

Journal of Cheminformatics

A Multiple Classifier System Identifies Novel Cannabinoid CB2 Receptor Ligands

--Manuscript Draft--

Manuscript Number:	CHIN-D-19-00037R1	
Full Title:	A Multiple Classifier System Identifies Novel Cannabinoid CB2 Receptor Ligands	
Article Type:	Research article	
Funding Information:	Dutch Scientific Council (NWO) (VENI 14410)	Dr Gerard Van Westen
	Consellería de Educación, Universidades e Formación Profesional (ED431C2018/55-GRC)	Not applicable
Abstract:	<p>Drugs have become an essential part of our lives due to their ability to improve people's health and quality of life. However, for many diseases, approved drugs are not yet available or existing drugs have undesirable side effects, making the pharmaceutical industry strive to discover new drugs and active compounds. The development of drugs is an expensive process, which typically starts with the detection of candidate molecules (screening) for an identified protein target. To this end, the use of high-performance screening techniques has become a critical issue in order to palliate the high costs. Therefore, the popularity of computer-based screening (often called virtual screening or in-silico screening) has rapidly increased during the last decade. A wide variety of Machine Learning (ML) techniques has been used in conjunction with chemical structure and physicochemical properties for screening purposes including (i) simple classifiers, (ii) ensemble methods, and more recently (iii) Multiple Classifier Systems (MCS). In this work, we apply an MCS for virtual screening (D2-MCS) using circular fingerprints. We applied our technique to a dataset of cannabinoid CB2 ligands obtained from the ChEMBL database. The HTS collection of Enamine (1.834.362 compounds), was virtually screened to identify 48.432 potential active molecules using D2-MCS. This list was subsequently clustered based on circular fingerprints and from each cluster, the most active compound was maintained. From these, the top 60 were kept, and 21 novel compounds were purchased. Experimental validation confirmed six highly active hits (>50% displacement at 10 μM and subsequent K_i determination) and an additional five medium active hits (>25% displacement at 10 μM). D2-MCS hence provided a hit rate of 29% for highly active compounds and an overall hit rate of 52%.</p>	
Corresponding Author:	Gerard Van Westen Leiden University Leiden, NETHERLANDS	
Corresponding Author E-Mail:	gerard@lacdr.leidenuniv.nl	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Leiden University	
Corresponding Author's Secondary Institution:		
First Author:	David Ruano-Ordás	
First Author Secondary Information:		
Order of Authors:	David Ruano-Ordás	
	Lindsey Burggraaff	
	Rongfang Liu	
	Cas van der Horst	
	Laura H. Heitman	
	Michael T.M. Emmerich	

	Jose R. Mendez
	Iryna Yevseyeva
	Gerard Van Westen
Order of Authors Secondary Information:	
Response to Reviewers:	<p>Journal of Cheminformatics Editorial board</p> <p>Dear Sir/Madam,</p> <p>Herewith we are resubmitting for publication as article a manuscript entitled: "A Multiple Classifier System Identifies Novel Cannabinoid CB2 Receptor Ligands"</p> <p>The manuscript submitted here demonstrates the usage of a multiple classifier system (D2-MCS) for usage in virtual screening. In the current work, it was applied to a data set consisting of cannabinoid CB2 receptor ligands. After training and validation, we virtually screened the enamine HTS library. From this library, we were able to identify 48.432 potential active molecules using D2-MCS.</p> <p>This list was subsequently clustered based on circular fingerprints and from each cluster, the most active compound was maintained. From these, the top 60 were kept (based on probability to be active), and 21 novel compounds were purchased. Experimental validation confirmed six highly active hits (>50% displacement at 10 μM and subsequent Ki determination) and an additional five medium active hits (>25% displacement at 10 μM). D2-MCS hence provided a hit rate of 29% for highly active compounds and an overall hit rate of 52%.</p> <p>We provide a list of all identified potential actives along with the calculated probability as supporting information.</p> <p>Following the pre-review we have provided a DOI which contains the data used: http://doi.org/10.5281/zenodo.2677650. Moreover the used scripts are available on GitHub : https://github.com/drordas/D2-MCS . Moreover, we have added an additional author (form included).</p> <p>We would like to suggest the following referees: Andreas Bender (University of Cambridge); Email: ab454@cam.ac.uk , Expert in cheminformatics</p> <p>Chris de Graaf (Heptares Therapeutics) ; Email: Chris.DeGraaf@heptares.com , Expert on G Protein Coupled Receptors</p> <p>Paulo Novais (University of Minho) ; Email: mpjon@di.uminho.pt , Expert in machine learning and ensemble methods.</p> <p>Boudewijn Lelieveldt (Leiden University Medical Center) ; Email: B.P.F.Lelieveldt@lumc.nl , Expert in machine learning based dimensionality reduction</p> <p>Sincerely,</p> <p>Gerard JP van Westen, PhD On behalf of all authors</p>
Additional Information:	
Question	Response

[Click here to view linked References](#)

A Multiple Classifier System Identifies Novel Cannabinoid CB2 Receptor Ligands

David Ruano-Ordás^{a,b,c,d,e}, Lindsey Burggraaff^f, Rongfang Liu^f, Cas van der Horst^f, Laura H. Heitman^f, Michael T.M. Emmerich^f, Jose R. Mendez^{a,b,d}, Iryna Yevseyeva^e, Gerard JP van Westen^f

^a Department of Computer Science, University of Vigo, ESEI - Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

^b CINBIO - Biomedical Research Centre, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain

^c Multicriteria Optimization and Decision Analysis (MODA) Research Group, LIACS, Leiden University, Niels Bohrweg 1, 2333-CA Leiden, The Netherlands

^d SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur). SERGAS-UVIGO

^e School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester LE1 9BH, UK

^f Drug Discovery and Safety, LACDR, Leiden University, Einsteinweg 55, 2333 CC, Leiden, The Netherlands

Email addresses: David Ruano-Ordás : drordas@uvigo.es , Lindsey Burggraaff : l.burggraaff@lacdr.leidenuniv.nl , Rongfang Liu : r.liu@lacdr.leidenuniv.nl , Cas van der Horst : c.van.der.horst@lacdr.leidenuniv.nl , Laura Heitman : l.h.heitman@chem.leidenuniv.nl , Michael Emmerich : m.t.m.emmerich@liacs.leidenuniv.nl , José Ramón Méndez : moncho.mendez@uvigo.es , Iryna Yevseyeva : iryna.yevseyeva@gmail.com , Gerard JP van Westen : gerard@lacdr.leidenuniv.nl

Abstract

Drugs have become an essential part of our lives due to their ability to improve people's health and quality of life. However, for many diseases, approved drugs are not yet available or existing drugs have undesirable side effects, making the pharmaceutical industry strive to discover new drugs and active compounds. The development of drugs is an expensive process, which typically starts with the detection of candidate molecules (screening) for an identified protein target. To this end, the use of high-performance screening techniques has become a critical issue in order to palliate the high costs. Therefore, the popularity of computer-based screening (often called virtual screening or *in-silico* screening) has rapidly increased during the last decade. A wide variety of Machine Learning (ML) techniques has been used in conjunction with chemical structure and physicochemical properties for screening purposes including (i) simple classifiers, (ii) ensemble methods, and more recently (iii) Multiple Classifier Systems (MCS). In this work, we apply an MCS for virtual screening (D2-MCS) using circular fingerprints. We applied our technique to a dataset of cannabinoid CB2 ligands obtained from the ChEMBL database. The HTS collection of Enamine (1.834.362 compounds), was virtually screened to identify 48.432 potential active molecules using D2-MCS. This list was subsequently clustered based on circular fingerprints and from each cluster, the most active compound was maintained. From these, the top 60 were kept, and 21 novel compounds were purchased. Experimental validation confirmed six highly active hits (>50% displacement at 10 μ M and subsequent K_i determination) and an additional five medium active hits (>25% displacement at 10 μ M). D2-MCS hence provided a hit rate of 29% for highly active compounds and an overall hit rate of 52%.

Keywords

Drugs discovery, clustering methods, measure-guided methodology, ensembling schemes

Introduction

In silico (or computational drug discovery) relies on different computer-based techniques to find a novel or improved bioactive compound, which should exhibit a strong affinity to a particular target. Although *in-silico* screening is present in the drug development process since the beginning of 90s [1, 2] its relevance has been progressively increasing until becoming an essential part of the drug-development process. This fact was mainly motivated by (i) a significant improvement in the performance of computer systems, (ii) the introduction of novel algorithms and more expressive molecular descriptors, and (iii) the advent of large-scale public bioactivity databases [3].

Limited processing capabilities of computer systems during the 90s led to *in silico* screening mainly focused on (i) building simple mathematical modelling approaches (often implemented as cellular automata) for large-scale simulations of complex systems [4], (ii) the development of big-data databases enables researchers to easily store and access the information [2] or (iii) the design of affinity fingerprints as novel descriptors for similarity searches in molecular databases and QSAR analyses [5]. As computers' performance increased, the use of simple Machine Learning (ML) classification schemes for screening purposes became popular. Concretely, the usage of Support Vector Machines (SVM) [6, 7],

1 Decision Trees (DT) [8], Naïve Bayes [9], K-Nearest Neighbourhoods (KNN) [10], Artificial
2 Neural Networks [11] and Self Organizing Maps (SOM) [12] were widely applied in the
3 domain.

4 However, during the last decade, the amount of public information available for screening
5 has increased rapidly with the introduction of resources such as ChEMBL and PubChem [3,
6 13]. This fact had a negative impact on the performance of simple ML approaches due to
7 their trend to build unstable classification models when handling a high volume of
8 information. In order to improve this situation and therefore, increase the predictive
9 performance, ML models were equipped with multiple layers (stacking, deep learning) and
10 identical ML algorithms were combined (ensemble of classifiers [14]). Specifically, in [15]
11 authors demonstrate the suitability of using of Deep Neural Networks (DNN) [16] and
12 Random Forests (RF) [17] methods against single ML models (such as Naïve Bayes or
13 SVM) to predict the bioactivity of molecules. Additionally, latest research work [18, 19]
14 applies several Boosting (such as AdaBoost or MultiBoost) and Fuzzy Forest approaches to
15 predict (*i*) bioactivity of molecules and (*ii*) toxicity of non-congeneric industrial chemicals,
16 respectively.
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 The usage of above-mentioned ensemble methods contributed to significant performance
2 improvements in the virtual screening domain. However, their introduction also brought
3 some important shortcomings such as: (i) the random selection of the information often used
4 to build each inner classifier, (ii) the common usage of weak classifiers such as C4.5 or
5 Decision Stumps to build up the classifier ensemble (although any ML classifier can be
6 used) and, (iii) the impossibility combining different inner classifiers and configurations for
7 them with concrete subsets of training information. These limitations are implicit to the
8 definition of ensemble classifiers and are the key features to distinguish them against the
9 great number of methods included in the Multiple Classifier Systems (MCS) [20] group.
10 Wozniak *et al.* [20] revealed interesting features of MCS, including (i) their good performance
11 when working in extreme situations such as scarcity of samples or information overload, (ii)
12 their ability to outperform inner individual classifiers, (iii) the increase of the probability of
13 finding an optimal model, and (iv) the reduction of the information (and hence the increase in
14 the performance and speed) used to build each inner classifier. Keeping into account the
15 above-mentioned issues we apply an MCS toolkit (called D2-MCS [21]) to increase the
16 performance of virtual screening.
17
18
19
20
21
22

23 Methods

24
25 This section evaluates the suitability of using MCS and its application in drug discovery
26 domain. It also introduces the dataset and measures used to perform the experimental
27 protocol. Finally, the methodology performed to carry out the virtual screening process is
28 explained in detail.
29
30

31 Datasets

32 **CB2 dataset**

33
34 The data was gathered from ChEMBL version 22 based on UniProt accession P34972 [22].
35 The activity data were filtered for potential duplicates, no activity or data validity comments
36 were allowed, and only data from binding assays with a pChEMBL value was kept. This led
37 to 3,925 compounds. Subsequently, compound fingerprints (FCFP_6) and physicochemical
38 properties were calculated (see supporting information in Additional File 1) [23]. The
39 FCFP_6 fingerprints were computed using the fingerprints to properties component from
40 Pipeline Pilot Version 2016.1.0 [24]; 2048 substructures/bits were selected based on their
41 occurrence frequency in the data set [24]. A presence of 50% was the optimum frequency.
42 Thereby, significant under- and overrepresentation were both avoided. Finally, the set was
43 made into a binary classification set where the activity cut-off was set at a pChEMBL value >
44 7 for active compounds and written to a tab-delimited text file using the InChIKey as unique
45 identifier [25]. The final set contained 1,977 active compounds and 1,948 inactive
46 compounds (CB2Set, supporting information [26]). The obtained dataset include 2,133
47 attributes (84 physicochemical properties, 2,048 chemical-structure features and the activity
48 class) to describe 3,925 compounds (instances). Table 1 shows the codification of each
49 feature grouped by type.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Feature characteristics and codification

<i>Feature type</i>	<i>Feature values</i>	<i>N° of features</i>
Chemical substructure fingerprints	Binary	2,048
	Discrete Values	50
	Continuous Values	34
Total		2,132

As can be observed from Table 1 each chemical substructure is codified using a binary representation to indicate its presence (1) or absence (0) for each chemical compound. Additionally, the physicochemical descriptors consist of continuous or discrete values depending on the descriptor type and metric representation.

Validation dataset

The high-throughput screening (HTS) set was downloaded from the Enamine website (containing 1,834,362 compounds without class information). Molecules were standardized and encoded using the same feature representation as was used for the CB2 dataset (2048 chemical substructure fingerprints and 84 physicochemical descriptors). This set will be referred to as ValidationSet.

Evaluation measures

Quite a few performance measures for assessing the accuracy and rank of different classification approaches exist in the drug discovery domain. Concretely, we select Matthews Correlation Coefficient (MCC) [27, 28] and Positive Predictive Values (PPV) [29–31] measures due to their demonstrated ability to minimize false negatives (FN) and false positives (FP) errors respectively.

MCC is a performance measure designed for binary classifiers that can be used in the case of imbalanced datasets (the distribution of instances in the classes is uneven). MCC can be easily computed from the values of the confusion matrix results (true positives or TP, true negatives or TN, false positives or FP and false negatives or FN) by using Equation 1.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (FN + TN) \times (FP + TN) \times (TP + FN)}} \quad 1$$

MCC is defined in the interval [-1,1], where 1 stand for no classification errors, -1 means that all input instances were misclassified and 0 reveals that the classification was absolutely uncorrelated with the real truth. As can be extrapolated from Equation 1, achieving a balanced number of positive and negative classification hits is mandatory to obtain higher MCC values. Additionally, the inclusion of the four quantiles (TP, TN, FP and FN) in the MCC formula allows giving a better summary of the performance of classification algorithms

1 regarding other well-known metrics (such as Accuracy [32] or F1-Score [33]). The benefits of
2 using MCC against other well-known measures commonly used to evaluate ML approaches
3 in the health domain has been demonstrated by Chicco [34].

4 From another perspective, PPV is a well-known measure in the drug discovery domain due
5 to its ability to assess the probability of having a positive outcome given a positive result
6 (also called a *posteriori* probability). Thus, PPV is an interesting measure since testing an
7 inactive molecule (due to an FP error) is very expensive. Due to this, some previous works
8 take advantage of it [35]. PPV can be computed by combining the values included in the
9 confusion matrix in the form defined by Equation 2.
10
11

$$12 \quad PPV = \frac{TP}{TP + FP} \quad 2$$

13
14
15
16
17
18 As could be noted, PPV is not able to accurately handle most situations if used in isolation.
19 In fact, a classifier could reach the maximum PPV score by identifying only one active
20 molecule. With regard to this, over a balanced dataset where the probability of finding one
21 active molecule is $\frac{1}{2}$, a classifier could randomly select one instance to classify it as active
22 and assign the inactive label to the remaining ones. This classifier could achieve a PPV
23 score of one in half of the experimentations (those which the instance classified as active
24 was really active). Therefore, PPV needs to be accompanied by other performance
25 indicators, such as MCC.
26
27
28

29 Modelling

30
31 To build our classification software we use D2-MCS due to its ability to easily build high-
32 performance *in silico* screening MCS models [21]. D2-MCS is an R-based toolkit that
33 provides an efficient and flexible MCS mechanism that can be highly customized to ensure
34 an adequate adaptation to the intrinsic characteristics of the target dataset. Particularly, D2-
35 MCS is able to handle high dimensional datasets by grouping the features of molecules
36 (dataset columns) into several groups (called feature-clusters) according to user-defined
37 criteria (i.e. type of chemical compounds, molecular weight, etc.). Then, for each feature-
38 cluster, the toolkit is able to automatically determine the most suitable classifier (simple or
39 ensemble) together with its best configuration. According to this information, D2-MCS builds
40 a set of classifiers (one per feature cluster) whose outputs will be combined to generate a
41 single solution. The set of selected trained classifiers (one for each dataset part) together
42 with a voting system comprises a whole D2-MCS instance. Figure 1 shows a global overview
43 of the D2-MCS operation.
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

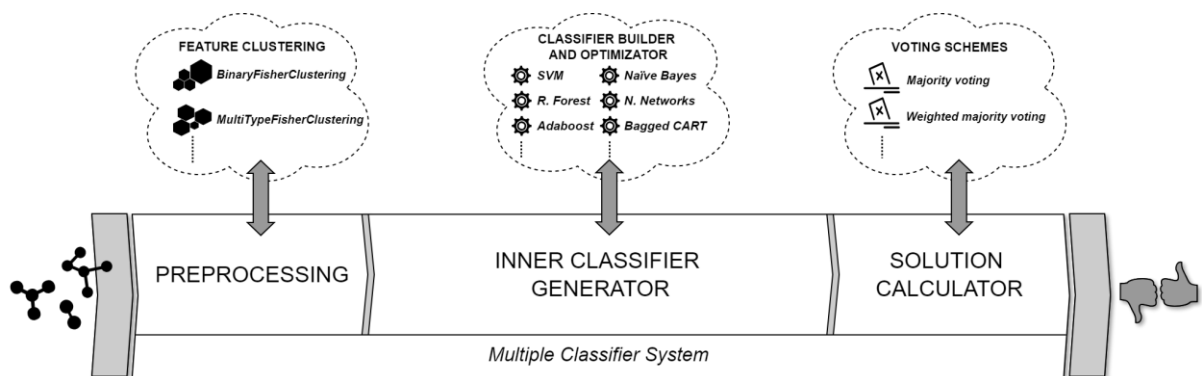


Figure 1. Structure and functionality of D2-MCS toolkit. D2-MCS builds a set of classifiers (one per feature cluster) whose outputs will be combined to generate a single solution. The set of selected trained classifiers (one for each dataset part) together with a voting system comprises a whole D2-MCS instance.

As shown in Figure 1, D2-MCS operation is divided into three different stages. The first stage (called PREPROCESSING in Figure 1) comprises the partitioning of training information based on a specific feature-clustering algorithm. Although D2-MCS provides by default several clustering methods (Fisher, Information Gain, etc.), it also allows users to define customized feature clustering methods in order to increase its compatibility regardless of the way of representing or encoding the information.

Then, for each split of the original dataset, D2-MCS is able to detect the most effective classifier (and its best configuration) from a wide variety of ML techniques (up to 236 different classifiers from 47 families [36]). The best classifiers for each knowledge partition together with their optimal configurations are compiled together to act as a set of individual experts whose outputs should be combined to generate a final result. To this end, D2-MCS implements two simple methods to combine the outputs of inner classifiers (see SOLUTION CALCULATOR part in Figure 1) and it provides an API to easily define new output aggregator methods [21].

Then, the third stage (see SOLUTION CALCULATOR part in Figure 1) is responsible for compiling the best classifier for each cluster (achieving highest performance values) together with their optimal configuration to act as a set of individual experts whose outputs should be combined to generate a final solution. To this end, D2-MCS implements two simple methods to combine the outputs of inner classifiers (one inner classifier per cluster): (i) a simple majority voting system where the final class is the one obtaining more than half of the votes and (ii) a weighted majority voting where the winner is the class achieving the highest overall value.

In order to execute our experimentation, the dataset instances (rows) were randomly divided into four homogeneous and evenly sized groups. Figure 2 represents the configuration of groups and their usage for evaluating MCSs and single ML techniques.

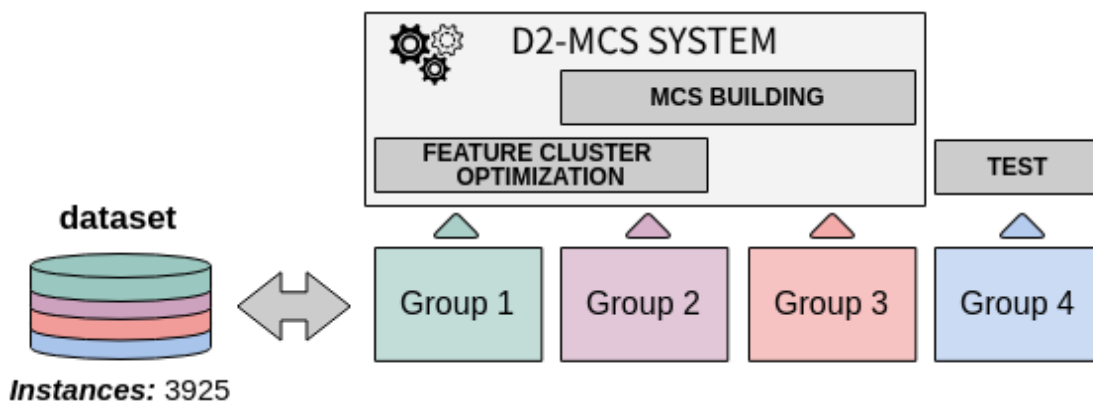


Figure 2. Dataset partitioning. The dataset instances (rows) were randomly divided into four homogeneous and evenly sized groups.

As shown in Figure 2, the first two groups were used to select an appropriate number of feature-clusters for MCS systems. Then, the second and third groups were used to build the MCS (select the most appropriate classifier for each dataset partition, build classifiers, and optimize their configurations). Finally, the fourth group has been reserved to assess the performance of the achieved MCS.

As previously stated, during the first stage of D2-MCS process (see Figure 1) the original dataset is divided into several groups of non-repeated features. Although the latest version D2-MCS provides several feature-clustering algorithms, we used the same clustering method as used in [21] (called *MultiTypeFisherClustering*) due to the good results achieved in this domain. Concretely, the experimentation carried out in [21] demonstrated the suitability of dividing the features into three clusters.

Once the best clustering configuration is obtained (three clusters), the MCS building stage is executed. In detail, this stage is responsible for determining the best ML models (and parameter configuration) for each cluster. Additionally, D2-MCS allows defining an objective function to customize the model parameter-optimization process. To follow the same criterion as previously commented, we use both PPV and MCC measures, which entails the generation of two different MCCs (PPV-based and MCC-based models).

Then, in order to test the final performance, both MCS (PPV-based and MCC-based) were executed over the remaining dataset (see Group 4 in Figure 2) composed by 982 instances (504 active and 478 inactive compounds). To compute the final class of each compound, the outputs of the inner classifiers included in each MCS are combined using a voting scheme where a compound is classified as Active whenever the number of positive outputs of each inner classifiers is greater or equal than the negative ones. Conversely, the compound is classified as Inactive.

To follow the evaluation criteria used during the optimization stage, the classification performance achieved for each experimental configuration was assessed using both MCC and PPV measures (see Figure 3). Additionally, Figure 3 represents (i) the final performance achieved for each measure and (ii) the confusion matrix values (TP, TN, FP and FN). For comparison purposes, Figure 3 also includes the performance achieved on each cluster during the training stage.

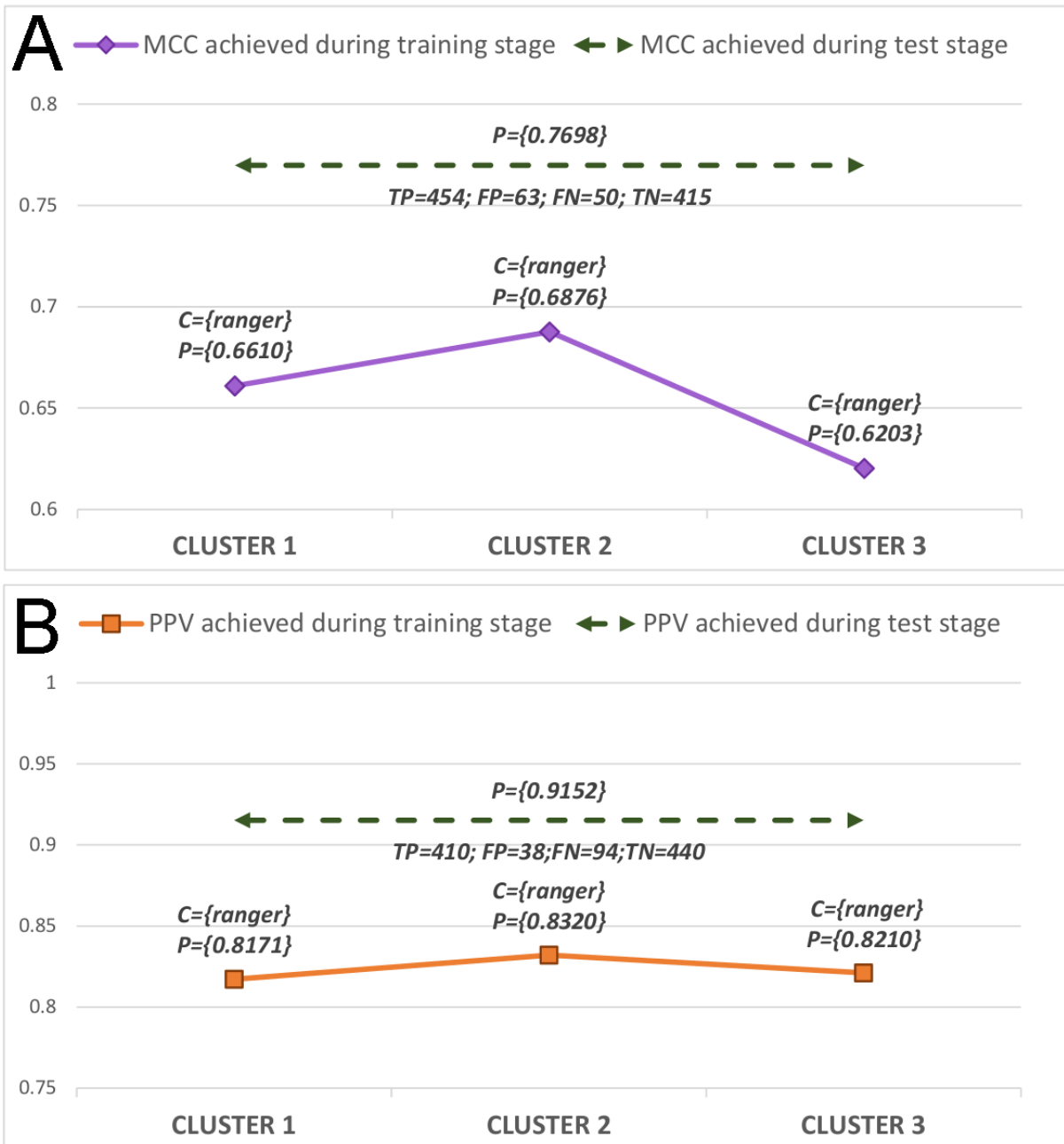
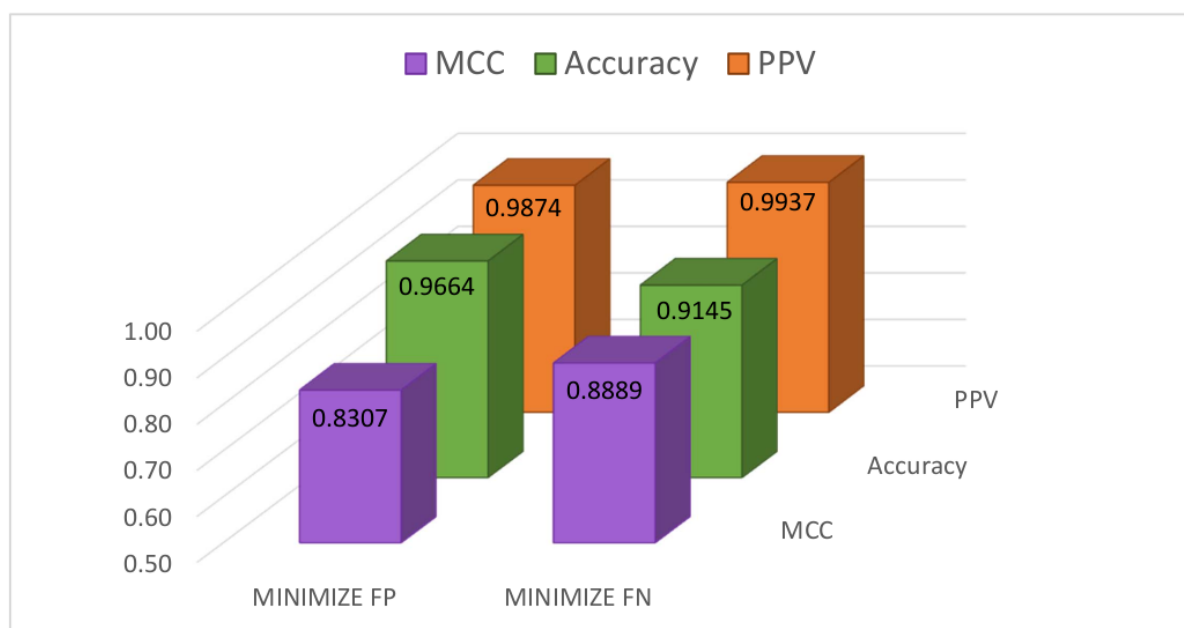


Figure 3. Performance comparison plot for the testing stage. A) indicates the performance obtained using the MCC measure and B) shows the performance obtained using the PPV measure.

As can be realised from Figure 3A, the performance achieved during the test stage slightly outperforms the individual outcomes obtained during the optimization stage. Additionally, the use of the MCC measure allows achieving a balanced number of misclassification errors ($FP \approx FN$). Furthermore, from Figure 3B it is easy to realise that using PPV as an objective function significantly reduces the number of FP errors at expenses of increasing FN errors.

1 Additionally, after performing a global overview of Figure 3 we can realise: (i) the ability to
2 build a suitable measure-guided knowledge-generalization models, and (ii) the importance of
3 using an adequate domain-oriented measure in order to minimize the number of
4 misclassification errors. In fact, as can be seen in Figure 3, MCC based models achieve
5 fewer error rates than the PPV measure (113 and 132 errors respectively). Despite this, the
6 results are quite promising (the rate of correctly classified compounds is very high), although
7 we are aware that can be increased even more by taking advantage of the intrinsic
8 characteristics of MCS.
9

10 In order to demonstrate this hypothesis, we generate two meta-models by combining the
11 predictions achieved by the MCS models trained using MCC and PPV measures (see
12 Minimize FP and Minimize FN in Figure 4). Concretely, Minimize FP is responsible for
13 labelling the target compound as Active whenever is predicted as 'Active' by both MCS (PPV
14 and MCC) while Minimize FN identifies the target compound as Active only if one of the
15 MCS (PPV and MCC) predicts the compound as 'Active'. For comparison purposes, both
16 meta-models were executed over the same testing dataset (see Group 4 in Figure 2) as
17 used by primitive MCS models (PPV-based and MCC-based).
18
19
20
21



22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43 Figure 4. Performance comparison achieved for *Minimize FP* and *Minimize FN* meta-models.

44
45 As can be seen in Figure 4, both meta-models clearly improve the performance achieved by
46 the primitive MCS models. Focusing on the first approximation (Minimize FP), the
47 performance is increased up to 6.45% (MCC) and 7.32% (PPV) regarding the original
48 models optimised for MCC and PPV respectively. On the other hand, the second meta-
49 model outperforms up to 11.91% (MCC) and 7.85% (PPV) the results achieved with regards
50 the corresponding primitive models. Although the second approximation seems the most
51 suitable alternative (best values of MCC and PPV), the Accuracy measure points *Minimize*
52 *FP* as the best model. The main reason for this circumstance can be easily explained
53 through the confusion matrix described in Table 2.
54
55
56
57
58
59
60
61
62
63
64
65

1 As can be seen in Table 2, the number of overall errors achieved by second approximation
2 is bigger than *Minimize FP* (84 vs 33 respectively). Taking into account that Accuracy
3 computes the overall probability of performing a correct classification, it is easy to conclude
4 that the low rate of misclassification errors motivates the good Accuracy level achieved by
5 first approximation.
6

7 Table 2. Confusion matrix achieved for both configurations.
8

	<i>TP</i>	<i>FP</i>	<i>TN</i>	<i>FN</i>
<i>MINIMIZE FP</i>	474	3	475	30
<i>MINIMIZE FN</i>	480	60	418	24

9
10
11
12
13
14
15
16 Additionally, as can be realised from Table 2, the ability to avoid discarding potential Active
17 compounds makes *Minimize FN* an adequate alternative for the research domain (where
18 discovering the whole spectrum of potential candidate drugs is more important than
19 minimizing trial costs). Conversely, *Minimize FP* approximation achieves a significant
20 reduction of FP errors (up to 95%) when compared with *Minimize FP*. This fact makes
21 *Minimize FP* a suitable approximation for the pharmaceutical industry where minimizing
22 unnecessary trial tests (reduce costs) is more important than losing potential Active
23 candidates.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Probabilistic-based ranking methodology

The designed probabilistic-based methodology was used to rank active-predicted compounds from the ValidationSet. As can be seen in Figure 5, the first stage is responsible for compiling from all inner individual classifiers included Minimize FP model (an MCS model comprising 3 classifiers for optimizing PPV and another one with 3 inner models for MCC) the class probability of each compound tagged as Active (48,232). As can be depicted from stage 1 in Figure 5, the achieved probabilities are always greater than 0.5 since only compounds previously labelled as Active (Active > 0.5, Inactive <= 0.5) were selected.

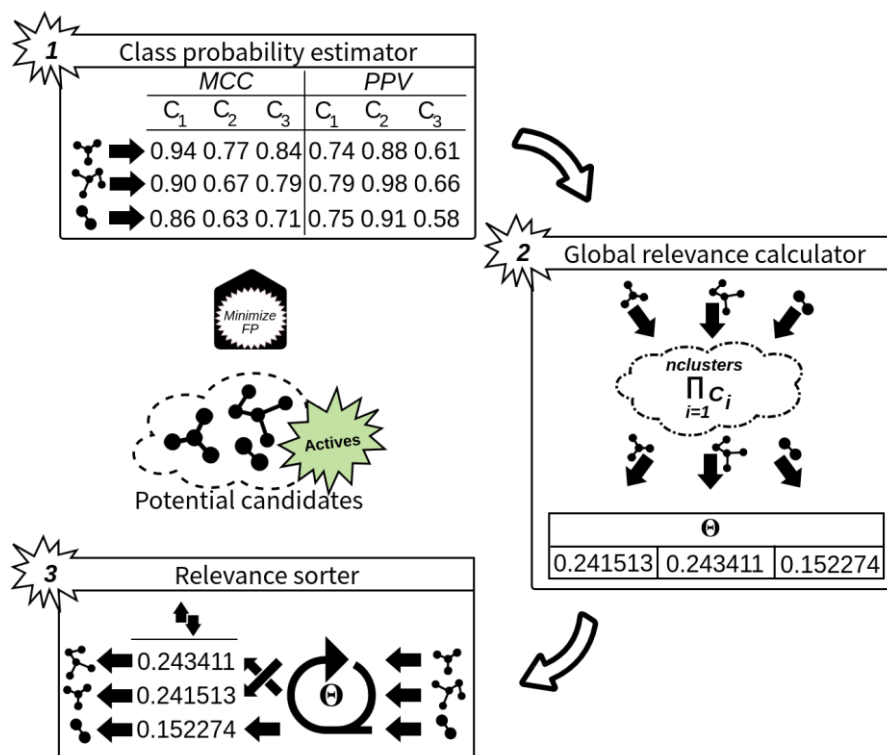


Figure 5. Workflow of three-stage potential candidate ranker methodology. Our ranking methodology comprised three main stages: (i) class probability estimator, (ii) global relevance calculator and (iii) relevance sorter.

Once all probabilities are obtained, during the second stage we compute the global relevance (denoted as Θ in Figure 5) of each candidate as a mathematical product of all its probabilities (see Equation 3).

$$\Theta = \prod_{i=1}^{numcluster} C_i \quad 3$$

where *numcluster* stands for the number of clusters comprising the used meta-model.

Combining these probabilities using the product operator allows achieving a wide variety of output values (and thereby improves compatibility) even when individual input values are very close. As an example, given two vectors of values [0.75, 0.75, 0.6], [0.6, 0.9, 0.6], the product operator (Π) is able to achieve 0.337 and 0.324 respectively, while the summation (Σ) and the arithmetic mean (\bar{X}) obtain the same values (2.1 and 0.7

1 respectively). Finally, the third stage entails the arrangement of the chemical compounds by
2 descendant according to its global relevance value (Θ). This ensures that the best
3 candidates are placed in the initial positions.
4

5 Chemical clustering 6

7 The 48,232 predicted actives were clustered based on the same binary features (FCFP_6)
8 that were used for model training using the cluster molecules component in Pipeline Pilot
9 version 2016 [24]. An average cluster population of 20 was selected and the maximum
10 Tanimoto distance between the cluster centre and members was set at 0.35 (forcing a
11 similarity of > 0.65 within clusters), resulting in 28,217 clusters.
12
13
14

15 In Vitro Experimental Techniques

16 Cell culture and membrane preparation

17 CHOK1hCB2_bgal cells (DiscoverRx, Fremont, CA, USA) were cultured in Dulbecco's
18 Modified Eagle's Medium/Nutrient Mixture F-12 Ham supplemented with 10% fetal calf
19 serum, 1 mM glutamine, 50 $\mu\text{g}/\text{mL}$ penicillin, 50 $\mu\text{g}/\text{mL}$ streptomycin, 300 mg/mL hygromycin
20 and 800 $\mu\text{g}/\text{mL}$ geneticin in a humidified atmosphere at 37°C and 5% CO₂. Cells were
21 subcultured twice a week at a ratio of 1:20 on 10-cm diameter plates by trypsinization. For
22 membrane preparation, the cells were subcultured with a ratio of 1:10 and transferred to 15-
23 cm diameter plates. The cells were collected by scraping in 5 mL phosphate-buffered saline
24 (PBS) and centrifuged at 1,000g for 5 min. Pellets derived from 30 plates were combined
25 and resuspended in 20 mL cold Tris-HCl, MgCl₂ buffer (50 mM Tris-HCl (pH 7.4), 5 mM
26 MgCl₂). The cell suspension was homogenized using an UltraTurrax homogenizer (Heidolph
27 Instruments Schwabach, Germany). Membranes and cytosolic fractions were separated by
28 centrifugation in a Beckman Optima LE-80K ultracentrifuge (Beckman Coulter Inc., Fullerton,
29 CA, USA) at 100,000 g for 20 min at 4°C. The supernatant was discarded. The pellet was
30 resuspended in 10 mL cold Tris-HCl, MgCl₂ buffer and homogenization and centrifugation
31 steps were repeated. The membranes were resuspended in 10 mL cold Tris-HCl, MgCl₂
32 buffer. Aliquots of 50 μL were stored at -80°C until further use. The protein concentration
33 was determined using the Pierce™ BCA Protein Assay Kit (ThermoFisher Scientific,
34 Waltham, MA, USA).
35
36
37
38
39
40
41
42
43

44 [³H]CP55940 Displacement assay

45 [3H]CP55940 displacement assays on 96-well plates were performed in 50 mM Tris-HCl (pH
46 7.4), 5 mM MgCl₂, and 0.1% BSA assay buffer. Membrane aliquots of CHOK1CB2_bgal
47 containing 1.5 μg membrane protein were incubated at 25°C for 2h in the presence of ~1.5
48 nM [3H]CP55940 (specific activity 149 Ci/mmol; PerkinElmer, Waltham, MA). At first, all
49 compounds were tested at a final concentration of 10 μM . When radioligand displacement
50 was greater than 50%, full curves were recorded to determine the affinity (pKi) values of the
51 compounds. Six different concentrations of the compounds were added by an HP D300
52 digital dispenser (Tecan Group Ltd, Männedorf, Switzerland). In order to determine the total
53 binding, a control without test compound was included. Nonspecific binding was determined
54 in the presence of 10 μM AM630. The total assay volume was 100 μL . The final
55 concentration of DMSO was $\leq 0.25\%$. The incubation was terminated by rapid vacuum
56
57
58
59
60
61
62
63
64
65

1 filtration through GF/C 96-well filter plates (PerkinElmer, Waltham, MA), to separate the
2 bound and free radioligand, using a PerkinElmer Filtermate-harvester (PerkinElmer,
3 Groningen, The Netherlands). Filters were subsequently washed twenty times with ice-cold
4 assay buffer. The filter-bound radioactivity was determined by scintillation spectrometry
5 using a Microbeta2® 2450 microplate counter (PerkinElmer, Boston, MA), after addition of
6 25µl MicroScint 20 (PerkinElmer, Groningen, The Netherlands) and 3h incubation.
7

8 Data Analysis 9

10 All experimental data were analyzed using GraphPad Prism 7 [37]. The data were
11 normalized to percentage specific radioligand binding, where the total binding is 100% and
12 nonspecific binding is 0%. Nonlinear regression for one-site was used to determine the IC_{50}
13 values from the full curve [3H]CP55940 displacement assays. The pK_i values were obtained
14 using Equation 4 proposed by Cheng-Prusoff [38].
15
16

$$17 K_i = \frac{IC_{50}}{\left(1 + \left(\frac{[L]}{K_D}\right)\right)} \quad 4$$

18 where [L] is the exact concentration [3H]CP55940 determined per experiment and the K_D is
19 the dissociation constant of [3H]CP55940, which is 1.24 nM as determined by Soethoudt et
20 al [38]. All data were obtained from three separate experiments performed in duplicate.
21
22
23

24 Results 25

26 This section presents the effectiveness achieved by our proposal. To this end, we get the
27 potential candidates by applying our MCS models over the Validation set. Moreover, most
28 promising candidates were used by executing our probabilistic-based ranking methodology.
29 Finally, in vitro analysis were performed over the previously achieved candidates to
30 determine their real activity.
31
32
33

34 Virtual Screening 35

36 We applied our MCS models in virtual screening in a non-controlled environment. Here, we
37 do not know the activities of the chemical compounds. Virtual screening refers to the use of
38 computational approaches to identify chemical structures that are predicted to have
39 particular properties. To this end, we designed new experimentation to analyse the
40 behaviour of both meta-models (*Minimize FP* and *Minimize FN*) in a realistic scenario. We
41 classified a list of chemical compounds included in the ValidationSet in order to determine
42 their activity. Below, Table 4 summarizes the outcomes achieved by each model grouped by
43 activity (Active or Inactive). As can be depicted for Table 4, the number of Active compounds
44 predicted by *Minimize FN* is higher than *Minimize FP* (representing 9.085% and 2.629% of
45 the whole dataset), while *Minimize FP* was able to classify more compounds as Inactive.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 3: Summary of predictions group by model.

<i>Meta-models</i>		<i>Predictions</i>
<i>Minimize FP</i>	<i>Minimize FN</i>	
48,232	166,664	<i>Active</i>
1,786,130	1,667,698	<i>Inactive</i>
1,834,362	1,834,362	<i>Total</i>

This scenario clearly fits the behaviour described in Table 3, where *Minimize FP* trends to reduce the FP rate despite sacrificing potential Active compounds while *Minimize FN* is focused on exploring all the potential candidate compounds at expenses of increasing the number of unnecessary trials (caused by FP errors).

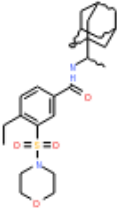
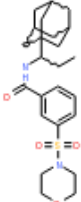
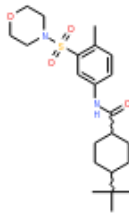
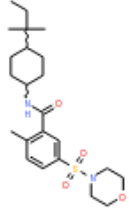
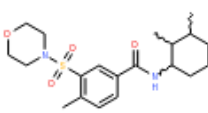
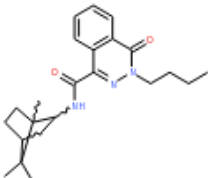
The high amount of potential Active components (48,232) makes it infeasible (in terms of human resources and trial cost) to perform an evaluation of all the predicted actives. Therefore, we selected candidates for experimental validation from the compounds classified as Active by *Minimize FP*. We address the importance of using an adequate candidate-selection method when dealing with a reduced set of compounds (representing only 0.083% of the potential candidates) to avoid obtaining unrepresentative information. To prevent random selection of candidates, we combined a chemical clustering method with a probabilistic-based ranking methodology. The designed probabilistic-based ranking methodology was used to rank each active-predicted compound (see Additional File 2). This ranking was subsequently used to select the most suitable candidates from chemical clusters. These clusters were constituted from the list of 48,232 predicted actives. Clustering of the predicted actives resulted in 28,217 chemical clusters. From each cluster, the top scoring member (based on the ranks generated by the probabilistic-based ranking methodology) was kept while the other cluster members were discarded. Finally, the top 60 scoring compounds (60 clusters) were selected and 21 novel and diverse compounds were purchased. The average similarity in the set based on Tanimoto similarity was 0.19 ± 0.11 , the average probability to be active was 0.77 ± 0.02 , and the average similarity to the training set was 0.74 ± 0.06 (distance 0.26). Hence, it can be concluded that the set selected was internally chemically diverse, highly probable to be active, and relatively close to the training set.

In Vitro Evaluation

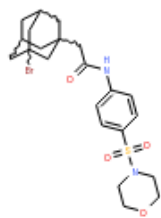
The affinities of the 21 purchased compounds for the human CB2 receptor were determined in a radioligand displacement assay using [³H]CP55940 as the radiolabeled competitor (Table 4). Six compounds were able to displace more than 50% of the radioligand at 10 μ M, and were thus further characterized for their affinity, where the compound with the highest affinity was **Z336532434** (pKi 7.67). Taken together, we were able to obtain 11 hits from the 21 novel compounds (representing a 52% hit rate). As can be seen from Table 4, four out of these 11 are in the top five based on probability. Moreover, the top 10 compounds based on probability contained 7 out of 11 actives. We conclude that our defined probability can be a good estimator of biological activity. Most notable is compound Z27680708, which was

measured to have a pKi of 7.46 while the Tanimoto similarity to the training set was the lowest of the 21 tested compounds at 0.69.

Table 4: Experimentally validated compounds

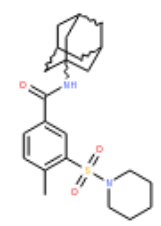
Data Image	IDnumber / InChiKey	Probability	Distance To Closest	pKi \pm SEM or % displ.
	Z336532434 / MQIUMQLPFGFWME-UHFFFAOYSA-N	0.82	0.32	7.67 \pm0.17
	Z28609248 / HXJYJTXXUOYRSB-UHFFFAOYSA-N	0.81	0.29	16%
	Z26476746 / VYCWCTZNPBMJFW-UHFFFAOYSA-N	0.80	0.21	6.54 \pm0.14
	Z91179667 / XGVYRTRSINEVTE-UHFFFAOYSA-N	0.78	0.15	29%
	Z32934509 / OLTBRMQFCQBIR-UHFFFAOYSA-N	0.78	0.28	6.47 \pm0.02
	Z28357657 / NPRYSOPFJOGFSA-UHFFFAOYSA-N	0.78	0.34	6.81 \pm0.29

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65



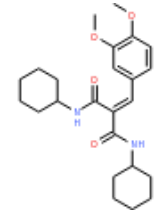
Z30007452
/
VBFKBSAAMKINJD-UHFFFAOYSA-N

0.77 0.24 -2%



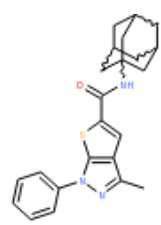
Z27687312
/
IHBHBQAPEZJCNM-UHFFFAOYSA-N

0.77 0.23 7.22 ±0.46



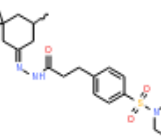
Z46091805
/
QKQCBVJKUBSZOR-UHFFFAOYSA-N

0.76 0.24 38%



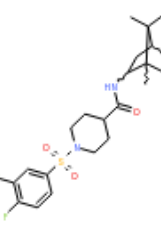
Z27687279
/
WTGACPGXOMAZFA-UHFFFAOYSA-N

0.76 0.22 38%



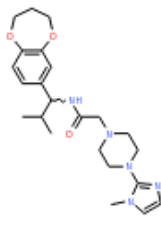
Z44866691
/
WPWBUEOMELTWOC-FCDQGJHFSAN

0.76 0.25 -1%



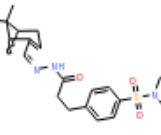
Z28357392
/
VBIMVPWQQQUESTK-UHFFFAOYSA-N

0.76 0.13 26%



Z1317886912
/
MEXULSRPIBCDQX-UHFFFAOYSA-N

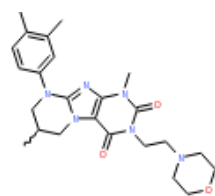
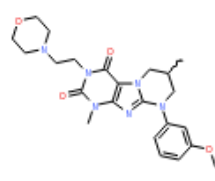
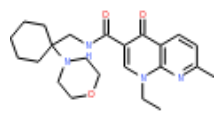
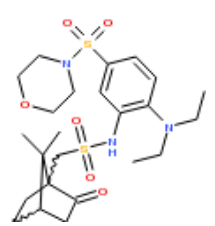
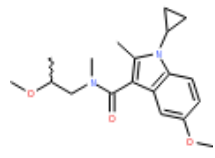
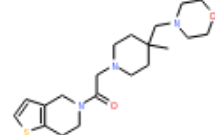
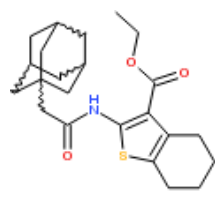
0.76 0.28 3%



Z44867007
/
PCCXRCZRNECAZ-JLPGSUDCSA-N

0.76 0.30 0%

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

	Z237484560 / LIGIHTRZFDNAN-UHFFFAOYSA-N	0.75	0.15	-1%
	Z223843850 / CVSSLUCDGDGHX-UHFFFAOYSA-N	0.75	0.32	-5%
	Z27019562 / WNXCAGCQOBOQMO-UHFFFAOYSA-N	0.75	0.33	30%
	Z55473655 / VDTRQSFAESBVFB-UHFFFAOYSA-N	0.75	0.26	7%
	Z2094674960 / RISCNDGLDMULEE-UHFFFAOYSA-N	0.75	0.29	0%
	Z1523102560 / IXASXIGZGJSBJT-UHFFFAOYSA-N	0.75	0.30	18%
	Z27680708 / HKWXDCJIBMAAFV-UHFFFAOYSA-N	0.74	0.31	7.46 ±0.32

Shown are the structure, enamine identifier (ID number), InChIKey, assigned probability, distance to the training set, and biological activity. Biological activity is shown as pKi (with a standard error of the mean) when available or % displacement of the radioligand by 10 uM of the compound. Identified novel hits are indicated in bold.

Conclusions

This work uses Multiple Classifier Systems (MCS) applied in early preclinical drug discovery. Concretely, we apply D2-MCS over a training dataset to build two measure-guided MCS (PPV and MCC). Furthermore, two meta-models (*Minimize FP* and *Minimize FN*) were generated by combining the predictions achieved by the previous MCS models.

Results achieved by both meta-models show the suitability of using *Minimize FP* due to its ability to avoid FP errors (only 3 from 477). To this end, we execute *Minimize FP* over a validation dataset (comprised of 1.834.362 compounds) together with our probabilistic-based ranking methodology to obtain the 21 most promising active compounds.

We have demonstrated that an appropriate combination of MCS can be successfully used for virtual screening (to predict the biological activity of chemical structures). The identified hits were chemically diverse while similar to the training set. We were successfully able to determine a probability of biological activity, which demonstrated a predictive performance for biological activity.

Despite the promising results achieved here (being 52.38% hits), further improvements should be addressed to increase the classification performance. Therefore, future work should be focused on two main aspects (*i*) dataset processing and (*ii*) the improvement of D2-MCS toolkit. Regarding data quality, the detection, and removal of irrelevant, noisy, or valueless features from the input dataset should be considered. Moreover, to increase the performance of D2-MCS new and efficient feature clustering methods should be implemented.

Additional files

Additional File 1. Physicochemical descriptors comprising CB2Set.

Additional File 2. List of potential candidates sorted by probability of being Active.

Abbreviations

HTS, High-Throughput Screening

FP, False Positives

FN, False Negatives

TP, True Positives

TN, True Negatives

MCS, Multiple Classifier Systems

DNN, Deep Neural Networks

SVM, Support Vector Machines

D2-MCS, Drugs Discovery for Multi-Clustering System

PPV, Positive Prediction Values

MCC, Matthews Correlation Coefficient

Declarations

Authors' contributions

DRO, JRM, ME, and GvW conceived the study. DRO and IY were responsible for designing and executing the *in silico* experiments. IY designed the three-stage candidate ranker methodology. JRM supervised the *in silico* experimentation; DRO, JRM, and GvW wrote the paper. LB generated the dataset. GvW performed clustering and compound selection. RL and CvdH performed the *in vitro* experimental validation. LHH supervised the *in vitro* experimentation. All authors proofread and agree with the manuscript.

Acknowledgements

D. Ruano-Ordás was supported by a post-doctoral fellowship from Xunta de Galicia (ED481B 2017/018). SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from the University of Vigo for hosting its IT infrastructure.

Competing interests

The authors declare no competing interests.

Availability of data and materials

The MCS framework is available on github: <https://github.com/drordas/D2-MCS> The data used/generated in this study is available from ChEMBL and is available here: <http://doi.org/10.5281/zenodo.2677650> The predicted probabilities for the virtual screening are included as supporting information.

Funding

This work was supported by the Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group.

Gerard JP van Westen was funded by the Dutch Scientific Council Applied and Engineering Sciences (NWO-TTW) for funding (VENI 14410).

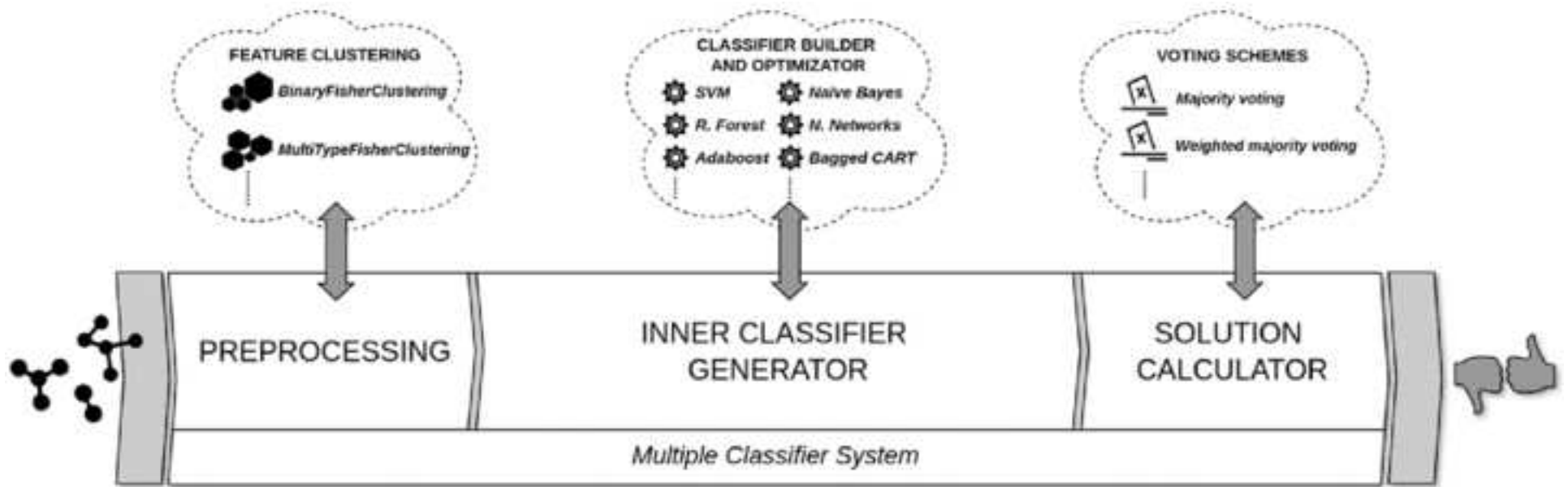
References

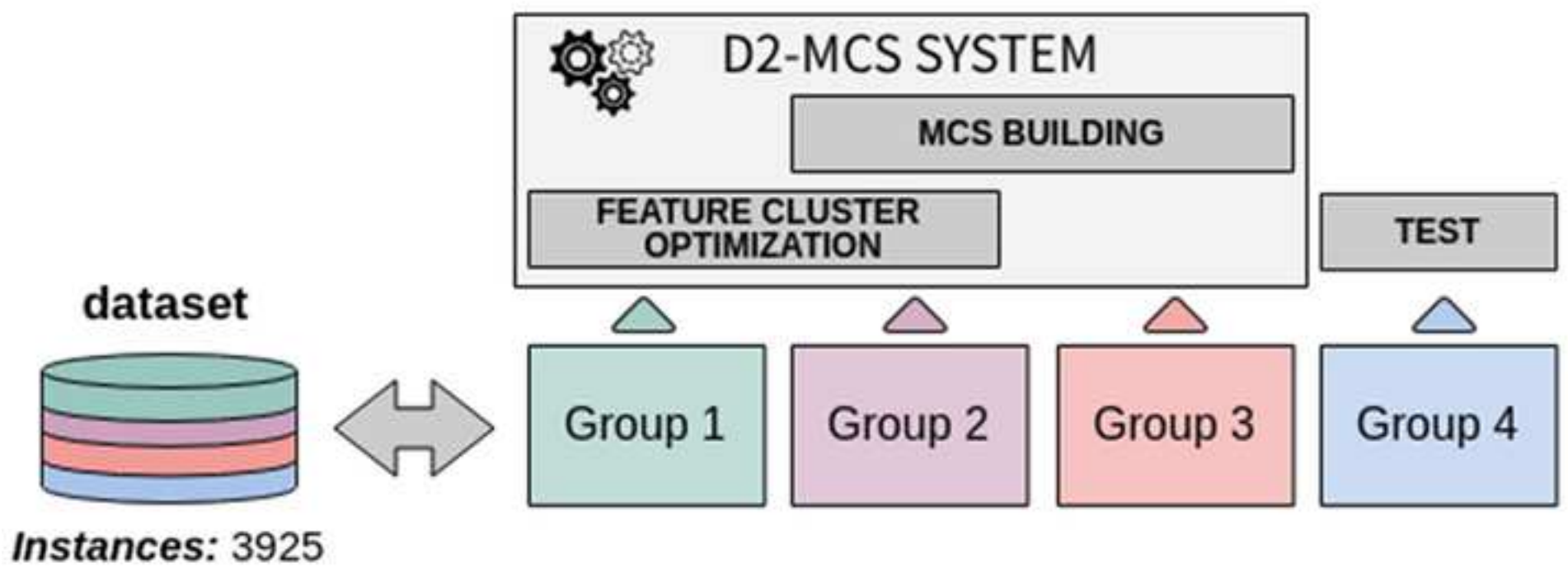
1. Sieburg HB (1990) Physiological studies in silico. *Stud Sci Complex* 12:321–342
2. Danchin A, Médigue C, Gascuel O, et al (1991) From data banks to data bases. *Res Microbiol* 142:913–916. [https://doi.org/10.1016/0923-2508\(91\)90073-J](https://doi.org/10.1016/0923-2508(91)90073-J)
3. Gaulton A, Bellis LJ, Bento AP, et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100–D1107.

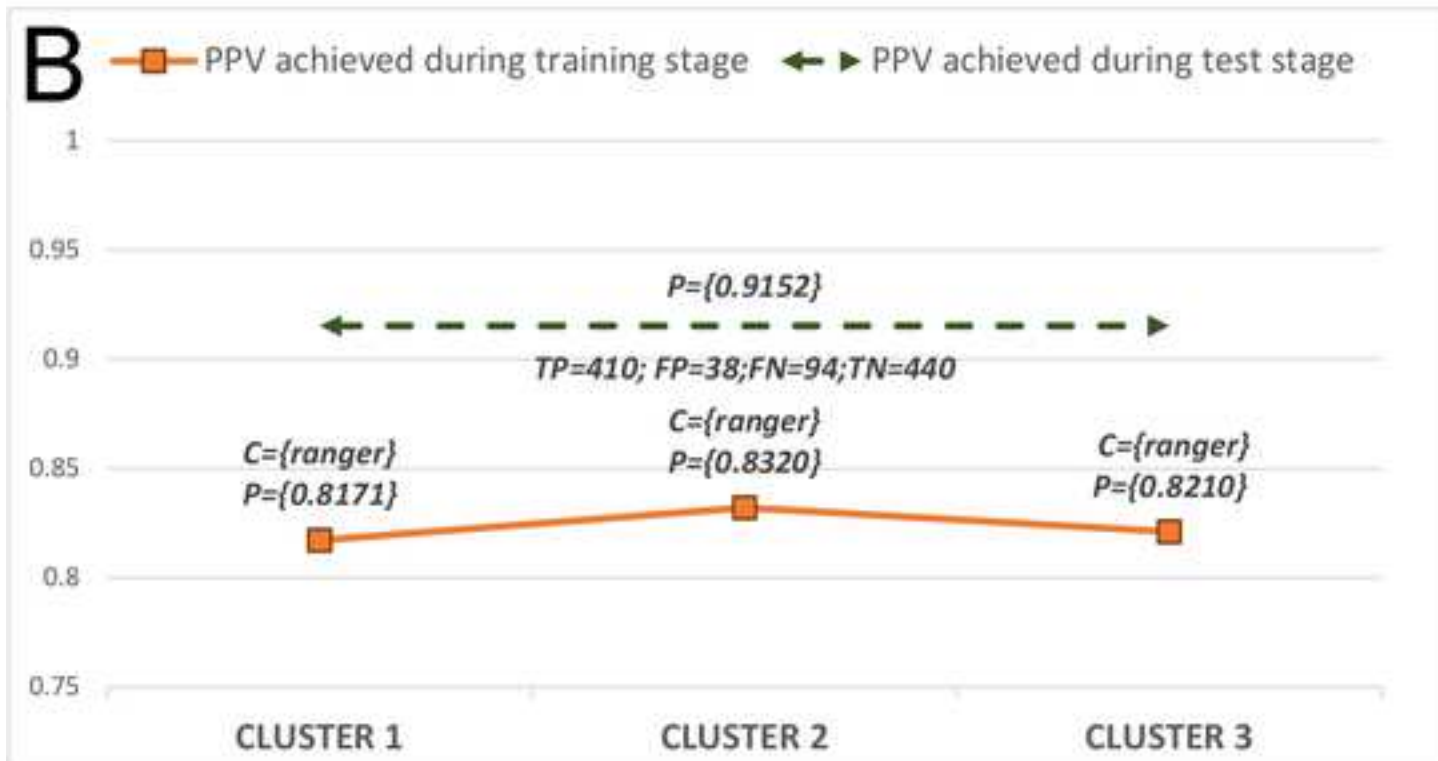
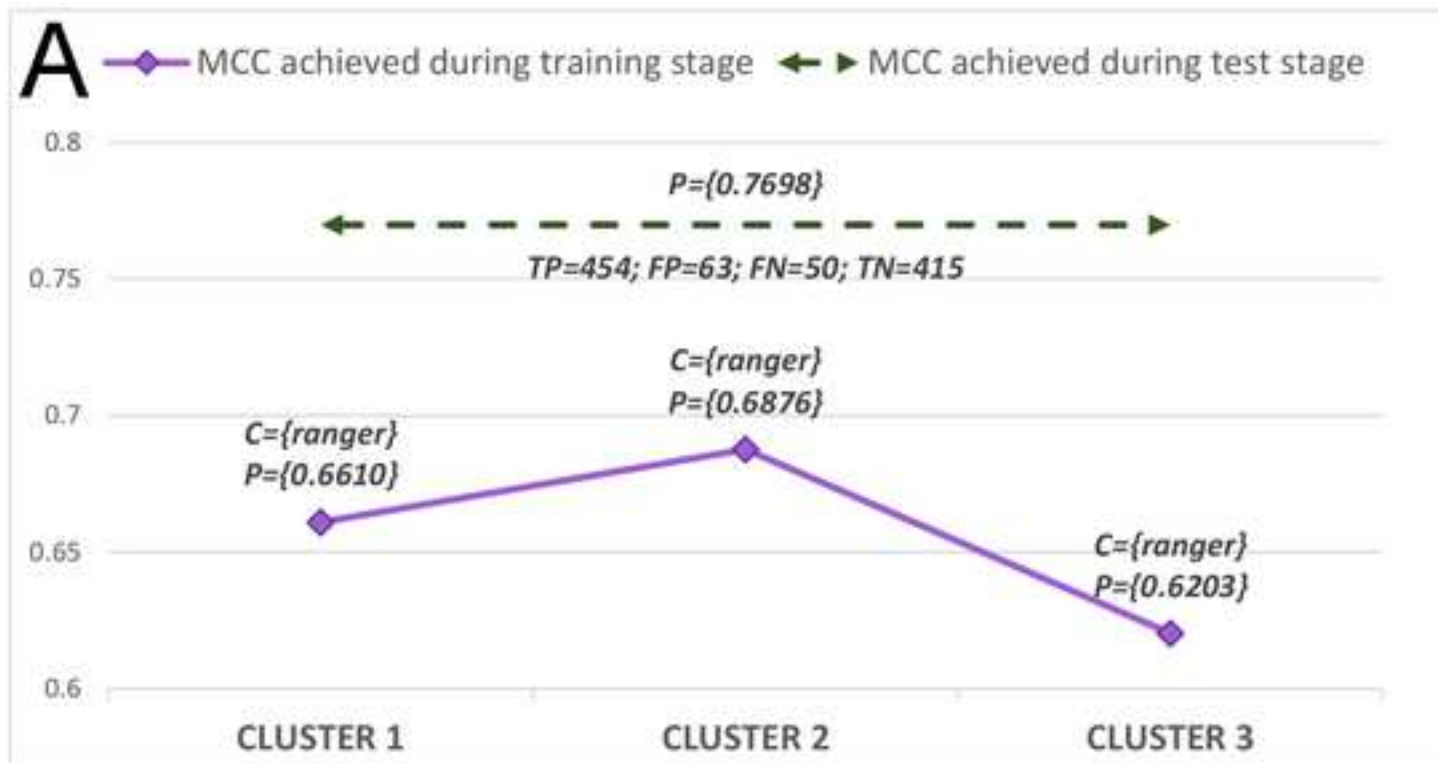
- <https://doi.org/10.1093/nar/gkr777>
4. Sieburg HB (1990) The cellular device machine: Point of departure for large-scale simulations of complex biological systems. *Comput Math with Appl* 20:247–267. [https://doi.org/10.1016/0898-1221\(90\)90332-E](https://doi.org/10.1016/0898-1221(90)90332-E)
 5. Briem H, Lessel UF (2000) In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect Drug Discov Des* 20:231–244. <https://doi.org/10.1023/A:1008793325522>
 6. Mahé P, Ralaivola L, Stoven V, Vert J-P (2006) The Pharmacophore Kernel for Virtual Screening with Support Vector Machines. *J Chem Inf Model* 46:2003–2014. <https://doi.org/10.1021/ci060138m>
 7. Azencott C-A, Ksikes A, Swamidass SJ, et al (2007) One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J Chem Inf Model* 47:965–974. <https://doi.org/10.1021/ci600397p>
 8. Schneider N, Jäckels C, Andres C, Hutter MC (2008) Gradual in Silico Filtering for Druglike Substances. *J Chem Inf Model* 48:613–628. <https://doi.org/10.1021/ci700351y>
 9. Watson P (2008) Naïve Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors. *J Chem Inf Model* 48:166–178. <https://doi.org/10.1021/ci7003253>
 10. Kauffman GW, Jurs PC (2001) QSAR and k -Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J Chem Inf Comput Sci* 41:1553–1560. <https://doi.org/10.1021/ci010073h>
 11. Patel J, Chaudharl C (2005) Introduction to the artificial neural networks and their applications in QSAR studies. In: 5th World Congress on Alternatives and Animal Use in the Life Sciences. pp 269–274
 12. Vracko M (2005) Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies. *Curr Comput Aided-Drug Des* 1:73–78. <https://doi.org/10.2174/1573409052952224>
 13. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities. pp 217–241
 14. Dietterich TG (2000) Ensemble Methods in Machine Learning. In: International Workshop on Multiple Classifier Systems. pp 1–15
 15. Lenselink EB, ten Dijke N, Bongers B, et al (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform* 9:45. <https://doi.org/10.1186/s13321-017-0232-0>
 16. Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural Networks* 61:85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
 17. Boulesteix A-L, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov* 2:493–507. <https://doi.org/10.1002/widm.1072>
 18. Hashim H, Saeed F (2017) Prediction of New Bioactive Molecules of Chemical Compound Using Boosting Ensemble Methods. In: International Conference on Soft Computing in Data Science. pp 255–262
 19. Acharya UR, Akter A, Chowriappa P, et al (2018) Use of Nonlinear Features for Automated Characterization of Suspicious Ovarian Tumors Using Ultrasound Images in Fuzzy Forest Framework. *Int J Fuzzy Syst* 20:1385–1402. <https://doi.org/10.1007/s40815-018-0456-9>
 20. Woźniak Michał and Graña M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. *Inf Fusion* 16:3–17. <https://doi.org/10.1016/j.inffus.2013.04.006>
 21. Ruano-Ordás D, Yevseyeva I, Fernandes VB, et al (2019) Improving the drug discovery process by using multiple classifier systems. *Expert Syst Appl* 121:292–303. <https://doi.org/10.1016/j.eswa.2018.12.032>
 22. Gaulton A, Bellis LJ, Bento AP, et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100–D1107.

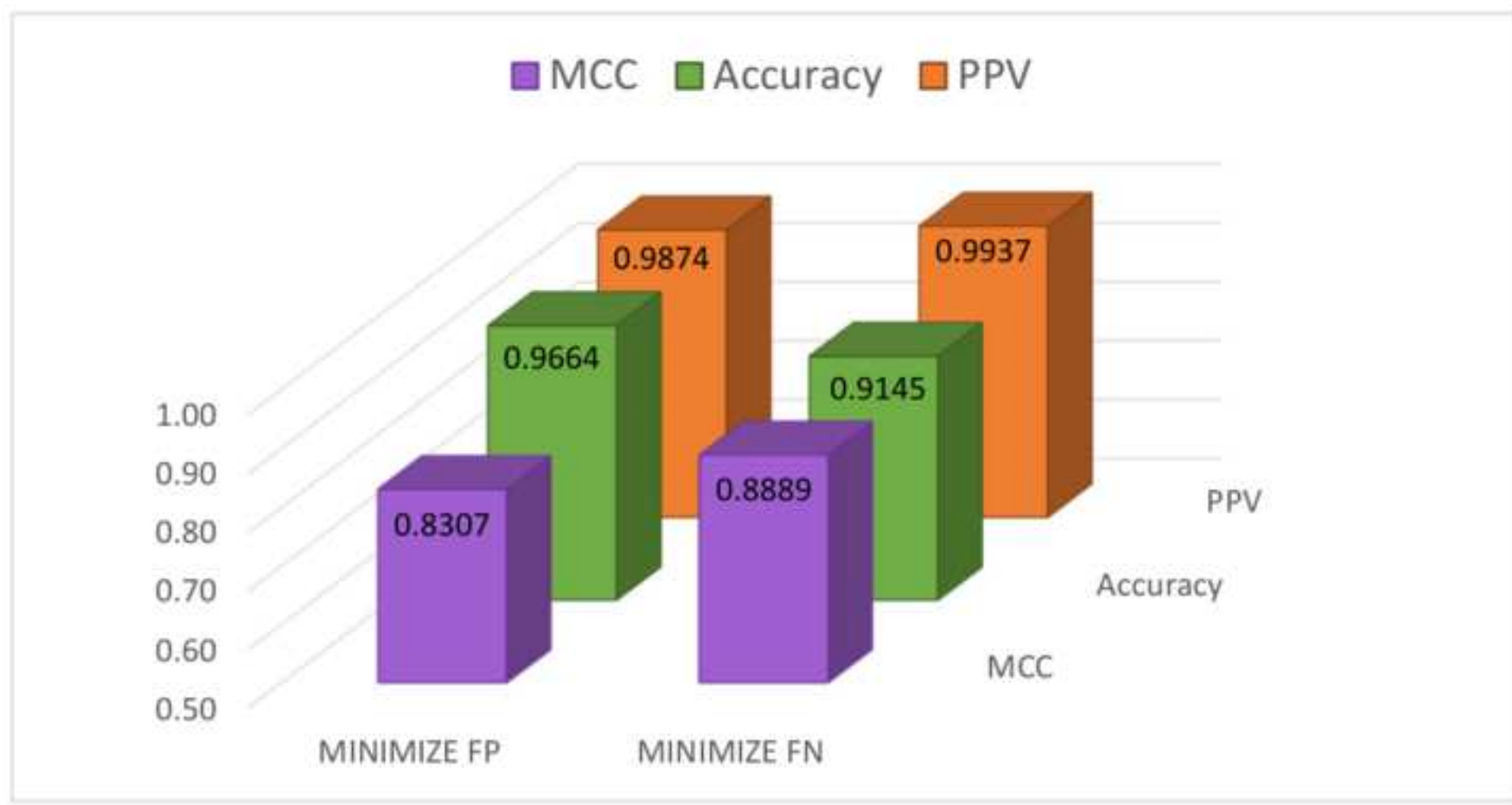
<https://doi.org/10.1093/nar/gkr777>

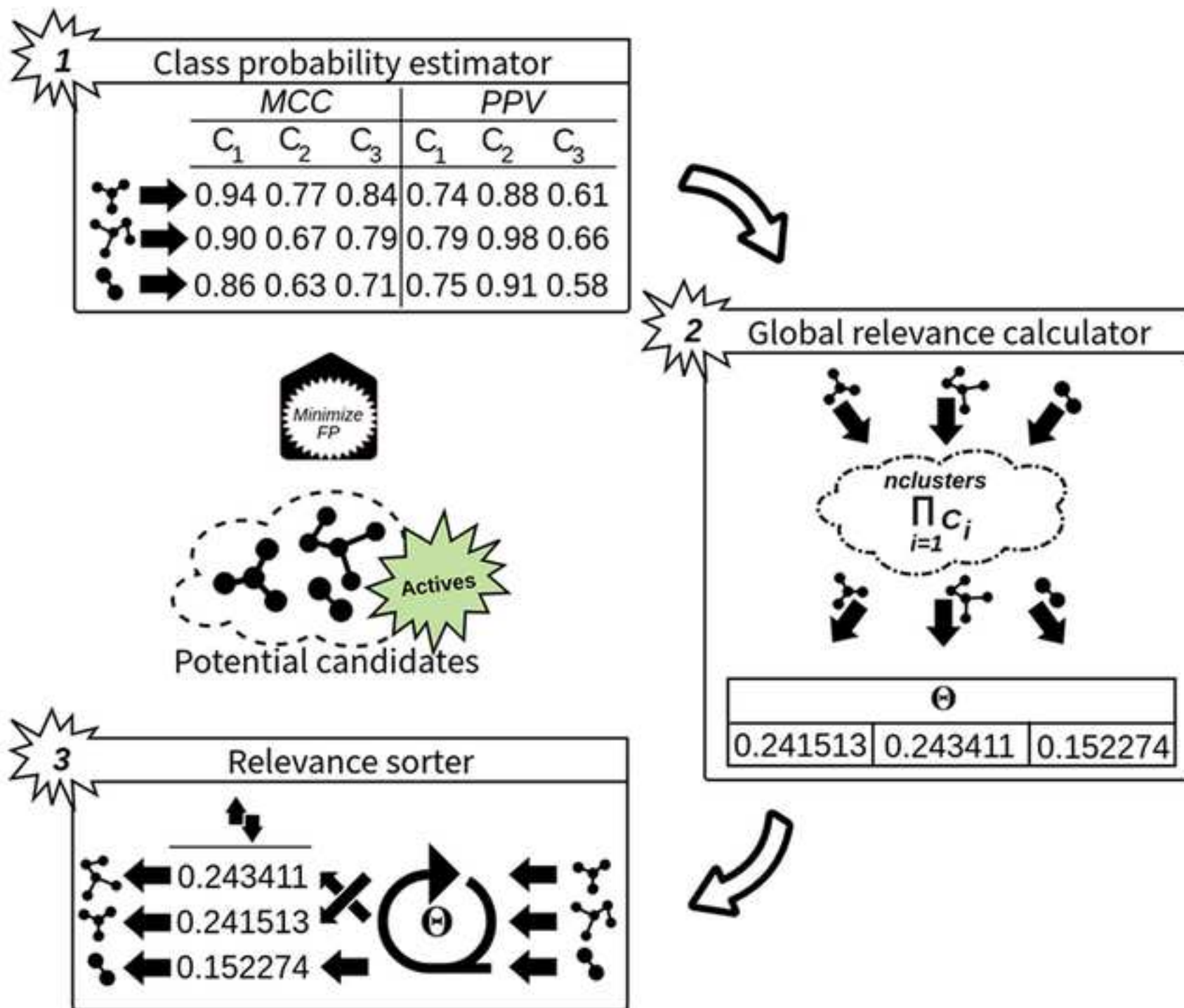
23. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. *J Chem Inf Model* 50:742–754. <https://doi.org/10.1021/ci100050t>
24. Dassault Systèmes BIOVIA (2016) Pipeline Pilot (version 2016)
25. Heller S, McNaught A, Stein S, et al (2013) InChI - the worldwide chemical structure identifier standard. *J Cheminform* 5:7. <https://doi.org/10.1186/1758-2946-5-7>
26. Burggraaff L (2018) CB2 Set Supporting Information. https://surfdrive.surf.nl/files/index.php/s/RAjHDCwZ3H3Lazr/download?path=%2FCB2&files=FCFP_6_Supporting_info_dataset.txt.gz. Accessed 15 Apr 2019
27. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* 12:e0177678. <https://doi.org/10.1371/journal.pone.0177678>
28. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta - Protein Struct* 405:442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
29. Lalkhen AG, McCluskey A (2008) Clinical tests: sensitivity and specificity. *Contin Educ Anaesth Crit Care Pain* 8:221–223. <https://doi.org/10.1093/bjaceaccp/mkn041>
30. Bewick V, Cheek L, Ball J (2004) Receiver operating characteristic curves. *Crit Care* 8:508. <https://doi.org/10.1186/cc3000>
31. Hajian-Tilaki K (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp J Intern Med* 4:627–635
32. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27:861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
33. Goutte C, Gaussier E (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. pp 345–359
34. Chicco D (2017) Ten quick tips for machine learning in computational biology. *BioData Min* 10:35. <https://doi.org/10.1186/s13040-017-0155-3>
35. Maxim LD, Niebo R, Utell MJ (2014) Screening tests: a review with examples. *Inhal Toxicol* 26:811–828. <https://doi.org/10.3109/08958378.2014.955932>
36. Kuhn M (2008) Building Predictive Models in R Using the caret Package. *J Stat Softw* 28:. <https://doi.org/10.18637/jss.v028.i05>
37. GraphPad Software Inc (2018) GraphPad Prism 7
38. Yung-Chi C, Prusoff WH (1973) Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharmacol* 22:3099–3108. [https://doi.org/10.1016/0006-2952\(73\)90196-2](https://doi.org/10.1016/0006-2952(73)90196-2)















Click here to access/download
Supplementary Material
Supplementary_File_2.xlsx



Click here to access/download
Supplementary Material
Author_change_form.pdf

