

# Differentiation of AmpC beta-lactamase binders vs. decoys using classification *k*NN QSAR modeling and application of the QSAR classifier to virtual screening

Jui-Hua Hsieh · Xiang S. Wang · Denise Teotico ·  
Alexander Golbraikh · Alexander Tropsha

Received: 31 October 2007 / Accepted: 18 February 2008 / Published online: 13 March 2008  
© Springer Science+Business Media B.V. 2008

**Abstract** The use of inaccurate scoring functions in docking algorithms may result in the selection of compounds with high predicted binding affinity that nevertheless are known experimentally not to bind to the target receptor. Such falsely predicted binders have been termed ‘binding decoys’. We posed a question as to whether true binders and decoys could be distinguished based only on their structural chemical descriptors using approaches commonly used in ligand based drug design. We have applied the *k*-Nearest Neighbor (*k*NN) classification QSAR approach to a dataset of compounds characterized as binders or binding decoys of AmpC beta-lactamase. Models were subjected to rigorous internal and external validation as part of our standard workflow and a special QSAR modeling scheme was employed that took into account the imbalanced ratio of inhibitors to non-binders (1:4) in this dataset. 342 predictive models were obtained with correct classification rate (CCR) for both training and test sets as high as 0.90 or higher. The prediction accuracy was as high as 100% (CCR = 1.00) for

the external validation set composed of 10 compounds (5 true binders and 5 decoys) selected randomly from the original dataset. For an additional external set of 50 known non-binders, we have achieved the CCR of 0.87 using very conservative model applicability domain threshold. The validated binary *k*NN QSAR models were further employed for mining the NCGC AmpC screening dataset (69653 compounds). The consensus prediction of 64 compounds identified as screening hits in the AmpC PubChem assay disagreed with their annotation in PubChem but was in agreement with the results of secondary assays. At the same time, 15 compounds were identified as potential binders contrary to their annotation in PubChem. Five of them were tested experimentally and showed inhibitory activities in millimolar range with the highest binding constant  $K_i$  of 135  $\mu$ M. Our studies suggest that validated QSAR models could complement structure based docking and scoring approaches in identifying promising hits by virtual screening of molecular libraries.

**Keywords** Binding decoys · AmpC beta-lactamase · *k*NN classification · PubChem · Virtual screening

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-008-9199-2) contains supplementary material, which is available to authorized users.

J.-H. Hsieh · X. S. Wang · A. Golbraikh · A. Tropsha (✉)  
Laboratory for Molecular Modeling, Division of Medicinal  
Chemistry and Natural Products and Carolina Exploratory  
Center for Cheminformatics Research, School of Pharmacy,  
University of North Carolina at Chapel Hill, CB #7360,  
Beard Hall, Chapel Hill, NC 27599-7360, USA  
e-mail: alex\_tropsha@unc.edu

D. Teotico  
Department of Pharmaceutical Chemistry and Graduate Group  
in Chemistry and Chemical Biology, University of California  
San Francisco, 1700 4th Street, San Francisco, CA 94158-2330,  
USA

## Introduction

Due to rapid advances in protein crystallography [1, 2], the number of x-ray characterized biological targets and their complexes with various low molecular weight ligands in the RCSB Protein Data Bank (PDB) [3] has been growing rapidly. This growth has been concurrent with the development of a vast array of structure-based virtual screening approaches [4–13]. These methods include two critical components, i.e., docking and scoring. It has been shown that multiple binding poses of putative receptor ligands

resulting from docking include those that are geometrically close to the native (i.e., experimental) ligand orientation in the binding site. However, identifying (the most) native-like binding poses among many alternatives (i.e., ‘geometrical decoys’) resulting from docking continues to present a universal problem to most scoring functions [14–16]. Furthering this problem is a demonstrated inability of many scoring functions to discriminate between ligands that are known to bind to the target receptor from those known to be non-binders yet predicted to bind by a docking/scoring method (so called ‘binding decoys’) [17, 18].

Many strategies have been proposed to improve scoring functions such as to decrease the number of false positives as well as improve the enrichment of true positives [18–23]. Nevertheless, in a recent study, Shoichet and coworkers [15, 24] reported docking of over 200,000 compounds into the active site of AmpC beta-lactamase that produced many binding decoys. These compounds were ranked highly by many scoring functions such as DOCK, ScreenScore and FlexX etc., but were found to be false positives as a result of experimental validation. Similar results have been observed for several other systems (available from the B. Shoichet’s laboratory website, <http://shoichetlab.compbio.ucsf.edu/take-away.php>).

From the traditional three-dimensional docking and scoring prospective, the existence of binding decoys illustrates the need for developing more robust scoring functions. However, the same results could be also approached from a cheminformatics prospective. Thus, the two groups, i.e., experimentally confirmed binders and binding decoys represent two classes of compounds that could be possibly discriminated by their chemical features, or descriptors. The problems of this type (i.e., discriminating binding from non-binding compounds based on their chemical structure descriptors only) are rather common in case of ligand based drug design approaches such as Quantitative Structure Activity Relationship (QSAR) modeling. In fact, the use of binary QSAR modeling towards the problem of discriminating true binders vs. decoys may be perhaps even more challenging than the standard binary QSAR modeling. Indeed, in this case both classes of compounds are apparently sufficiently similar to each other to fit into the same receptor binding site to be indistinguishable by well-defined and validated scoring functions. Thus, being able to discriminate binders vs. (similar) decoys should be a difficult exercise. On the other hand the successful structural models could potentially inform protein structure based scoring functions about specific functional groups that are primarily responsible for the discriminatory power of the QSAR models but most likely are not adequately scored by the traditional scoring functions. Furthermore structural models could be potentially used for mining external compound libraries to

identify novel putative binders providing a potential alternative to structure based virtual screening methods [25, 26].

The goal of this study was to develop robust binary classification QSAR models that would have high predictive power to differentiating binders vs. non-binding ‘decoys’ for AmpC beta-lactamase. We have employed a rigorous validated QSAR modeling workflow that has been developed in our laboratory in recent years. This workflow that incorporates a virtual screening module was applied successfully to several ligand datasets leading to the identification of experimentally confirmed novel hits for different biological targets [27–31] (see recent review [25]). Herein, we report on classification QSAR models that are capable of discriminating binders from decoys with the external classification accuracy exceeding 90%. Furthermore, we have used these models to screen the compound library tested earlier in the AmpC assay and available from PubChem [32]. We have identified 15 molecules as putative AmpC ligands and demonstrated in subsequent experimental studies that five compounds chosen from these hits were millimolar binders. It worth emphasizing that in all studies reported in this paper we did not use any information on the crystallographic structure of AmpC-ligand complexes and moreover, chemical descriptors were generated from two-dimensional rendering of molecular structures.

## Methods

### Datasets

#### *Compounds used for QSAR model building*

The AmpC beta-lactamase inhibitors and binding decoys were downloaded from Dr. Brian Shoichet’s laboratory web site [33]. This dataset contains 21 confirmed inhibitors (cf. Table 1 of Supporting Information) and 80 decoys. The inhibitors were shown to be non-covalent, reversible AmpC beta-lactamase inhibitors [15, 24, 34]. All decoys were shown to have no binding to AmpC at 1 mM concentration but falsely predicted to bind by multiple scoring functions [24, 35].

#### *Library for virtual screening*

We used the dataset of 69653 compounds that was screened in the HTS assays for AmpC beta-lactamase inhibition by the National Center for Chemical Genomics (NCGC). The screening results are reported in PubChem as Bioassays AID584 [36] and AID585 [37]. The experimental protocols are described in [38] as well as in the

PubChem database. AID584 and AID585 were designed for screening of specific and promiscuous AmpC beta-lactamase inhibitors, respectively. Compounds are classified as having full titration curves, partial modulation, partial curve (weaker actives), single point activity (at highest concentration only), or inactive. Compounds that showed activity in both AID584 and AID585 assays were considered ‘true’ positives. However, if compounds were only found active in AID585 but inactive in AID584, they were categorized as ‘aggregators’. Thus, 64 compounds were identified as ‘true inhibitors of’ the AmpC beta-lactamase that could be used to test the ability of QSAR model based virtual screening to recover known hits.

#### *AmpC beta-lactamase competitive inhibitor assay*

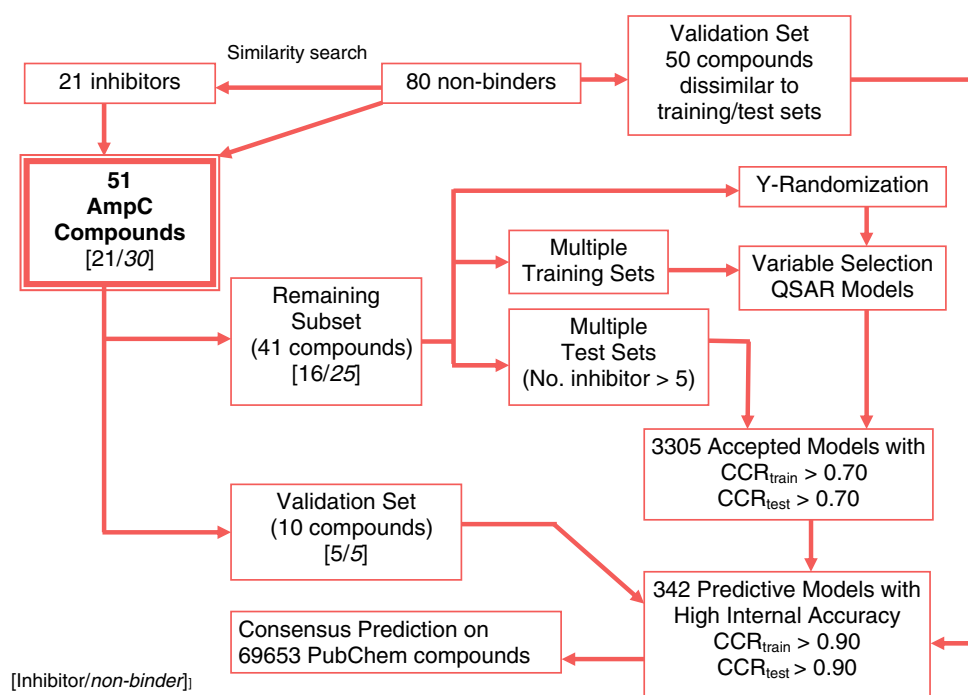
The details of enzymatic assays to measure the efficiency of AmpC beta-Lactamase inhibitors were described in detail elsewhere [26, 39]. Briefly, the change in initial rate of substrate hydrolysis at increasing concentrations of the inhibitor was monitored and the  $IC_{50}$  was obtained using the resulting dose-response curve. The inhibition constant,  $K_i$ , was derived from the  $IC_{50}$  value using the Cheng-Prusoff equation.

#### Training, test, and external validation set selection

We have followed the rigorous QSAR workflow for model building, validation and database mining (Fig. 1)

established in our laboratory (see [25] for recent overview). For classification QSAR modeling, it would be ideal to have the balanced ratio between different compound classes in the modeling dataset. However, the AmpC beta-lactamase binding dataset included 21 inhibitors and 80 decoys, i.e., it is imbalanced with the inhibitors to non-binders ratio of 1:4. In the absence of special statistical treatment, such ratio would skew the prediction accuracy of the classification models. Thus, the distance matrix was calculated in the multidimensional descriptor space for all 101 compounds and similarity search was carried out using 21 inhibitors as queries against the remaining 80 non-binders. 30 compounds were selected from the original 80 non-binders as most similar to 21 inhibitors using Euclidean distance as similarity metric (we note that this treatment makes the task of building the discriminatory binary QSAR models even more challenging. Consequently, these 30 non-binders combined with 21 true inhibitors formed a new balanced dataset for QSAR model building. The remaining 50 “dissimilar” non-binders were retained as an external validation set. Furthermore, 10 compounds (five binders and five decoys) were randomly excluded from the balanced dataset of 51 compounds and formed a second external validation set. The remaining 41 compounds were considered a modeling dataset that was divided into multiple diverse and representative training and test sets using the Sphere Exclusion approach developed in our laboratory earlier [39, 40].

**Fig. 1** The workflow of QSAR model building, validation and virtual screening as applied to the AmpC beta-lactamase dataset of 21 inhibitors and 80 non-binding decoys



## Generation of 2D molecular descriptors

The SMILES [41] strings of each compound in AmpC beta-lactamase dataset were converted to 2D chemical structures using the Unity module of the SYBYL software package [42]. The MolConnZ [43] software (version 4.09) was used to calculate a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices [44–46], kappa molecular shape indices [47, 48], topological and electrotopological state indices [49–51], differential connectivity indices, graph's radius and diameter [52], Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, counts of paths and edges between different kinds of vertices.

Overall, MolConnZ produced over 770 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, only 644 chemically relevant descriptors were initially calculated and 340 descriptors were eventually used for AmpC beta-lactamase binding dataset after deleting descriptors with zero value or zero variance. MolConnZ descriptors were range-scaled prior to distance calculations since the absolute scales for MolConnZ descriptors can differ by orders of magnitude [53]. Accordingly, our use of range-scaling avoided giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional MolConnZ descriptor space.

## *k* Nearest neighbors (*k*NN) classification method

The *k*NN classification QSAR method [53, 54] is based on the idea that the class that a compound belongs to can be defined by the class membership of its nearest neighbors (i.e., most similar compounds) taking into account weighted similarities between a compound and its nearest neighbors. Since our implementation of *k*NN approach includes variable selection, the similarity is evaluated using only a subset of all descriptors (*nvar*). The similarity is characterized by weighted Euclidean distance between compounds in multidimensional descriptor space. Thus, the class membership of compound *i* can be predicted from the following equation:

$$\hat{y}_i = \sum_{j=1}^k \left[ \frac{\exp(-d_{ij})}{\sum_{j'=1}^k \exp(-d_{ij'})} y_j \right] \quad (1)$$

where *k* is the number of nearest neighbors (*k* = 1–5) of compound *i*, *y<sub>j</sub>* is the class membership of compound *j* and

*d<sub>ij</sub>* is the Euclidean distance between compound *i* and its *j*th nearest neighbors. In practice, the value of  $\hat{y}_i$  is rounded to determine the class membership of compound *i*:

$$\hat{y}'_i = \text{round}(\hat{y}_i) \quad (2)$$

The model is internally validated by leave-one-out cross-validation (LOO-CV) where each compound is eliminated from the training set and its class membership is predicted as the class the majority of its *k* nearest neighbors belongs to. The descriptor set is optimized by simulated annealing approach with the Metropolis-like acceptance criterion to achieve the best CCR value. The CCR is defined as [28]:

$$\text{CCR} = 0.5(\text{TP}/N_1 + \text{TN}/N_0) \quad (3)$$

where *N<sub>1</sub>* and *N<sub>0</sub>* are the number of inhibitors and non-binders in the dataset, TP and TN are the number of known inhibitors predicted as inhibitors (true positives) and the number of non-binders predicted as non-binders (true negatives). The statistical significance of the training and test set models is characterized by the LOO-CV CCR<sub>train</sub> and predictive CCR<sub>test</sub>, respectively. In summary, the variable selection *k*NN classification method generates a model with the highest value of CCR that is characterized by the optimal *k* value, the number of nearest neighbors, and a subset of selected descriptors. Additional details of this approach can be found elsewhere [53, 55].

## Applicability domain of *k*NN models

When developing *k*NN QSAR models, each compound is represented as a point in *M*-dimensional descriptor space (where *M* is the total number of selected descriptors); thus, the molecular similarity between any two molecules can be characterized by the Euclidean distance between their representative points. The Euclidean distance *d<sub>ij</sub>* between two points *i* and *j* (which correspond to compounds *i* and *j*) in *M*-dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (4)$$

Compounds with the smallest distance between one another are considered to have the highest similarity.

Theoretically, for any compound that can be represented by its MZ descriptors one should be able to predict its class membership using classification *k*NN approach. However, if the distance between the query compound and its *k* nearest neighbors in the training set is large, then the query compound is too dissimilar to the training set compounds, and the prediction of its activity using *k*NN approach appears meaningless. Therefore, a similarity threshold (or model applicability domain) should be introduced to avoid

making predictions for compounds, which differ substantially from the training set molecules [56]. The similarity threshold is defined as follows:

$$D_T = \bar{y} + Z\sigma \quad (5)$$

Here,  $\bar{y}$  is the average Euclidean distance of the  $k$  nearest neighbors of each compound within the training set (where the value of  $k$  is the same as in predictive  $k$ NN QSAR models),  $\sigma$  is the standard deviation of these Euclidean distances, and  $Z$  is an arbitrary parameter to control the significance level. Typically, we set  $Z$  to 0.5, which places the boundary for deciding whether a compound is within or outside of the applicability domain at one-half of the standard deviation. It is important to notice that increasing the value of  $Z$  would increase the number of compounds in the external set that are considered within the applicability domain but could decrease the accuracy of prediction due to inclusion of dissimilar nearest neighbors.

#### Y-randomization test

Y-randomization test is widely used to ensure model robustness [57]. It includes rebuilding the training set models using randomized activities (Y-vector) of the training set and comparing the resulting model statistics with that for the original test set. It is expected that models built with randomized activities should have significantly lower CCR value for both the training and test sets. In the model building process, it is possible that sometimes, though infrequently, high CCR values may be obtained due to a chance correlation or structural redundancy of the training set. If QSAR models obtained in the Y-randomization test have relatively high LOO-CV  $CCR_{train}$  as well as predictive  $CCR_{test}$ , it implies that acceptable QSAR models cannot be obtained for the given dataset by the current modeling method. In this study, the Y-randomization test was performed twice for each training/test set splits.

#### Virtual screening using $k$ NN models

As mentioned above, the screening database included 69653 compounds tested by the NCGC against AmpC beta lactamase. The primary HTS screening assay identified 64 “true” hits. Thus, we chose to screen the same database *in silico* using QSAR models as predictors. Only QSAR models that passed both internal and external validation tests were used. For each model we retained its parameters established in the process of external validation, i.e., the number of nearest neighbors  $k$ , selected descriptors, and  $Z_{cutoff}$  value for the applicability domain.

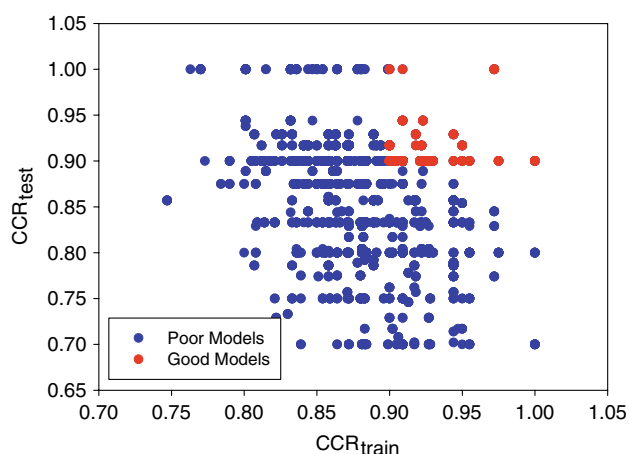
## Results and discussion

### $k$ NN Binary classification models

As shown in Fig. 2, the  $k$ NN QSAR method with variable selection afforded multiple models with optimal accuracy characterized as CCR for both training and test sets. In total, there were 3305 models with both  $CCR_{train}$  and  $CCR_{test}$  equal or higher than 0.70. Most models with  $CCR_{test} \geq 0.70$  also had corresponding  $CCR_{train} \geq 0.70$ , but the opposite was not always true. The models with high values of both  $CCR_{train}$  and  $CCR_{test}$  ( $\geq 0.70$ ) were considered acceptable. 342 predictive models with the highest values of CCR ( $CCR_{train}$  and  $CCR_{test} \geq 0.90$ , red dots in Fig. 2) were selected for consensus prediction. Table 1 summarizes the detailed confusion matrix and statistical parameters for the best  $k$ NN binary classification models. The  $CCR_{train}$  and  $CCR_{test}$  were found to be as high as 0.91 and 1.00, respectively, which implies that the models could identify correctly all 18 nonbinders and 9 out of 11 inhibitors ( $SE = 0.82$ ,  $SP = 1.00$ ,  $EN(1) = 2.00$ , and  $EN(0) = 1.69$ ) in the training set and all binders and nonbinders in the test set. This remarkably high internal accuracy and the large number of acceptable models imply that the  $k$ NN classification method was generally successful in correctly distinguishing binders *vs.* decoys using MolConnZ chemical descriptors of compounds only.

### QSAR Model validations

In addition to the internal validation of  $k$ NN models using test sets, Y-randomization and external validation are the critical steps of the entire QSAR workflow (Fig. 1). Only models that have been validated by these two steps can be utilized for external prediction and database mining [56].



**Fig. 2** The plot of  $k$ NN classification QSAR model accuracy for test ( $CCR_{test}$ ) vs. training ( $CCR_{train}$ ) sets for AmpC beta-lactamase dataset



**Table 1** Ten best *k*NN QSAR classification models with highest CCR values for all test sets using MolconnZ descriptors

Model no.	Nearest neighbors no.	CCR <sub>train</sub>	Confusion matrix						Statistics for the models				
			N(1)	N(2)	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)	CCR <sub>test</sub>
1	5	0.91	11	18	9	18	0	2	0.82	1.00	2.00	1.69	1.00
2	5	0.90	10	18	8	18	0	2	0.80	1.00	2.00	1.67	1.00
3	1	0.97	11	18	11	17	1	0	1.00	0.94	1.89	2.00	1.00
4	5	0.91	11	16	9	16	0	2	0.82	1.00	2.00	1.69	0.94
5	4	0.92	11	16	10	15	1	1	0.91	0.94	1.87	1.82	0.94
6	4	0.92	9	19	8	18	1	1	0.89	0.95	1.89	1.79	0.93
7	1	0.94	10	18	10	16	2	0	1.00	0.89	1.80	2.00	0.93
8	5	0.92	10	19	9	17	1	1	0.90	0.89	1.89	1.80	0.92
9	5	0.92	9	19	8	18	1	1	0.89	0.95	1.89	1.79	0.92
10	5	0.90	10	17	8	17	0	2	0.80	1.00	2.00	1.67	0.92

N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate

### Y-randomization test

In Y-randomization test, the binary annotations of AmpC beta-lactamase as inhibitors or non-binders were randomly shuffled and *k*NN classification models were built with the same parameter setting. The test was performed twice and both runs of Y-randomization tests showed that there were relatively small numbers of 330 and 429 models having both CCR<sub>train</sub> and CCR<sub>test</sub> higher than 0.70. However, there were no models with both CCR value higher than 0.90. It implied that the *k*NN models obtained with real binding affinities and CCR greater than 0.90 are robust.

### External validation

Two datasets were employed for external validation, i.e. the 10 compounds randomly excluded from modeling sets and 50 non-binders, which were relatively dissimilar in their structure from the 21 inhibitors in the original dataset. Consensus predictions were carried out using 342 predictive models with CCR<sub>train</sub> and CCR<sub>test</sub> greater than 0.9 under different Z value cutoffs (Z = 0.5–3.0, Table 2). The prediction accuracy for the 10-compound external validation set was 100% for both 5 inhibitors and 5 non-binders under Z<sub>cutoff</sub> = 0.5, leading to CCR = 1.00, SE = 1.00, SP = 1.00, EN(1) = 2.00, and EN(0) = 2.00.

**Table 2** Consensus predictions under different Z value cutoffs for two external validation sets, the randomly-excluded 10 compounds from modeling sets and 50 non-binders which were dissimilar in structure to 21 inhibitors in the original dataset

External validation sets	Z <sub>cutoff</sub>	Prediction CCR	Confusion matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)
10 Randomly-excluded compounds	0.5	1.00	5	5	5	5	0	0	1.00	1.00	2.00	2.00
	0.5	0.87	0	47	0	41	6	0	N/A	0.87	N/A	N/A
50 Non-binders	1.5	0.87	0	47	0	41	6	0	N/A	0.87	N/A	N/A
	3.0	0.86	0	49	0	42	7	0	N/A	0.86	N/A	N/A
64 HTS 'hits'	0.5	0.20	25	0	5	0	0	20	0.20	N/A	N/A	N/A
	1.5	0.10	41	0	4	0	0	37	0.10	N/A	N/A	N/A
	3.0	0.15	55	0	8	0	0	47	0.15	N/A	N/A	N/A

N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate

<sup>a</sup> Many N(1) inhibitors of 64 HTS 'hits' and N(2) non-binders of 50 non-binders were out of application domain of all consensus models, thus having no prediction. Only data for compounds found within the AD were used for statistical summaries

The accuracy of prediction for 50 non-binders was also high, ranging from  $CCR = 0.87$  under  $Z_{\text{cutoff}} = 0.5$  to  $CCR = 0.86$  under  $Z_{\text{cutoff}} = 3.0$  (Table 2). Because of the applicability domain inherent to individual  $kNN$  QSAR models, the consensus prediction usually cannot cover the whole dataset. By increasing the  $Z_{\text{cutoff}}$  from 0.5 to 3.0, the prediction coverage for 50 non-binders increased from 94% to 98% whereas the prediction accuracy decreased. Figure 3 shows the consensus scores and the coverage of predictive models for each of the 50 non-binders. The consensus score, in terms of the average class number in classification QSAR, was calculated by the fraction of models that predicted a compound as non-binder over the total number of models used for prediction plus 1. Under  $Z_{\text{cutoff}} = 0.5$ , six falsely predicted inhibitors (average class number  $<1.5$ ) were within the applicability domain of only 70 models (i.e., approximately 20% of all models), i.e., the model coverage was as low as 20%. In general, the prediction with such a low coverage is viewed as of low confidence level. The higher  $Z_{\text{cutoff}}$  significantly raised the model coverage for both inhibitor and non-binder prediction because of the extended applicability domain for individual models. In Fig. 3(b, c), the model coverage for predicting inhibitors jumped up to 53% for  $Z_{\text{cutoff}} = 1.5$  and up to 94% for  $Z_{\text{cutoff}} = 3.0$ . However, the prediction with extended applicability domain for consensus models also comes with lower confidence level. Generally speaking, in order to have the reliable and

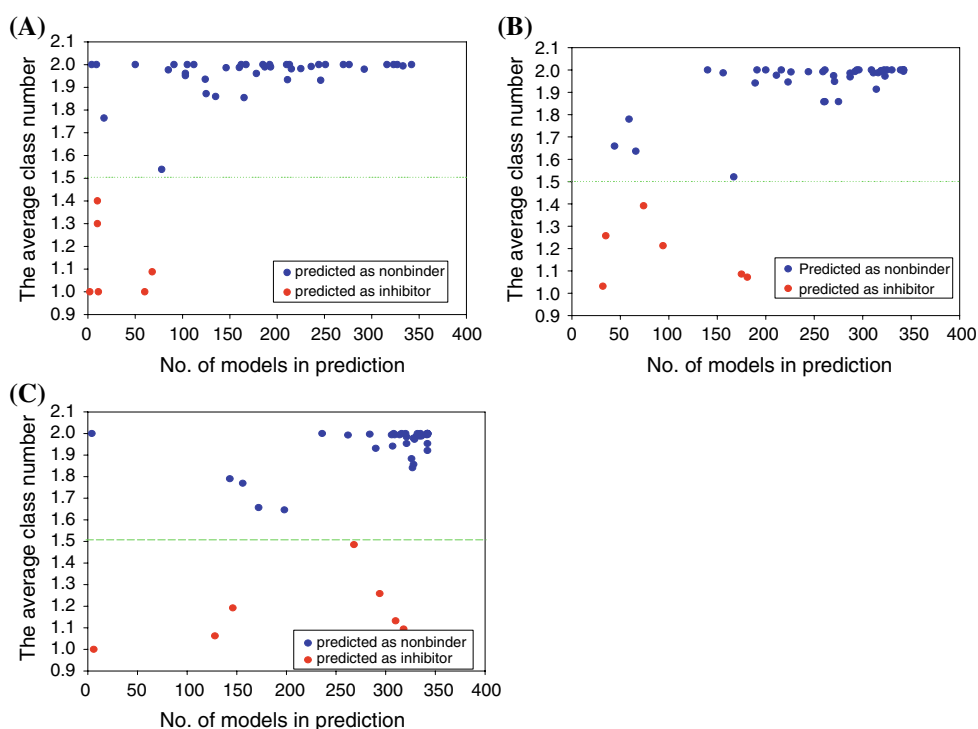
accurate prediction, one has to have the broader model coverage and a smaller  $Z_{\text{cutoff}}$  value.

In summary, 342 models with both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  equal to or greater than 0.90 could be applied for consensus prediction and database mining. The models chosen for the prediction had relatively small  $Z_{\text{cutoff}}$  ( $=0.5$ ) and relatively broad coverage for compounds in external datasets ( $\geq 50\%$ ).

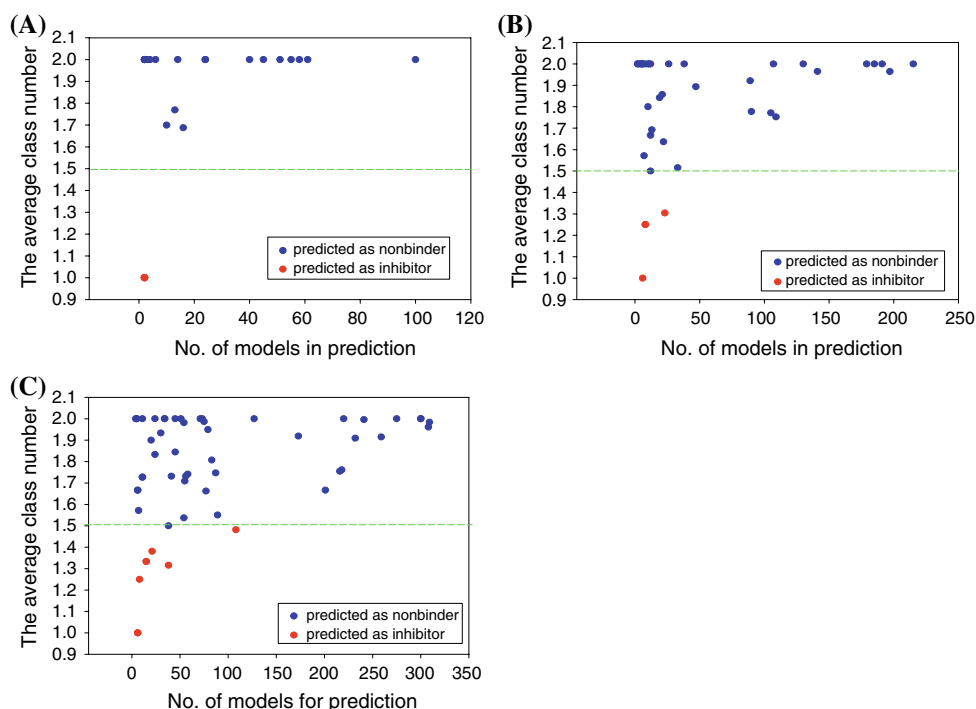
### External prediction

We used models built from 41 AmpC inhibitor/nonbinder dataset to verify the 64 “actives” from AID 584 and AID 585 screening. Under  $Z_{\text{cutoff}} = 0.5$ , we could only generate predictions for 25 compounds out of 64 “actives” whereas the remaining compounds were found to be outside of the applicability domain. As shown in Table 2, five out of these 25 compounds were predicted as true inhibitors. However, the predictions were based on only two models (out of 342 models with both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  higher than 0.90, cf. Fig. 4a). Thus, the coverage for both compounds and consensus models was extremely low and as a result these predictions should not be viewed as reliable. Even under higher  $Z_{\text{cutoff}} = 3.0$ , the model coverage was still low such that “actives” were predicted by only 110 models (32% of all models, cf. Fig. 4c). Furthermore, the formal prediction accuracy (assuming that the 64 hits were true inhibitors) was extremely low, e.g.  $CCR = 0.20$

**Fig. 3** The consensus scores and the coverage of predictive models for the 50 non-binding decoys dissimilar to the modeling dataset. Three  $Z_{\text{cutoff}}$  values were used: (a)  $Z_{\text{cutoff}} = 0.5$ ; (b)  $Z_{\text{cutoff}} = 1.5$ ; (c)  $Z_{\text{cutoff}} = 3.0$



**Fig. 4** The consensus scores and the coverage of predictive models for the 64 HTS hits identified from the primary HTS screening assays reported in PubChem. Three Z cutoff values were used: (a)  $Z_{\text{cutoff}} = 0.5$ ; (b)  $Z_{\text{cutoff}} = 1.5$ ; (c)  $Z_{\text{cutoff}} = 3.0$



( $Z_{\text{cutoff}} = 0.5$ ),  $\text{CCR} = 0.10$  ( $Z_{\text{cutoff}} = 1.5$ ) and  $\text{CCR} = 0.15$  ( $Z_{\text{cutoff}} = 3.0$ ) (Table 2). Thus, based on our modeling results none of the 64 compounds in the NCGC set was predicted reliably as a non-covalent and reversible inhibitor.

Notably, the independent experimental verification of those 64 “actives” hits appears to confirm the results of our consensus prediction based on recent results obtained in Dr. B. Shoichet’s lab. These studies [35] have shown that 25 of these active compounds are beta-lactam-based irreversible inhibitors of beta-lactamase. Five to ten additional actives are believed to be aggregators. The data on the remaining 35 compounds have not been confirmed yet but preliminary data indicate that none of them act as true reversible inhibitors of beta-lactamase (Dr. Shoichet, personal communications). These recent results confirm that our models are both accurate and robust.

#### Descriptor interpretation

A summary of descriptors ranked as top 20 based on their frequency of occurrences in 342 consensus models are given in Table 3. The frequency of occurrence is defined as percentage of models where a descriptor is present. For instance, the highest frequency of 32.2% means that a particular descriptor type occurs in about 110 out of the total of 342 models. The descriptor class and the structural illustration of individual descriptor types are shown in this Table as well. It should be noted that molecular connectivity descriptors are predominant in all models, i.e., over

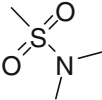
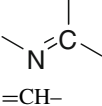
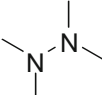
50% of descriptors in 20 top-ranked most frequent descriptor types belong to this class. The remaining descriptor types are mostly related to class of electrotopological state (E-state) indices which reflect the electronic environment of each atom due to its intrinsic electronic properties and the influence of other atoms in the molecule. By mapping the frequent descriptors to the inhibitors and non-binders in the dataset, the sulfonamide group was found to be a common feature in both inhibitors and non-binders. Importantly, all the nitrile groups could only be found in the structure of non-binders. Thus, conventional structure based scoring functions appear to be insensitive to (the presence or absence of) this group in chemical structures. This result illustrates a potential power of QSAR models in informing conventional scoring functions of their possible deficiencies that probably could be corrected with ease.

#### Virtual screening using predictive QSAR models

Instead of using only one single and best model for virtual screening, the consensus prediction approach was applied that relies on averaging predictions from all qualified models, i.e. 342 models with both  $\text{CCR}_{\text{train}}$  and  $\text{CCR}_{\text{test}}$  equal to or greater than 0.90. The complete modeling set (i.e., including training and test sets) was used for the prediction using each model as opposed to using only the corresponding training set. Initially, as many as 4565 compounds in the NCGC dataset of 69653 compounds were predicted as inhibitors by at least one of 342 models.



**Table 3** The 20 most frequent MolConnZ descriptors found in acceptable *k*NN QSAR models

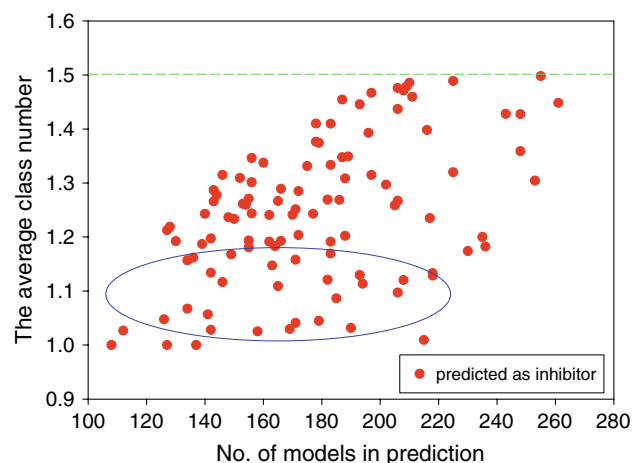
Rank <sup>a</sup>	Descriptor ID	Frequency <sup>b</sup>	Descriptor class	Illustration
1	nHCsatu	32.2	Atom-type counts	CH <sub>n</sub> (unsaturated)
2	Hsulfonamide	28.4	Group-type Hydrogen E-State values	
3	nnitrile	27.5	Group-type counts	–C≡N
4	Hmin	27.2	Minimum H E-State	
5	naaO	26.3	Atom-type counts	:O: (aromatic)
6	naaS	26.3	Atom-type counts	aSa (aromatic)
7	SaaCH	26.0	Atom-type EState sums	:CH:
8	n3Pad24	26.0	Vertex alpha-delta counts	
9	SssCH2	26.0	Atom-type EState sums	–CH <sub>2</sub> –
10	SHBint5	25.4	Internal H-Bond counts and EStates	
11	Xvch5	24.3	Valence cluster/chain Chi indices	
12	n2Pag23	24.3	Vertex alpha-gamma counts	
13	IDW	24.0	Bonchev-Trinajstic information indices	
14	htets2	23.7	Total topological state indices based on H E-State indices	
15	nimine	23.7	Group-type counts	
16	ndsCH	23.4	Atom-type counts	=CH–
17	IDC	23.4	Bonchev-Trinajstic information indices	
18	tets3	23.1	Total topological state indices based on E-State indices	
19	n3Pad13	23.1	Vertex alpha-delta counts	
20	nhydrazine	22.8	Group-type counts	

<sup>a</sup> *k*NN rank is based on the frequency of each descriptor occurred

<sup>b</sup> Frequency is the number of times each descriptor occurred in 342 validated models

To narrow the hit list and obtain the higher confidence level for each prediction, we took both the consensus score (average class number) and model coverage into account. In particular, only the hits with average class number between 1.0 and 1.2 and the model coverage over 50% (171 out of 342 models) were selected (Fig. 5). Furthermore, we restricted ourselves to the most conservative applicability domain for each model using  $Z_{\text{cutoff}} = 0.5$ . We found that there were only 15 compounds that satisfied both criteria (Table 4).

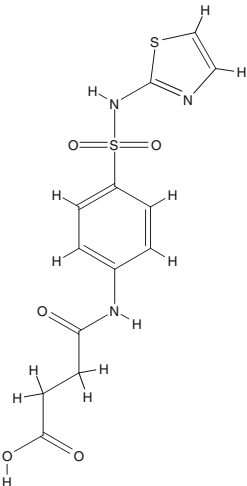
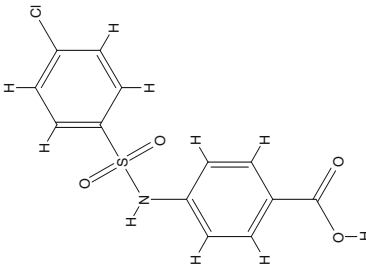
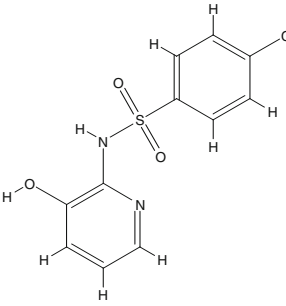
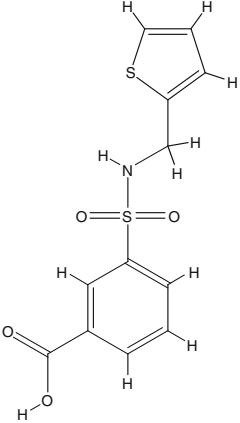
We have clustered these 15 compounds together with 16 competitive inhibitors from the training set using tools available in PubChem [58]. Each compound was represented by a fingerprint of 881 substructure keys, indicating the presence or absence of a particular chemical substructure. The pairwise similarity between compounds was measured by the Tanimoto coefficients (TC), which were used for hierarchical clustering of hits. The most chemically different pair of structures had TC = 0.522 (Fig. 6). Several structural classes were observed depending on the



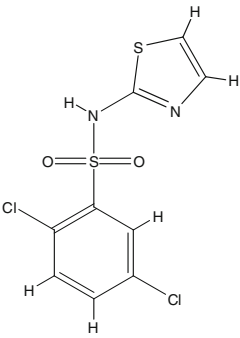
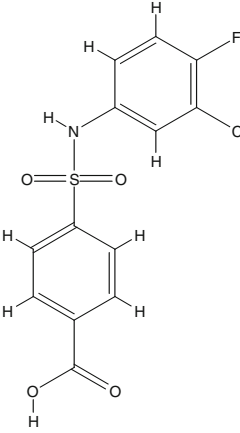
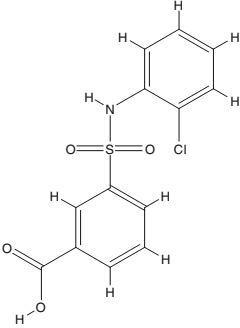
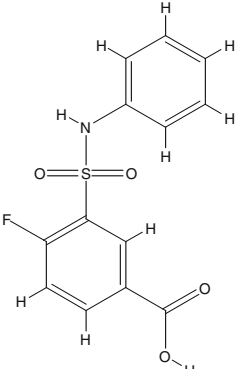
**Fig. 5** The consensus scores and the coverage of predictive models for the mining hits in the NCGC database ( $Z_{\text{cutoff}} = 0.5$ )

TC thresholds, e.g. there were four clusters at TC = 0.70. Notably, many of the 15 computational hits were found to be structurally similar to inhibitors used in model building.

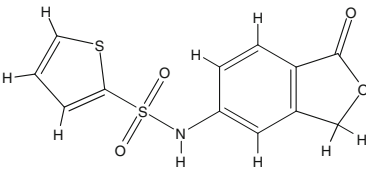
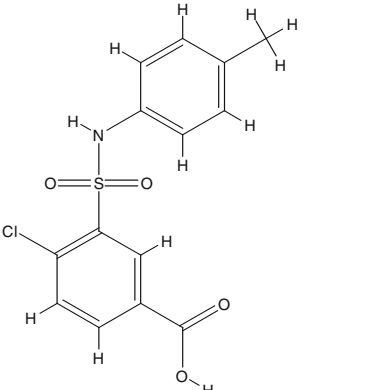
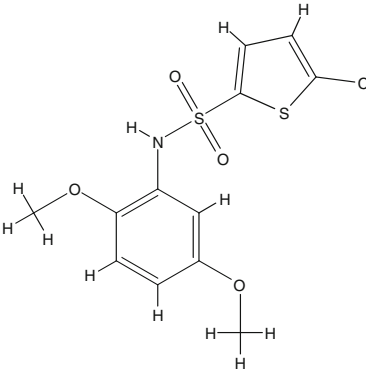
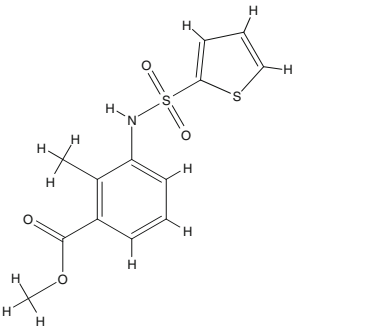
**Table 4** The 15 computational hits predicted as AmpC beta-lactamase inhibitors as a result of mining the NCGC AmpC screening library

Structure	Serial no.	PubChem CID	No. of models predicted as inhibitor	No. of models predicted as non-binder	No. of models in prediction	Average class num.	Exp. IC <sub>50</sub> (mM) <sup>a</sup>
	1	5315	213	2	215	1.01	Unknown
	2	39854	186	20	206	1.10	Unknown
	3	573009	190	40	230	1.17	Unknown
	4	647810	193	43	236	1.18	3.0

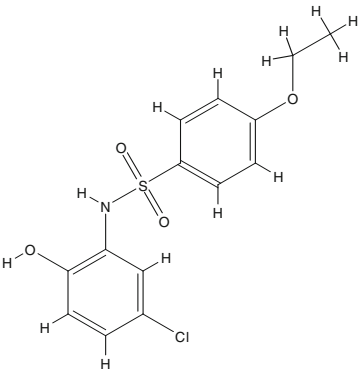
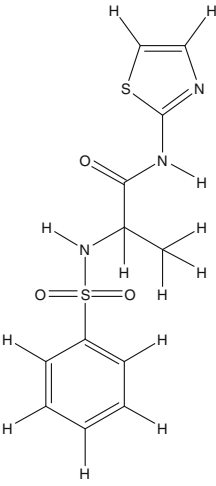
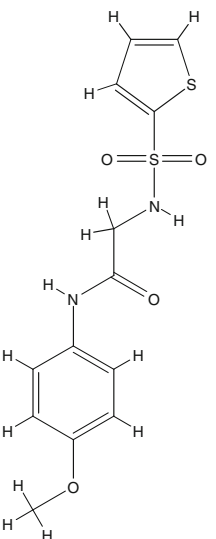
**Table 4** continued

Structure	Serial no.	PubChem CID	No. of models predicted as inhibitor	No. of models predicted as non-binder	No. of models in prediction	Average class num.	Exp. IC <sub>50</sub> (mM) <sup>a</sup>
	5	661093	172	22	194	1.11	Unknown
	6	665205	171	8	179	1.04	9.0
	7	699751	189	29	218	1.13	0.7 <sup>b</sup>
	8	699907	184	6	190	1.03	1.8

**Table 4** continued

Structure	Serial no.	PubChem CID	No. of models predicted as inhibitor	No. of models predicted as non-binder	No. of models in prediction	Average class num.	Exp. IC <sub>50</sub> (mM) <sup>a</sup>
	9	713179	169	16	185	1.09	Unknown
	10	793725	168	25	193	1.13	Unknown
	11	843845	152	31	183	1.17	Unknown
	12	970871	183	25	208	1.12	Unknown

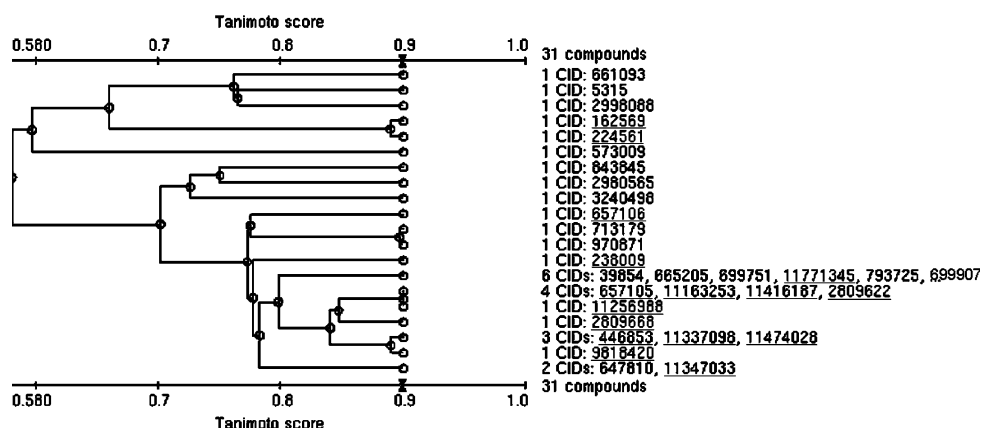
**Table 4** continued

Structure	Serial no.	PubChem CID	No. of models predicted as inhibitor	No. of models predicted as non-binder	No. of models in prediction	Average class num.	Exp. IC <sub>50</sub> (mM) <sup>a</sup>
	13	2980565	190	28	218	1.13	7.0
	14	2998088	160	22	182	1.12	Unknown
	15	3240498	148	35	183	1.19	Unknown

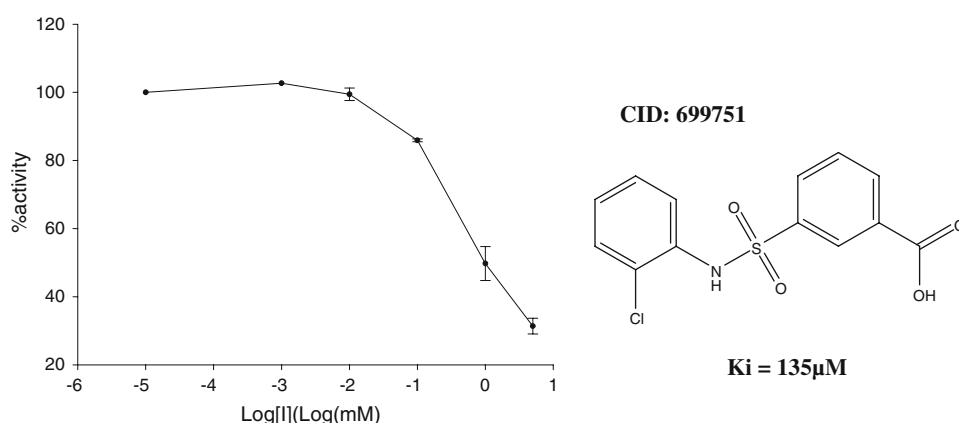
<sup>a</sup> Only five hit compounds were sent to experimental validation upon the commercial availability<sup>b</sup> The full IC<sub>50</sub> curve was generated in further experiment and the Ki value was determined



**Fig. 6** The structural clustering of 15 mining hits from NCGC database combined with 16 AmpC beta-lactamase competitive inhibitors (underlined) based on the Tanimoto score. The computations were carried out at the PubChem server



**Fig. 7** The full dose response curve for compound 699751. The experimental studies were conducted in B. Shoichet's laboratory



There were five hits that were highly similar ( $TC \geq 0.90$ ; CID: 39854, 665205, 699751, 793725 and 699907) to the competitive inhibitor 11771345. More often than not, hit compound 647810 was in close proximity to inhibitor 11347033 with the Tanimoto coefficients over 0.90. Several computational hits were selected for the experimental validation in Dr. Shoichet's laboratory as potential AmpC beta-lactamase inhibitors.

We should emphasize that our model validation is a critical inherent feature of our QSAR modeling workflow. This issue of model validation has been given a lot of attention by the QSAR research community [59]. Until recently, most practitioners merely presumed that internally cross-validated models built from available training set data should be externally predictive. We and others have demonstrated that internal validation techniques such as leave-one-out (LOO) or even leave-many-out (LMO) cross-validation applied to the training set is insufficient to ensure the external predictive power of QSAR models [56, 60]. Thus, we used two external validation sets in this study as well as the Y-randomization test to ensure the robustness and predictive power of *k*NN models. Needless to say, the use of externally validated models and applicability

domains is especially critical when the models are employed in virtual screening.

Another important feature of many current biomolecular datasets, especially generated as a results of HTS campaigns is the imbalance between “actives” and “inactives”, obviously in favor of inactives. For example, the hit rates in assays deposited in PubChem by the NIH screening centers forming the Molecular Library Screening Center Network (MLSCN) are very low, in most cases not exceeding 0.5% [61]. The imbalanced datasets pose a significant problem for classification QSAR modeling because models that predict correctly the same fraction of objects in each class will have different objective function values. To circumvent this problem in this study, we conducted the similarity search between the members of the underrepresented class (inhibitors) vs. another one (non-binders). A subset of the original dataset that was relatively balanced (2:3) was formed and utilized for model building. The 50 non-binders that were less similar to binders were retained as one of the external validation datasets. The classification models built for the balanced subset were shown to predict compounds in this external dataset as non-binders with very high accuracy. Among the 47

non-binders (3 were outside of the applicability domain), 41 were accurately annotated by consensus prediction (CCR = 0.87, cf. Table 2). The success of this strategy suggests that it could be applied to the analysis of many imbalanced datasets.

### Experimental validation

Of the 15 computational hits from mining the NCGC AmpC screening library, five compounds were selected based on their chemical similarity (measured by Euclidean distance in the MZ descriptor space) to the 21 inhibitors and commercial availability. We should stress that binary QSAR models were used for prediction so no quantitative estimate of binding affinity could be made. All five hits (CID: 647810, 665205, 699751, 699907 and 2980565; Table 4) did show the inhibitory activities at millimolar level at the single concentration. Among them, compound 699751 had the highest inhibitory activity at 0.7 mM. For this compound, a full dose-response curve was obtained and the inhibition constant,  $K_i$ , was calculated by the Cheng-Prusoff equation using the  $IC_{50}$  and  $K_d$ , the dissociation constant of AmpC for the substrate measured in a separate assay. Thus, compound 699751 yielded the  $K_i$  value of 135  $\mu$ M (Fig. 7). In summary, the above results did prove the predictive power of our binary  $k$ NN classification QSAR models built for AmpC beta-lactamase inhibitors. These studies illustrate that the validated QSAR workflow, as employed in this paper, could be used as a general tool for identifying promising hits by the means of virtual screening of chemical libraries.

### Conclusions

Our studies demonstrate that binary  $k$ NN classification QSAR models built with MolconnZ descriptors can accurately differentiate true AmpC beta-lactamase inhibitors from non-binding decoys. A special QSAR modeling scheme was employed for this imbalanced dataset and the models were rigorously validated using both internal (multiple training/test set divisions and Y-randomization) as well as external (two external validation sets) validation approaches. We have demonstrated that this strategy afforded multiple QSAR models with high internal and external predictive power. As part of our QSAR modeling workflow, the predictors were further utilized for mining the NCGC dataset (69653 compounds tested for AmpC beta-lactamase binding). We found that our validated models disagreed with the experimental annotation of 64 compounds as AmpC binders as reported in PubChem BioAssays AID584 [36] and AID585 [37]. Interestingly, our negative predictions for these compounds appear to be

in agreement with the preliminary results of the confirmatory secondary assays conducted in B. Shoichet's lab (B. Shoichet, personal communications). On the other hand, our models used in the most conservative way (i.e., in consensus fashion and with the strictest applicability domain criteria) did identify 15 putative AmpC inhibitors among compounds annotated as experimental non-binders in the NCGC assays reported in PubChem. Five of them showed inhibition activities at the millimolar concentration, and one compound (compound 699751) was found to have the highest  $K_i$  of 135  $\mu$ M. The results of our studies suggest that at least in some cases when a sufficient amount of data on true binders vs. nonbinding compounds is available simple QSAR modeling approaches could be used successfully to complement (and possibly educate based on QSAR model interpretation) the conventional scoring functions used in three-dimensional docking studies. Furthermore, as we have demonstrated in this paper, QSAR models can be successfully used not only to discriminate binders vs. binding decoys but most importantly, for finding promising hits by the means of virtual screening of chemical libraries.

**Acknowledgements** We would like to thank Drs. Brian Shoichet and John Irwin for providing the AmpC dataset and fruitful discussions. We also acknowledge the access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The studies reported in this paper were supported in part by the NIH research grant GM066940 and the RoadMap Center planning grant P20-HG003898. Denise Teotico was supported by NIH grants GM71630 and GM59957.

### References

1. Sharff A, Jhoti H (2003) High-throughput crystallography to enhance drug discovery. *Curr Opin Chem Biol* 7:340–345
2. Blundell TL, Jhoti H, Abell C (2002) High-throughput crystallography for lead discovery in drug design. *Nat Rev Drug Discov* 1:45–54
3. RCSB. PDB. <http://www.rcsb.org/>. Accessed 2007
4. Dessalew N, Bharatam PV (2007) Identification of potential glycogen kinase-3 inhibitors by structure based virtual screening. *Biophys Chem* 128:165–175
5. Lu IL, Huang CF, Peng YH, Lin YT, Hsieh HP, Chen CT et al (2006) Structure-based drug design of a novel family of PPAR gamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. *J Med Chem* 49: 2703–2712
6. Zhou Y, Peng H, Ji Q, Qi J, Zhu Z, Yang C (2006) Discovery of small molecule inhibitors of integrin  $\alpha$ v $\beta$ 3 through structure-based virtual screening. *Bioorg Med Chem Lett* 16: 5878–5882
7. Du L, Li M, You Q, Xia L (2007) A novel structure-based virtual screening model for the hERG channel blockers. *Biochem Biophys Res Commun* 355:889–894
8. Kellenberger E, Springael JY, Parmentier M, Hachet-Haas M, Galzi JL, Rognan D (2007) Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J Med Chem* 50:1294–1303

9. Zhao L, Brinton RD (2005) Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases. *J Med Chem* 48:3463–3466
10. Evers A, Klabunde T (2005) Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J Med Chem* 48:1088–1097
11. Oh M, Im I, Lee YJ, Kim YH, Yoon JH, Park HG et al (2004) Structure-based virtual screening and biological evaluation of potent and selective ADAM12 inhibitors. *Bioorg Med Chem Lett* 14:6071–6074
12. Christmann-Franck S, Bertrand HO, Goupil-Lamy A, der Garabedian PA, Mauffret O, Hoffmann R et al (2004) Structure-based virtual screening: an application to human topoisomerase II alpha. *J Med Chem* 47:6840–6853
13. Kim YG, Thai KM, Song J, Kim KK, Park HJ (2007) Identification of novel ligands for the Z-DNA binding protein by structure-based virtual screening. *Chem Pharm Bull (Tokyo)* 55:340–342
14. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH et al (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49:5912–5931
15. Graves AP, Brenk R, Shoichet BK (2005) Decoys for docking. *J Med Chem* 48:3714–3728
16. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) On evaluating molecular-docking methods for pose prediction and enrichment factors. *J Chem Inf Model* 46:401–415
17. Park H, Lee J, Lee S (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* 65:549–554
18. Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* 46:2287–2303
19. Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP (2006) eHiTS: an innovative approach to the docking and scoring function problems. *Curr Protein Pept Sci* 7:421–435
20. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB (2002) Consensus scoring for ligand/protein interactions. *J Mol Graph Model* 20:281–295
21. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* 42:5100–5109
22. Wang R, Wang S (2001) How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci* 41:1422–1426
23. Yang JM, Chen YF, Shen TW, Kristal BS, Hsu DF (2005) Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model* 45:1134–1146
24. Powers RA, Morandi F, Shoichet BK (2002) Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* 10:1013–1023
25. Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des* 13:3494–3504
26. Tropsha A (2005) Application of predictive QSAR models to database mining. In: Oprea T (ed) *Cheminformatics in drug discovery*. Wiley-VCH, pp 437–455
27. Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A (2005) Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J Comput Aided Mol Des* 19:229–242
28. de Cerqueira LP, Golbraikh A, Oloff S, Xiao Y, Tropsha A (2006) Combinatorial QSAR modeling of P-glycoprotein substrates. *J Chem Inf Model* 46:1245–1254
29. Oloff S, Mailman RB, Tropsha A (2005) Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J Med Chem* 48:7322–7332
30. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A (2004) Application of predictive QSAR models to database mining: identification and experimental validation of novel anti-convulsant compounds. *J Med Chem* 47:2356–2364
31. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng W, Wolschann P et al (2004) Combinatorial QSAR of ambergris fragrance compounds. *J Chem Inf Comput Sci* 44:582–595
32. NCBI. PubChem. <http://pubchem.ncbi.nlm.nih.gov/>. Accessed 2007
33. Shoichet BK. Dr. Brian Shoichet Take-away Webpage. <http://shoichetlab.combio.ucsf.edu/take-away.php>. Accessed 2007
34. Tondi D, Morandi F, Bonnet R, Costi MP, Shoichet BK (2005) Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture. *J Am Chem Soc* 127:4632–4639
35. Feng BY, Simeonov A, Jadhav A, Babaoglu K, Inglese J, Shoichet BK et al (2007) A high-throughput screen for aggregation-based inhibition in a large compound library. *J Med Chem* 50:2385–2390
36. PubChem. PubChem Bioassay AID 584. <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=584>. Accessed 2007
37. PubChem. PubChem Bioassay AID 585. <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=585>. Accessed 2007
38. Feng BY, Shelat A, Doman TN, Guy RK, Shoichet BK (2005) High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* 1:146–148
39. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers* 5:231–243
40. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* 17:241–253
41. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36
42. Sybyl 7.2. (2007) Tripos, Inc.
43. MolconnZ 4.09. (2007) eduSoft, LC.
44. Kier LB, Hall LH (1976) *Molecular connectivity in chemistry and drug research*. Academic Press, New York
45. Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis*. Wiley, New York
46. Randi M (1975) On characterization on molecular branching. *J Am Chem Soc* 97:6609–6615
47. Kier LB (1985) A shape index from molecular graphs. *Quant Struct-Act Relat* 4:109–116
48. Kier LB (1987) Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant Struct-Act Relat* 6:8–12
49. Kier LB, Hall LH (1990) An electrotopological state index for atoms in molecules. *Pharm Res* 7:801
50. Kier LB, Hall LH (1991) An Index of Electrotopological State of Atoms in Molecules. *J Math Chem* 7:229
51. Kier LB, Hall LH (1999) *Molecular structure description: the electrotopological state*. Academic Press
52. Petitjean M (1992) Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J Chem Inf Comput Sci* 32:331–337
53. Zheng W, Tropsha A (2000) Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci* 40:185–194
54. Tropsha A (2003) Recent trends in quantitative structure-activity relationships. In: Abraham D (ed) *Burger's medicinal chemistry and drug discovery*. Wiley, New York, pp. 49–77

55. Itskowitz P, Tropsha A (2005) kappa Nearest neighbors QSAR modeling as a variational problem: theory and applications. *J Chem Inf Model* 45:777–785
56. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar Comb Sci* 22:69–77
57. Wold S, Eriksson L (1995): Statistical validation of QSAR results. In: Waterbeemd Hvd (ed) *Chemometrics methods in molecular design* (Methods and principles in medicinal chemistry, Vol 2). Wiley-VCH Verlag GmbH, Weinheim (Germany), pp 309–318
58. PubChem. Structural Clustering. <http://pubchem.ncbi.nlm.nih.gov/assay/assaycluster.cgi>. Accessed 2007
59. Jorgensen WL, Tirado-Rives J (2006) QSAR/QSPR and proprietary data. *J Chem Inf Model* 46:937
60. Golbraikh A, Tropsha A (2002) Beware of  $q(2)!$ . *J Mol Graph Model* 20:269–276
61. Oprea TI, Tropsha A, Faulon JL, Rintoul MD (2007) Systems chemical biology. *Nat Chem Biol* 3:447–450