Mol Divers (2014) 18:637–654
DOI 10.1007/s11030-014-9513-y

FULL-LENGTH PAPER

Toward the computer-aided discovery of FabH inhibitors. Do predictive QSAR models ensure high quality virtual screening performance?

Yunierkis Pérez-Castillo · Maykel Cruz-Monteagudo · Cosmin Lazar · Jonatan Taminau · Mathy Froeyen · Miguel Ángel Cabrera-Pérez · Ann Nowé

Received: 21 November 2013 / Accepted: 28 February 2014 / Published online: 27 March 2014

Abstract Antibiotic resistance has increased over the past two decades. New approaches for the discovery of novel antibacterials are required and innovative strategies will be necessary to identify novel and effective candidates. Related to this problem, the exploration of bacterial targets that remain unexploited by the current antibiotics in clinical use is required. One of such targets is the β-ketoacyl-acyl carrier protein synthase III (FabH). Here, we report a ligand-based modeling methodology for the virtual-screening of large collections of chemical compounds in the search of potential FabH inhibitors. QSAR models are developed for a diverse dataset of 296 FabH inhibitors using an in-house modeling framework. All models showed high fitting, robustness, and generalization capabilities. We further investigated the performance of the developed models in a virtual screening scenario. To carry out this investigation, we implemented a desirability-based algorithm for decoys selection that was shown effective in the selection of high quality decoys sets. Once the QSAR models were validated in the context of a virtual screening experiment their limitations arise. For this reason, we explored the potential of ensemble modeling to overcome the limitations associated to the use of single classifiers. Through a detailed evaluation of the virtual screening performance of ensemble models it was evidenced, for the first time to our knowledge, the benefits of this approach in a virtual screening scenario. From all the obtained results,

Electronic supplementary material The online version of this article (doi:10.1007/s11030-014-9513-y) contains supplementary material, which is available to authorized users.

Y. Pérez-Castillo (✉) · C. Lazar · J. Taminau · A. Nowé (✉)
Computational Modeling Lab (CoMo), Department of Computer Sciences, Faculty of Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: yunierkis@gmail.com

A. Nowé
e-mail: ann.nowe@vub.ac.be

Y. Pérez-Castillo · M. Cruz-Monteagudo · M. Á. Cabrera-Pérez
Molecular Simulations and Drug Design Group, Centro de Bioactivos Químicos, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba

Y. Pérez-Castillo · M. Froeyen
Laboratory for Medicinal Chemistry, Rega Institute for Medical Research KU Leuven, Minderbroedersstraat 10, 3000 Leuven, Belgium

M. Cruz-Monteagudo
CIQ, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

M. Cruz-Monteagudo
REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

M. Cruz-Monteagudo
Centro de Estudios de Química Aplicada (CEQA), Faculty of Chemistry and Pharmacy, Central University of Las Villas, 54830 Santa Clara, Cuba

J. Taminau
Cartagenia N.V. Technologielaan 3, 3001 Leuven, Belgium

M. Á. Cabrera-Pérez
Department of Engineering, Area of Pharmacy and Pharmaceutical Technology, Miguel Hernández University, Sant Joan d'Alacant, 03550 Alicante, Spain

M. Á. Cabrera-Pérez
Department of Pharmacy and Pharmaceutical Technology, University of Valencia, Burjassot, 46100 Valencia, Spain

we could arrive to a significant main conclusion: at least for FabH inhibitors, virtual screening performance is not guaranteed by predictive QSAR models.

Keywords Ligand-based drug design · Virtual screening · FabH inhibitors · Ensemble modeling · QSAR

Introduction

Bacterial diseases have not yet been overcome. It is estimated that two million people die every year as a consequence of bacterial infections [1] and an increased resistance of pathogens to the antibiotics in clinical use has been reported [2]. This situation is aggravated by the fact that only two new classes of antibiotics have been introduced in the marked during the past 30 years [3]. In addition, known antibiotics are directed against a small subset of bacterial targets. To overcome this critical situation, it is urgent to develop innovative approaches and strategies for the identification of novel antibiotic candidates [4].

Traditionally, diverse targets in key processes of bacterial cell cycle have been investigated in antibiotics research. Some of these targets participate in critical cellular processes such as cell wall biosynthesis, transcription and translation of the genetic code, cell division, metabolic pathways, resistance mechanism, and virulence factors [5]. Among them, one of the biosynthetic pathways that can be effectively employed as target for the development of new antibiotics is that of fatty acid biosynthesis (FAS). The FAS pathway displays some key features that make it attractive for the development of new potential broad-spectrum antibacterial agents. Two of these features are: (1) its essentiality for bacteria survival and (2) its divergence from the analog human process [6]. The major difference between fatty acids synthesis in humans and bacteria is that in humans this process is carried out by a single multifunctional enzyme, while in bacteria the pathway involves various enzymes which can be separately considered putative targets [7].

Among the enzymes composing the type II FAS in bacteria, β-ketoacyl-acyl carrier protein synthase III (FabH) plays two important roles: (1) the initialization of the fatty acids synthesis cycle and (2) the regulation of the whole pathway [8]. This enzyme is present in many different important human pathogens and has been shown to be essential [9]. Also, compounds derived from different chemical scaffolds have been shown to inhibit FabH from a wide range of microorganisms, including multi-drug resistant bacterial strains [10–24]. From all these observations, it is inferred that FabH can be used as a molecular target for novel antibiotic candidates [9].

Apart from few modeling studies, the amount of data available for FabH and its inhibitors has not been exploited to develop computational models to perform computer-aided design of FabH inhibitors [13,25–27]. Specifically, no model for the ligand-based virtual screening (VS) of large collections of chemical compounds to search for potential FabH inhibitors has been reported.

One of the ligand-based modeling methodologies routinely employed in drug discovery projects is quantitative structure–activity relationships (QSAR). The objective of this type of methodology is to correlate the structures of a set of molecules (encoded by molecular descriptors) with their biological activity. Studies reporting the experimental corroboration of QSAR modeling predictions have been published before [28]. However, obtaining predictive QSAR models, and more important, models suitable for virtual screening tasks is not straightforward.

There are several issues that have to be taken into consideration during the whole process of QSAR model development. As has been previously proposed, the quality of the data used to train the models can have a huge influence on their performance [29,30]. For this reason, before any calculation the datasets must undergo a curation step including but not limited to: the removal of salts/fragments; the use of one unique representation of every functional group; the adequate treatment of tautomeric forms; the removal of duplicate compounds and the visual inspection of as much data as possible. This step is critical to develop predictive and reliable high quality QSAR models.

One important consideration to make during the development of classification models is the number of samples in each group and how the dataset is split into the training, selection, and external validation subsets. Regarding the problem of the balance between the number of samples in each category for binary classification, it is necessary to balance the actives and inactives groups to avoid models that would be biased against the majority class and hence useless for virtual screening. One of the options to balance the groups before deriving the classification models is to reduce the number of samples in the larger group [30]. On the other hand, to obtain reliable QSAR models, the representative compounds of the test set should be close to those of the training set and vice versa. Furthermore, the representative compounds of the training set should be distributed within the whole area covered by the entire dataset. Several methodologies have been proposed to obtain an adequate partition of the datasets such as Self-Organizing Maps, clustering, maximum dissimilarity, fractional design, D-optimal design, the Kennard–Stone method, and sphere exclusion algorithms [31–33].

Another issue to take into account is the performance of the developed models in a virtual screening scenario. It is also important that a model for virtual screening is able to position active compounds at the beginning of the ranked list [34,35]. In other words, a method achieving a good classification performance but ranking the active compounds at the middle of the ordered list has no practical utility for vir-

tual screening since they will never be selected for experimental assays. Besides, the validation of a model for virtual screening should be accomplished using a validation dataset resembling a real virtual screening scenario. In this sense, given that the amount of known inactive compounds for any specific problem is limited, it is important to combine a set of known active compounds with decoy molecules [35–37]. This means that the validation set for a virtual screening tool should contain molecules that resemble the actives' physicochemical properties and at the same time are structurally dissimilar from them [34,35]. The ratio of active compounds to inactive ones is also important during the validation process since hit rates in real world problems range from 0.01 to 0.14 % [37]. Special attention should be paid to this last issue since it has been demonstrated that this rate influence the values of the metrics used to evaluate the performance of virtual screening tools [34].

Here, we present the first study aimed to obtain a ligand-based modeling tool suitable for the VS of large collections of chemical compounds in the search for potential FabH inhibitors. We provide a detailed description of the steps involved in the obtaining of predictive QSAR models using our in-house QSAR modeling framework. One of the objectives of this research is to make a thorough validation of the obtained QSAR models, and specifically evaluate their virtual screening performance. Since a set of decoy molecules is necessary for the correct assessment of the performance of the models in the context of a virtual screening experiment, we also present a novel strategy for the selection of target-specific decoys sets. Although ensemble modeling has been previously employed in improving classification performance of QSAR models [38,39], to our knowledge their worth for virtual screening has not been extensively evaluated using large collections of decoy molecules. Here, we provide evidences supporting the idea that ensemble modeling can be successfully employed to obtain reliable virtual screening tools capable of overcoming the limitations associated to the use of single models.

Computational methods

Dataset compilation

A dataset of 296 compounds was compiled from 11 literature sources that report the antibacterial and inhibitory activities of FabH inhibitors against *E. coli* and ecFabH [14–24]. In all the papers the authors report the minimum inhibitory concentration (MIC) obtained via microdilution experiments for all compounds and the FabH inhibitory activity of a subset of the most active compounds determined by enzyme inhibition assays. For that reason, it was decided to use MIC as property to model. Since the MIC a discrete value, the compiled dataset was used for classification studies. It is important to highlight that all the compounds were synthesized and tested in the same lab and that in all the papers the authors report the antibacterial activity of Kanamycin B. Considering that the experiment outcome can change depending on the experimental conditions being used, even when the experiments are carried out in the same lab following the same protocol, the assessment of the same reference molecule in all the literature reports makes a consistent inter-experiments comparison possible. The activity threshold to consider a compound in the active group was set to up to four times the MIC of Kanamycin B. The remaining compounds were regarded as inactives. This resulted in 125 active and 171 inactive compounds.

Representation, curation, standardization, and codification of molecular structures

The FabH inhibitors dataset was initially represented using the SMILES notation and then converted to a multiple two-dimensional SDF file using the MOLCONVERT tool of the JChem v5.9 software package [40]. This dataset as well as the decoy compounds candidates (see the next sections) was subject to a curation process to ensure a uniform structural representation across all the compounds. The dataset was standardized using the STANDARDIZER module of the JChem v5.9 software package [40]. During the standardization step, several filters were applied in the following sequence: fragment removal, removal of explicit hydrogen atoms, neutralization, taumerization, transformation of functional groups such as nitro to one unique representation, aromatization, and addition of explicit hydrogen atoms. The next step for the curation of the dataset is the identification of duplicate structures. Duplicate structures were detected with the EdiSDF tool of the ISIDA/QSPR package [41].

Molecular descriptors were calculated with the DRAGON v.6 software [42]. Initially, all the 1D and 2D descriptors were computed (3763 in total) and constant features were removed. Next, all pairs of descriptors with correlation greater than 0.9 were identified and only one feature from each pair was kept.

Dataset splitting

For the development of the QSAR models, the FabH inhibitors dataset was split into training, selection, and external subsets. The training set was used for feature selection and classifiers training, the selection subset was reserved for use with the training set for the model validation and selection steps while the external set is only used to evaluate the final generalization capabilities of the optimal models. It should be pointed out that for the calculations here described all the molecular descriptors were scaled to the interval [0,1]

to avoid features in larger numeric ranges dominate those in smaller ranges.

Before splitting the dataset into the three subsets, to avoid models biased toward the majority class the number of samples in each group was balanced. The first step in balancing the dataset is the calculation of the pairwise matrix of Euclidean distances in the descriptors space. From this matrix, the distance of the samples in the majority group to the minority one was calculated as the mean distance of it to each sample on the minority subset. Then, the size of the two groups was balanced by removing the compounds in the majority group that are furthest from the minority one.

From the balanced dataset, 20 % of the compounds in each group were randomly selected as the external validation set. The remaining data, regarded as the learning data subset, were split into the training and selection subsets using the three sphere exclusion algorithms proposed by Golbraikh et al. [32] and that have been previously used for the rational selection of training and test sets for QSAR modeling. This method uses the distance between the samples in the N-dimensional descriptor space as a measure of their similarity. In brief, the algorithm starts by selecting one compound from the data set and constructing a hypersphere around it. Then the compound closer to hypersphere center is assigned to the selection set while the hypersphere center and the rest of the compounds inside it are assigned to the training set. Then, the compounds inside the hypersphere are removed from the initial pool of samples and the process is repeated until the initial pool of compounds is empty. These calculations are performed with the features scaled to the [0,1] interval to avoid higher influence of the features with bigger numeric ranges during the distance computations and comparisons as well as during the hyperspheres construction.

The three variants 1M, 2M, and 3M of the sphere exclusion algorithm proposed by the developers were implemented in MATLAB R2009a [43].The radius of the constructed spheres is defined as $R = c(V/N)^{1/K}$, c is an adjustable parameter named dissimilarity level, V is the volume occupied by one sample that is set to 1, N is the number of samples in the complete dataset and K is the number of molecular descriptors. The dissimilarity level c was varied between 0.1 and 5 with a step of 0.1 to obtain different partitions of the learning data into training and selections subsets.

To evaluate the quality of the obtained learning set partitions at each dissimilarity level, three parameters were evaluated: the diversity index of the selection set with respect to the training set $M_{\text{sel,train}}(c)$, the diversity index of the training set with respect to the selection set $M_{\text{train,sel}}(c)$ and the diversity index of the training set $I_{\text{train}}(c)$. These parameters characterize the closeness of the compounds in the selection set to those in the training one, the closeness of the samples in the training set to those in the selection set and the diversity of the training data subset, respectively. To compute $M_{\text{sel,train}}(c)$ a sphere of radius $R = c(V/N)^{1/K}$, is constructed taking as center each selection set sample, then the number of points of the selection set for which the sphere contains no point of the training set is determined to be N_{a}, finally $M_{\text{sel,train}}(c) = N_{\text{a}}/N_{\text{sel}}$ where $N\text{ok}_{\text{sel}}$ is the number of compounds in the selection set. In a similar way is computed $M_{\text{train,sel}}(c) = N_{\text{b}}/N_{\text{train}}$ where N_{b} is the number of training samples for which the sphere contains no points of the selection set and N_{train} is the number of samples in the training set. To calculate the diversity index of the training set $I_{\text{train}}(c)$ spheres of radius $R = c\,(V/N)^{1/K}$, are constructed around each training sample and the number of points that contains no other training sample inside its sphere N_{c} is determined. Then, $I_{\text{train}}(c)$ is computed as $I_{\text{train}}(c) = N_{\text{c}}/N_{\text{train}}$. According to these definitions an ideal learning set partition should yield values of $M_{\text{sel,train}}(c) = 0$, $M_{\text{train,sel}}(c) = 0$ and $I_{\text{train}}(c) = 1$.

QSAR modeling

QSAR modeling was carried out using an in-house framework that relies on the classification algorithms that we previously implemented and validated [44]. This QSAR modeling framework is based in the combination of three different feature selection and classification algorithms. To perform feature selection, we employed three algorithms that differ on their underlying concepts: Genetic Algorithms (GA) which is a well-known type of bio-inspired stochastic search engine; bagged trees (BT) which are based on the manipulation of the input data and classification trees; and features ranking (FR) that is a stable feature selection technique, this means that each time it is run with the same input data the subset of selected features will be the same. The same diversity-based idea was fallowed for the selection of the classification algorithms to be combined with the feature selection methods. We selected: Adaboost which is an ensemble-based classification algorithm; Linear Discriminant Analysis that is a well-established linear classifier; and the effective Least Squares Support Vector Machines that is a non-linear modeling approach.

The combination of these feature selection and classification algorithms provides us with nine different pools of classifiers from which we can select the best performing model to solve any particular QSAR modeling problem. As pointed out before, in a typical QSAR modeling, there are many statistically equivalent solutions and a researcher won't know a priori whether a specific modeling approach will be effective in finding a good enough solution to its problem or not. Thus, to explore the structure-activity landscape using a diverse set of modeling tools increases the chances of finding high quality QSAR models. The procedure used for training each model type set and select one optimal model is similar for all modeling approaches. The overall QSAR modeling work-

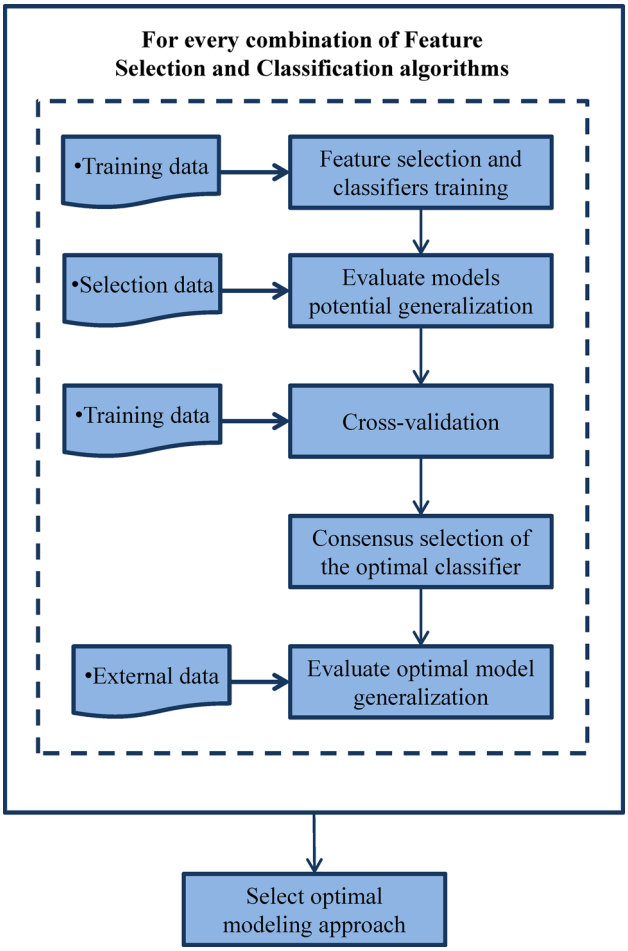

Fig. 1 Overall workflow followed to obtain the optimal model for each classification strategy. For each of the nine classification approaches used, we train a pool of models from which we select the optimal one that combines accuracy, robustness, and potential predictive power. The evaluation of the predictive power of the optimal model is accomplished using the reserved external validation set. Finally, we select the optimal classification approach among these nine optimal models

flow is shown in Fig. 1, and its pseudocode is provided in Chart 1.

For every one of the three feature selection techniques, the modeling process starts with the use of the training dataset to generate a pool of potential solutions (or models) $\mathbf{S_{i,j}}$ by combining the feature selection algorithm FS_i and the classification algorithm C_j. In our QSAR modeling framework GA are used for a wrapper while BT and FR are used as filter methods. For this reason, and to save computation time, the feature selection process for the BT and FR algorithms are carried out only once, the selected features are saved, and they are loaded later to train each type of classifiers based on these filter methods.

The number of models generated by a particular modeling approach depends on the classification tool being used. Once this pool of models is obtained, the generalization capability of each potential solution is evaluated using the selection dataset. After this step, the pool of obtained models can be optionally reduced to keep only those combining high fitting and generalization capabilities $\mathbf{S^*_{i,j}}$. Afterward, the models in $\mathbf{S^*_{i,j}}$ are cross-validated using the Leave-One-Out and Bootstrapping strategies to evaluate their robustness. Next, a consensus ranking approach is applied to the whole set of models $\mathbf{S^*_{i,j}}$ to retrieve the one that better combines accuracy, robustness and generalization capabilities $\mathbf{S^{opt}_{i,j}}$.

At this point, we have selected a model that combines fitting, robustness and potential predictive power among all potential solutions. Since the prediction of the selection data subset is used as decision maker during the model selection step it cannot be considered as the final estimator of the predictive power of the model. Because of this, and to correctly evaluate the generalization capabilities of the optimal model, its real predictive power has to be corroborated using the external validation data as the last step of the modeling process. We should highlight that if the model selected as optimal fails in achieving a good performance when predicting the external validation subset, then it will be regarded as having no real predictive power and hence it cannot be used to predict new data.

Once one optimal model is selected for every one of the nine implemented classification approaches, another consensus ranking procedure is carried out to select among these nine models the optimal classification technique S^{opt} to solve the classification problem under investigation. This QSAR modeling framework was implemented in MATLAB R2009a [43]. The detailed description of the steps involved in each block of Fig. 1 is provided as Supporting Information in Sect. S1. This section includes the algorithms setup and parameters optimization procedures as well as the steps followed to validate and select the models. In addition, we provide the description of the metric employed to compute the model distances in the descriptors space in the Supporting Information Sect. S2.

Applicability domain

To define the applicability domain of the models, two definitions were used. Both applicability domain definitions are based on the descriptors range method [45]. For the first definition of the applicability domain (ADD-1) each feature included in the model is used to build a hyper-rectangle defined by the maximum and minimum values of the features and to each dimension the standard deviation of each feature is added in both directions. A sample is considered to be inside the model's applicability domain if it is contained within the defined hyper-rectangle.

The second applicability domain definition (ADD-2) is based on the Principal Component Analysis (PCA) of the molecular descriptors included in the model as described in

Chart 1 QSAR modeling framework pseudocode

1. //
2. // Require: Dataset $\mathbf{Z} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and vector of observed classifications $\mathbf{y}$
3. //
4. For each feature selection technique FS_i in $\{GA, BT, FR\}$
5. For each classifier C_j in $\{AB, LDA, LS-SVM\}$
6. If $[(FS_i = 'BT' || FS_i = 'FR') \&\& (j = 1)]$
7. Select features subset $\mathbf{FS_i^*}$ using FS_i
8. Save $\mathbf{FS_i^*}$
9. Use training data to find set of potential solutions $\mathbf{S_{i,j}}$ by combining features subset $\mathbf{FS_i^*}$ and C_j
10. Elseif $(FS_i = 'BT' || FS_i = 'FR')$
11. Load $\mathbf{FS_i^*}$
12. Use training data to find potential solutions $\mathbf{S_{i,j}}$ by combining features subset $\mathbf{FS_i^*}$ and C_j
13. Else
14. Use training data to find potential solutions $\mathbf{S_{i,j}}$ by combining FS_i and C_j
15. End If
16. Estimate the potential predictive power of models in $\mathbf{S_{i,j}}$ using the selection data subset
17. Optionally, filter solutions to keep only fitted and potentially generalizable ones $\mathbf{S_{i,j}^*}$
18. Cross-validate models in $\mathbf{S_{i,j}^*}$
19. Consensus selection of the optimal model $S_{i,j}^{opt}$ from the pool of potential solutions $\mathbf{S_{i,j}^*}$
20. Use external dataset to evaluate the predictive power of $S_{i,j}^{opt}$
21. Return optimal model $S_{i,j}^{opt}$ for the FS_iC_j combination
22. End For
23. End For
24. Consensus selection of the optimal classification technique S^{opt} among the $\mathbf{S_{i,j}^{opt}}$ set of optimal models
25. Return optimal classification model S^{opt}

our previous publication [44]. In brief, a PCA is performed for the training data set and the Principal Components (PCs) explaining 99 % of the observed variance are employed to build a hyper-rectangle defining the AD of the model. Next, the samples in the selection and external sets are projected into the PC transformed space. The samples inside the previously defined hyper-rectangle are considered to be inside the model's AD. In other words, a sample is considered to be inside the model's AD if its PC transformed coordinates are within the range of the corresponding training set PC transformed coordinates. This second applicability domain definition is more restrictive than the first one.

Virtual screening performance

To evaluate the performance of the developed models in a virtual screening scenario the following metrics were computed: Area Under the Accumulation Curve (AUAC); Area under the Receiver Operating Characteristic Curve (ROC); Enrichment factor (EF), and Boltzmann-enhanced discrimination of ROC (BEDROC) [34,36]. It is important to note that AUC and ROC don't discriminate the early part of the rank-ordered list from the last part and hence these metrics are not appropriate to address the early recognition problem. On the other hand, although correctly ranking virtual screening methods, EF has the disadvantage that it equally weights all the active compounds within the considered cutoff. On the other hand, BEDROC is able of overcoming the above difficulties. The definitions of these metric are provided in the Supporting Information Sect. S3

Decoys selection

The number of decoy molecules to select per known inhibitor was selected following the guidelines provided by Truchon et al. [34]. Specifically, the number of decoys was selected using the empirical relationship proposed by Truchon et al. between the maximum relative deviation Δ_{max}, the number of active compounds n and the enrichment parameter α:

$$N_{\min} = \frac{\alpha n}{2\Delta_{\max}} \quad (1)$$

The source of the compounds for the decoys selection was the ZINC database and two decoys subsets were generated [46]. To build the first decoys subset the candidate molecules were

first filtered to ensure that the decoy candidates are inside the applicability domain of each model to be validated. In this case, the first applicability domain definition was used since it is less restrictive. To build the second decoys subset, no applicability domain consideration was made.

As proposed by Huang et al. [47] a good decoys set should be as topologically different as possible from the active compounds and at the same time resemble their drug-like properties. Here, a desirability function that takes into account both characteristics was defined. This "decoy-likeness" function was expressed as the geometric mean of the topological dissimilarity and the physicochemical similarity. Five physicochemical descriptors were calculated to encode the drug-like properties: Molecular weight, Number of H-bond donors, Number of H-bond acceptors, Number of rotatable bonds, and LogP. These molecular descriptors were calculated with the Dragon v6.0 software [42]. The physicochemical similarity of two compounds was defined by their Euclidean distance on the 5-dimensional descriptors space and is denoted as $D\_PQ(i, j)$.

The computation of the topological dissimilarity between the decoy candidates and the active compounds was based on chemical fingerprints. The chemical fingerprints were calculated with the GENERATEMD tool of the JChem v5.9 software package [40]. The dissimilarity of any pair of compounds was defined as $D\_T(i, j) = 1 - T(i, j)$ being T the Tanimoto coefficient between molecules i and j. The topological dissimilarity $D\_T(i, j)$ was computed with the COMPR tool of the JChem v5.9 software package [40]. Since the diversity of the selected decoys is also important, N candidates were selected based on the decoy likeness metric defined above and from them $n < N$ decoys were selected using clustering, N is selected according to Eq. 1. The distance for the clustering process was computed based on both physicochemical similarity and topological dissimilarity using the COMPR tool of the JChem v5.9 software package [40]. The pseudocode of the decoys selection process is given in Chart 2.

Models ensemble for virtual screening

The models ensemble was build considering the applicability domain of each individual model and the relative ranking of each compound on the model's ranked list as shown in Chart 3. The assembling of the models starts with the search of the compounds inside the applicability domain of each model. Since considering predictions of samples that out of the models' applicability domain can lead to unreliable predictions, the applicability domain evaluation is a critical step of the process. Next, the aggregated scores produced by each model for the compounds inside its applicability domain are used to obtain a ranking for those compounds. In the next step, all the samples that have the same scoring value are assigned the same ranking value. This particular step takes into account that if more than one sample has the same scoring value then they are indistinguishable by the model producing the ranking. The last step in the generation of the individual ranking of each model is to compute the relative ranking of each sample in **S** as its rank value divided by the number of unique score values in **S**.

To illustrate the importance of using the relative ranking instead of the absolute one when comparing the performance of two models, let's suppose that, as is the case in a real virtual screening experiment, two models A and B have different coverage of the data being screened by their applicability domains. Let's also consider that the coverage of the dataset by model A is greater than the coverage by model B. In such scenario is obvious that to be ranked on the same position by model A has more merit than to be ranked in the same position by model B. The use of the relative ranking when comparing the ranks produced by two models correctly addresses this problem. Once the relative rank of every sample in each model considering the model's applicability domain is determined, these relative rank values are averaged over the models the compound is inside their applicability domains to obtain the final aggregated score. Finally, the compounds are sorted according to these aggregated scores in ascending order to obtain the final ensemble ranking.

Results and discussion

Dataset preparation

The collection, curation, class assignment, representation, codification, and dataset splitting processes were carried out following the procedures described in the Computational Methods section. The compilation of the dataset conducted to 296 FabH inhibitors spanning 11 different chemical scaffolds. The whole dataset, including the discarded isomers, is available as Electronic Supporting Information in SDF format. The SDF file, additionally to the compound structures, includes: the compounds code; the MIC of each compound; their activity relative the Kanamycin B according to the data reported in each source experiment; their observed classification; the link to the source references and their names. The detailed results of this process are presented in the Supporting Information Sect. S4 and they are summarized below.

From a total of 3763 1D and 2D DRAGON molecular descriptors only a subset of 654 was used during the modeling process. The balancing of the dataset resulted in 121 compounds in each group. Afterward, 24 compounds of each group representing 20 % of total samples were randomly selected for the external validation set. The remaining 194 compounds were regarded as the learning set that was split

Chart 2 Pseudocode of the decoys selection process

```
//
// Requires a set of known active compounds and a set of decoy candidates
//
Calculate the physicochemical descriptors of the active molecules
Calculate the physicochemical descriptors of the decoy candidates
Calculate the chemical fingerprints of the active molecules
Calculate the chemical fingerprints of the decoy candidates
For each inhibitor
Calculate D_PQ of each decoy candidate to the inhibitor
Calculate D_T of each decoy candidate to the inhibitor
Put D_PQ and D_T in a [0 1] scale. The distances scaling was performed in such a way that the untransformed min(D_PQ) and the
  untransformed max(D_T) were set to 1
Compute the decoy likeness of each candidate
Select the top N candidates according to their decoys likeness
Cluster the N top candidates and select n representative decoys
Remove the selected n decoys from the pool of candidates
End For
```

Chart 3 Pseudocode of the models ensemble process

```
//
// Requires a set of compounds S predicted by M models
//
// Generate each model individual ranking
//
For each model in M
Find the set of compounds inside its applicability domain S'.
Sort the scores of S' in descending order
If there are samples with the same score in S', then find the unique scoring values and assign identical ranking values to samples having the
  identical score. These ranking values go from 1 to the number of unique scoring values in S'
Calculate the relative rank of each sample in S' as its rank order divided by the number of unique scoring values in S'
End For
//
// Aggregate the individual rankings into one ensemble
//
For each sample in S
Find the set of models the sample is inside their applicability domain M'.
Compute the final aggregated rank as the sum over M' of the sample's relative rank divided by the number of models in M'
End For
```

into training and selection subsets using the three sphere exclusion algorithms 1M, 2M, and 3M. This partition scheme leads to the selection of 24–25 % of the learning data for the selection subset and at the same time it guaranties the representativeness of each classification group in the selection set. The results obtained for such optimal learning dataset partition are summarized in the Supporting Information Table TS2, while the composition of the obtained datasets is provided in the Supporting Information Table TS3.

Classification performance

The three learning dataset partitions obtained with the three sphere exclusion algorithms 1M, 2M, and 3M were modeled using the QSAR modeling framework described in the Computational Methods section. It should be remarked that the external dataset that was randomly selected from the balanced dataset is never used during the training or selection of the optimal models and it is reserved to evaluate the

Table 1 Accuracy estimation parameters for the optimal model derived from each classification approach

Method[a]	Size[b]	Train (%)[c]	Sel. (%)[d]	LOO (%)[e]	Boot (%)[f]	Ext (%)[g]
Sphere exclusion 1M						
GA-AB	8	82.43(88.16/76.39)	69.57(90.48/52.00)	81.08	73.75	75.00(83.33/66.67)
GA-LDA	17	89.19(88.16/90.28)	76.09(76.19/76.00)	86.49	78.54	68.75(58.33/79.17)
GA-LSSVM*	21	97.30(100.00/94.44)	78.26(90.48/68.00)	91.22	84.09	77.08(75.00/79.17)
BT-LSSVM	24	100.00(100.00/100.00)	80.43(95.24/68.00)	76.35	72.37	75.00(70.83/79.17)
BT-AB	13	70.95(71.05/70.83)	69.57(80.95/60.00)	65.54	63.17	77.08(75.00/79.17)
BT-LDA	15	75.00(82.89/66.67)	71.74(80.95/64.00)	65.54	61.85	70.83(70.83/70.83)
R-LSSVM	24	100.00(100.00/100.00)	78.26(90.48/68.00)	77.03	69.31	75.00(66.67/83.33)
R-AB	1	70.95(78.95/62.50)	58.70(71.43/48.00)	70.95	70.95	70.83(75.00/66.67)
R-LDA	7	73.65(86.84/59.72)	58.70(80.95/40.00)	72.97	68.57	68.75(79.17/58.33)
Mean[h]	14.44	84.38	71.26	76.35	71.40	73.15
Sphere exclusion 2M						
GA-AB	5	78.38(86.49/70.27)	78.26(95.65/60.87)	78.38	72.36	72.92(83.33/62.50)
GA-LDA	15	85.14(85.14/85.14)	80.43(91.30/69.57)	83.78	73.59	70.83(66.67/75.00)
GA-LSSVM*	18	96.62(98.65/94.59)	86.96(86.96/86.96)	87.16	80.57	72.92(70.83/75.00)
BT-LSSVM	19	100.00(100.00/100.00)	84.78(91.30/78.26)	73.65	70.01	85.42(87.50/83.33)
BT-AB	9	70.27(78.38/62.16)	80.43(82.61/78.26)	64.19	63.85	70.83(79.17/62.50)
BT-LDA	16	76.35(82.43/70.27)	80.43(86.96/73.91)	70.27	64.63	79.17(79.17/79.17)
R-LSSVM	25	97.97(98.65/97.30)	84.78(95.65/73.91)	70.95	66.42	70.83(66.67/75.00)
R-AB	14	69.59(77.03/62.16)	73.91(95.65/52.17)	68.92	63.72	62.50(58.33/66.67)
R-LDA	18	70.27(78.38/62.16)	78.26(86.96/69.57)	63.51	61.34	64.58(66.67/62.50)
Mean[h]	15.44	82.73	80.92	73.42	68.50	72.22
Sphere exclusion 3M						
GA-AB	11	81.38(84.93/77.78)	75.51(83.33/68.00)	78.62	72.59	79.17(83.33/75.00)
GA-LDA	12	86.21(86.30/86.11)	73.47(79.17/68.00)	84.83	78.45	68.75(62.50/75.00)
GA-LSSVM*	17	86.90(86.30/87.50)	77.55(83.33/72.00)	83.45	76.88	68.75(62.50/75.00)
BT-LSSVM	12	97.93(100.00/95.83)	73.47(83.33/64.00)	79.31	72.37	68.75(75.00/62.50)
BT-AB	11	73.10(86.30/59.72)	65.31(79.17/52.00)	70.34	67.92	75.00(83.33/66.67)
BT-LDA	12	77.24(79.45/75.00)	67.35(79.17/56.00)	71.72	70.79	70.83(70.83/70.83)
R-LSSVM	17	100.00(100.00/100.00)	75.51(83.33/68.00)	78.62	70.66	72.92(70.83/75.00)
R-AB	7	71.72(83.56/59.72)	65.31(83.33/48.00)	70.34	65.99	72.92(79.17/66.67)
R-LDA	16	77.24(80.82/73.61)	71.43(83.33/60.00)	59.31	62.45	75.00(79.17/70.83)
Mean[h]	12.78	83.52	71.66	75.17	70.90	72.45

* The optimal model per experiment better combining robustness and generalization capabilities is highlighted
[a] Modeling approach
[b] Number of feature or single feature LDA models that the model contains
[c] Accuracy in the prediction of the training dataset represented as Accuracy (Sensitivity/Specificity)
[d] Accuracy in the prediction of the selection dataset represented as accuracy (sensitivity/specificity)
[e] LOO cross-validation accuracy
[f] Bootstrap cross-validation accuracy
[g] Accuracy in the prediction of the external dataset represented as accuracy (sensitivity/specificity)
[h] Mean value of the accuracy estimation parameters
GA Genetic Algorithm feature selection, *BT* Bagged Tress feature selection, *R* Ranking feature selection, *LS-SVM* Least Square Support Vector Machine, *AB* Adaboost Ensemble, *LDA* linear discriminant analysis

generalization capabilities of the selected optimal models to correctly asses the real predictive power of the selected optimal models. The accuracy estimation parameters for the optimal model derived from each classification approach are presented in Table 1. The subset of features each model is trained from and the models' parameters are provided in the Electronic Supporting Information Table TS4.

As can be seen from Table 1, it is possible to obtain models with high values of the accuracy estimators for the FabH inhibitors dataset regardless the training/selection partition-

ing schema used. Moreover, all the models show high generalization capabilities as seen from the accuracy in predicting the external dataset. There are not significant differences between the mean accuracy estimators values when the learning data partitioning algorithm changes, however, the data training/selection partition obtained with the 1M algorithm yields slightly robust and generalizable models as shown by the mean LOO, Bootstrap and External Set accuracy values.

Nevertheless, since there are not large differences in the mean accuracy of the prediction of the external test set, the above results suggest stability and statistical equivalence on the quality of the models independently of the learning set partitioning algorithm. It is also interesting to note that the best performing QSAR modeling approach, selected according to the procedure described in [44] (and also in the Supporting Information Sect. S1) for selecting the optimal classification approach, is the GA–LSSVM wrapper for all the three data partitioning algorithms.

Hitherto, all the evidence suggests that most of the 27 classifiers analyzed have a similar performance according to the investigated statistical parameters. To further explore the equivalence of these models regarding the information used for training and their outputs, their distance matrices in the features and models' outputs spaces were examined. The pairwise distances in the descriptors space were calculated as described in our previous publication [44] (see Supporting Information Sect. S2). To compute the distance between the model's outputs, the Hamming distance was employed [48]. A representation of both pairwise distance matrices is provided as Supporting Information in Fig. FS2 and they are computed using the learning (training+selection) dataset and the predictions obtained for the external dataset, respectively. We also examined the frequency each sample in the external dataset is misclassified by all models. The detailed results of these analyses are provided in the Supporting Information Sect. S5.

The information provided by these experiments confirms two main facts: biological activity is a complex process involving many variables; and a computational model cannot capture all the relevant information in a SAR landscape without losing generalization capabilities. Therefore, every classification method can only partly explain the observed structure–activity relationship which leads to obtain statistical equivalent solutions.

Virtual screening performance

As shown above, using nine different classification approaches and three algorithms to split the learning dataset it is possible to obtain accurate, robust, and generalizable QSAR models for the classification of FabH inhibitors. The practical relevance of QSAR models for drug discovery relies on their prospective application as virtual screening tools. Thus, a thorough evaluation of each model obtained is needed before deciding whether they can be used for virtual screening experiments or not.

In a real virtual screening scenario the dataset contains thousands of inactive compounds and only a few actives. To evaluate the performance of the trained models in a scenario closer to a real virtual screening experiment some additional validations are needed. The most important condition that a model has to fulfill to be considered for virtual screening applications is to be able to rank the actives molecules at the beginning of its ranked list. That is, it should correctly address the early recognition problem which means that it has to be able to rank active compounds at the very beginning of its ranked list. Since there are, not only for the problem being investigated but also for every molecular target, only a limited number of confirmed non-active molecules, a plausible solution to realistically estimate the VS performance of any algorithm intended to be used as such is to collect a large enough set of decoy molecules. Decoys are molecules that are likely to be inactive since they are structurally dissimilar from the known inhibitors. Besides, to avoid an easy discrimination by the models based solely on the physicochemical properties, the decoys are selected to have physicochemical properties as similar as possible to those of the known inhibitors. To study the performance of the models in virtual screening experiments, the 24 compounds in the external dataset were selected and decoys were extracted from the ZINC database (more than 18 million of compounds) following the protocol described in the Computational Methods section.

The minimum number of molecules to be used for validating the models for virtual screening was calculated using Eq. 1. It was estimated for the case when the 24 inhibitors are considered ($n = 24$), the 80 % of the total screening score comes from the top 1% of the ranked list ($\alpha = 160.9$) and the maximum relative deviation is set to 5 % ($\Delta_{\max} = 0.05$). Solving Eq. 1 with these parameters values yields that the minimum number of molecules (actives+decoys) to use in the virtual screening validation set is 38616. This means that for each active molecule 1608 decoys should be selected and that the validation dataset will contain 0.06 % of active compounds.

Before selecting the decoy molecules two more issues need to be taken into account. First, when comparing the performance of different classifiers for virtual screening the number of samples considered by each model as well as the actives rate should be approximately the same. Second, only those predictions made for samples inside a model's applicability domain can be considered reliable. To make the performance of each model comparable, the ZINC database was first filtered to keep only those compounds inside each model's applicability domain. Since this filtering step can drastically reduce the number of available decoy molecule candidates, the less strict applicability domain definition

ADD-1 (see the Computational Methods section) was used. From here, on this virtual screening validation set where each decoy is inside every model applicability domain is referenced as Virtual Screening Validation Set 1 (VSVS-1) and the decoy molecules contained on it is referenced as Decoys Set 1.

On the other hand, the assumption that all molecules are inside a model's applicability domain is far from a real life scenario. For this reason a second decoys subset was selected without making any prior consideration about whether they fall within the individual models' applicability domains. The validation set containing the 24 active compounds from the external dataset and these decoys is referenced as Virtual Screening Validation Set 2 (VSVS-2) from here on and the decoys set it contains is referenced as Decoys Set 2. The decoy molecules were selected using the procedure described in the Computational Methods section and the 1608 decoys per active compound were the centroids of the same number of clusters whose members are the 5,000 decoy candidates with the highest "decoy-likeness+." The physicochemical properties of the selected decoys sets are summarized in the Supporting Information Table TS6 and it can be seen that they cover the range of these properties for the active compounds. The SDF files of both decoys sets are provided as Supporting Information.

To further asses the quality of both Virtual Screening Validation Sets we examined the mean physicochemical similarity, the mean structural (topological) dissimilarity and the mean decoy-likeness of each subset of decoys selected for every known inhibitor to all the active compounds. The detailed results of these analyses are provided in the Supporting Information Sect. S6 and based on them, we can be confident that both decoys sets are adequate for evaluating the performance of the classifiers previously developed in the context of virtual screening experiments.

To study how the applicability domain of the models can influence their virtual screening performance, we computed the percent of coverage of these virtual screening validation sets by the applicability domain of each model. The results of these calculations are summarized in the Supporting Information Table TS7. As expected the whole VSVS-1 is within the applicability domain of each of the 27 classification models when it is defined by the ADD-1. On the other hand, when no prior assumption is made for the selection of the decoys set the coverage of the resulting VSVS-2 by each model applicability domain is variable. The data shown in Table TS7 are consistent with the fact that the ADD-2 is stricter than ADD-1. As discussed before, to use the stricter ADD-2 considerably limit the number of molecules that can be used as the source for decoys.

However, to use such flexible applicability domain definition in a real virtual screening experiment can be dangerous since the predictions made for samples close to the hyper-rectangle boundaries could exhibit low reliability. Based on this, the VSVS-1 is used to accomplish the comparison of the different classifiers and the final evaluation of the virtual screening tools is made using the combination of VSVS-2 and ADD-2. The accumulation curves obtained for each model when the VSVS-1 is evaluated are shown in Fig. 2 and the values obtained for ROC, BEDROC, and EF are summarized in Table 2.

Before analyzing the results of the virtual screening simulations, one additional consideration needs to be made. For the Adaboost-based classifiers, it should be taken into account that the output of an Adaboost ensemble is the weighted sum of LDA model's outputs. Hence, they do not produce continue score values but levels of scores defined by the number of LDA models used to train the ensemble. If a subset of top ranked samples coming from a set of samples with equal score values were to be selected, then a random selection process should be followed since such samples are indistinguishable. Following this reasoning, to obtain the rankings for this type of models, the compounds were sorted according to their scores, and all the samples having the same score were randomly distributed inside the unique score bin.

A general observation when the accumulation curves and their virtual screening performance metrics are analyzed is that, despite most of them show ROC values away of that corresponding to a random distribution, only a few models can actually achieve good initial enrichment of active compounds. It is worth noting that the model showing the best classification performance for each learning dataset partition (the GA-LSSVM model in all the three cases, see Table 1) is unable to provide initial enrichments greater than what is expected from a uniform distribution of the active samples in the ranked list. More important, the results obtained confirm what has been proposed before: high values of ROC or predictive QSAR models don't guarantee a good virtual screening performance. When a bigger subset of the ranked list is considered, more models yield enrichments of actives above random but most of them merely outperform the behavior of the uniform distribution. Some other important conclusions can be derived from the analysis of the performance of these models. For example, only 7 out the 27 models (27 %) are able to yield enrichment values superior to what is expected from a uniform distribution of the actives for the top 1 % of the ranked list. If the number of samples to be selected by the virtual screening tool increases, for instance up to 8 % of the screened data, a model such as the BT-LDA model derived when the learning dataset is split with the sphere exclusion algorithm 3M could be considered for virtual screening.

It can also be seen that only one model per data partition: BT-LSSVM for the partition obtained with the sphere exclusion algorithm 1M and R-LSSVM for the data partitions derived with the sphere exclusion algorithms 2M and 3M are able to achieve high initial enrichments even when only the

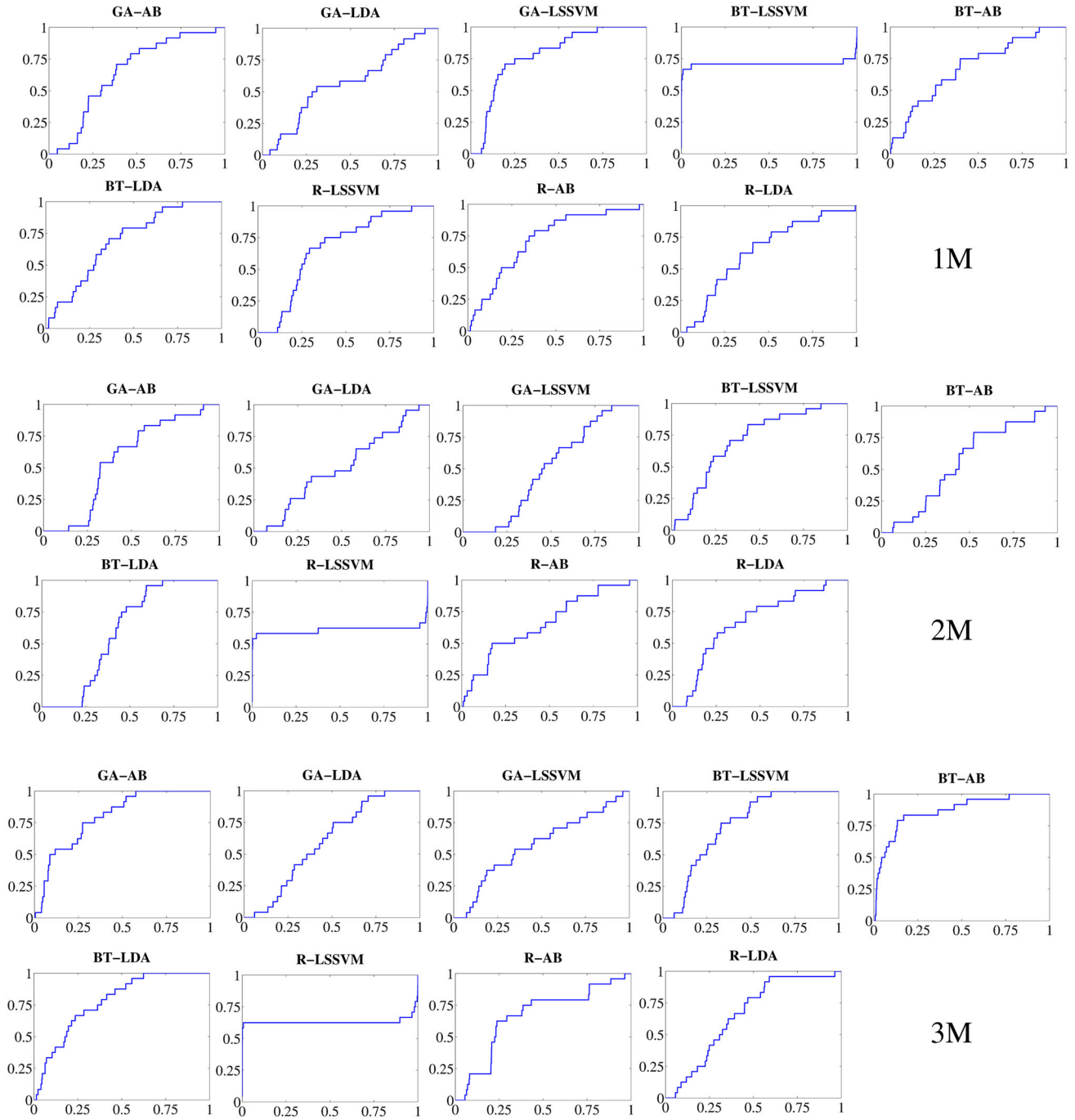

Fig. 2 Accumulation curves (fraction of actives retrieved vs. fraction of data screened) obtained with the VSVS-1 for each of the 27 trained classifiers. The results are presented for the dataset partitions obtained with the three sphere exclusion algorithms 1M, 2M, and 3M. Despite all models being predictive QSAR models, not all of them are good virtual screening tools

top 1 % of the ranked list is considered. One could consider that any of these three models can be used for virtual screening, however, if the values of the coverage of the VSVS-2 by these models' applicability domains are analyzed (see the Supporting Information Table TS7), it can be noticed that it is very poor: only 5.98, 2.92 and 8.22 % of the VSVS-2 is covered by the BT-LSSVM (SE 1M), R-LSSVM (SE 2M) and R-LSSVM (SE 3M), respectively. Since the VSVS-2 is closer to what is expected to be found in a real virtual screening problem, the poor coverage of its members by these models' applicability domains can limit their worth for virtual screening. The effect of the limitations that their applicability domains can impose to a virtual screening campaign are more marked when a stricter definition of the applicability

Table 2 Metrics used to evaluate the virtual screening performance of the trained models

Model	ROC	BEDROC[a]			EF[b]		
		$\alpha = 160.9$	$\alpha = 32$	$\alpha = 20$	1 %	5 %	8 %
Perfect ranking[c]	1.00	1.00	1.00	1.00	100	20	12.5
Random Ranking[c]	0.50	0.00	0.01	0.03	1	1	1
Sphere exclusion 1M							
GA-AB	0.65	0.000	0.011	0.029	0.00	0.83	0.52
GA-LDA	0.57	0.000	0.018	0.042	0.00	0.83	0.52
GA-LSSVM	0.78	0.000	0.031	0.092	0.00	0.00	1.04
BT-LSSVM	0.71	0.609	0.656	0.669	66.56	13.32	8.85
BT-AB	0.68	0.048	0.111	0.147	8.31	2.50	2.08
BT-LDA	0.69	0.006	0.071	0.108	0.00	1.66	2.60
R-LSSVM	0.67	0.000	0.004	0.020	0.00	0.00	0.00
R-AB	0.72	0.007	0.084	0.130	0.00	3.33	2.60
R-LDA	0.64	0.000	0.019	0.045	0.00	0.83	1.04
Sphere exclusion 2M							
GA-AB	0.57	0.000	0.000	0.004	0.00	0.00	0.00
GA-LDA	0.51	0.000	0.005	0.016	0.00	0.00	0.54
GA-LSSVM	0.50	0.000	0.000	0.002	0.00	0.00	0.00
BT-LSSVM	0.71	0.005	0.054	0.087	0.00	1.66	1.04
BT-AB	0.56	0.000	0.010	0.025	0.00	0.00	1.04
BT-LDA	0.60	0.000	0.000	0.002	0.00	0.00	0.00
R-LSSVM	0.61	0.529	0.558	0.565	54.1	11.65	7.29
R-AB	0.66	0.014	0.092	0.137	4.15	2.50	3.12
R-LDA	0.67	0.000	0.009	0.036	0.00	0.00	0.00
Sphere Exclusion 3M							
GA-AB	0.80	0.020	0.106	0.180	4.15	3.33	5.20
GA-LDA	0.60	0.000	0.007	0.020	0.00	0.00	0.52
GA-LSSVM	0.57	0.000	0.009	0.033	0.00	0.00	0.52
BT-LSSVM	0.73	0.000	0.011	0.041	0.00	0.00	0.52
BT-AB	0.86	0.044	0.288	0.382	4.15	9.99	7.29
BT-LDA	0.78	0.005	0.093	0.157	0.00	4.16	4.16
R-LSSVM	0.63	0.595	0.616	0.619	62.40	12.49	7.81
R-AB	0.66	0.000	0.024	0.059	0.00	0.00	2.08
R-LDA	0.66	0.000	0.016	0.041	0.00	0.00	1.04

[a] BEDROC calculated for three different values of α. Each value means that 80 % of the total score comes from 1, 5 and 8 % of the ranked list, respectively

[b] Enrichment Factor at 1, 5 and 8 % of the screened data

[c] Value of each metric for a perfect ranking (all actives at the beginning of the list) and for a random distribution (actives uniformly distributed along the ranked list)

domain is used. Given that a too narrow applicability domain is closely related to highly specialized models, a too limited applicability domain would also have a negative impact on the ability of a screening tool to discover new structural scaffolds.

The finding that all the models previously developed have limited or none value for virtual screening can be disappointing. A possible solution to these limitations is the use of ensemble modeling. It has been previously shown that ensemble modeling is effective in improving the classification performance of the base classifiers as well as their coverage of chemical space [38,39,44]. However, to the best of our knowledge, no work has been devoted to study an analogous positive effect of this approach on virtual screening performance. Theoretically, by using ensemble modeling it would be possible to take advantage of each model's positive characteristics and overcome their individual limitations in virtual screening. To test this hypothesis a few ensembles were

Table 3 Virtual screening performance metrics for the different ensembles built with the VSVS-1

Model	Size[a]	ROC	BEDROC[b]			EF[c]		
			160.9	32	20	1 %	5 %	8 %
Perfect Ranking[d]		1.00	1.00	1.00	1.00	100	20	12.5
Random Ranking[d]		0.50	0.00	0.01	0.03	1	1	1
Ensemble 1	9	0.79	0.234	0.455	0.512	33.28	11.66	8.33
Ensemble 2	9	0.68	0.030	0.164	0.234	4.16	5.83	5.20
Ensemble 3	9	0.79	0.212	0.393	0.450	29.12	11.66	7.29
Ensemble 4	27	0.76	0.275	0.453	0.503	37.44	12.49	7.81
Ensemble 5	7	0.73	0.403	0.516	0.545	49.92	11.66	7.81
Ensemble 6	14	0.78	0.320	0.496	0.540	37.44	12.49	7.81
Ensemble 7	2	0.72	0.108	0.210	0.267	12.48	4.99	4.68
Ensemble 8	2	0.63	0.088	0.204	0.261	16.65	5.83	4.68
Ensemble 9	3	0.81	0.278	0.446	0.486	37.44	10.82	7.29

[a] Number of models the ensemble is composed of
[b] BEDROC calculated for three different values of α. Each used value means that 80 % of the total score comes from 1, 5 and 8 % of the ranked list, respectively
[c] Enrichment Factor at 1, 5 and 8 % of the screened data
[d] Value of each metric for a perfect ranking (all actives at the beginning of the list) and for a random distribution (actives uniformly distributed along the ranked list)

trained for the VSVS-1 using the procedure described in the Computational Methods section. These ensembles differed in the way their members were selected and they are defined as provided in the Supporting Information Table TS8. Their performance statistics are shown in Table 3 and the obtained accumulation curves are plotted in the Supporting Information Fig. FS5. The detailed list of the members of each ensemble is provided in the Electronic Supporting Information in Table TS9.

The results provided in Table 3 and in Fig. FS5 clearly show that the three ensembles built with models from the three different learning data partition algorithms (Ensembles 4, 5 and 6) outperform those trained with models coming from only one of the three sphere exclusion algorithms. The importance of excluding the models with a low performance on the very first fraction of the ranked list can be illustrated by the comparison of Ensembles 4 (formed by all the 27 classifiers) and Ensemble 9 (three models from the 3M data partition). The first of these ensembles contains nine times more models than the second one, and they have about the same performance as can be seen from Table 3 and Supporting Information Fig. FS5.

It is true that the ensembles built with models coming from any data partition algorithm and filtered to discard those with a low performance at the very beginning of their ranked lists outperform the ensembles trained with all the models. However, this observation cannot be extrapolated to the ensembles trained with models coming from only one particular learning dataset partition. These observations are not surprising and they find a simple explanation in two main facts. First, it is well known that one of the key characteristics of a good ensemble is the diversity of its members and as was previously shown there is equivalent inter- and intra-dataset partition models' diversity in both the input and output spaces. Second, diversity is not enough to have a good ensemble, it is also important to have a number of models that can cover as much as possible the input space. The diversity of the models in the cases above discussed is guaranteed as was previously discussed. A comprehensive comparison of these ensembles is available in the Supporting Information Sect. S7.

From all these ensemble comparisons, it can be concluded that to select the models from any data partition that are able to retrieve at least a compound at the top 1 % of the ranked list and to combine them in an ensemble is an effective strategy to train a virtual screening tool. We should highlight the effectiveness and simplicity of this proposed heuristic to effectively select the members of the ensemble. Despite this being true, it should be kept in mind that the VSVS-1, which decoys are inside each model's applicability domain, is not representative of a real virtual screening scenario. For this reason a second decoy molecules set was selected without making any prior assumption about the model's applicability domain and it was combined with the 24 active compounds of the external dataset to form the Virtual Screening Validation Set 2 (VSVS-2). As shown in the Supporting Information Table TS7 the coverage of this second decoys set by the models' applicability domains changes from one model to other. As previously discussed, since the models' applicability domains are different the metrics used to evaluate each

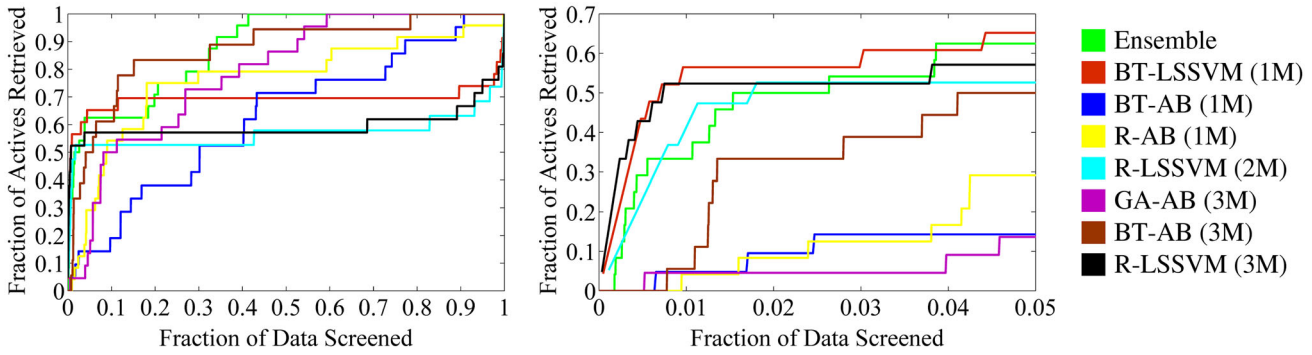

Fig. 3 Accumulation curves (fraction of actives retrieved vs. fraction of data screened) obtained for the ensemble trained with the VSVS-2 and their members. The curves are plotted for whole screened data (left) and for the top 5 % of the ranked lists (right)

model quality for virtual screening are not comparable. The performance of each individual model when they are used to predict the VSVS-2 was investigated and the obtained accumulative curves are presented in the Supporting Information Fig. FS6.The models that are able to identify at least one active compound on the first 1 % of their ranked lists are shown in red since they will be used to train an ensemble following the same procedure as before for Ensemble 5.

According to the results obtained for the ensemble modeling using the VSVS-1, the models that are able to retrieve at least one active compound on the top 1 % of their respective ranked list were selected to build the final ensemble to be employed for the virtual screening of databases of chemical compounds for the identification of potential FabH inhibitors. These models resulted to be: BT-LSSVM (1M), BT-AB (1M), R-AB (1M), R-LSSVM (2M), GA-AB (3M), BT-AB (3M), and R-LSSVM (3M). It is interesting to note that each of the three feature selection techniques (Genetic Algorithms, Bagged Trees and Features Ranking) as well as two classification algorithms (Adaboost and Least-Square Support Vector Machines) are represented in this set of seven models the ensemble will be built from. According to the Supporting Information Table TS7, all these models have an uneven coverage of the VSVS-2 by their applicability domains ranging from 2.92 to 96.84 %. In contrast, the ensemble formed by these models covers 98.97 % of the VSVS-2. It can be thought that most of this ensemble applicability domain coverage is provided by the R-AB model derived from the 1M dataset partition that covers by itself 96.84 % of the VSVS-2; however, only 28.29 % of the VSVS-2 is exclusively covered by this model applicability domain. This means that the remaining models together cover 70.58 % of the VSVS-2 which represents a 14.84 % increment relative to the next model with the highest coverage of the VSVS-2 (BT-AB (3M)). The accumulative curves of the ensemble formed by these models as well as that of its members are plotted in Fig. 3. This ensemble will be referred as VSVS-2 Ensemble from here on.

It is worth noting that when the results obtained with both the VSVS-1 and the VSVS-2 are analyzed, seven models are found in each case that are able to retrieve at least one active compound on the top 1 % of their respective ranked lists. Even more interesting, six of those models are the same in both experiments: BT-LSSVM (1M), BT-AB (1M), R-LSSVM (2M), GA-AB (3M), BT-AB (3M), and R-LSSVM (3M). This last observation highlights the robustness of the developed methodology. It might seem from Fig. 3 that the performance of the trained ensemble is not as good as that of the BT-LSSVM (1M), R-LSSVM (2M) and R-LSSVM (3M) models, nevertheless as previously discussed the applicability domain of these models only cover 5.98, 2.92 and 8.22 % of the VSVS-2, respectively. In other words, these three models can be regarded as very specialized ones, meaning that working by themselves they have very strict limitations on the data they can predict, but if they are used together as members of an ensemble they can provide it with very useful information. The virtual screening performance estimators for this ensemble (VSVS-2 Ensemble) are summarized in Table 4.

For the modeling experiment using the VSVS-2 that is closer to a real virtual screening scenario, the performance statistics also show that the trained ensemble (VSVS-2 Ensemble) is able to perform much better than what is expected from a uniform distribution of the active compounds in the ranked list. Despite being lower than for Ensemble 5, the performance metrics of the VSVS-2 Ensemble show that it is able to retrieve 33 % of the active compounds in the first 1 % of the ranked list and up to 60 % of them in the first 5 %. This means that the top 1 % of the ranked list contains 33.29 times more actives than expected from a uniform distribution of the inhibitors in the whole dataset. The fact that no prior consideration about the models' applicability domains is made before screening the database and it is only considered to select which ensemble members will contribute to the prediction of each sample, adds an extra value to the proposed virtual screening setup. Granted that for real virtual screen-

Table 4 Virtual screening performance estimators for the ensemble built using the VSVS-2

Model	Size[a]	ROC	BEDROC[b]			EF[c]		
			160.9	32	20	1 %	5 %	8 %
Perfect Ranking[d]		1	1	1	1	100	20	12.5
Random Ranking[d]		0.5	0	0.01	0.03	1	1	1
VSVS-2 Ensemble	7	0.88	0.234	0.457	0.510	33.29	12.49	7.81

[a] Number of models the ensemble is composed of
[b] BEDROC calculated for three different values of α. Each used value means that 80 % of the total score comes from 1, 5 and 8 % of the ranked list, respectively
[c] Enrichment Factor at 1, 5 and 8 % of the screened data
[d] Value of each metric for a perfect ranking (all actives at the beginning of the list) and for a random distribution (actives uniformly distributed along the ranked list)

ing campaigns the fraction of the database to be selected for further analysis depends on the number of screened compounds and the available experimental resources and that the VSVS-2 Ensemble achieves a good performance at the analyzed dataset fractions, this ensemble can be considered the best performing tool for the ligand-based virtual screening of databases to search for potential novel FabH inhibitors.

Concluding remarks

Here, we presented the results of the training and validation of binary classification models aimed to identify potential FabH inhibitors. QSAR models were trained using the combination of different feature selection and classification techniques as implemented in our in-house QSAR modeling framework and different partitions of the learning dataset into training and selection subsets. One optimal model per classification experiment was selected based on a consensus ranking approach and these optimal models showed high accuracy, robustness, and generalization capabilities. Besides, the set of optimal classifiers was shown to be diverse in both the information they are trained from and their outputs.

During the modeling process, we paid attention to factors known to negatively affect the performance of QSAR models. By considering these factors in our modeling experiments, we were able to obtain predictive QSAR models. However, when these models were validated in the context of a real virtual screening scenario they had very limited or none applicability for real virtual screening campaigns.

It was shown, for the first time, that using ensemble modeling can overcome the virtual screening limitations associated to the application of single models, complementing the well-known capabilities of ensemble modeling to improve classification performance as well as chemical space coverage. The final virtual screening tool, the VSVS-2 Ensemble, was shown to combine the positive features of the individual classifiers and to provide high initial enrichment of active compounds at the very first part of the ranked list. The results obtained herein lead us to our main "cheminformatically relevant" finding: a predictive QSAR model is not necessarily a good virtual screening tool. For this reason, we propose that the retrospective validation of the QSAR-derived virtual screening tools using problem-specific decoy molecules should become a regular practice in ligand-based drug discovery.

We are currently combining the VSVS-2 Ensemble and our previously reported structure-based modeling workflow of FabH inhibitors [27] in the virtual screening of commercially available collections of chemical compounds.

Acknowledgments Pérez-Castillo Y. thanks the Flemish Interuniversity Council (VLIR) for financial support through the project: "Strengthening research and PhD formation in Computer Sciences and its applications" in the framework of the VLIR-UCLV collaborative program. Cabrera-Pérez M. A. thanks to theprojects 1- A1/036687/11: Montaje de un laboratorio de Química Computacional, con fines académicos y científicos, para el diseño de potenciales candidatos a fármacos, en enfermedades de alto impacto social and 2- DCI-ALA/19.09.01/10/21526/245-297/ALFA 111(2010)29: Red-Biofarma. Red para el desarrollo de metodologías biofarmacéuticas racionales que incrementen la competencia y el impacto social de las Industrias Farmacéuticas Locales. Cruz-Monteagudo M. acknowledges FCT for the grant [SFRH/BPD/90673/2012] co-financed by the European Social Fund.

References

1. Monaghan RL, Barrett JF (2006) Antibacterial drug discovery-then, now and the genomics future. Biochem Pharmacol 71:901–909. doi:10.1016/j.bcp.2005.11.023
2. Nikaido H (2009) Multidrug resistance in bacteria. Annu Rev Biochem 78:119–146. doi:10.1146/annurev.biochem.78.082907.145923
3. Fischbach MA, Walsh CT (2009) Antibiotics for emerging pathogens. Science 325:1089–1093. doi:10.1126/science.1176667
4. Moir DT, Opperman TJ, Butler MM, Bowlin TL (2012) New classes of antibiotics. Curr Opin Pharmacol 12:535–544. doi:10.1016/j.coph.2012.07.004
5. Yoneyama H, Katsumata R (2006) Antibiotic resistance in bacteria and its future for novel antibiotic development. Biosci Biotechnol Biochem 70:1060–1075. doi:10.1271/bbb.70.1060

6. Heath RJ, Rock CO (2004) Fatty acid biosynthesis as a target for novel antibacterials. Curr Opin Investig Drugs 5:146–153
7. Jayakumar A, Tai MH, Huang WY, al-Feel W, Hsu M, Abu-Elheiga L, Chirala SS, Wakil SJ (1995) Human fatty acid synthase: properties and molecular cloning. Proc Natl Acad Sci USA 92:8695–8699. doi:10.1073/pnas.92.19.8695
8. Heath RJ, Rock CO (1996) Regulation of fatty acid elongation and initiation by acyl–acyl carrier protein in *Escherichia coli*. J Biol Chem 271:1833–1836. doi:10.1074/jbc.271.4.1833
9. Perez-Castillo Y, Perez MA (2008) Bacterial beta-ketoacyl-acyl carrier protein synthase III (FabH): an attractive target for the design of new broad-spectrum antimicrobial agents. Mini Rev Med Chem 8:36–45. doi:10.2174/138955708783331559
10. Daines RA, Pendrak I, Sham K, Van Aller GS, Konstantinidis AK, Lonsdale JT, Janson CA, Qiu X, Brandt M, Khandekar SS, Silverman C, Head MS (2003) First X-ray cocrystal structure of a bacterial FabH condensing enzyme and a small molecule inhibitor achieved using rational design and homology modeling. J Med Chem 46:5–8. doi:10.1021/jm025571b
11. He X, Reeve AM, Desai UR, Kellogg GE, Reynolds KA (2004) 1,2-Dithiole-3-ones as potent inhibitors of the bacterial 3-ketoacyl acyl carrier protein synthase III (FabH). Antimicrob Agents Chemother 48:3093–3102. doi:10.1128/AAC.48.8.3093-3102.2004
12. Alhamadsheh MM, Waters NC, Sachdeva S, Lee P, Reynolds KA (2008) Synthesis and biological evaluation of novel sulfonylnaphthalene-1,4-diols as FabH inhibitors. Bioorg Med Chem Lett 18:6402–6405. doi:10.1016/j.bmcl.2008.10.097
13. Nie Z, Perretta C, Lu J, Su Y, Margosiak S, Gajiwala KS, Cortez J, Nikulin V, Yager KM, Appelt K, Chu S (2005) Structure-based design, synthesis, and study of potent inhibitors of beta-ketoacyl-acyl carrier protein synthase III as potential antimicrobial agents. J Med Chem 48:1596–1609. doi:10.1021/jm049141s
14. Li HQ, Shi L, Li QS, Liu PG, Luo Y, Zhao J, Zhu HL (2009) Synthesis of C(7) modified chrysin derivatives designing to inhibit beta-ketoacyl-acyl carrier protein synthase III (FabH) as antibiotics. Bioorg Med Chem 17:6264–6269. doi:10.1016/j.bmc.2009.07.046
15. Lv PC, Wang KR, Yang Y, Mao WJ, Chen J, Xiong J, Zhu HL (2009) Design, synthesis and biological evaluation of novel thiazole derivatives as potent FabH inhibitors. Bioorg Med Chem Lett 19:6750–6754. doi:10.1016/j.bmcl.2009.09.111
16. Li HQ, Luo Y, Lv PC, Shi L, Liu CH, Zhu HL (2010) Design and synthesis of novel deoxybenzoin derivatives as FabH inhibitors and anti-inflammatory agents. Bioorg Med Chem Lett 20:2025–2028. doi:10.1016/j.bmcl.2010.01.032
17. Lv PC, Sun J, Luo Y, Yang Y, Zhu HL (2010) Design, synthesis, and structure–activity relationships of pyrazole derivatives as potential FabH inhibitors. Bioorg Med Chem Lett 20:4657–4660. doi:10.1016/j.bmcl.2010.05.105
18. Shi L, Fang RQ, Zhu ZW, Yang Y, Cheng K, Zhong WQ, Zhu HL (2010) Design and synthesis of potent inhibitors of beta-ketoacyl-acyl carrier protein synthase III (FabH) as potential antibacterial agents. Eur J Med Chem 45:4358–4364. doi:10.1016/j.ejmech.2010.05.033
19. Li HQ, Luo Y, Zhu HL (2011) Discovery of vinylogous carbamates as a novel class of beta-ketoacyl-acyl carrier protein synthase III (FabH) inhibitors. Bioorg Med Chem 19:4454–4459. doi:10.1016/j.bmc.2011.06.048
20. Li ZL, Li QS, Zhang HJ, Hu Y, Zhu DD, Zhu HL (2011) Design, synthesis and biological evaluation of urea derivatives from *o*-hydroxybenzylamines and phenylisocyanate as potential FabH inhibitors. Bioorg Med Chem 19:4413–4420. doi:10.1016/j.bmc.2011.06.049
21. Zhang HJ, Qin X, Liu K, Zhu DD, Wang XM, Zhu HL (2011) Synthesis, antibacterial activities and molecular docking studies of Schiff bases derived from *N*-(2/4-benzaldehyde-amino) phenyl-*N*'-phenyl-thiourea. Bioorg Med Chem 19:5708–5715. doi:10.1016/j.bmc.2011.06.077
22. Zhang HJ, Zhu DD, Li ZL, Sun J, Zhu HL (2011) Synthesis, molecular modeling and biological evaluation of beta-ketoacyl-acyl carrier protein synthase III (FabH) as novel antibacterial agents. Bioorg Med Chem 19:4513–4519. doi:10.1016/j.bmc.2011.06.021
23. Li Y, Luo Y, Hu Y, Zhu DD, Zhang S, Liu ZJ, Gong HB, Zhu HL (2012) Design, synthesis and antimicrobial activities of nitroimidazole derivatives containing 1,3,4-oxadiazole scaffold as FabH inhibitors. Bioorg Med Chem 20:4316–4322. doi:10.1016/j.bmc.2012.05.050
24. Yang YS, Zhang F, Gao C, Zhang YB, Wang XL, Tang JF, Sun J, Gong HB, Zhu HL (2012) Discovery and modification of sulfur-containing heterocyclic pyrazoline derivatives as potential novel class of beta-ketoacyl-acyl carrier protein synthase III (FabH) inhibitors. Bioorg Med Chem Lett 22:4619–4624. doi:10.1016/j.bmcl.2012.05.091
25. Ashek A, San Juan AA (2007) HQSAR study of beta-ketoacyl-acyl carrier protein synthase III (FabH) inhibitors. J Enzyme Inhib Med Chem 22:7–14. doi:10.1080/14756360600920149
26. Singh S, Soni LK, Gupta MK, Prabhakar YS, Kaskhedikar SG (2008) QSAR studies on benzoylaminobenzoic acid derivatives as inhibitors of beta-ketoacyl-acyl carrier protein synthase III. Eur J Med Chem 43:1071–1080. doi:10.1016/j.ejmech.2007.06.018
27. Perez-Castillo Y, Froeyen M, Cabrera-Perez MA, Nowe A (2011) Molecular dynamics and docking simulations as a proof of high flexibility in *E. coli* FabH and its relevance for accurate inhibitor modeling. J Comput Aided Mol Des 25:371–393. doi:10.1007/s10822-011-9427-z
28. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, Clark J, Connelly MC, Sigal M, Hodges D, Guiguemde A, Guy RK, Tropsha A (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model 53:475–492. doi:10.1021/ci300421n
29. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50:1189–1204. doi:10.1021/ci100176x
30. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29:476–488. doi:10.1002/minf.201000061
31. Potter T, Matter H (1998) Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. J Med Chem 41:478–488. doi:10.1021/jm9700878
32. Golbraikh A, Tropsha A (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers 5:231–243. doi:10.1023/A:1021372108686
33. Hou X, Yan A (2013) Classification of *Plasmodium falciparum* glucose-6-phosphate dehydrogenase inhibitors by support vector machine. Mol Divers 17:489–497. doi:10.1007/s11030-013-9447-9
34. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508. doi:10.1021/ci600426e
35. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martinez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model. doi:10.1021/ci200528d
36. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–What can we learn from earlier mistakes? J Comput Aided Mol Des 22:213–228. doi:10.1007/s10822-007-9163-6
37. Bender A, Bojanic D, Davies JW, Crisman TJ, Mikhailov D, Scheiber J, Jenkins JL, Deng Z, Hill WA, Popov M, Jacoby E,

Glick M (2008) Which aspects of HTS are empirically correlated with downstream success? Curr Opin Drug Discov Devel 11:327–337
38. Helguera AM, Pérez-Garrido A, Gaspar A, Reis J, Cagide F, Vina D, Cordeiro MNDS, Borges F (2013) Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. Eur J Med Chem 59:75–90. doi:10.1016/j.ejmech.2012.10.035
39. Fernandez M, Caballero J, Fernandez L, Sarai A (2011) Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). Mol Divers 15:269–289. doi:10.1007/s11030-010-9234-9
40. ChemAxon (2012) J Chem. 5.9.0 edn.
41. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov, ev V, Hoonakker F, Tetko IV, Marcou G (2008) ISIDA—platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 4:191–198. doi:10.2174/157340908785747465
42. Talete (2010) DRAGON (Software for Molecular Descriptor Calculation). 6.0 edn.
43. MATLAB (2009). R2009a edn. The MathWorks Inc.
44. Perez-Castillo Y, Lazar C, Taminau J, Froeyen M, Cabrera-Perez MA, Nowe A (2012) GA(M)E-QSAR: a novel, fully automatic genetic-algorithm-(meta)-ensembles approach for binary classification in ligand-based drug design. J Chem Inf Model 52:2366–2386. doi:10.1021/ci300146h
45. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim 33:445–459
46. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768. doi:10.1021/ci3001277
47. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801. doi:10.1021/jm0608356
48. Hamming R (1950) Error detecting and error correcting codes. Bell Syst Tech J 26:147–160. doi:10.1002/j.1538-7305.1950.tb00463.x