582867	0.06665
Precision	0.239579	0.042457	0.871429	0.112031	1	0	1	0	0.003575	0.000719
f1	0.380694	0.050952	0.927473	0.063925	1	0	1	0	0.007114	0.001427
fa	0.978643	0.03561	0.999833	0.000149	1	0	1	0	0.641763	0.052824
AUC	0.979656	0.032796	0.999883	0.000125	1	0	1	0	0.645211	0.065429
Accuracy	0.99348	0.001691	0.999667	0.000298	1	0	1	0	0.583167	0.06647
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	2.654305	2.482875

	5 folds/859 – 102 (5%)		5 folds /859 – 512 (25%)		5 folds /859 – 1024 (50%)		5 folds /859 – 1536 (75%)		5 folds /859 – 2048 (100%)	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Sensitivity	1	0	1	0	1	0	1	0	0.6	0.133333
Specificity	0.998333	0.000869	1	0	1	0	1	0	0.4216	0.047446
Precision	0.575641	0.128957	1	0	1	0	1	0	0.002109	0.000606
f1	0.72203	0.10581	1	0	1	0	1	0	0.004204	0.001208
fa	0.999166	0.000435	1	0	1	0	1	0	0.492162	0.070469
AUC	0.9989	0.000331	1	0	1	0	1	0	0.475083	0.09424
Accuracy	0.998337	0.000867	1	0	1	0	1	0	0.421956	0.047535
EF5%	19.90728	0	19.90728	0	19.90728	0	19.90728	0	3.317881	3.634556

SUPPLEMENTARY D

Table D.1 - Metrics using Random Forest, without Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	0	0.998004	0	0	1	0	0.5
Target_548_ECFP4	0	0.998004	0	0	1	0	0.5
Target_600_ECFP4	0	0.998004	0	0	1	0	0.5
Target_644_ECFP4	0	0.998004	0	0	1	0	0.5
Target_652_ECFP4	0	0.998004	0	0	1	0	0.5
Target_689_ECFP4	0	0.998004	0	0	1	0	0.5
Target_692_RDKIT	0	0.998004	0	0	1	0	0.5
Target_712_ECFP4	0	0.998004	0	0	1	0	0.5
Target_713_Atom Pair	0	0.998004	0	0	1	0	0.5
Target_733_ECFP4	0	0.998004	0	0	1	0	0.5
Target_737_Atom Pair	0	0.998004	0	0	1	0	0.5
Target_810_ECFP4	0	0.998004	0	0	1	0	0.5
Target_832_ECFP4	0	0.998004	0	0	1	0	0.5
Target_846_ECFP4	0	0.998004	0	0	1	0	0.5
Target_852_RDKIT	7.962914	0.998137	0.4	0.066667	1	0.114286	0.533333
Target_858_ECFP4	0	0.998004	0	0	1	0	0.5
Target_859_Atom Pair	0	0.998004	0	0	1	0	0.5

Table D.2 - Metrics using SVM, without Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	15.92583	0.99654	0.028571	0.033333	0.998467	0.030769	0.5159
Target_548_ECFP4	19.90728	0.997738	0.346667	0.133333	0.999467	0.174315	0.5664
Target_600_ECFP4	15.92583	0.997073	0.095238	0.066667	0.998933	0.075214	0.5328
Target_644_ECFP4	19.90728	0.99654	0.181905	0.133333	0.998267	0.139103	0.5658
Target_652_ECFP4	19.90728	0.997139	0.268571	0.1	0.998933	0.124276	0.549467
Target_689_ECFP4	15.92583	0.997605	0.116667	0.066667	0.999467	0.084444	0.533067
Target_692_RDKIT	0	0.998004	0	0	1	0	0.5
Target_712_ECFP4	11.94437	0.997472	0.122222	0.066667	0.999333	0.076667	0.533
Target_713_Atom Pair	15.92583	0.997738	0.1	0.033333	0.999667	0.05	0.5165
Target_733_ECFP4	11.94437	0.996806	0.021053	0.066667	0.998667	0.032	0.532667
Target_737_Atom Pair	11.94437	0.997871	0.1	0.033333	0.9998	0.05	0.516567
Target_810_ECFP4	19.90728	0.997538	0.166667	0.066667	0.9994	0.094444	0.533033
Target_832_ECFP4	19.90728	0.997605	0.655294	0.366667	0.998867	0.401741	0.682767
Target_846_ECFP4	19.90728	0.997073	0.324603	0.266667	0.998533	0.283419	0.6326
Target_852_RDKIT	15.92583	0.99847	0.68	0.333333	0.9998	0.423203	0.666567
Target_858_ECFP4	0	0.998004	0	0	1	0	0.5
Target_859_Atom Pair	0	0.998004	0	0	1	0	0.5

Table D.3 - Metrics using Naive Bayes, without Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	15.92583	0.997538	0.1	0.033333	0.999467	0.05	0.5164
Target_548_ECFP4	7.962914	0.997804	0.1	0.033333	0.999733	0.05	0.516533
Target_600_ECFP4	15.92583	0.997671	0.2	0.033333	0.9996	0.057143	0.516467
Target_644_ECFP4	3.981457	0.997871	0	0	0.999867	0	0.499933
Target_652_ECFP4	7.962914	0.997937	0.1	0.033333	0.999867	0.05	0.5166
Target_689_ECFP4	7.962914	0.997804	0	0	0.9998	0	0.4999
Target_692_RDKIT	8.286675	0.790685	0.001564	0.166667	0.791933	0.00309	0.4793
Target_712_ECFP4	0	0.998004	0	0	1	0	0.5
Target_713_Atom Pair	18.13515	0.957152	0.00782	0.1	0.958867	0.014287	0.529433
Target_733_ECFP4	11.94437	0.997472	0	0	0.999467	0	0.499733
Target_737_Atom Pair	18.72211	0.970259	0.003175	0.033333	0.972133	0.005797	0.502733
Target_810_ECFP4	7.962914	0.997804	0	0	0.9998	0	0.4999
Target_832_ECFP4	7.962914	0.997937	0.1	0.033333	0.999867	0.05	0.5166
Target_846_ECFP4	15.92583	0.998137	0.5	0.233333	0.999667	0.3	0.6165
Target_852_RDKIT	11.01702	0.89501	0.012301	0.6	0.8956	0.024024	0.7478
Target_858_ECFP4	7.962914	0.997804	0	0	0.9998	0	0.4999
Target_859_Atom Pair	18.58828	0.951963	0	0	0.953867	0	0.476933

Table D.4 - Metrics using Logistic Regression, without Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	11.94437	0.997405	0.066667	0.033333	0.999333	0.044444	0.516333
Target_548_ECFP4	15.92583	0.997472	0.18	0.133333	0.9992	0.152727	0.566267
Target_600_ECFP4	11.94437	0.997804	0.2	0.033333	0.999733	0.057143	0.516533
Target_644_ECFP4	19.90728	0.997472	0.166667	0.066667	0.999333	0.094444	0.533
Target_652_ECFP4	19.90728	0.997605	0.2	0.033333	0.999533	0.057143	0.516433
Target_689_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_692_RDKIT	0	0.998004	0	0	1	0	0.5
Target_712_ECFP4	7.962914	0.997538	0.153333	0.1	0.999333	0.113889	0.549667
Target_713_Atom Pair	11.94437	0.997871	0.1	0.033333	0.9998	0.05	0.516567
Target_733_ECFP4	11.94437	0.997804	0.2	0.033333	0.999733	0.057143	0.516533
Target_737_Atom Pair	7.962914	0.997871	0.2	0.033333	0.9998	0.057143	0.516567
Target_810_ECFP4	19.90728	0.997671	0.2	0.033333	0.9996	0.057143	0.516467
Target_832_ECFP4	19.90728	0.997804	0.662857	0.366667	0.999067	0.409567	0.682867
Target_846_ECFP4	19.90728	0.997472	0.425714	0.266667	0.998933	0.312005	0.6328
Target_852_RDKIT	15.92583	0.998337	0.68	0.266667	0.9998	0.363203	0.633233
Target_858_ECFP4	0	0.998004	0	0	1	0	0.5
Target_859_Atom Pair	0	0.998004	0	0	1	0	0.5

SUPPLEMENTARY E

Table E.1 - Metrics using Random Forest, with Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	11.94437	0.998204	0.6	0.1	1	0.171429	0.55
Target_548_ECFP4	15.92583	0.998337	0.8	0.166667	1	0.271429	0.583333
Target_600_ECFP4	7.962914	0.998137	0.4	0.066667	1	0.114286	0.533333
Target_644_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_652_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_689_ECFP4	7.962914	0.998204	0.4	0.1	1	0.157143	0.55
Target_692_RDKIT	0	0.998004	0	0	1	0	0.5
Target_712_ECFP4	0	0.998004	0	0	1	0	0.5
Target_713_Atom Pair	19.90728	0.999468	1	0.733333	1	0.838788	0.866667
Target_733_ECFP4	3.981457	0.998137	0.2	0.066667	1	0.1	0.533333
Target_737_Atom Pair	19.90728	0.9998	1	0.9	1	0.933333	0.95
Target_810_ECFP4	7.962914	0.998137	0.4	0.066667	1	0.114286	0.533333
Target_832_ECFP4	0	0.998004	0	0	1	0	0.5
Target_846_ECFP4	15.92583	0.998669	0.8	0.333333	1	0.457143	0.666667
Target_852_RDKIT	15.92583	0.998403	0.8	0.2	1	0.314286	0.6
Target_858_ECFP4	0	0.998004	0	0	1	0	0.5
Target_859_Atom Pair	19.90728	0.999667	1	0.833333	1	0.89697	0.916667

Table E.2 - Metrics using SVM, with Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	19.90728	1	1	1	1	1	1
Target_548_ECFP4	19.90728	1	1	1	1	1	1
Target_600_ECFP4	19.90728	1	1	1	1	1	1
Target_644_ECFP4	19.90728	1	1	1	1	1	1
Target_652_ECFP4	19.90728	0.9998	1	0.9	1	0.941818	0.95
Target_689_ECFP4	19.90728	1	1	1	1	1	1
Target_692_RDKIT	19.90728	0.999268	1	0.633333	1	0.75697	0.816667
Target_712_ECFP4	19.90728	1	1	1	1	1	1
Target_713_Atom Pair	19.90728	0.999933	1	0.966667	1	0.981818	0.983333
Target_733_ECFP4	19.90728	0.999933	1	0.966667	1	0.981818	0.983333
Target_737_Atom Pair	19.90728	0.999867	1	0.933333	1	0.96	0.966667
Target_810_ECFP4	19.90728	0.999933	1	0.966667	1	0.981818	0.983333
Target_832_ECFP4	19.90728	1	1	1	1	1	1
Target_846_ECFP4	19.90728	1	1	1	1	1	1
Target_852_RDKIT	19.90728	0.999468	1	0.733333	1	0.835152	0.866667
Target_858_ECFP4	19.90728	1	1	1	1	1	1
Target_859_Atom Pair	19.90728	1	1	1	1	1	1

Table E.3 - Metrics using Naive Bayes, with Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_548_ECFP4	7.962914	0.998137	0.4	0.066667	1	0.114286	0.533333
Target_600_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_644_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_652_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_689_ECFP4	0	0.998004	0	0	1	0	0.5
Target_692_RDKIT	19.52583	0.983433	0.273011	0.933333	0.983533	0.361573	0.958433
Target_712_ECFP4	0	0.998004	0	0	1	0	0.5
Target_713_Atom Pair	15.92583	0.998403	0.8	0.2	1	0.314286	0.6
Target_733_ECFP4	0	0.998004	0	0	1	0	0.5
Target_737_Atom Pair	11.94437	0.99827	0.6	0.133333	1	0.214286	0.566667
Target_810_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_832_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_846_ECFP4	19.90728	0.998603	1	0.3	1	0.431429	0.65
Target_852_RDKIT	19.90728	0.993546	0.262161	0.933333	0.993667	0.400788	0.9635
Target_858_ECFP4	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667
Target_859_Atom Pair	3.981457	0.998071	0.2	0.033333	1	0.057143	0.516667

Table E.4 - Metrics using Logistic Regression, with Stratified Feature Selection

	EF5%	SPE	RECALL	PREC	ACCU	f1	AUC
Target_466_ECFP4	19.90728	1	1	1	1	1	1
Target_548_ECFP4	19.90728	1	1	1	1	1	1
Target_600_ECFP4	19.90728	1	1	1	1	1	1
Target_644_ECFP4	19.90728	1	1	1	1	1	1
Target_652_ECFP4	19.90728	0.999933	1	0.966667	1	0.981818	0.983333
Target_689_ECFP4	19.90728	1	1	1	1	1	1
Target_692_RDKIT	19.90728	0.9998	1	0.9	1	0.941818	0.95
Target_712_ECFP4	19.90728	1	1	1	1	1	1
Target_713_Atom Pair	19.90728	1	1	1	1	1	1
Target_733_ECFP4	19.90728	1	1	1	1	1	1
Target_737_Atom Pair	19.90728	1	1	1	1	1	1
Target_810_ECFP4	19.90728	0.999933	1	0.966667	1	0.981818	0.983333
Target_832_ECFP4	19.90728	1	1	1	1	1	1
Target_846_ECFP4	19.90728	1	1	1	1	1	1
Target_852_RDKIT	19.90728	0.999601	1	0.8	1	0.883636	0.9
Target_858_ECFP4	19.90728	1	1	1	1	1	1
Target_859_Atom Pair	19.90728	1	1	1	1	1	1

4.4. Patent: BR 10 2020 007050 9 - Uso Do Tetraclorodecaoxido Para Produzir Medicamentos Para Tratar Pacientes Com Covid-19

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

08/04/2020 870200044848 15:38

29409161911880135

Pedido nacional de Invenção, Modelo de Utilidade, Certificado de Adição de Invenção e entrada na fase nacional do PCT

Número do Processo: BR 10 2020 007050 9

Dados do Depositante (71)

Depositante 1 de 1

Nome ou Razão Social: UNIVERSIDADE FEDERAL DE MINAS GERAIS

Tipo de Pessoa: Pessoa Jurídica

CPF/CNPJ: 17217985000104

Nacionalidade: Brasileira

Qualificação Jurídica: Instituição de Ensino e Pesquisa

Endereço: Av. Antônio Carlos, 6627 - Unidade Administrativa II - 2º andar- sala 2011

Cidade: Belo Horizonte

Estado: MG

CEP: 31270-901

País: Brasil

Telefone: (31) 3409-6430

Fax:

Email: patentes@ctit.ufmg.br

Dados do Pedido

Natureza Patente: 10 - Patente de Invenção (PI)

Título da Invenção ou Modelo de Utilidade (54): USO DO TETRACLORODECAOXIDO PARA PRODUZIR MEDICAMENTOS PARA TRATAR PACIENTES COM COVID-19

Resumo: A presente tecnologia trata da seleção de um fármaco para reposicionamento, para produção de medicamentos para o tratamento de pacientes infectados pelo coronavírus SARS-CoV-2, agente etiológico da COVID-19. Por recuperação de informação de semântica latente, foram recuperados fármacos que poderão integrar a terapêutica farmacológica para COVID-19. Dentre eles, o tetraclorodecaoxido, que pode ser utilizado como único princípio ativo, ou pode estar associado a outros fármacos, preferencialmente antivirais.

Dados do Inventor (72)

Inventor 1 de 2

Nome: CARMELINA FIGUEIREDO VIEIRA LEITE

CPF: 01818638630

Nacionalidade: Brasileira

Qualificação Física: Pesquisador

Endereço: Rua Cláudio Manoel, 210, apt 101. Funcionários

Cidade: Belo Horizonte

Estado: MG

CEP: 30140-100

País: BRASIL

Telefone: (31) 340 93932

Fax:

Email: patentes@ctit.ufmg.br

Inventor 2 de 2

Nome: MARCOS AUGUSTO DOS SANTOS

CPF: 27456510644

Nacionalidade: Brasileira

Qualificação Física: Professor do ensino superior

Endereço: Alameda Serra da Canastra, 197. Vila Del Rei

Cidade: Nova Lima

Estado: MG

CEP: 34007-209

País: BRASIL

Telefone: (31) 340 93932

Fax:

Email: patentes@ctit.ufmg.br

Documentos anexados

Tipo Anexo	Nome
Comprovante de pagamento de GRU 200	1 - Comprovante pagamento GRU 29409161911880135.pdf
Portaria	2 - Portaria 010-2019 - Prof. Gilberto UFMG.pdf
Relatório Descritivo	3 - Relatório descritivo.pdf
Reivindicação	4 - Reivindicações.pdf
Resumo	5 - Resumo.pdf

Acesso ao Patrimônio Genético

☑ Declaração Negativa de Acesso - Declaro que o objeto do presente pedido de patente de invenção não foi obtido em decorrência de acesso à amostra de componente do Patrimônio Genético Brasileiro, o acesso foi realizado antes de 30 de junho de 2000, ou não se aplica.

Declaração de veracidade

☑ Declaro, sob as penas da lei, que todas as informações acima prestadas são completas e verdadeiras.

INSTRUÇÕES:

A data de vencimento não prevalece sobre o prazo legal. O pagamento deve ser efetuado antes do protocolo. Órgãos públicos que utilizam o sistema SIAFI devem utilizar o número da GRU no campo Número de Referência na emissão do pagamento. Serviço: 200-Pedido nacional de Invenção, Modelo de Utilidade, Certificado de Adição de Invenção e entrada na fase nacional do PCT

Clique aqui e pague este boleto através do Auto Atendimento Pessoa Física.

Clique aqui e pague este boleto através do Auto Atendimento Pessoa Jurídica.

Recibo do Pagador

BANCO DO BRASIL | **001-9** | **00190.00009 02940.916196 11880.135170 4 80820000007000**

Nome do Pagador/CPF/CNPJ/Endereço
UNIVERSIDADE FEDERAL DE MINAS GERAIS CPF/CNPJ: 17217985000104
AV ANTONIO CARLOS 6627 UNIDADE ADMINISTRATIVA II 2 ANDAR SALA 2011, BELO HORIZONTE -MG CEP:31270901
Sacador/Avalista

Nosso-Número	Nr. Documento	Data de Vencimento	Valor do Documento	(=) Valor Pago
29409161911880135	29409161911880135	23/11/2019	70,00	

Nome do Beneficiário/CPF/CNPJ/Endereço
INSTITUTO NACIONAL DA PROPRIEDADE INDUST CPF/CNPJ: 42.521.088/0001-37
RUA MAYRINK VEIGA 9 24 ANDAR ED WHITE MARTINS , RIO DE JANEIRO - RJ CEP: 20090910

Agência/Código do Beneficiário
2234-9 / 333028-1

Autenticação Mecânica

BANCO DO BRASIL | **001-9** | **00190.00009 02940.916196 11880.135170 4 80820000007000**

Local de Pagamento
PAGÁVEL EM QUALQUER BANCO ATÉ O VENCIMENTO

Data de Vencimento
23/11/2019

Nome do Beneficiário/CPF/CNPJ
INSTITUTO NACIONAL DA PROPRIEDADE INDUST CPF/CNPJ: 42.521.088/0001-37

Agência/Código do Beneficiário
2234-9 / 333028-1

Data do Documento	Nr. Documento	Espécie DOC	Aceite	Data do Processamento
25/10/2019	29409161911880135	DS	N	25/10/2019

Nosso-Número
29409161911880135

Uso do Banco	Carteira	Espécie	Quantidade	xValor
29409161911880135	17	R$		

(=) Valor do Documento
70,00

Informações de Responsabilidade do Beneficiário

A data de vencimento não prevalece sobre o prazo legal.
O pagamento deve ser efetuado antes do protocolo.
Órgãos públicos que utilizam o sistema SIAFI devem utilizar o número da GRU n
o campo Número de Referência na emissão do pagamento.
Serviço: 200-Pedido nacional de Invenção, Modelo de Utilidade, Certificado de
Adição de Invenção e entrada na fase nacional do PCT

(-) Desconto/Abatimento

(+) Juros/Multa

(=) Valor Cobrado

Nome do Pagador/CPF/CNPJ/Endereço
UNIVERSIDADE FEDERAL DE MINAS GERAIS CPF/CNPJ: 17217985000104
AV ANTONIO CARLOS 6627 UNIDADE ADMINISTRATIVA II 2 ANDAR SALA 2011,
BELO HORIZONTE-MG CEP:31270901

Código de Baixa

Sacador/Avalista

Autenticação Mecânica - **Ficha de Compensação**

__ SIAFI2019-DOCUMENTO-CONSULTA-CONGRU (CONSULTA GUIA DE RECOLHIMENTO DA UNIAO
05/11/19 16:44 USUARIO : CELTON
DATA EMISSAO : 04Nov19 TIPO : 1 - PAGAMENTO NUMERO : 2019GR800932
UG/GESTAO EMITENTE : 153254 / 15229 - ADMINISTRACAO GERAL/UFMG
UG/GESTAO FAVORECIDA : 183038 / 18801 - INSTITUTO NACIONAL DA PROPRIEDADE INDU
RECOLHEDOR : 153254 GESTAO : 15229
CODIGO RECOLHIMENTO : 72200 - 6 COMPETENCIA: OUT19 VENCIMENTO: 08Nov19
DOC. ORIGEM: 153254 / 15229 / 2019NP002451 PROCESSO :
RECURSO : 1
(=)VALOR DOCUMENTO : 70,00
(-)DESCONTO/ABATIMENTO:
(-)OUTRAS DEDUCOES :
(+)MORA/MULTA :
(+)JUROS/ENCARGOS :
(+)OUTROS ACRESCIMOS :
(=)VALOR TOTAL : 70,00
NOSSO NUMERO/NUMERO REFERENCIA : 00029409161911880135
CODIGO DE BARRAS : 89610000000 0 70000001010 3 95523127220 9 00360640000 4
OBSERVACAO
Serviço: 200 - Pedido nacional de invenção, modelo de utilidade, certificado d
e adição de invenção e entrada na fase nacional do PCT.
LANCADO POR : 09663457627 - LUDMILA UG : 153254 04Nov2019 16:51
PF1=AJUDA PF3=SAI PF2=DADOS ORC/FIN PF4=ESPELHO PF12=RETORNA

CTIT

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Gabinete da Reitora

PORTARIA Nº 010, DE 24 DE JANEIRO DE 2019

A REITORA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, no uso de suas atribuições legais e estatutárias, considerando o disposto nos artigos 11 e 12 do Decreto-Lei n° 200, de 25 de fevereiro de 1967, bem como no Decreto nº 7.689, de março de 2012; na Portaria nº 249, de 13 de junho de 2012, do Ministério do Planejamento, Orçamento e Gestão (MPOG), no Decreto nº 9.189, de 1º de novembro de 2017, e no inciso II do art. 2º, combinado com o art. 3º, ambos da Portaria MEC nº 36, de 18 de janeiro de 2018, do Ministério da Educação (MEC),

RESOLVE:

Art. 1º. Tornar sem efeito a portaria nº 174, de 8 de agosto de 2018;

Art. 2º Delegar competência ao Diretor da Coordenadoria de Transferência e Inovação Tecnológica (CTIT), professor GILBERTO MEDEIROS RIBEIRO, inscrição UFMG nº 247405, matrícula SIAPE nº 1964486, e a seu substituto eventual, para, no âmbito da CTIT:

a) assinar, por meio eletrônico ou físico, documentos ou instrumentos jurídicos concernentes ao exercício das atividades de competência da CTIT, no âmbito da Lei 10.973/04 – Lei de Inovação Tecnológica, da Política de Inovação da UFMG e suas resoluções específicas, tais como Contrato de Transferência de *Know-How*, Contrato de Licenciamento de Tecnologia, Contrato de Partilhamento de Titularidade de Tecnologia, Acordos de Confidencialidade e Termos de Sigilo, Termos de Autorização de Teste e documentos afins;

b) assinar, por meio eletrônico ou físico, documentação necessária para depósito, processamento, adição, retificação, substituição, modificação, ampliação e resposta de relatórios referentes a objeto de proteção de propriedade intelectual junto aos órgãos competentes, em âmbito nacional e internacional;

c) autorizar a realização de despesas dentro dos limites orçamentários da CTIT;

b) autorizar a concessão de suprimento de fundos a servidores da Unidade, bem como determinar a baixa de responsabilidade;

c) requisitar passagens e transportes em geral, por quaisquer vias, nos limites da dotação orçamentária da CTIT;

(...)

Avenida Antônio Carlos, 6627 · 31270-901 · Belo Horizonte · Minas Gerais
+55 31 3409-4127 · Fax: 31 3409-4130 · chefia@gabinete.ufmg.br · www.ufmg.br

ORDM nº 3921/2019

PORTARIA Nº 010, DE 24 DE JANEIRO DE 2019

d) autorizar viagens de servidores, a serviço da Unidade, arbitrando-lhes as respectivas diárias, obedecidas as disposições legais pertinentes;

e) assinar contratos, decorrentes de licitação, de sua dispensa ou inexigibilidade, no âmbito da CTIT;

f) prover arrecadação de receitas em geral no âmbito da CTIT;

g) apurar dívidas de terceiros para com a Universidade, oriundas de contratos de cotitularidade, licenciamento, transferência, dentre outros, adotando as medidas necessárias à regularização delas, no âmbito da CTIT.

Art. 3º Subdelegar competência ao Diretor da CTIT para:

a) celebrar novos contratos administrativos decorrentes de licitação, de sua dispensa ou de inexigibilidade ou prorrogar contratos em vigor relativos a atividades de custeio cujos valores sejam inferiores a R$500.000,00 (quinhentos mil reais); e

b) autorizar a realização de despesas relativas a atividades de custeio cujos valores sejam inferiores a R$500.000,00 (quinhentos mil reais).

Art. 4º A presente Portaria entra em vigor nesta data.

Belo Horizonte, 24 de janeiro de 2019.

Profa. Sandra Regina Goulart Almeida
Reitora

CONFERE COM O ORIGINAL
EM 28 / 01 / 19
Juliana Campideli Neves Garcia
Secretária Executiva
Inscrição nº 22547-9

SCG/jcng

L:Document/Portaria/p19-010

“USO DO TETRACLORODECAOXIDO PARA PRODUZIR MEDICAMENTOS PARA TRATAR PACIENTES COM COVID-19”

[01] A presente tecnologia trata da seleção de um fármaco para reposicionamento, para produção de medicamentos para o tratamento de pacientes infectados pelo coronavírus SARS-CoV-2, agente etiológico da COVID-19. Por recuperação de informação de semântica latente, foram recuperados fármacos que poderão integrar a terapêutica farmacológica para COVID-19. Dentre eles, o tetraclorodecaoxido, que pode ser utilizado como único princípio ativo, ou pode estar associado a outros fármacos, preferencialmente antivirais.

[02] A busca por medicamentos para tratar pacientes acometidos pela síndrome respiratória aguda severa (SARS) após infecção por SARS-CoV-2 tem se voltado para o reposicionamento de fármacos já aprovados para uso humano, tendo em vista a urgente necessidade de aplicação do tratamento.

[03] A presente tecnologia propõe o reposicionamento do fármaco tetraclorodecaoxido,(TCDO), o qual foi selecionado após análise *in silico,* criada aqui para esse fim. A análise *in silico* aqui realizada utiliza mecanismos para recuperar os resultados de uma pesquisa de consulta de palavras únicas ou múltiplas em um banco de dados de documentos. Nesse contexto, conceitos como indexação semântica latente (LSI) (DOI:10.1017/S0962492906240017), decomposição por valores singulares (SVD) e distância do cosseno/euclideana são introduzidos. Na presente tecnologia, os alvos teóricos para a SARS-CoV-2 são representados em um modelo de espaço vetorial (VSM), em que proteínas do vírus são tratadas como documentos e as entradas do InterPro (seus respectivos descritores) são tratados como palavras em um documento, em analogia ao VSM em informações de sistemas de recuperação. A aplicação de um LSI que usa SVD no conjunto de dados

de alvos de medicamentos levou à criação de um espaço-alvo reduzido para drogas que revelou uma correlação latente entre esses alvos. A aplicação da métrica da distância do cosseno e euclideana com um ponto de corte de valor definido sinalizou as proteínas que estavam localizadas topologicamente perto de alvos teóricos.

[04] Vários fármacos vêm sendo testados para o tratamento de pacientes infectados por SARS-CoV-2. Dentre eles, podem-se citar: Pirfenidona; ASC-09/ritonavir; Darunavir; Cobicistat; Cloroquina; Hidroxicloroquina; Interferon beta; Remdesivir; Ritonavir; Lopinavir; Sarilumab; Oseltamivir; Arbidol (umifenovir); Losartana de potássio. Nenhum desses fármacos apresenta semelhança com o TCDO.

[05] O fármaco tetraclorodecaoxido (TCDO) tem sido descrito, no estado da técnica, para vários usos, mas nenhum deles relacionado ao tratamento de pacientes infectados por coronavirus.

[06] O fármaco tetraclorodecaoxido (TCDO) foi proposto para o tratamento de feridas, tendo em vista sua atividade como bactericida, apresentando, como mecanismo de ação, a ativação direta do sistema de macrófagos e o aumento da pressão parcial de oxigênio, além de propriedades mitogênicas em fibroblastos e em novos vasos sanguíneos. Além disso, sua decomposição não gera metabólitos tóxicos (Parikh R et al., “The Efficacy and Safety of Tetrachlorodecaoxide in Comparison with Super-oxidised Solution in Wound Healing”, Arch Plast Surg. 2016 Sep;43(5):395-401. doi: 10.5999/aps.2016.43.5.395. Epub 2016 Sep 21).

[07] O TCDO, também denominado WF10, foi proposto para tratamento de úlceras de pé diabético (Yingsakmongkol N., “Clinical outcomes of WF10 adjunct to standard treatment of diabetic foot ulcers”. J Wound Care. 2013 Mar;22(3):130-2, 134-6. PMID: 23665731; Yingsakmongkol N, Maraprygsavan P, Sukosit P., “Effect of WF10 (immunokine) on diabetic foot ulcer therapy: a double-blind, randomized, placebo-

controlled trial", J Foot Ankle Surg. 2011 Nov-Dec;50(6):635-40. doi: 10.1053/j.jfas.2011.05.006. Epub 2011 Jul 1. PMID: 21723750); além de ser proposto no tratamento de pacientes com câncer (Kühne L et al., "WF10 stimulates NK cell cytotoxicity by increasing LFA-1-mediated adhesion to tumor cells", J Biomed Biotechnol. 2011;2011:436587. doi: 10.1155/2011/436587. Epub 2011 May 3. PMID: 21629753).

[08] O documento de patente DE3625867, cuja data de prioridade é 31/07/1986, intitulado "Use of tetrachlorodecaoxide in ophthalmology", descreve o uso de tetrachlorodecaoxide para preparar medicamento oftalmológico com atividade bactericida.

[09] O documento de patente US 5877222, cuja data de prioridade é 02/03/1999, intitulado "Method for treating aids-associated dementia", descreve o uso de TCDO para tratar demência associada à SIDA, cujo mecanismo de ação seria a inibição da produção de TNF-alfa.

[010] A presente invenção propõe o uso de TCDO para produzir medicamentos para tratamento de pacientes com COVID-19.

DESCRIÇÃO DETALHADA DA TECNOLOGIA

[011] A presente tecnologia trata da seleção de um fármaco para reposicionamento, para produção de medicamentos para o tratamento de pacientes infectados pelo SARS-CoV-2, agente etiológico da COVID-19. Por recuperação de informação de semântica latente, foram recuperados fármacos que poderão integrar a terapêutica farmacológica para COVID-19. Dentre eles, o tetraclorodecaoxido, que pode ser utilizado como único princípio ativo, ou pode estar associado a outros fármacos, preferencialmente antivirais.

[012] A presente invenção pode ser mais bem compreendida através dos exemplos que se seguem, não limitantes.

EXEMPLO 1 – ANÁLISE *IN SILICO* PARA BUSCA DE FÁRMACOS PARA REPOSICIONAMENTO

[013] O método para extração da informação latente foi baseado em fármacos aprovados, descritos por Interpro (DOI:10.1093/bioinformatics/btu792; Estrela, B., Repurposing approved drugs using search engine idea, 2018).

[014] Foi realizada a decomposição por valores singulares (Estrela, B.; Repurposing approved drugs using search engine idea; 2018), para retirar o ruído da matriz, evidenciando melhor as relações entre os alvos, nomeadamente, as latentes. No caso da presente tecnologia, foram utilizados 334 valores singulares para representar a matriz. Com a técnica de indexação por semântica latente, foi possível relacionar alvos que não estão diretamente próximos (relacionados). No caso da presente tecnologia, os alvos considerados diretamente relacionáveis foram aqueles que partilham InterPros. Alvos latentes foram aqueles que estão próximos, porém, não partilham InterPros. Tendo em vista todas as técnicas disponíveis na Indústria Farmacêutica, os alvos diretamente relacionáveis foram descartados. Os alvos latentes, que provêm de técnicas de engenharia de busca mais recentes, foram considerados.

[015] A proteína estudada e projetada para o caso do SARS_CoV-2 foi a Replicase polyprotein 1ab (UniProtKB - P0C6X3). Adicionalmente, foi projetada a proteína Orf1a polyprotein (UniProtKB - A7J8L3). Esta foi escolhida pela comparação do genoma do SARS-CoV-2 com a vizinhança filogenética. Os resultados dos alvos latentes encontrados estão representados na **Tabela 1**.

[016] Aplicando a regressão logística, foi possível identificar o fragmento do genoma que difere dos genomas vizinhos. Com essa informação, foi possível localizar o *locus* e a proteína que é traduzida. Conforme mostrado na **Tabela 2**, o fármaco comercializado que modula essa proteína é o ácido benzóico (DrugBank ID -DB08748).

Tabela 1 – Resultados da Indexação de Semântica latente

Alvo	**Alvo latente**	**InterPro**	**Fármaco aprovado que atue no alvo latente**	**Distância Euclideana, em relação ao alvo**
A7J8L3	Q86VB7	IPR001190 IPR017448 IPR036772	DB05389	0.597246508
P0C6X3	P26676	IPR024352 IPR039736 IPR026890 IPR014023 IPR025786 IPR016269 IPR002877	DB00811	1.1331
			DB06408	

Tabela 2 – Resultado da utilização da regressão logística, com fragmentos do genoma

Genoma SARS-CoV-2	**n.º sequências próximas filogeneticamrne**	**Fragmento (tamanho 9) identificado**	***locus***	**Proteina codificada**	**Fármaco que modula esta proteina**
NC_045512.2/ MN908947.3	403	GTACGGTCG	*orf1ab*	Orf1ab polyprotein	DB08748

[017] Dessa forma, a análise dos resultados permitiu a seleção dos seguintes fármacos para reposicionamento: ácido benzoico (DrugBank ID -DB08748); ribavirina (DrugBank ID - DB00811); taribavirina (DrugBank ID - DB06408); tetraclorodecaoxido (DrugBank ID - DB05389).

[018] O método proposto para seleção de fármacos para reposicionamento se mostrou eficaz, tendo em vista que os fármacos ácido benzoico, ribavirina e seu pró-fármaco taribavirina já apresentam evidência de eficácia para uso no tratamento de pacientes com coronavírus (Verschueren KH et al., "A structural view of the inactivation of the SARS coronavirus main proteinase by benzotriazole esters", Chem Biol. 2008 Jun;15(6):597-606. doi: 10.1016/j.chembiol.2008.04.011;

NEWS 27 FEBRUARY 2020 Coronavirus puts drug repurposing on the fast trac Existing antivirals and knowledge gained from the SARS and MERS outbreaks gain traction as the fastest route to fight the current coronavirus epidemic. disponível em: https://www.nature.com/articles/d41587-020-00003-1, acessado em 30/03/2020).

[019] Diante das evidências de eficácia do método de seleção *in silico* proposto, conclui-se que o fármaco tetraclorodecaoxido, ainda não descrito para esse uso, apresenta-se como forte candidato para o tratamento de pacientes infectados por SARS-CoV-2.

REIVINDICAÇÕES

1. USO DO TETRACHLORODECAOXIDE, caracterizado por ser para produzir medicamentos para o tratamento de pacientes infectados por COVID-19.

2. USO DO TETRACHLORODECAOXIDE, de acordo com a reivindicação1, caracterizado por ser em combinação com outros fármacos, preferencialmente antivirais.

RESUMO

“USO DO TETRACLORODECAOXIDO PARA PRODUZIR MEDICAMENTOS PARA TRATAR PACIENTES COM COVID-19”

A presente tecnologia trata da seleção de um fármaco para reposicionamento, para produção de medicamentos para o tratamento de pacientes infectados pelo coronavírus SARS-CoV-2, agente etiológico da COVID-19. Por recuperação de informação de semântica latente, foram recuperados fármacos que poderão integrar a terapêutica farmacológica para COVID-19. Dentre eles, o tetraclorodecaoxido, que pode ser utilizado como único princípio ativo, ou pode estar associado a outros fármacos, preferencialmente antivirais.

5. Integrative Discussion

The Milk-Way algorithm was built in a heterogeneous way, using and aggregating different techniques so that the final classification had high performance. It was made possible by *ad-hoc* selection for both active and inactive compounds, which respects the chemical subtleties of each test and dealing with imbalanced data. Another critical step is the stratified selection of features, with the possibility of using all descriptors, even if they are more than the number of compounds. This last detail was possible because of the modification in the system resolution of the *logit* transformation.

The regression method is helpful to any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The main focus of logistic regression analysis is a classification of individuals in different groups. A standard regression equation consists of the true values of a few independent variables and weights produced by the model to predict the value of the dependent variable. The dependent variable is the predicted variable, while the independent variables are constants or categorical. The term logistic regression analysis comes from the *logit* transformation, which is applied to the dependent variable. Simple and multiple linear regression analysis is used to evaluate the mathematical correlation between dependent variables and the independent variable(s) (COKLUK, 2010).

The patent BR 10 2019 027703 3 (Item 4.1) describes the Milk-Way algorithm using a study of NRTIs, with and without stratified feature selection. The blind test was evident that the Milk-Way algorithm overcame the results of SVM and RF.

The paper published (Item 4.2) described the Milk-Way algorithm, but it had a study case of CDK-2. Using a docking protocol (DOCK6.8) (ALLEN; BALIUS; MUKHERJEE; BROZELL *et al.*, 2015), we were able to evaluate our proposed drugs to repurpose inside of the target. Despite the importance of the CDK-2 protein, not many commercial drugs act against it. Thus, we investigated the use of drug repurposing as an aid to CDK-2 drug development. Drug repurposing is the strategy of discovering new uses or conditions for approved drugs to not only assess the effects of the drug on a new target but also to reduce the cost of developing a new drug. The screening was performed using commercialized drugs retrieved from the DrugBank (WANG; BOLTON; DRACHEVA; KARAPETYAN *et al.*, 2010). The drugs which

obtain a probability (P(*x*)) of 0.98 or higher were selected. In total, five drugs were selected: pramocaine, prochlorperazine, trifluoperazine, methionine, and pergolide. These first three drugs have already been mentioned to repurpose CDK-2 (FENG; XIA; GAO; XU *et al.*, 2018; HARDT; BEBER; RASCHE; KAMBUROV *et al.*, 2019; QI; DING, 2013).

The draft paper (Item 4.3) used a benchmark dataset, with 17 targets from the MUV database. Just with the stratified feature selection that metrics reach AUC above 0.90. There is one observation that has a significant impact - the stratified feature selection made through the Milk-way algorithm also improved the performance of SVM e LR. Eventually, with this high metrics, we could hypothesize that it occurs overfitting. However, it is improbable because all metrics, with these three algorithms, reached better metrics. Overfitting occurs when a machine learning model is trained to predict the training data too well, such that it does not generalize to new data sets (LIU; CHEN; KRAUSE; PENG, 2019).

The patent BR 10 2020 007050 9 (Item 4.4) results in the exploration of vector space, suppressing the limitation of Milk-Way, which needs a ligand to be built. It was a suggestion during the pandemic time.

Nevertheless, all the *in silico* suggestions have to be confirmed *in vitro* assays.

6. Final Considerations and Perspectives

Our holistic approach can classify ligands with the support of the selected case studies. The use of literature data and datasets were appropriate for testing the algorithm and measuring the results. It is essential to notice, the cases investigated throughout the papers and patents are unrelated to each other, demonstrating a practical way to prove the efficiency of the proposed algorithm for LBVS. Nevertheless, it is essential to highlight the acceptance of a higher number of attributes (ligands' features) than entities (ligands), without a problem of rank deficiency. This added factor is opposed to the classical logistic regression in which is obligatorily to have more entities than attributes. The proposed mathematical modulation does not require a massive infra-structure apparatus to be performed and constitutes a good strategy for selecting promising compounds. With the analysis of alpha value, by the logic and rigor of the equation, it is expected that the higher alpha values belong to the essential parts of the reference ligand. This technique indicates that analysis was not only a classifier but also an indicator of critical components in the ligands. By including docking protocol results, it was possible to infer the potential complementarity of the selected ligands in a protein environment. It will always be an approximation, but, as we have a benchmark (a commercial drug approved), the comparations were made with more confidence and consistency.

The Milk-Way algorithm is non-recursive and demonstrated excellent performance when compared to well-known methods like SVM and RF, and with less computational infrastructure. Since there is an attempt to continually improve the efficiency of computational processes to develop new and repositioning drugs, we proposed a robust approach that provides a general classifier to separate actives from inactives compounds in a dataset of ligands for any data-driven LBVS.

Our proposed classification algorithm will undoubtedly help in the process of discovering new drugs.

I am pleased and grateful to reach the main aim of the thesis, from a personal and academic point of view.

The next step of the evolution of the Milk-Way algorithm is to applicate it to other targets and validated them with *in vitro* assays.

Another work is in development, and negotiation is the repurposing of drugs to integrate the pharmacological therapy of COVID-19 and to ZIKA virus infection. Based on the theoretical therapeutic target of SARS-CoV-2 (ZHANG; KUTATELADZE, 2020) and ZIKA virus (YUAN; CHAN; DEN-HAAN; CHIK *et al.*, 2017; ZHAO; YI; DU; CHUANG *et al.*, 2017), we used the Milk-Way and the exploration of vector space to propose potential drugs to lead this pandemic. For a suggestion of Coordenadoria de Transferência e Inovação Tecnológica, as the drugs are being negotiated, we could not mention which are.

An example of the exploration of vector space, it is the study case of CDK-2. The first target that does not have InterPro in common, O15431 (High-affinity copper uptake protein 1). The drugs which interact in this target are (MAGRANE; CONSORTIUM, 2011): Carboplatin; Cisplatin; Copper; Oxaliplatin. We can confirm, with literature, that the target is related to CDK-2.

The cisplatin toxicity can produce hearing loss, but there are no approved drugs to prevent it. Assay *in vitro* and *in vivo* demonstrated CDK-2 inhibitors prevent cisplatin-induced ototoxicity (HAZLITT; TEITZ; BONGA; FANG *et al.*, 2018; TEITZ; FANG; GOKTUG; BONGA *et al.*, 2018). It showed reduced cisplatin-induced mitochondrial production of reactive oxygen species and caspase 3/7-mediated cell death (TEITZ; FANG; GOKTUG; BONGA *et al.*, 2018). Another presumed relation is with oxaliplatin. The adapalene is a synthetic retinoid which it is mainly used for topical therapy of acne vulgaris. Shi *et al*. (2015) showed that adapalene inhibits CDK-2, potentially using as a drug for treating human colorectal cancer, particularly with oxaliplatin. The carboplatin, cytoplasmatic CDK-2 was used like a protein-based biomarker, in the combination of carboplatin and eribulin, in phase II, in women with triple negative early-stage breast cancer (KAKLAMANI; JERUSS; HUGHES; SIZIOPIKOU *et al.*, 2015). Until this moment, I did not find a relation with CDK-2 and copper, just with CDK-1 (MIGUEL; PETERSEN; GONZALES-ZUBIATE; OLIVEIRA *et al.*, 2015).

7. Other works

We performed a website (http://bioinfo.dcc.ufmg.br/kardexplore/), using data-mining to analyze a philosophical paper written by Allan Kardec (KARDEC, 1867), *first-of-its-kind*.

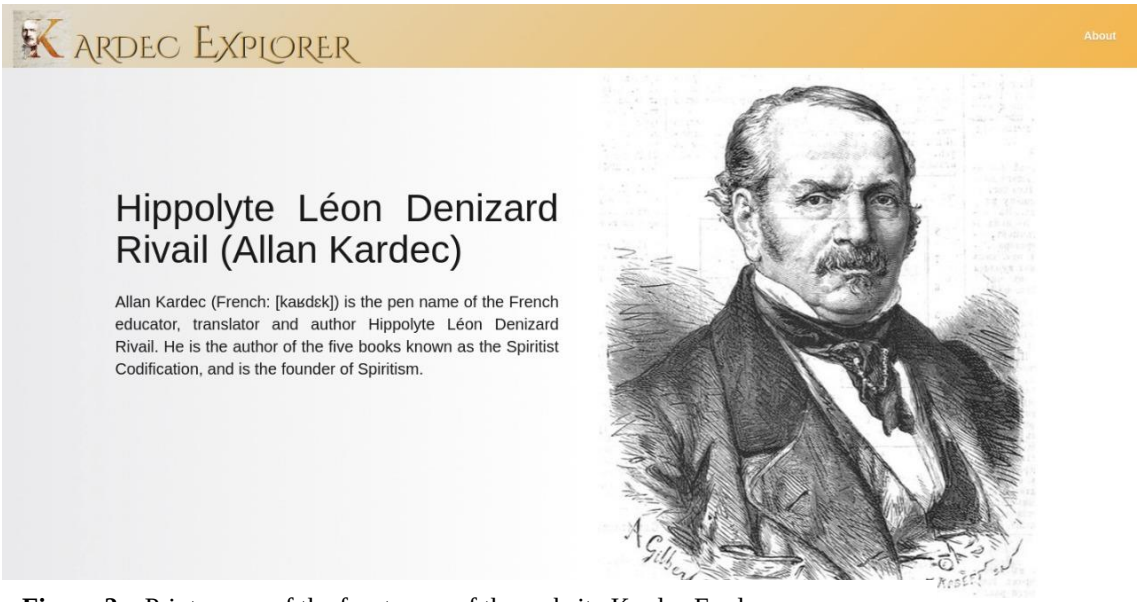


Figure 2 – Printscreen of the front page of the website Kardec Explorer

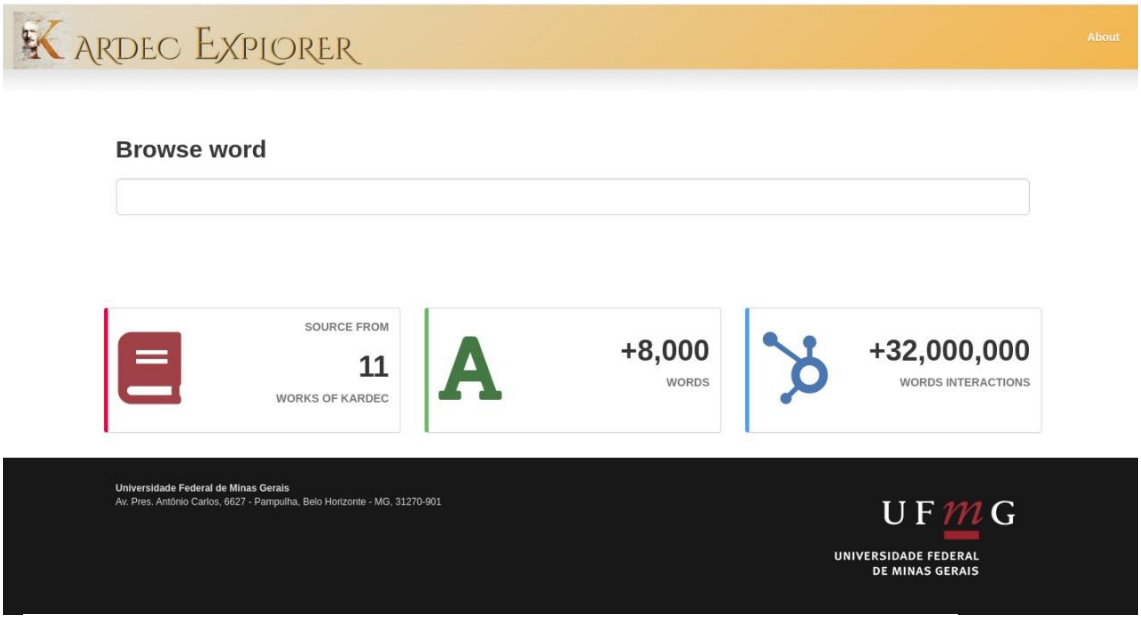


Figure 3 – Printscreen of the informative page of the website Kardec Explorer

I also had an opportunity to do a peer-review of an article in the Journal of Biomolecular Structure & Dynamics, which I learned much more than I expected.

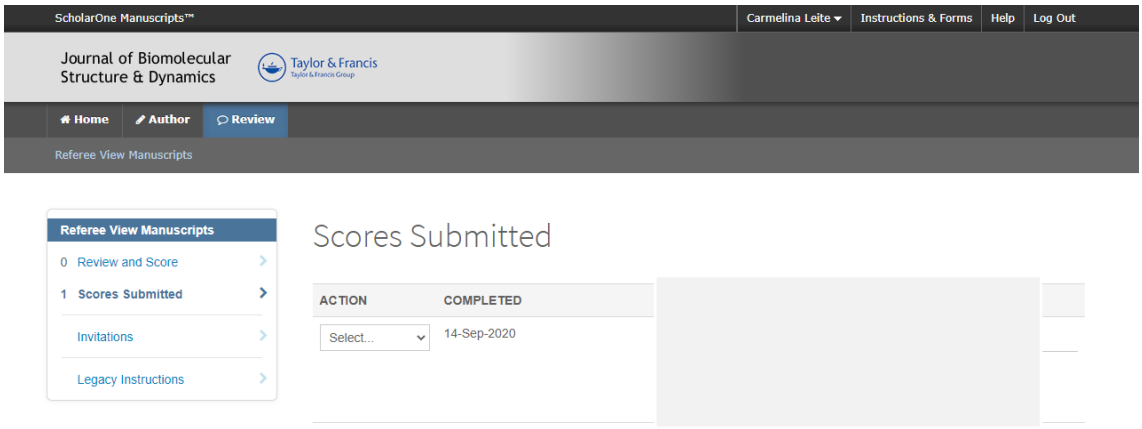


Figure 4 – Printscreen of the proof of peer-review

A guide to performing systematic literature reviews in bioinformatics (MARIANO; LEITE; SANTOS; E O ROCHA *et al.*, 2017) is a technical report on a systematic review of the literature bioinformatics proposal a protocol for this area. My contribution was at all stages of the work.

Characterization of glucose-tolerant β-glucosidases used in biofuel production under the bioinformatics perspective: A systematic review (MARIANO; LEITE; SANTOS; MARINS *et al.*, 2017) already published and consists of the characterization of glucose tolerant and non-tolerant β-glucosidases. My contribution was in the initial part and the systematic literature review.

And the last contribution was in a process of submission, *Using modified logistic regression for taxonomic classification of β-glucosidases based on structural signatures.* My contribution was in the method part.

8. References

ALLEN, W. J.; BALIUS, T. E.; MUKHERJEE, S.; BROZELL, S. R. *et al.* DOCK 6: impact of new features and current docking performance. **Journal of computational chemistry**, 36, n. 15, p. 1132-1156, 2015.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics surveys**, 4, p. 40-79, 2010.

CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom pairs as molecular features in structure-activity studies: definition and applications. **Journal of Chemical Information and Computer Sciences**, 25, n. 2, p. 64-73, 1985.

CARPENTER, K. A.; HUANG, X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *In*: **Curr Pharm Des**, 2018. v. 24, p. 3347-3358.

CHEN, B.; WILD, D.; GUHA, R. PubChem as a source of polypharmacology. **Journal of chemical information and modeling**, 49, n. 9, p. 2044-2055, 2009.

COKLUK, O. Logistic Regression: Concept and Application. **Educational Sciences: Theory and Practice**, 10, n. 3, p. 1397-1407, 2010.

DAI, W.; GUO, D. A Ligand-Based Virtual Screening Method Using Direct Quantification of Generalization Ability. **Molecules**, 24, n. 13, p. 2414, 2019-06-30 2019. Article.

DEGTYARENKO, K.; DE MATOS, P.; ENNIS, M.; HASTINGS, J. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. **Nucleic acids research**, 36, n. suppl 1, p. D344-D350, 2008.

EVERITT, B. S. D.; EVERITT, G. B. S.; DUNN, G. **Applied multivariate data analysis**. 1991.

FENG, Z.; XIA, Y.; GAO, T.; XU, F. *et al.* The antipsychotic agent trifluoperazine hydrochloride suppresses triple-negative breast cancer tumor growth and brain metastasis by inducing G0/G1 arrest and apoptosis. **Cell Death & Disease**, 9, n. 10, p. 1006, 2018-09-26 2018. OriginalPaper.

GAULTON, A.; BELLIS, L. J.; BENTO, A. P.; CHAMBERS, J. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. **Nucleic Acids Research**, 40, n. D1, p. D1100-D1107, 2011.

GEPPERT, H.; VOGT, M.; BAJORATH, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. January 20, 2010 2010. review-article.

GIMENO, A.; OJEDA-MONTES, M. J.; TOMÁS-HERNÁNDEZ, S.; CERETO-MASSAGUÉ, A. *et al.* The Light and Dark Sides of Virtual Screening: What Is There to Know? **International journal of molecular sciences**, 20, n. 6, p. 1375, 2019.

GRISONI, F.; CONSONNI, V.; TODESCHINI, R. Impact of molecular descriptors on computational models. *In*: **Computational Chemogenomics**: Springer, 2018. p. 171-209.

HAN, L.; WANG, Y.; BRYANT, S. H. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in PubChem. **BMC bioinformatics**, 9, n. 1, p. 1, 2008.

HAO, M.; WANG, Y.; BRYANT, S. H. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. **Analytica chimica acta**, 806, p. 117-127, 2014.

HARDT, C.; BEBER, M. E.; RASCHE, A.; KAMBUROV, A. *et al.* ToxDB. Germany: Vertebrate Genomics Department at the Max Planck Institute for Molecular Genetics in Berlin 2019.

HAZLITT, R. A.; TEITZ, T.; BONGA, J. D.; FANG, J. *et al.* Development of Second-Generation CDK2 Inhibitors for the Prevention of Cisplatin-Induced Hearing Loss. **Journal of medicinal chemistry**, 61, n. 17, p. 7700-7709, 2018.

HRIPCSAK, G.; ROTHSCHILD, A. S. Agreement, the f-measure, and reliability in information retrieval. **Journal of the American Medical Informatics Association : JAMIA**, 12, n. 3, p. 296-298, May-Jun 2005.

HUNTER, S.; APWEILER, R.; ATTWOOD, T. K.; BAIROCH, A. *et al.* InterPro: the integrative protein signature database. **Nucleic acids research**, 37, n. suppl 1, p. D211-D215, 2009.

IRWIN, J. J. Community benchmarks for virtual screening. **Journal of Computer-Aided Molecular Design**, 22, n. 3, p. 193-199, March 01 2008. journal article.

IRWIN, J. J.; SHOICHET, B. K. ZINC-a free database of commercially available compounds for virtual screening. **Journal of chemical information and modeling**, 45, n. 1, p. 177-182, 2005.

KAKLAMANI, V. G.; JERUSS, J. S.; HUGHES, E.; SIZIOPIKOU, K. *et al.* Phase II neoadjuvant clinical trial of carboplatin and eribulin in women with triple negative early-stage breast cancer (NCT01372579). **Breast Cancer Res Treat**, 151, n. 3, p. 629-638, Jun 2015.

KANEHISA, M.; GOTO, S.; FURUMICHI, M.; TANABE, M. *et al.* KEGG for representation and analysis of molecular networks involving diseases and drugs. **Nucleic acids research**, 38, n. suppl 1, p. D355-D360, 2010.

KARDEC, A. Caractères de la révélation spirite. **Revue Spirite Journal D' Etudes Psychologiques**, Septembre, X, 1867.

KNOX, C.; LAW, V.; JEWISON, T.; LIU, P. *et al.* DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. **Nucleic acids research**, 39, n. suppl 1, p. D1035-D1041, 2011.

LI, Q.; SHAH, S. Structure-based virtual screening. *In*: **Protein Bioinformatics**: Springer, 2017. p. 111-124.

LIU, Y.; CHEN, P. C.; KRAUSE, J.; PENG, L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. **JAMA**, 322, n. 18, p. 1806-1816, 11 2019.

MAGRANE, M.; CONSORTIUM, U. UniProt Knowledgebase: a hub of integrated protein data. **Database: The Journal of Biological Databases and Curation**, 2011, p. bar009,

MARIANO, D.; LEITE, C.; SANTOS, L.; E O ROCHA, R. *et al.* **A guide to performing systematic literature reviews in bioinformatics**. 2017.

MARIANO, D.; LEITE, C.; SANTOS, L.; MARINS, L. F. *et al.* **Characterization of glucose-tolerant β-glucosidases used in biofuel production under the bioinformatics perspective: A systematic review**. 2017.

MIGUEL, R. B.; PETERSEN, P. A.; GONZALES-ZUBIATE, F. A.; OLIVEIRA, C. C. *et al.* Inhibition of cyclin-dependent kinase CDK1 by oxindolimine ligands and corresponding copper and zinc complexes. **J Biol Inorg Chem**, 20, n. 7, p. 1205-1217, Oct 2015.

MISHRA, N.; BASU, A. Exploring different virtual screening strategies for acetylcholinesterase inhibitors. **BioMed research international**, 2013, 2013.

NEVES, B. J.; BRAGA, R. C.; MELO-FILHO, C. C.; MOREIRA-FILHO, J. T. *et al.* QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. **Frontiers in Pharmacology**, 9, n. 1275, 2018-November-13 2018. Mini Review.

PAOLINI, G. V.; SHAPLAND, R. H. B.; VAN HOORN, W. P.; MASON, J. S. *et al.* Global mapping of pharmacological space. **Nature biotechnology**, 24, n. 7, p. 805-815, 2006.

PLEWCZYNSKI, D.; SPIESER, S. A. H.; KOCH, U. Assessing Different Classification Methods for Virtual Screening. **Journal of Chemical Information and Modeling**, 46, n. 3, p. 1098-1106, 2006/05/01 2006.

POWERS, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

QI, L.; DING, Y. Potential antitumor mechanisms of phenothiazine drugs. **Science China Life Sciences**, 56, n. 11, p. 1020-1027, 2013.

RAJU, T. N. K. The Nobel Chronicles. **The Lancet**, 355, n. 9208, p. 1022, 2000.

RDKit: Cheminformatics and Machine Learning Software. 2019.

RINIKER, S.; FECHNER, N.; LANDRUM, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. **Journal of chemical information and modeling**, 53, n. 11, p. 2829-2836, 2013.

RINIKER, S.; LANDRUM, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. **J Cheminform**, 5, n. 1, p. 26, May 2013.

RINIKER, S.; WANG, Y.; JENKINS, J. L.; LANDRUM, G. A. Using information from historical high-throughput screens to predict active compounds. **Journal of chemical information and modeling**, 54, n. 7, p. 1880-1891, 2014.

ROHRER, S. G.; BAUMANN, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. **Journal of Chemical Information and Modeling**, 49, n. 2, p. 169-184, 2009/02/23 2009.

SCHIERZ, A. C. Virtual screening of bioassay data. **Journal of cheminformatics**, 1, p. 21, 2009.

SEIFERT, M. H.; WOLF, K.; VITT, D. Virtual high-throughput in silico screening. **Biosilico**, 1, n. 4, p. 143-149, 2003.

SHARMAN, J. L.; MPAMHANGA, C. P.; SPEDDING, M.; GERMAIN, P. *et al.* IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. **Nucleic acids research**, 39, n. suppl 1, p. D534-D538, 2011.

SHI, X. N.; LI, H.; YAO, H.; LIU, X. *et al.* Adapalene inhibits the activity of cyclin-dependent kinase 2 in colorectal carcinoma. **Mol Med Rep**, 12, n. 5, p. 6501-6508, Nov 2015.

STOCKWELL, B. R. Exploring biology with small organic molecules. **Nature**, 432, n. 7019, p. 846-854, Dec 2004.

TEITZ, T.; FANG, J.; GOKTUG, A. N.; BONGA, J. D. *et al.* CDK2 inhibitors as candidate therapeutics for cisplatin- and noise-induced hearing loss. **J Exp Med**, 215, n. 4, p. 1187-1203, Apr 2 2018.

TRUNK, G. V. A problem of dimensionality: A simple example. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, n. 3, p. 306-307, 1979.

WANG, Y.; BOLTON, E.; DRACHEVA, S.; KARAPETYAN, K. *et al.* An overview of the PubChem BioAssay resource. **Nucleic acids research**, 38, n. suppl 1, p. D255-D266, 2010.

YIN, L.; GE, Y.; XIAO, K.; WANG, X. *et al.* Feature selection for high-dimensional imbalanced data. **Neurocomputing**, 105, p. 3-11, 2013.

YUAN, S.; CHAN, J. F.-W.; DEN-HAAN, H.; CHIK, K. K.-H. *et al.* Structure-based discovery of clinically approved drugs as Zika virus NS2B-NS3 protease inhibitors that potently inhibit Zika virus infection in vitro and in vivo. **Antiviral Research**, 145, p. 33-43, 2017/09/01/ 2017.

ZHANG, Y.; KUTATELADZE, T. G. Molecular structure analyses suggest strategies to therapeutically target SARS-CoV-2. **Nature Communications**, 11, n. 1, p. 2920, 2020/06/10 2020.

ZHAO, B.; YI, G.; DU, F.; CHUANG, Y.-C. *et al.* Structure and function of the Zika virus full-length NS5 protein. **Nature Communications**, 8, n. 1, p. 14762, 2017/03/27 2017.

ÉLDEN, L. Numerical linear algebra in data mining. **Acta Numerica**, 15, 2006.