Discovery of new β-D-glucosidase inhibitors via pharmacophore modeling and QSAR analysis followed by in silico screening.

ARTICLE *in* JOURNAL OF MOLECULAR MODELING · MARCH 2011

Impact Factor: 1.74 · DOI: 10.1007/s00894-010-0737-1 · Source: PubMed

4 AUTHORS, INCLUDING:

Reema Abu Khalaf
Al-Zaytoonah University of Jordan
12 PUBLICATIONS **80** CITATIONS
SEE PROFILE

Mohammad S Mubarak
University of Jordan
159 PUBLICATIONS **1,082** CITATIONS
SEE PROFILE

Mutasem Omar Taha
University of Jordan
121 PUBLICATIONS **1,173** CITATIONS
SEE PROFILE

Discovery of New β-D-Galactosidase Inhibitors via Pharmacophore Modeling and QSAR Analysis Followed by *In Silico* Screening

AHMED MUTANABBI ABDULA,[1] REEMA ABU KHALAF,[2] MOHAMMAD S. MUBARAK,[1] MUTASEM O. TAHA[3*]
[1]*Department of Chemistry, Faculty of Science, University of Jordan, Amman, Jordan*
[2]*Department of Pharmaceutical Sciences, Faculty of Pharmacy, Al Zaytoonah Private University of Jordan, Amman, Jordan*
[3]*Drug Discovery Unit, Department of Pharmaceutical Sciences, Faculty of Pharmacy, University of Jordan, Amman, Jordan*

Received 11 January 2010; Revised 14 May 2010; Accepted 23 June 2010
DOI 10.1002/jcc.21635
Published online 20 August 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Glycosidases, including β-D-galactosidase, are involved in a variety of metabolic disorders, such as diabetes, viral or bacterial infections, and cancer. Accordingly, we were prompted to find new β-D-galactosidase inhibitors. Towards this end, we scanned the pharmacophoric space of this enzyme using a set of 41 known inhibitors. Genetic algorithm and multiple linear regression analyses were used to select an optimal combination of pharmacophoric models and physicochemical descriptors to yield self-consistent and predictive quantitative structure-activity relationship (QSAR). Five pharmacophores emerged in the QSAR equations suggesting the existence of more than one binding mode accessible to ligands within β-D-galactosidase pocket. The successful pharmacophores were complemented with strict shape constraints in an attempt to optimize their receiver-operating characteristic curve profiles. The validity of the QSAR equations and the associated pharmacophoric models were experimentally established by the identification of several β-D-galactosidase inhibitors retrieved via *in silico* search of two structural databases: the National Cancer Institute list of compounds and our in house built structural database of established drugs and agrochemicals.

 J Comput Chem 32: 463–482, 2011

Key words: β-D-galactosidase inhibitors; quantitative structure-activity relationships; *in silico* screening; pharmacophore modeling; shape constraints; receiver-operating characteristic

Introduction

Glycosidases are widespread enzymes in living organisms.[1,2] They catalyze the hydrolysis of glycosidic bonds in carbohydrates and glycoconjugates resulting in low-molecular weight monosaccharides and oligosaccharides. Glycosidases, including β-D-galactosidase, are involved in a variety of metabolic disorders and diseases, such as diabetes, viral or bacterial infections, and cancer. Therefore, the inhibition of glycosidases, including β-D-galactosidase, has many potential applications, for example, antidiabetic, antiviral (HIV, influenza), and anticancer agents.[3–8]

The intense interest, during the last decade, in the chemistry, biochemistry, and pharmacology of glycosidase inhibitors has lead to many types of natural and synthetic glycosidase inhibitors.[9–14] All previously reported glycosidase inhibitors were sugar-mimics (azasugars and carbasugars) and their analogs; however, no attempts were made to discover new glycosidase inhibitory scaffolds of better chemical stabilities, pharmacokinetic profiles, and higher potencies.[9–14]

More than 20,000 glycosidase sequences are now known and have been classified into more than 100 families based on sequence similarities.[15] β-glycosidases hydrolyze glycosides with overall retention of anomeric configuration using a double-displacement mechanism via a covalent glycosyl-enzyme intermediate. The covalent intermediate is flanked by highly dissociative transition states which possess substantial oxocarbenium ion-like character (Fig. 1). Such transition states feature extensive sp^2 hybridization and partial positive charge (predominantly along the bond between the anomeric carbon and endocyclic oxygen) and likely involve pyranoside distortion to half-chair or boat conformations.[16]

Some of the most powerful glycosidase inhibitors contain imidazole moieties,[17–21] such as the cyclohexylethyl-substituted

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Mutasem O. Taha; e-mail: mutasem@ju.edu.jo

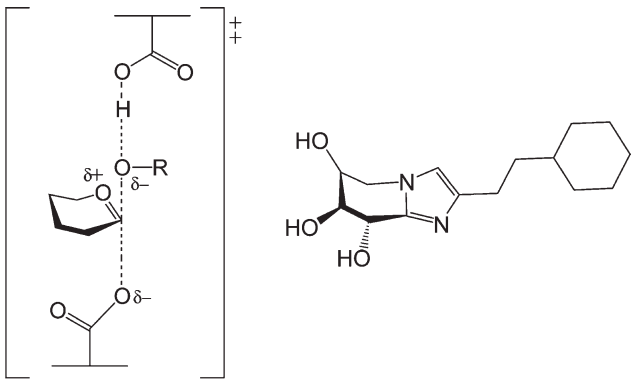

Figure 1. (A) Structure of the oxocarbenium ion-like transition state, and (B) structure of the cyclohexylethyl-substituted glucoimidazole **12**.

glucoimidazole **12** (Supporting Information Figs. 1 and 2, Table A), which display a flattened transition state-mimicking conformation resulting from fusion of the planar imidazole ring to the "glycon." This compound also possesses a vacant lone pair of electrons for lateral "protonation" by the acid/base residue.[22,23]

The current interest in the development of new β-D-galactosidase inhibitors combined with the lack of adequate computer-aided drug discovery efforts in this area, as well as the absence of any crystallographic structures for β-D-galactosidase prompted us to explore the possibility of developing ligand-based three-dimensional (3D) pharmacophore(s) integrated within self-consistent quantitative structure-activity relationship (QSAR) model for β-D-galactosidase inhibitors. The pharmacophore model(s) can be used as 3D search query(ies) to mine 3D libraries for new β-D-galactosidase inhibitors, whereas the QSAR model helps to predict the biological activities of the captured compounds and therefore prioritize them for *in vitro* evaluation.

The fact that all reported β-D-galactosidase inhibitors are slow binding/transition state analog (TSA) sugar-mimics[9–14] should complicate pharmacophore modeling and subsequent *in silico* search. TSAs resemble the substrate at its postulated transition to products. Such inhibitors require stringent steric and 3D provisions to dock into the enzymatic binding site during its sterically demanding high-energy transition state. TSAs are known to be much more tightly bound to the targeted enzyme than their ground state counterparts (i.e., substrate analogues), which further supports the notion about the pronounced sensitivity of TSA-enzyme complexes to slight misalignments among their complementary attractive groups.[24–26] This conduct is expected to cause rugged structure-activity surface, which limit the ability of the pharmacophore theory to explain activity/inactivity variations among training compounds. In fact, pharmacophore modeling requires continuous bioactivity variation attributable to the presence or absence of certain chemical features, that is, smooth SAR surface.

The pronounced sensitivity of TSA to slight structural modifications should also complicate the subsequent use of pharmacophore models as 3D search queries to mine for new hits. Pharmacophore models would be too lax, and therefore, promiscuous in capturing TSAs as *in silico* hits, that is, they may identify many inactive hits (false positives).[27]

In fact, no previous pharmacophore modeling efforts have been reported for β-D-galactosidase inhibitors, probably as a consequence of their rugged SAR. Accordingly, we were prompted to hybridize our QSAR-based pharmacophore models with tight ligand shapes and to use the combination as 3D search queries.

We previously reported the use of this innovative approach towards the discovery of new inhibitory leads against glycogen synthase kinase 3β,[28] dipeptidyl peptidase,[29] hormone sensitive lipase,[30] bacterial MurF,[31] protein tyrosine phosphatase 1B,[32] influenza neuraminidase,[33] and cholesteryl ester transfer protein.[34]

We used the HYPOGEN module from the CATALYST software package to construct numerous plausible binding hypotheses for β-D-galactosidase inhibitors.[35] Subsequently, genetic function algorithm (GFA) and multiple linear regression (MLR) analysis were used to search for optimal QSAR that combines high-quality binding pharmacophores with other molecular descriptors and capable of explaining bioactivity variation across a collection of diverse β-D-galactosidase inhibitors.

The optimal pharmacophores were further validated by evaluating their ability to successfully classify a list of compounds as actives or inactives by assessing their receiver-operating characteristic (ROC) curves. Subsequently, the optimal pharmacophores were complemented with tight shape constraints to enhance their ROC profiles. Thereafter, the resulting shape-complemented pharmacophores were used as 3D search queries to screen several available virtual molecular databases for new β-D-galactosidase inhibitors.

CATALYST models drug–receptor interactions using information derived from the ligand structures.[35–43] HYPOGEN identifies a 3D array of a maximum of five chemical features common to active training ligands that provides relative alignment for each input molecule consistent with binding to a proposed common receptor site. The conformational flexibility of training ligands is modeled by creating multiple conformers that cover a specified energy range for each input molecule.[32,38–40,44–48]

The SHAPE module in CATALYST is a shape-based similarity searching method. The Van der Waals surface of a molecule (in a certain conformation) is calculated and represented as a set of points of uniform average density on a grid. The surface

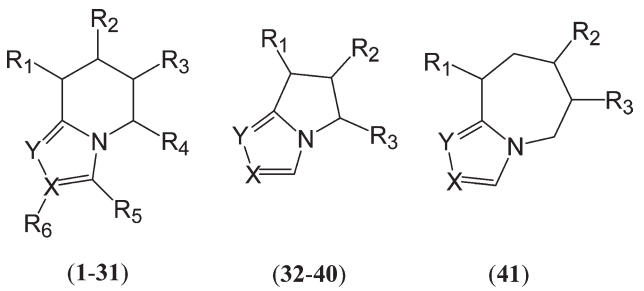

Figure 2. The chemical scaffolds of β-D-galactosidase training compounds, the detailed structures are given in Table A (Supporting Information).

points enclose a volume on the grid. The geometric center of the set of points is computed along with the three principal component vectors passing through the center. The maximum extents along each principal axis and the total volume are calculated. These provide shape indices that can be compared with the query and used in an initial screening step to eliminate poor matches from further consideration.[49] CATALYST pharmacophores, with or without shape constraints, have been used as 3D queries for database searching and in 3D-QSAR studies.[38,40,44,49]

Materials and Methods

Molecular Modeling

Software and Hardware

The following software packages were used in the present research.

- CATALYST (Version 4.11), Accelrys (www.accelrys.com), USA.
- CERIUS2 (Version 4.10), Accelrys (www.accelrys.com), USA.
- CS ChemDraw Ultra 6.0, Cambridge Soft (http:// www.cambridgesoft.Com), USA.
- Pharmacophore and QSAR modeling studies were accomplished using CATALYST (HYPOGEN module) and CERIUS2 software suites from Accelrys (San Diego, California, www.accelrys.com) installed on a Silicon Graphics Octane2 desktop workstation equipped with a dual 600 MHz MIPS R14000 processor (1.0 GB RAM) running the Irix 6.5 operating system. Structure drawing was accomplished using ChemDraw Ultra 6.0 installed on a Pentium 4 PC.

Data Set of β-D-Galactosidase Inhibitors

The structures of 41 β-D-galactosidase inhibitors **1**–**41** (Supporting Information Fig. 2 and Table A) were collected from recently published literature.[17–21] The *in vitro* bioactivities of the collected inhibitors were determined by using identical bioassay conditions and were expressed as K_i values (inhibition constants, μM), which allowed us to pool them for pharmacophore and QSAR analysis. The logarithm of measured $1/K_i$ values were used in the three-dimensional quantitative structure activity analysis (3D-QSAR), thus correlating the data linear to the free energy change. Inactive collected compounds (e.g., **21**, **25**, **26**, **27**, **28**, and **31**, Supporting Information Fig. 2 and Table A) were assumed to have K_i values of 2000 μM, which is six logarithmic cycles from the most potent compound (**12**, $K_i = 0.002$ μM). These assumptions are necessary to allow statistical correlation and QSAR analysis.[30,31] The two-dimensional (2D) chemical structures of the inhibitors were sketched using ChemDraw Ultra and saved in MDL-molfile format. Subsequently, they were imported into CATALYST, converted into corresponding standard 3D structures and energy minimized to the closest local minimum using the molecular mechanics CHARMm force field implemented in CATALYST. The resulting 3D structures were used as starting conformers for CATALYST conformational analysis.

Conformational Analysis

The molecular flexibilities of the collected compounds were taken into account by considering each compound as a collection of conformers representing different areas of the conformational space accessible to the molecule within a given energy range. Accordingly, the conformational space of each inhibitor (**1**–**41**, Supporting Information Fig. 2 and Table A) was explored adopting the "best conformer generation" option within CATALYST based on the generalized CHARMm force field implemented in the program. Default parameters were used in the conformation generation procedure, that is, a conformational ensemble was generated with an energy threshold of 20 kcal/mol from the local minimized structure, which has the lowest energy level and a maximum limit of 250 conformers per molecule.[35]

Pharmacophoric Hypotheses Generation

All 41 molecules with their associated conformational models were regrouped into a spreadsheet. The biological data of the inhibitors were reported with uncertainty values of 2 or 3, which means that the actual bioactivity of a particular inhibitor is assumed to be situated somewhere in an interval ranging from 1/2 to 2 or 1/3 to 3 times the reported bioactivity value of that inhibitor, respectively.[39,41,43] The uncertainty value is of great impact on the qualities of the resulting pharmacophores, as it controls the number of training compounds within the "most potent category" (see eq. (1) below).

Subsequently, a structurally diverse training set was selected for pharmacophore modeling of β-D-galactosidase: **1**, **9**, **12**, **13**, **15**, **19**, **21**, **22**, **24**, **25**, **26**, **27**, **28**, **29**, **30**, and **31** (Supporting Information Fig. 1 and Table A). Typically, CATALYST requires informative training sets that include at least 16 compounds of evenly spread bioactivities over at least three and a half logarithmic cycles. Lesser training lists could lead to chance correlation and thus faulty models.[39,41,43] The selected training set was used to conduct eight modeling runs (Supporting Information Table B) to explore the pharmacophoric space of β-D-galactosidase inhibitors. The exploration process included an altering interfeature spacing parameter (100 and 300 picometers) and the maximum number of allowed features in the resulting pharmacophore hypotheses, that is, they were allowed to vary from 4 to 5 for the first, third, fifth, and seventh runs and from 4 to 4 for the second, fourth, sixth and eighth runs (Supporting Information Table B). Pharmacophore modeling using CATALYST proceeds through three successive phases: Constructive phase, subtractive phase, and optimization phase.[39,41,43] During the constructive phase, CATALYST generates common conformational alignments among the most-active training compounds. Only molecular alignments based on a maximum of five chemical features are considered. The program identifies a particular compound as being within the most active category if it satisfies eq. (1).[39,41,43]

$$(\mathrm{MAct} \times \mathrm{UncMAct}) - (\mathrm{Act}/\mathrm{UncAct}) > 0.0 \qquad (1)$$

Where "MAct" is the activity of the most active compound in the training set, "Unc" is the uncertainty of the compounds, and "Act" is the activity of the training compounds under question. However, if there are more than eight most-active inhibi-

tors, only the top eight are used. Therefore, in this case, the most potent category of training compounds included **9** and **12**.

In the subsequent subtractive phase, CATALYST eliminates some hypotheses that fit inactive training compounds. A particular training compound is defined as being inactive if it satisfies eq. (2):[39,41,43]

$$\text{Log (Act)} - \text{log (MAct)} > 3.5 \quad (2)$$

Accordingly, compounds **21**, **22**, **24**, **25**, **26**, **27**, **28**, **29**, **30**, and **31** are considered least active and therefore used in the subtractive phase.

However, in the optimization phase, CATALYST applies fine perturbations in the form of vectored feature rotation, adding new feature and/or removing a feature, to selected hypotheses that survived the subtractive phase, in an attempt to find new models of enhanced bioactivity/mapping correlation, that is, improved 3D-QSAR properties. Eventually, CATALYST selects the highest-ranking models (10 by default) and presents them as the optimal pharmacophore hypotheses resulting from the particular automatic modeling run.

Our pharmacophore exploration efforts culminated in 80 pharmacophore models from eight automatic runs.

Assessment of the Generated Hypotheses

During pharmacophore modeling, CATALYST attempts to minimize a cost function consisting of three terms: Weight cost, error cost, and configuration cost.[35,39–43] Weight cost is a value that increases as the feature weight in a model deviates from an ideal value of 2. The deviation between the estimated activities of the training set and their experimentally determined values adds to the error cost. The activity of any compound can be estimated from a particular hypothesis through eq. (3).[35]

$$\text{Log (Estimated Activity)} = \text{I} + \text{Fit} \quad (3)$$

Where, I = the intercept of the regression line obtained by plotting the log of the biological activity of the training set compounds against the fit values of the training compounds. The fit value for any compound is obtained automatically using eq. (4).[35]

$$\text{Fit} = \Sigma\text{mapped hypothesis features} \times \text{W}[1 - \Sigma\ (\text{disp/tol})^2] \quad (4)$$

Where, Σmapped hypothesis features represents the number of pharmacophore features that successfully superimpose (i.e., map or overlap with) corresponding chemical moieties within the fitted compound, W is the weight of the corresponding hypothesis feature spheres. This value is fixed to 1.0 in CATALYST-generated models, disp is the distance between the center of a particular pharmacophoric sphere (feature centroid) and the center of the corresponding superimposed chemical moiety of the fitted compound, tol is the radius of the pharmacophoric feature sphere (known as Tolerance, equals to 1.6 Å by default), and $\Sigma(\text{disp/tol})^2$ is the summation of $(\text{disp/tol})^2$ values for all pharmacophoric features that successfully superimpose corresponding chemical functionalities in the fitted compound.[35]

The third term, that is, the configuration cost, penalizes the complexity of the hypothesis. This is a fixed cost, which is equal to the entropy of the hypothesis space. The more the numbers of features (a maximum of five) in a generated hypothesis, the higher is the entropy with subsequent increase in this cost. The overall cost (total cost) of a hypothesis is calculated by summing over the three cost factors, however, error cost is the main contributor to total cost. CATALYST also calculates the cost of the null hypothesis, which presumes that there is no relationship in the data and that experimental activities are normally distributed about their mean. Accordingly, the greater the difference from the null hypothesis cost, the more likely that the hypothesis does not reflect a chance correlation. In a successful automatic modeling run, CATALYST ranks the generated models according to their total costs.[35]

An additional approach to assess the quality of CATALYST-HYPOGEN pharmacophores is to cross-validate them using the Cat-Scramble program implemented in CATALYST. This validation procedure is based on Fisher's randomization test.[50] In this validation test, we selected a 95% confidence level, which instructs CATALYST to generate 19 random spreadsheets by the Cat-Scramble command. Subsequently, CATALYST-HYPOGEN is challenged to use these random spreadsheets to generate hypotheses using exactly the same features and parameters used in generating the initial unscrambled hypotheses.[51] Success in generating pharmacophores of comparable cost criteria to those produced by the original unscrambled data reduces the confidence in the training compounds and the unscrambled original pharmacophore models.

Clustering of the Generated Pharmacophore Hypotheses

The generated pharmacophoric models (80) were clustered using the hierarchical average linkage method available in CATALYST. Subsequently, the highest-ranking representatives, as judged based on the correlation (r) of their best-fit values against $\log(1/K_i)$ values of the training compounds, were selected to represent their corresponding clusters in subsequent QSAR modeling.

QSAR Modeling

A set of training inhibitors was selected from the collected list (**1**–**41**, Supporting Information Fig. 2 and Table A) for QSAR modeling. However, as it is essential to evaluate the predictive power of the resulting QSAR models on an external set of inhibitors, eight molecules (ca. 20% of the dataset) were used as external test set to validate the QSAR models. The test molecules were selected as follows: The inhibitors were ranked according to their K_i values, subsequently; every fifth compound was selected for the test set starting from the high-potency end. This selection considers the fact that the test molecules must represent a range of biological activities similar to that of the training set. The selected test compounds were: **3**, **8**, **13**, **17**, **25**, **31**, **34**, and **39** for QSAR modeling (numbers are as in Fig. 2 and Table A of Supporting Information).

The logarithm of measured $1/K_i$ (μM) values was used in QSAR, thus correlating the data linear to the free energy change.

The chemical structures of the inhibitors were imported into CERIUS2 as standard 3D single conformer representations in

SD format. Subsequently, different descriptor groups were calculated for each compound using the C2.DESCRIPTOR module of CERIUS2. The calculated descriptors included various simple and valence connectivity indices, electro-topological state indices and other molecular descriptors, such as Charge, Apol, Dipole-mag, RadOfGyration, area, MW, V_m, density, PMI-mag, Rotlbonds, Hbond acceptor, and Hbond donor.[52] Furthermore, the training compounds were fitted (using the Best-fit option in CATALYST) against the representative pharmacophores and their fit values were added as additional descriptors. The fit value for any compound is obtained automatically via eq. (4).[35]

Genetic function approximation (GFA) was used to search for the best possible QSAR regression equation capable of correlating the variations in biological activities of the training compounds with variations in the generated descriptors, that is, multiple linear regression modeling (MLR). GFA techniques rely on the evolutionary operations of "crossover and mutation" to select optimal combinations of descriptors (i.e., chromosomes) capable of explaining bioactivity variation among training compounds from a large pool of possible descriptor combinations, that is, chromosomes population. However, to avoid overwhelming GFA-MLR with large number of poor descriptor populations, we removed lowest-variance descriptors (20%) before QSAR analysis. Each chromosome is associated with a fitness value that reflects how good it is compared with other solutions. The fitness function used herein is based on Friedman's "lack-of-fit" (LOF).[52]

Our preliminary diagnostic trials suggested the following optimal GFA parameters: Explore linear, quadratic, and spline equations at mating and mutation probabilities of 50%; population size = 500; number of genetic iterations = 30,000, and lack-of-fit (LOF) smoothness parameter = 1.0. However, to determine the optimal number of explanatory terms (QSAR descriptors), it was decided to scan and evaluate all possible QSAR models resulting from three to eight explanatory terms.

All QSAR models were validated using leave one-out cross-validation (r^2_{LOO}), bootstrapping (r^2_{BS}), and predictive r^2 (r^2_{PRESS}) calculated from the test set. The predictive r^2_{PRESS} is defined as:[52]

$$r^2_{PRESS} = SD - PRESS/SD \quad (5)$$

Where SD is the sum of the squared deviations between the biological activities of the test set and the mean activity of the training set molecules, PRESS is the squared deviations between predicted and actual activity values for every molecule in the test set.

Addition of Shape Constraints

Our QSAR-based pharmacophoric hypotheses were complemented with ligand-shapes and were used as 3D search queries. The shape components were introduced using the CatShape module of CATALYST.[53] Each pharmacophore was mapped against the most potent training inhibitor **12**. The best fitted conformers were used to generate shape constraints (with tolerance values ranging from 95–105%) that were subsequently merged with the corresponding pharmacophores.[35]

In Silico Screening for New β-D-Galactosidase Inhibitors

Each shape-complemented pharmacophore model was used as 3D search query to screen the National Cancer Institute list (NCI, 238,819 compounds) and our in-house built database of known drugs and agrochemicals (DAC, 3005 compounds). *In silico* screening was performed using the "Best Flexible Database Search" option implemented within CATALYST. The NCI hits were subsequently filtered based on Lipinski's and Veber's rules.[54,55] However, DAC hits were processed without subsequent postfiltering. Surviving hits were fitted against each corresponding pharmacophore model using the "best fit" option implemented within CATALYST. The fit values together with the relevant molecular descriptors of each hit were substituted in the corresponding QSAR equations. The highest ranking molecules based on QSAR predictions were acquired and tested *in vitro*.

In Vitro Experimental Studies

Materials

β-D-galactosidase (EC 3.2.1.23) from *Escherichia coli* and its corresponding substrate *p*-nitrophenyl-β-D-galactopyranoside were purchased from Sigma-Aldrich company. NCI hits were kindly donated from the NCI.

Preparation of Substrate and Enzyme Solution

p-nitrophenyl-β-D-galactopyranoside aqueous solution (4 mM) was prepared in phosphate buffer (pH = 7) as substrate solution for β-D-galactosidase. Although β-D-galactosidase was dissolved in phosphate buffer (pH = 7) to obtain final solution of 0.05 units/μL.

Preparation of Hit Compounds for In Vitro Assay

The tested compounds were initially dissolved in DMSO to yield stock solutions of 0.01 M. Subsequently, they were diluted to the required concentrations (7, 10, 40, and 100 μM) with the buffer solution.

β-D-Galactosidase Inhibition by Hit Compounds

The inhibitory potentials of the hit compounds against β-D-galactosidase were evaluated using spectrophotometric assay of the released *p*-nitrophenol. The β-D-galactosidase solution (0.25 units) was preincubated with 7, 10, 40, and 100 μM of each particular hit compound for 1 h at 25°C. The final concentration of DMSO did not exceed 1.0%. Subsequently, the substrate solution (0.4 mM) was added to the reaction mixture and the concentration of released *p*-nitrophenol was monitored at 405 nm every minute within 5-min period. Substrate concentrations were selected to approximate K_m values.[17–21] The percentage inhibition was determined from the residual activity for each compound by comparing the β-D-galactosidase activity with and without hit compound. The concentration required to give 50% inhibition (IC_{50}) was determined for the most potent hit.[31]

Results and Discussion

CATALYST enables automatic pharmacophore construction by using a collection of molecules with activities ranging over a

number of orders of magnitude. CATALYST pharmacophores (hypotheses) explain the variability of activity of the molecules with respect to the geometric localization of the chemical features present in the molecules used to build it. The pharmacophore model consists of a collection of features necessary for the biological activity of the ligands arranged in 3D space (e.g., hydrogen bond acceptors and donors, hydrophobic regions, etc.). Different hypotheses were generated for a series of β-D-galactosidase inhibitors. The biological activities of the training compounds spanned around 6.0 orders of magnitude. Genetic algorithm and multiple linear regression statistical analysis were subsequently used to select an optimal combination of complementary pharmacophores capable of explaining bioactivity variations among all collected inhibitors.

Data Mining and Conformational Coverage

The literature was extensively surveyed to identify as many reported structurally diverse β-D-galactosidase inhibitors as possible (**1–41**, Supporting Information Fig. 2 and Table A). The 2D structures of the inhibitors were imported into CATALYST and converted automatically into plausible 3D single conformer representations. The resulting single conformer 3D structures were used as starting point for conformational analysis and in the determination of various molecular descriptors for QSAR modeling.

The conformational space of each inhibitor was extensively sampled using the poling algorithm used within the CONFIRM module of CATALYST.[40] Efficient conformational coverage guarantees minimum conformation-related noise during pharmacophore generation and validation stages.[56]

Exploration of the Pharmacophoric Space of β-D-Galactosidase Inhibitors

CATALYST-HYPOGEN enables automatic pharmacophore construction by using a collection of at least 16 molecules with bioactivities spanning over 3.5 orders of magnitude.[37,39–43] HYPOGEN implements an optimization algorithm that evaluates large number of potential models for a particular target through fine perturbations to hypotheses that survived the subtractive and constructive phases.[39] The extent of the evaluated space is reflected by the configuration (Config.) cost calculated for each modeling run. It is generally recommended that the Config. cost of any HYPOGEN run not to exceed 17 (corresponding to 2^{17} hypotheses to be assessed by CATALYST) to guarantee thorough analysis of all models.[40]

The size of the investigated pharmacophoric space is a function of training compounds, selected input chemical features, and other CATALYST control parameters.[40] Restricting the extent of explored pharmacophoric space should improve the efficiency of optimization via allowing effective evaluation of limited number of pharmacophoric models. On the other hand, extensive restrictions imposed on the pharmacophoric space may reduce the possibility of discovering optimal pharmacophoric hypotheses, as they may occur outside the "boundaries" of the pharmacophoric space.

Therefore, we decided to explore the pharmacophoric space of β-D-galactosidase inhibitors under reasonably imposed "boundaries" through eight HYPOGEN automatic runs and a carefully selected training set (See Data Set of β-D-Galactosidase Inhibitors under Materials and Methods). The training set was selected in such away to guarantee maximal 3D diversity and continuous bioactivity spread over more than 3.5 logarithmic cycles. Furthermore, they were selected in such a way that differences in their inhibitory bioactivities are primarily attributable to the presence or absence of pharmacophoric features [e.g., hydrogen bond acceptor (HBA) or hydrogen bond donor (HBD) or hydrophobic (Hbic) or ring aromatic (RingArom)] rather than steric shielding and/or bioactivity-enhancing or -reducing auxiliary groups (e.g., electron donating or withdrawing groups). Special emphasis were given to the 3D diversity of the most active compounds in the training set due to their significant influence on the extent of the evaluated pharmacophore space through the Constructive Phase of HYPOGEN algorithm.

Guided by our reasonably restricted pharmacophore exploration concept, we restricted the software to explore pharmacophoric models containing from zero to three HBA, HBD, Hbic, and RingArom features and from zero to one PI features, that is, instead of the default range of zero to five (Supporting Information Table B). Furthermore, we instructed HYPOGEN to explore only four- and five-featured pharmacophores, that is, ignore models of lesser number of features (Supporting Information Table B). The later restriction has the advantage of narrowing the investigated pharmacophoric space and representing the feature-rich nature of known β-D-galactosidase inhibitors.

In each run, the resulting binding hypotheses were automatically ranked according to their corresponding "total cost" value, defined as the sum of error cost, weight cost, and configuration cost.[35,39–43] Error cost provides the highest contribution to total cost and it is directly related to the capacity of the particular pharmacophore as 3D-QSAR model, that is, in correlating the molecular structures to the corresponding biological responses.[35,39–43] HYPOGEN also calculates the cost of the null hypothesis, which presumes that there is no relationship in the data and that experimental activities are normally distributed about their mean. Accordingly, the greater the difference from the null hypothesis cost (residual cost, Table 1), the more likely that the hypothesis does not reflect a chance correlation.[35,39–43]

The pharmacophore models were further validated using Cat.Scramble, which is based on Fisher's randomization test.[35,50] In this test, the biological data and the corresponding structures are scrambled several times, and the software is challenged to generate pharmacophoric models from the randomized data. The confidence in the parent hypotheses (i.e., generated from unscrambled data) is lowered proportional to the number of times the software succeeds in generating binding hypotheses from scrambled data of apparently better cost criteria than the parent hypotheses.

Eventually, 80 pharmacophore models emerged from eight automatic HYPOGEN runs. All models illustrated Fisher's randomization test-based confidence levels $\geq$95%. Clearly from Table 1, all models shared comparable features and acceptable statistical success criteria. Emergence of several comparable pharmacophore models suggests the ability of β-D-galactosidase

Table 1. The Performances of the Best Representative Binding Hypotheses Generated for β-D-Galactosidase Inhibitors.

Run[a]	Hypotheses[b]	Pharmacophoric features in generated hypotheses	Total cost	Cost of null hypothesis	Residual cost[c]	R[d]	Global R[e]
1	**3**[f,g]	**HBA, 2xHBD, Hbic**	**106.4**	**231.6**	**125.2**	**0.93**	**0.69**
	8[g]	**2xHBA, HBD, Hbic**	**107.5**	**231.6**	**124.1**	**0.92**	**0.75**
	10[g]	**HBA, HBD, Hbic, RingArom**	**107.7**	**231.6**	**123.9**	**0.93**	**0.56**
2	2	HBA, HBD, Hbic, RingArom	103.4	231.6	128.2	0.93	0.58
	6	2xHBA, Hbic, RingArom	104.9	231.6	126.7	0.93	0.65
	8	2x HBD, Hbic, RingArom	105.2	231.6	126.4	0.94	0.59
3	2	2xHBA, Hbic, RingArom	104.7	231.6	126.9	0.92	0.52
	4	2xHBD, Hbic, RingArom	105.6	231.6	126.0	0.93	0.57
	10	HBA, 2xHBD, Hbic	107.0	231.6	124.6	0.92	0.59
4	5	HBA, 2xHBD, Hbic	106.3	231.6	125.3	0.92	0.72
	6	3xHBA, Hbic	107.3	231.6	124.3	0.92	0.73
	10	HBA, 2xHBD, Hbic	108.6	231.6	123.0	0.93	0.81
5	**4**[g]	**2xHBA, HBD, Hbic**	**119.6**	**493.0**	**373.4**	**0.98**	**0.43**
	5[g]	**3xHBA, Hbic**	**121.6**	**493.0**	**371.4**	**0.97**	**0.43**
	6	HBA, 2xHBD, Hbic	122.1	493.0	370.9	0.97	0.44
6	3	HBA, 2xHBD, Hbic	119.0	493.0	374.0	0.98	0.44
	6	2xHBA, HBD, Hbic	122.6	493.0	370.4	0.98	0.42
	7	3xHBA, Hbic	125.2	493.0	367.8	0.97	0.43
7	5	3xHBA, Hbic	135.4	493.0	357.6	0.93	0.69
	7	2xHBD, Hbic, RingArom	136.1	493.0	356.9	0.93	0.54
	10	2xHBA, HBD, Hbic	136.8	493.0	356.2	0.93	0.69
8	2	2xHBD, Hbic, RingArom	130.5	493.0	362.5	0.94	0.60
	3	HBA, 2xHBD, Hbic	131.6	493.0	361.4	0.94	0.65
	7	2xHBD, Hbic, RingArom	133.1	493.0	359.9	0.94	0.66

[a]Correspond to runs in Table B, Supporting Information.
[b]Best models from their respective clusters, as judged based on their Global R values (see below).
[c]The difference between the total cost and the cost of the corresponding null hypotheses.
[d]The correlation coefficients between bioactivity estimates and bioactivities of the training set compounds.
[e]Global R is the correlation coefficient calculated based on the linear regression between the fit values of collected inhibitors [**1–41**, Fig. 2 and Table A (Supporting Information)] against the particular pharmacophore hypothesis [using the "best fit" option and eq. (4)] and their respective β-D-galactosidase bioactivities.
[f]Rank of each hypothesis in each particular run by CATALYST.
[g]Bolded pharmacophores emerged in the best QSAR equations (bolded).

inhibitors to assume multiple pharmacophoric binding modes within the binding pocket for each enzyme. Therefore, it is quite challenging to select any particular pharmacophore hypothesis as a sole representative of the binding process.

QSAR Modeling

Pharmacophoric hypotheses are important tools in drug design and discovery as they provide excellent insights into ligand-macromolecule recognition; moreover, they can be used as 3D search queries to look for new biologically interesting scaffolds. However, their predictive value as 3D-QSAR models is usually limited by steric shielding and auxiliary groups (electron-donating and -withdrawing moieties).[42] This point combined with the fact that pharmacophore modeling of β-D-galactosidase inhibitors furnished many binding hypotheses of comparable statistical criteria (Table 1), prompted us to use classical QSAR analysis to search for the best combination of pharmacophore(s) and other 2D descriptors capable of explaining bioactivity variation across the whole list of collected inhibitors **1–41** (Supporting Information Fig. 1 and Table A). We used genetic function approximation and multiple linear regression QSAR (GFA-MLR-QSAR) analysis to search for an optimal QSAR equation(s).

GFA-MLR-QSAR selects optimal descriptor combinations based on the Darwinian concept of genetic evolution whereby the statistical criteria of regression models from different descriptor combinations (chromosomes) are used as fitness criteria.[52] GFA-MLR-QSAR analysis was used to explore various combinations of pharmacophores and other structural descriptors and to evaluate their statistical properties as predictive QSAR models.

The fit values obtained by mapping the representative hypotheses against all collected inhibitors **1–41** (Supporting Information Fig. 1 and Table A) were enrolled as independent variables (genes) in a cycle of GFA-MLR-QSAR analysis over 30,000 iterations using Friedman's LOF fitness criterion.[52,57] However, as it is essential to access the predictive power of the resulting QSAR models on an external set of inhibitors, we randomly selected eight inhibitors (Supporting Information Fig. 1 and Table A) and used them as external test set for validating the QSAR models (i.e., r^2_{PRESS}). Moreover, all QSAR models

were cross-validated automatically using the leave-one-out cross-validation in CERIUS2.[52,57]

Equations 6–10 show the details of the optimal QSAR models. Figures A–E (Supporting Information) show the corresponding scatter plots of experimental versus calculated bioactivities.

$$\begin{aligned}\text{Log } (1/K_i) &= -6.10 + 0.01\ (\text{Hypo3}/1)^2 \\ &\quad + 3.60\ (\ ^3\chi^C)^2 - 0.17\ (\text{SaasC})^2 \\ r^2_{33} &= 0.85, \text{F} - \text{statistic} = 53.72,\ r^2_{\text{LOO}} = 0.77, \\ &\qquad r^2_{\text{BS}} = 0.85,\ r^2_{\text{PRESS}} = 0.86 \quad (6)\end{aligned}$$

$$\begin{aligned}\text{Log } (1/K_i) &= -3.03 + 0.34\ (\text{Hypo8}/1 - 4.91) \\ &\quad + 11.98\ (\ ^3\chi^C - 0.98) - 1.33\ (\text{SaasC} - 0.70) \\ r^2_{33} &= 0.88, \text{F} - \text{statistic} = 65.66,\ r^2_{\text{LOO}} = 0.84, \\ &\qquad r^2_{\text{BS}} = 0.88,\ r^2_{\text{PRESS}} = 0.89 \quad (7)\end{aligned}$$

$$\begin{aligned}\text{Log } (1/K_i) &= -6.23 + 0.01\ (\text{Hypo10}/1)^2 - 4\times10^{-6}\ (\text{PMImag})^2 \\ &\quad + 14.85\ (\ ^3\chi^C)^2 - 0.02\ [\text{JursPPSA1} - 232.08] \\ &\quad + 1.52[\text{ShadowXlength} - 9.97] \\ &\quad + 4.31[\text{JursRNCS} - 4.10] \\ r^2_{33} &= 0.95,\ \text{F} - \text{statistic} = 80.90, r^2_{\text{LOO}} = 0.89, \\ &\qquad r^2_{\text{BS}} = 0.95, r^2_{\text{PRESS}} = 0.64 \quad (8)\end{aligned}$$

$$\begin{aligned}\text{Log } (1/K_i) &= -2.26 + 0.50\ [\text{Hypo4}/5 - 11.51] \\ &\quad + 26.21\ [\ ^3\chi^C - 0.46] + 1.0\ [2 - \text{AtypeH46}] \\ &\quad - 0.99\ [1.89 - \text{DipoleMag}] - 1.72\ [\text{Kappa2AM} \\ &\quad - 3.55] - 0.64\ [13.8 - \text{ShadowXlength}] \\ r^2_{33} &= 0.96,\ \text{F} - \text{statistic} = 91.03,\ r^2_{\text{LOO}} = 0.92, \\ &\qquad r^2_{\text{BS}} = 0.96,\ r^2_{\text{PRESS}} = 0.65 \quad (9)\end{aligned}$$

$$\begin{aligned}\text{Log } (1/K_i) &= -7.40 + 4.00\times10^{-3}\ (\text{Hypo5}/5)^2 \\ &\quad + 0.05\ (\text{ShadowXlength})^2 \\ &\quad - 4.9 \times 10^{-5}\ (\text{JursPPSA1})^2 + 19.01\ (\ ^3\chi^C)^2 \\ &\quad - 4 \times 10^{-6}\ (\text{PMImag})^2 - 0.06\ (\text{DipoleMag})^2 \\ &\quad - 0.09\ (\text{SsssCH})^2 \\ r^2_{33} &= 0.93,\ \text{F} - \text{statistic} = 49.98,\ r^2_{\text{LOO}} = 0.90, \\ &\qquad r^2_{\text{BS}} = 0.94,\ r^2_{\text{PRESS}} = 0.74 \quad (10)\end{aligned}$$

Where r^2_{33} is the correlation coefficient against 33 training compounds, r^2_{LOO} is the leave-one-out correlation coefficient, r^2_{BS} is the bootstrapping regression coefficient, and r^2_{PRESS} is the predictive r^2 determined for the eight test compounds.[52,57,58] Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5 represent the fit values of the training compounds against the five corresponding pharmacophores as calculated from eq. (4). $^3\chi^C$ is the third-order cluster connectivity index. Kappa2AM is one of the Kier and Hall Kappa shape indices encoding for the count of molecular branching. ShadowXlength is a shape descriptor that represents the *X*-axis component of the molecular shadow. JursPPSA1 and JursRNCS are Jurs descriptors. JursPPSA1 encodes for the Partial Positive Surface Area defined as the sum of solvent-accessible surface areas of all positively charged atoms, while JursRNCS encodes for the relative negative charge surface area calculated as: SaasC is electrotopological sum descriptor for monosubstituted aromatic carbon atoms. PMImag (magnitude of the principle moments of inertia) encodes for information about the spatial distribution of mass and molecular rotational properties. DipoleMag (magnitude of the dipole moment) is a 3D electronic descriptor that indicates the strength and orientation behavior of a molecule in an electrostatic field.[52,59]

Several descriptors emerged in eqs. (7)–(9) in spline format. The spline terms used herein are "truncated power splines" and are denoted by bolded brackets ([ ]). For example, $[f(x) - a]$ equals zero if the value of $(f(x) - a)$ is negative; otherwise, it equals $(f(x) - a)$.[52]

Figures 3–7 show the pharmacophoric features of the binding pharmacophores in QSAR eqs. (6)–(10) and how they map the most potent training compound **12** and the most potent discovered hit **42**, whereas Table C (Supporting Information) shows the corresponding *X*, *Y*, and *Z* coordinates of the pharmacophores.

Interestingly, Hypo8/1 and Hypo4/5 emerged in eqs. (7) and (9) in spline format, indicating that each binding mode contributes to ligand/β-D-galactosidase affinity only if the fit value of the particular ligand exceeds the corresponding spline threshold. For example, the ability of a certain ligand to map Hypo8/1 will impact its actual affinity to β-D-galactosidase only if its fit value exceeds 4.91, which correspond to the spline intercept associated with this pharmacophore in eq. (7). Because the two spline cutoffs (of both pharmacophores) resemble moderate to high overall ligand/pharmacophore mapping (the maximum value is 14.0), it appears that ligand binding to β-D-galactosidase is sensitive to misalignments among the attracting moieties within the complex such that lowering the fits value below 4.91 and 11.51 for Hypo8/1 and Hypo4/5, respectively, nullifies any affinity gains from mapping the pharmacophores.

A similar trend is also seen in eqs. (6), (8), and (10), albeit in quadratic format. Emergence of quadratic terms corresponding to Hypo3/1, Hypo10/1, and Hypo5/5 in eqs. (6), (8), and (10), suggests that minor misalignment among complementary moieties within Ligand/β-D-galactosidase complex causes drastic reduction in affinities.

Extreme sensitivity to misalignment among complementary groups within ligand/macromolecule complexes is probably related to the fact that most of our training compounds are transition state analogues, which require stringent three-dimensional distribution of binding features to complement, and therefore, fit the enzymatic catalytic site during transition state.

Emergence of connectivity ($^3\chi^C$), shape (Kappa2AM), electrotopological state indices (SaasC), moments of inertia (PMImag), dipole moment (DipoleMag), and shadow descriptors

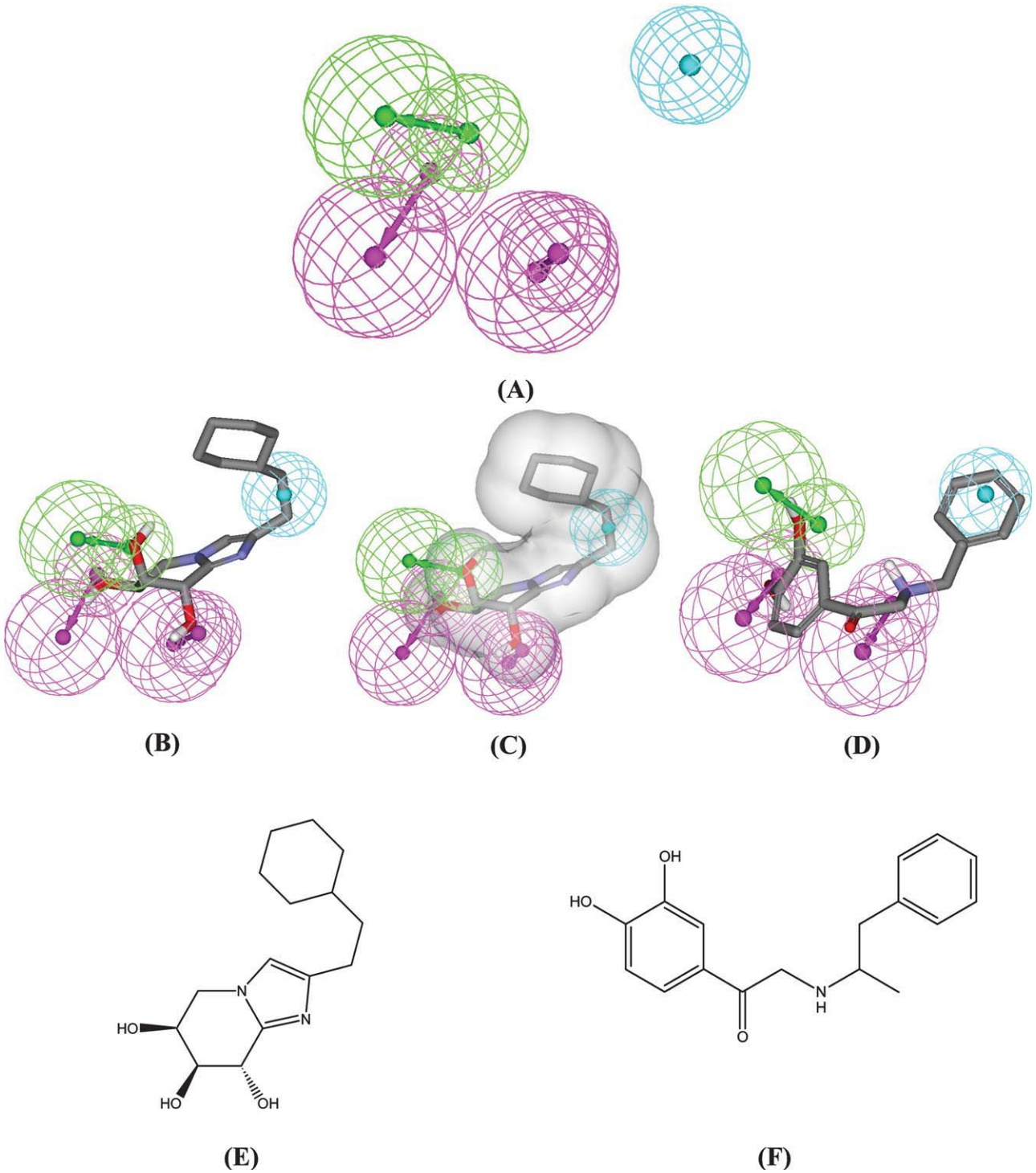

Figure 3. (A) The pharmacophoric features of the binding model Hypo3/1 (Table 1), HBA as green vectored spheres, HBD as a violet vectored spheres, and Hbic as blue spheres; (B) Hypo3/1 mapping the most potent training inhibitor **12** (Supporting Information Fig. 2 and Table A), K_i = 0.002 μM); (C) with shape constraints; (D) Hypo3/1 fitted against potent NCI hit **42** (Fig. 10, IC50 = 2.4 μM); (E) and (F) the chemical structures of **12** and **42**, respectively.

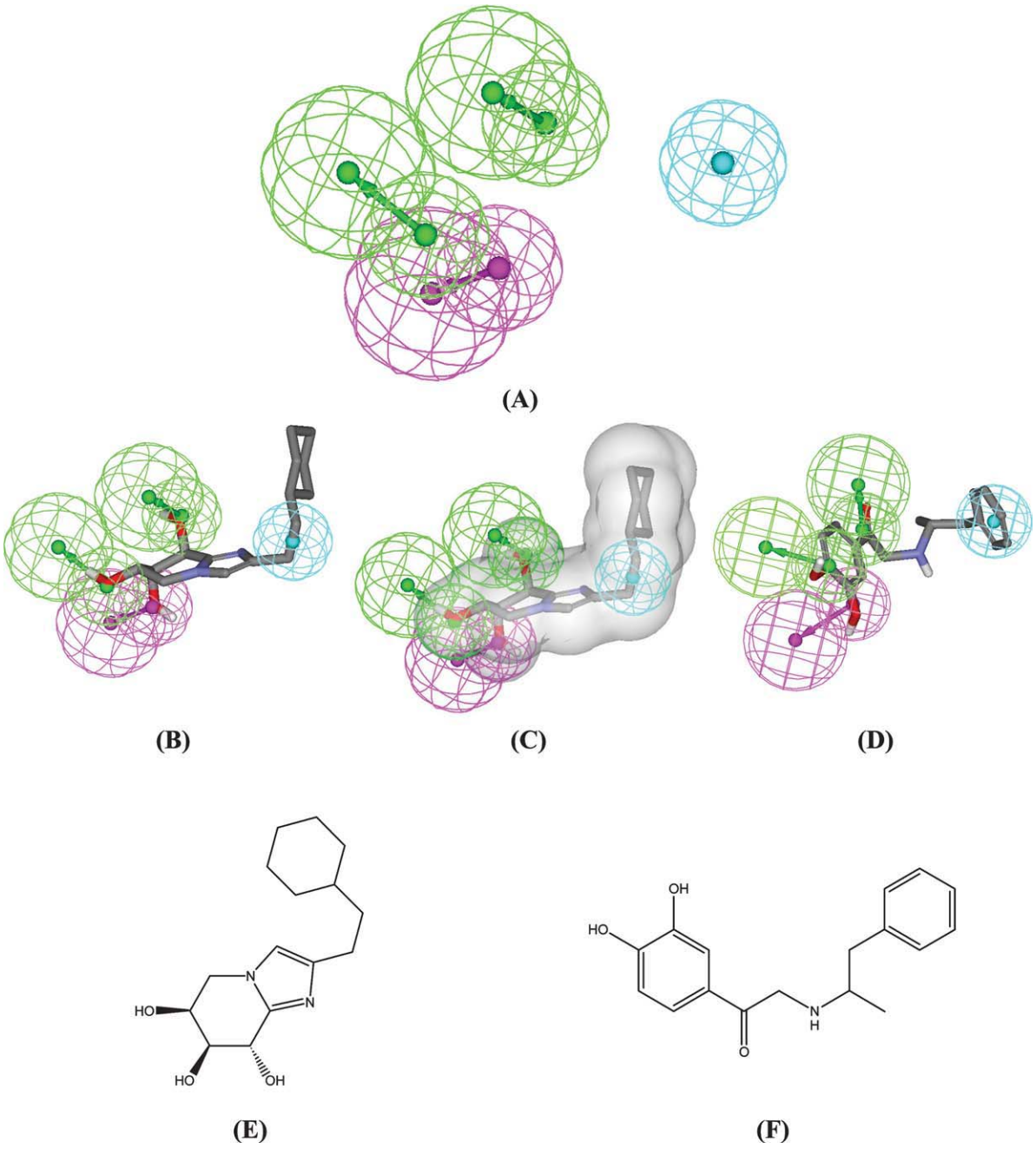

Figure 4. (A) The pharmacophoric features of the binding model Hypo8/1 (Table 1), HBA as green vectored spheres, HBD as a violet vectored spheres and Hbic as blue spheres; (B) Hypo8/1 mapping the most potent training inhibitor **12** (Supporting Information Fig. 2 and Table A), K_i = 0.002 μM); (C) with shape constraints; (D) Hypo8/1 fitted against potent NCI hit **42** (Fig. 10, IC50 = 2.4 μM); (E) and (F) the chemical structures of **12** and **42**, respectively.

(ShadowXlength) in eqs. (6)–(10) is suggestive of certain role played by ligands' topologies in the binding process. However, the information content of these descriptors is quite obscure. On the other hand, emergence of JursPPSA1 and JursRNCS in eqs. (8) and (10) combined with negative and positive regression coefficients, respectively, suggests direct relationship connecting ligand/β-D-galactosidase affinity and ligands' negative surface charges, which points to certain complementary electrophilic (positive) area(s) within the binding site, probably the ionized imidazoles of HIS150 and HIS342 (Figure 9a).

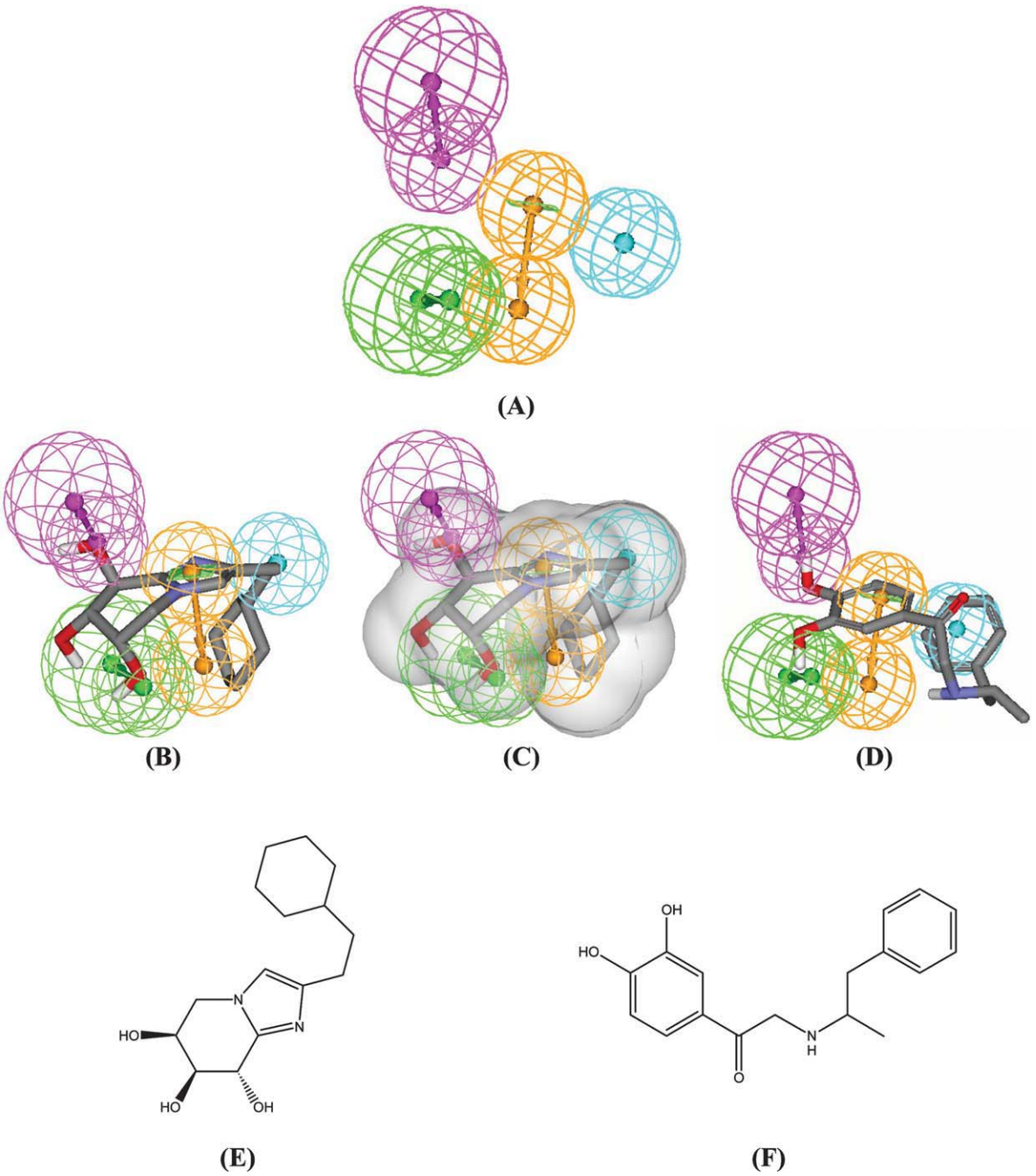

Figure 5. (A) The pharmacophoric features of the binding model Hypo10/1 (Table 1), HBA as green vectored spheres, HBD as a violet vectored spheres, Hbic as blue spheres and RingArom as orange vectored spheres; (B) Hypo10/1 mapping the most potent training inhibitor **12** (Supporting Information Fig. 2 and Table A), K_i = 0.002 μM); (C) with shape constraints; (D) Hypo10/1 fitted against potent NCI hit **42** (Fig. 10, IC50 = 2.4 μM); (E) and (F) the chemical structures of **12** and **42**, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ROC Curve Analysis and Shape Constraints

To further validate the resulting models (both QSAR and pharmacophores), we subjected our QSAR-selected pharmacophores to ROC analysis (see the experimental part of Supporting Information).[60–65] The number of actives in the ROC testing list was limited to five, whereas the decoys were extended to 352 (i.e., each active compound was challenged with 70 decoys) to provide a proper challenge for the pharmacophoric models, particularly as they are required to select sterically

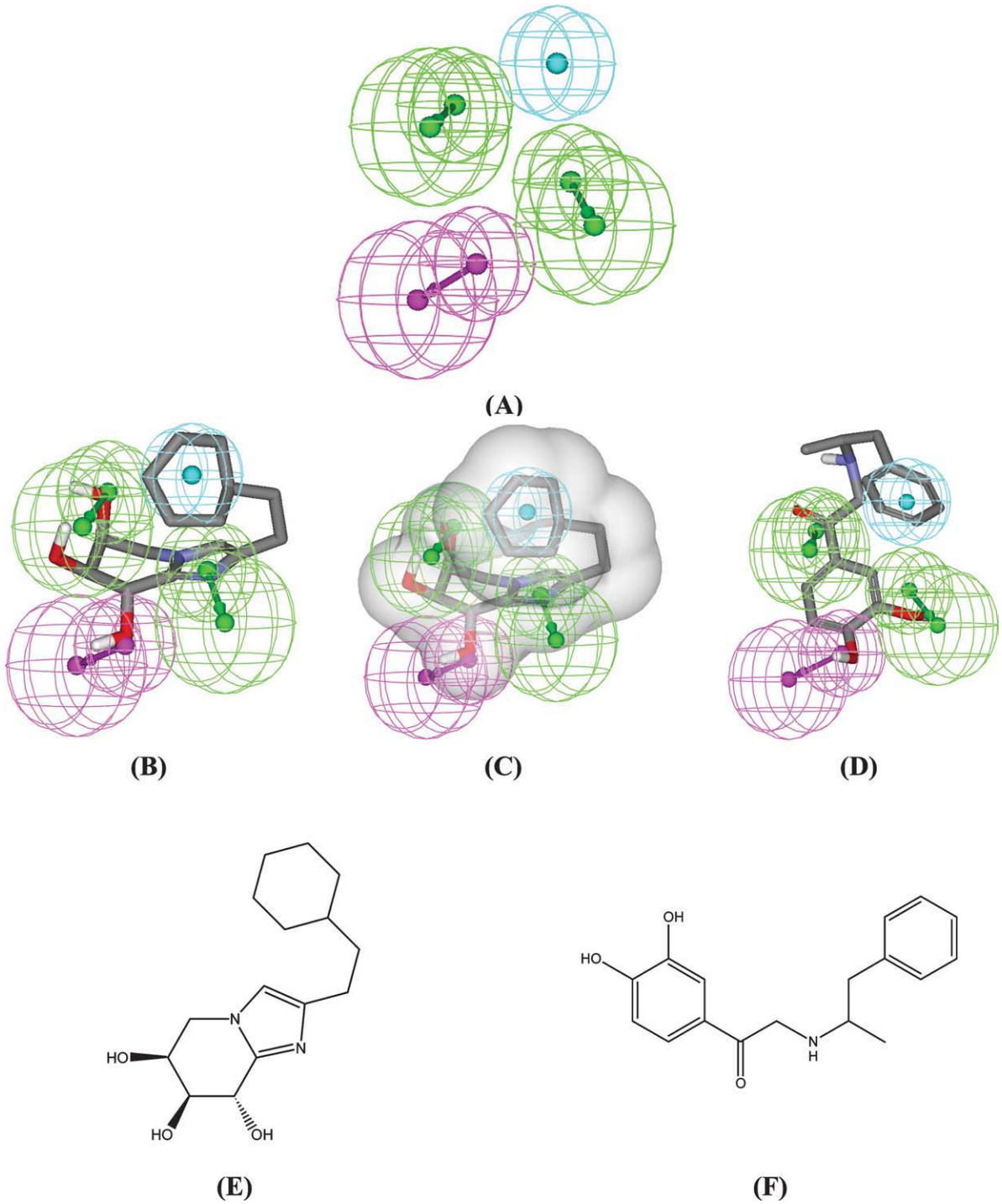

Figure 6. (A) The pharmacophoric features of the binding model Hypo4/5 (Table 1), HBA as green vectored spheres, HBD as a violet vectored spheres, and Hbic as blue spheres; (B) Hypo4/5 mapping the most potent training inhibitor **12** (Supporting Information Fig. 2 and Table A), K_i = 0.002 μM); (C) with shape constraints; (D) Hypo4/5 fitted against hit **42** (Fig. 10, IC50 = 2.4 μM); (E) and (F) the chemical structures of **12** and **42**, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

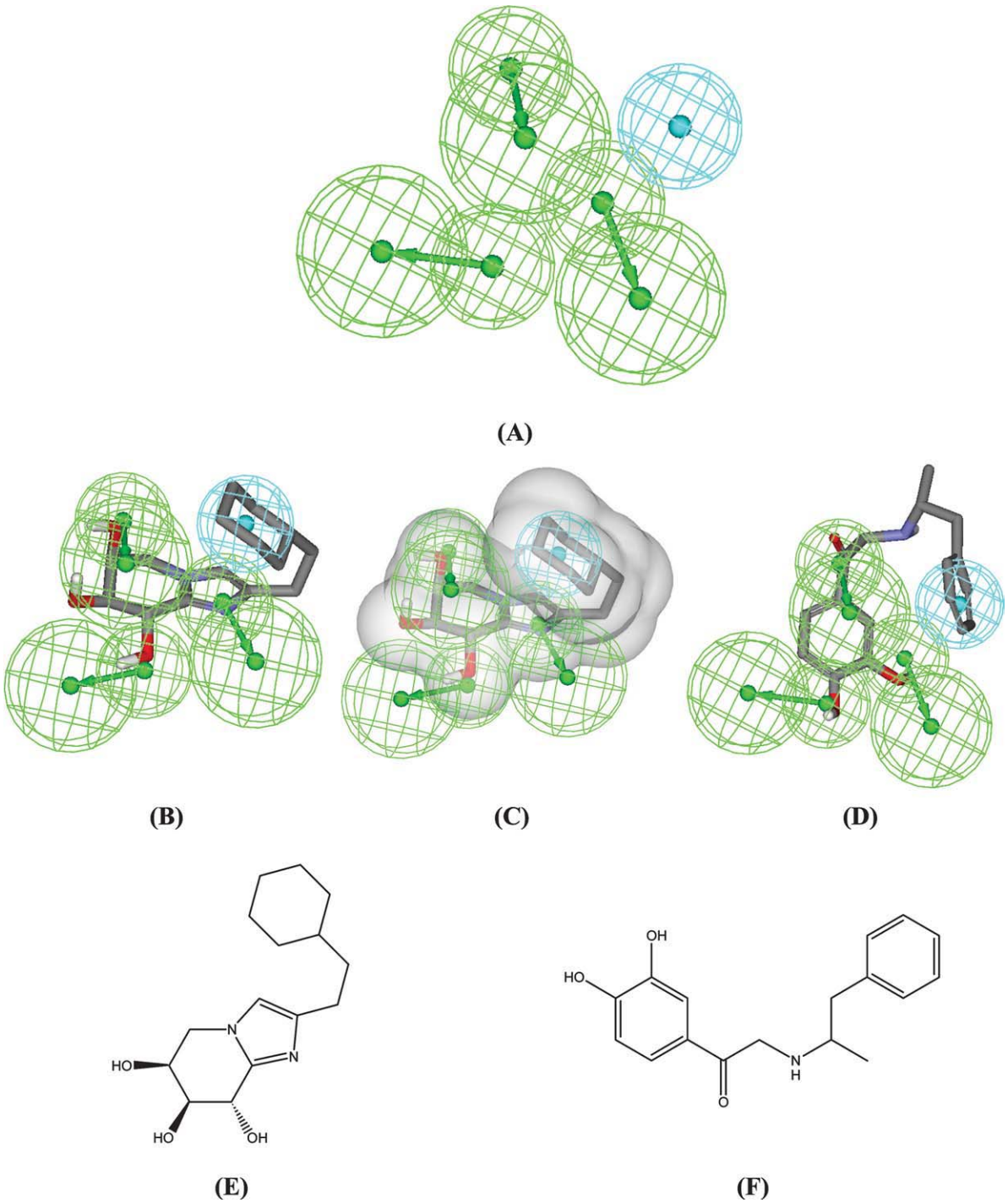

Figure 7. (A) The pharmacophoric features of the binding model Hypo5/5 (Table 1), HBA as green vectored spheres and Hbic as blue spheres; (B) Hypo5/5 mapping the most potent training inhibitor **12** (Supporting Information Fig. 2 and Table A), K_i = 0.002 μM); (C) with shape constraints; (D) Hypo5/5 fitted against hit **42** (Fig. 10, IC50 = 2.4 μM); (E) and (F) the chemical structures of **12** and **42**, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 2. Performance of QSAR-Selected Pharmacophores and Their Shape Complemented Versions as 3D Search Queries.

Pharmacophore model	ROC[a]-AUC[b]	ACC[c]	SPC[d]	TPR[e]	FNR[f]
Hypo3/1	64.36	98.30	98.96	0.6	0.01
Hypo8/1	60.35	98.30	98.96	0.6	0.01
Hypo10/1	60.00	98.30	98.96	0.6	0.01
Hypo4/5	46.37	98.30	98.96	0.6	0.01
Hypo5/5	43.81	98.30	98.96	0.6	0.01
Shape-Complemented Hypo3/1	86.07	98.31	98.97	0.6	0.01
Shape-Complemented Hypo8/1	82.21	98.31	99.66	0.2	0.003
Shape-Complemented Hypo10/1	85.38	98.31	98.97	0.6	0.01
Shape-Complemented Hypo4/5	72.48	98.31	98.97	0.6	0.01
Shape-Complemented Hypo5/5	69.10	98.31	98.97	0.6	0.01

[a]ROC: receiver operating characteristic.
[b]AUC: area under the curve.
[c]ACC: overall accuracy.
[d]SPC: overall specificity.
[e]TPR: overall true positive rate.
[f]FNR: overall false negative rate.

and pharmacophorically challenging TSAs during *in silico* screening.

Table 2 and Figure 8 show the ROC results of our QSAR-selected pharmacophores. Clearly from the figure and table, all QSAR-selected models illustrated weak to mediocre overall performance with AUC values ranging from 43.81 to 64.36%.

To enhance the ROC profiles of the QSAR–selected models, we decided to decorate them with shape-constraints derived from the most potent training inhibitor **12** (k_i = 0.002 μM). Shape constraints encode for the degree of 3D spatial similarity between screened compounds and the template ligand used to build the shape limitations.[35,49,53] To generate merged shape-pharmacophore queries, a selected potent training compound **12** was first fitted against the corresponding pharmacophore model; thereafter, the best-fitted conformer of the inhibitor was used to generate shape constraints that were subsequently merged with the pharmacophore. Figures 3–7 show the shape-complemented versions of β-D-galactosidase pharmacophores.

The generated merged pharmacophore-shape queries were used as 3D search queries against the NCI, DAC databases. Table 3 shows the performances of our shape-complemented pharmacophore queries.

Fig. 8 and Table 2 show the ROC result of the shape-decorated versions of the QSAR-selected models. Clearly, the performance of shape-complemented models improved significantly as reflected by their ROC-AUC, which ranged from 69.10 to 86.07%.

The pronounced ROC enhancements upon addition of shape constraints points to the fact that β-D-galactosidase inhibition requires precisely tailored slow binding/transition state analogs (TSA). Such inhibitors must possess stringent steric 3D properties best encoded *via* shape constraints, which exclude molecules that lack the ability to fill the binding pocket in the same way achieved by the template molecule (i.e., from which the shape was derived).

Comparing Pharmacophore Models with Crystallographic Complex

To further emphasize the validity of our pharmacophore/QSAR modeling approach, we compared the crystallographic structure of β-D-galactosidase/ligand complex[66] (PDB codes: 2CEQ and 2CER, resolutions: 2.14 and 2.2 Å, respectively) with Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5. Figure 9 shows the chemical structure of the ligand and compares its β-D-galactosidase complex with the way it maps Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5 using rigid mapping, that is, fitting the ligand's bound state against corresponding pharmacophores without conformational adjustments.

Pharmacophore mapping against the five models suggests that two to three of the glucoimidazole hydroxyls are involved in hydrogen-bonding within the binding pocket, as in Figures 9c–9g, which seems to correlate with hydrogen-bonding interactions tying these hydroxyls with the side chains of GLU387, ASN205, HIS150, TRP433, and GLU432, as in Figure 9a. Furthermore, mapping the imidazol ring of the phenethyl-glucoimidazole against RingArom feature in Hypo10/1, and HBA features in Hypo4/5 and Hypo5/5 agrees nicely with π–π stacking involving this ring and the phenolic side chain of TYR322, and hydrogen-bonding with carboxylic side chain of GLU206, as in Figure 9a. Finally, the terminal phenyl ring of the phenethyl-glucoimidazole seems

Table 3. Number of Captured Compounds by Shape-Complemented β-D-Galactosidase Pharmacophores.

		Pharmacophore models[c]				
3D Database[a]	Post screening filtering[b]	Hypo3/1	Hypo8/1	Hypo10/1	Hypo4/5	Hypo5/5
NCI	Before	385	215	221	290	1690
	After	283	48	164	203	1250
DAC[d]		6	7	5	1	23

[a]NCI: National Cancer Institute list of available compounds (238,819 structures), DAC: the list of established drugs and agrochemicals (3005 structures).
[b]Postscreening filtering employing Lipinski's and Veber's rules. One Lipinski's violation was tolerated.
[c]The number of captured hits by the sterically refined versions of the pharmacophore models.
[d]This list of compounds was *in silico* scanned without post-screening filtering.

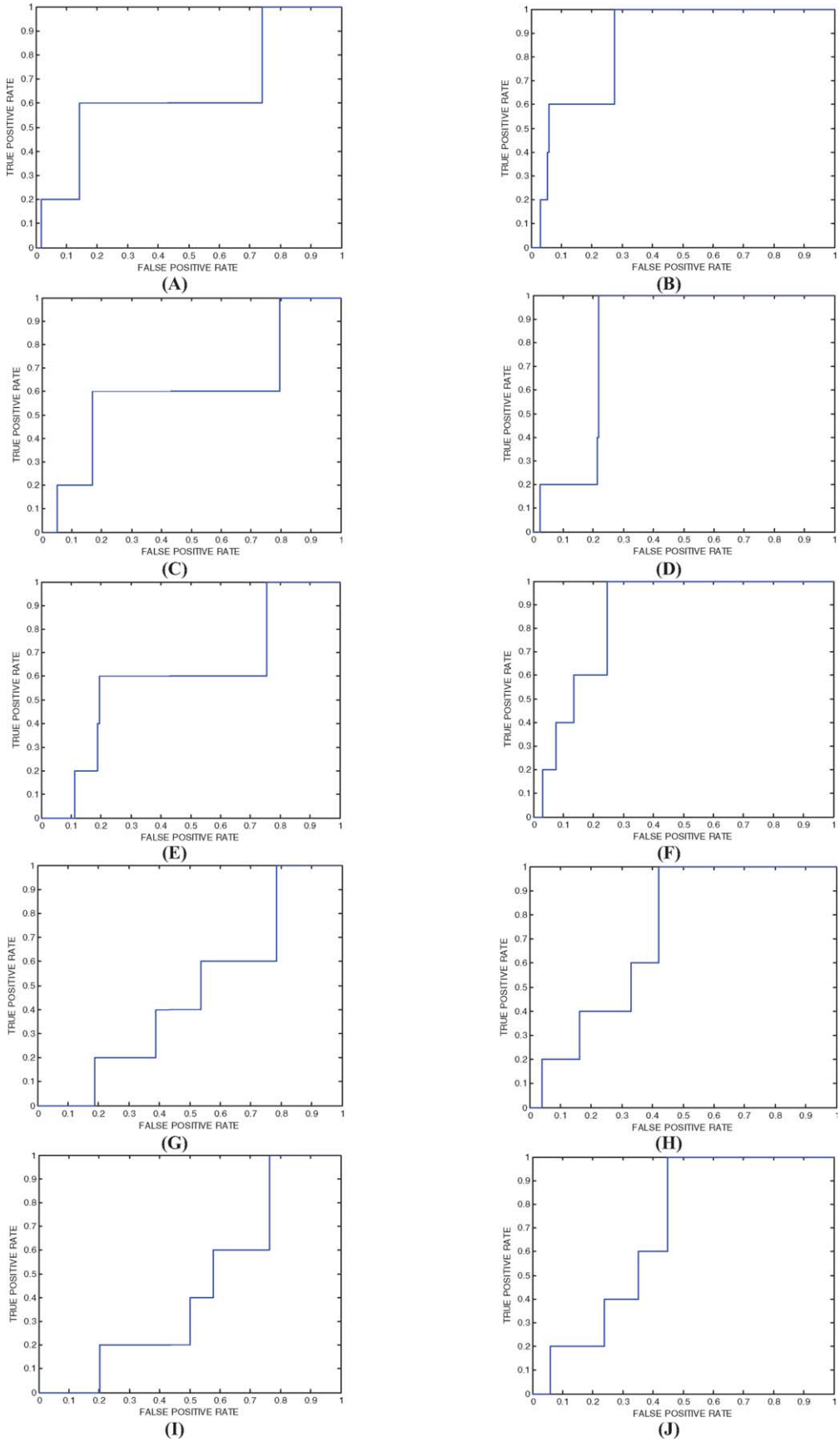

Figure 8. Receiver operating characteristic curves (ROCs) conducted for QSAR-selected models: (A) Hypo3/1, (B) shape-complemented Hypo3/1 (Shape-Hypo3/1), (C) Hypo8/1, (D) Shape-Hypo8/1, (E) Hypo10/1, (F) Shape-Hypo10/1, (G) Hypo4/5, (H) Shape-Hypo4/5, (I) Hypo5/5, and (J) Shape-Hypo5/5. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

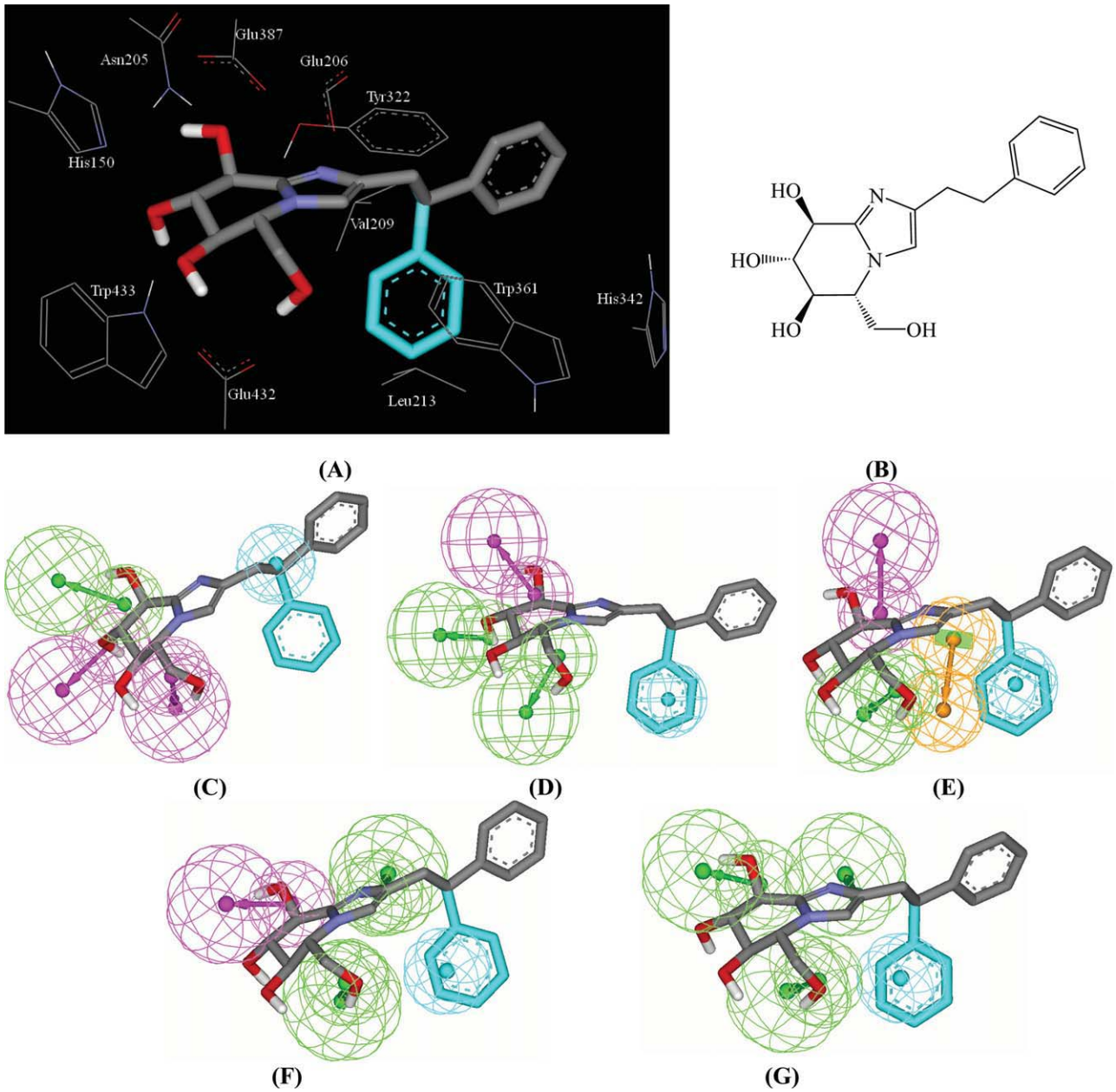

Figure 9. (A) Crystallographic structure of phenethylglucoimidazole cocrystallized with β-D-galactosidase (PDB code: 2CER and 2CEQ, resolutions: 2.2 and 2.14 Å, respectively), (B) the chemical structure of phenethyl-glucoimidazole, (C), (D), (E), (F), and (G) mapping the phenethyl-glucoimidazole against Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5, respectively. Blue-colored aromatic ring points to the dangling of this ring as reflected by the wide distribution of corresponding electron density. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to dangle in hydrophobic/aromatic pocket comprised of VAL209, TRP361, HIS342, and LEU213, as in Figure 9a, which correlates with hydrophobic features in the five pharmacophores positioned in the same region.

Clearly from the above discussion, the five models, that is, Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5, represent close binding modes assumed by the ligand within β-D-galactosidase. These models point to limited number of critical interactions required for high ligand/β-D-galactosidase affinity. In contrast, crystallographic complex reveal many bonding interactions without highlighting critical ones. Incidentally, Figures 9a only shows interactions corresponding to pharmacophoric features in

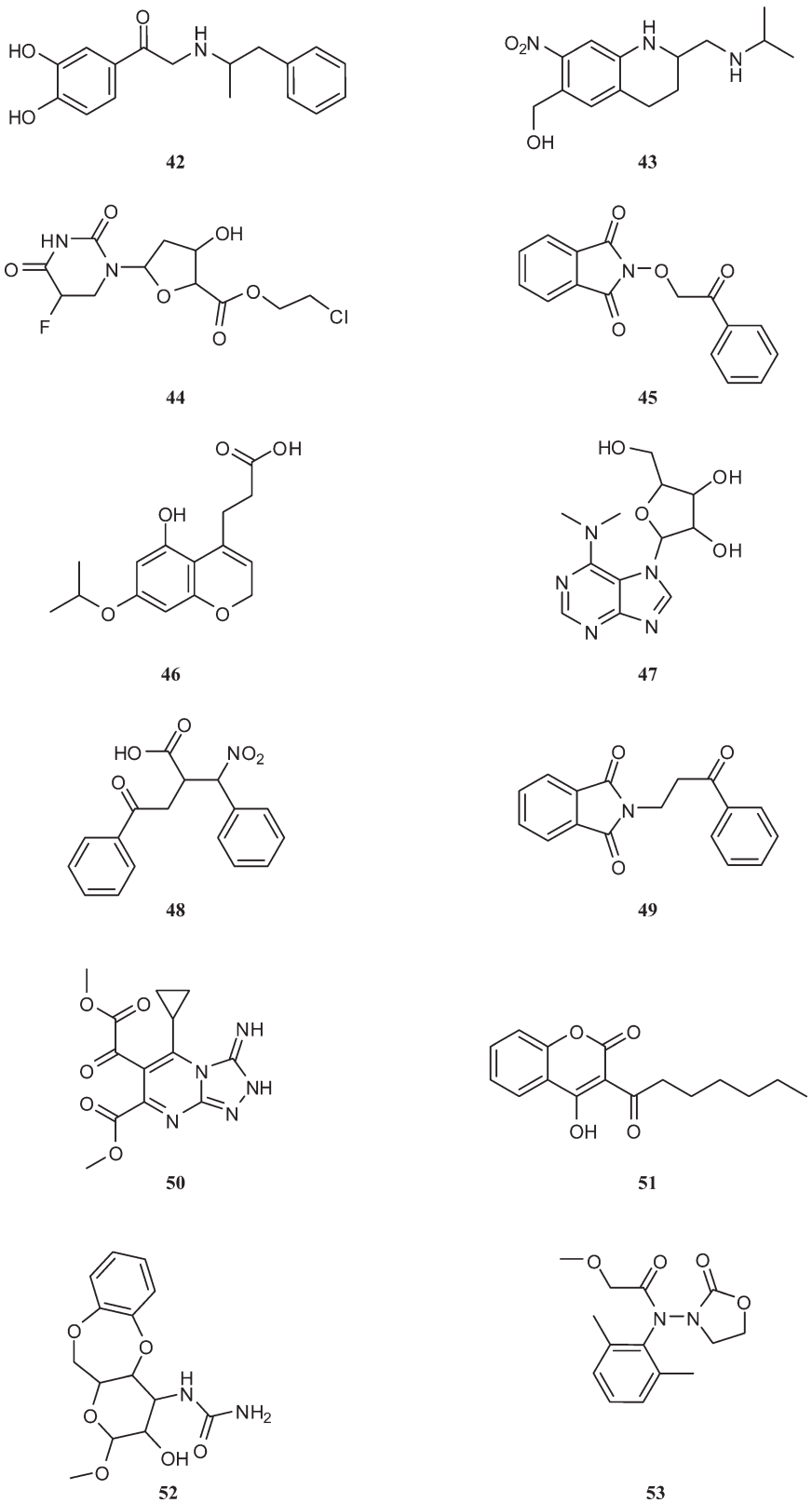

Figure 10. The chemical structures of the tested highest-ranking β-D-galactosidase hits (as suggested by the best QSAR models).

Table 4. The QSAR-Estimated and *In-vitro* Bioactivities of Highest Ranking β-D-Galactosidase Inhibitory Hit Molecules Captured by Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5 and Hypo5/5.

		Fit values against[b]					QSAR-estimates		Actual affinities[c]
No.[a]	Name or NCI code	Hypo3/1	Hypo8/1	Hypo10/1	Hypo4/5	Hypo5/5	Log ($1/K_i$)	K_i (nM)	% Inhibition at 10 μM or (IC_{50})
42	4354	11.7	10.4	8.9	15.8	15.8	0.907	0.124[h]	IC_{50} = 2.4 μM[i]
43	352888	–	10.0	–	–	–	0.876	0.133[f]	15[d]
44	119173	–	–	–	–	17.2	−1.939	86.88[h]	NI [d]
45	280997	–	–	–	–	17.1	−0.698	4.990[h]	10[d]
46	118770	–	–	–	–	16.3		0.150[h]	13[d]
47	131660	–	–	9.1	–	–	6.135	0.163[g]	2[d]
48	408695	–	–	–	–	17.3	−2.408	0.256[h]	6[d]
49	190759	–	–	–	–	17.1	−0.961	9.139[h]	9[d]
50	382708	–	–	–	–	17.0	−0.829	6.748[h]	NI [d]
51	372921	–	–	–	–	17.1	−0.782	6.047[h]	NI[d]
52	210933	11.3	–	–	–	–	0.983	0.104[e]	33[d]
53	Oxadixyl	–	–	–	–	15.1	−3.972	9.37×10^{3}[h]	32[j]

[a]Hits shown in Fig. 10.
[b]Best-fit values against corresponding binding hypothesis according to eq. (4).
[c]*In vitro* enzyme inhibition. Each value (% inhibition or IC_{50}) represents the average of at least two measurements.
NI: No inhibition.
[d]% inhibition at 10 μM.
[e]These values estimated from eq. (6).
[f]These values estimated from eq. (7).
[g]These values estimated from eq. (8).
[h]These values estimated from eq. (10).
[i]41% inhibition at 1 μM.
[j]% inhibition at 7 μM.

the pharmacophores, whereas other binding interactions were hidden for clarity.

In Silico Screening for β-D-Galactosidase Inhibitors and Subsequent In Vitro Evaluation

Our QSAR-selected shape-complemented pharmacophores, that is, Hypo3/1, Hypo8/1, Hypo10/1, Hypo4/5, and Hypo5/5 [eqs. (6)–(10)], were used as 3D search queries against two available 3D flexible structural databases, namely, the NCI list of compounds (238,819 compounds) and our in-house built database of known DAC (3005 compounds), to discover new inhibitory leads of alternative scaffolds against β-D-galactosidase.

Table 3 shows the number of captured hits by each pharmacophore model. Hits are defined as those compounds that have their chemical groups spatially overlap (map) with corresponding features within the particular pharmacophoric model. NCI hits were subsequently filtered based on Lipinski's and Veber's rules.[54,55] Surviving hits were fitted against corresponding pharmacophores (without shape constraints) and their fit values against particular hypothesis were substituted in the corresponding appropriate QSAR eqs. (6)–(10) to determine their predicted bioactivities. A particular QSAR was considered to be appropriate for predicting the bioactivity of certain hit molecule if it includes the pharmacophore model that captured the molecule, for example, eq. (7) was used to determine the bioactivity of hit **43** because it includes Hypo8/1, which captured **43**. However, hit **42** was captured by all pharmacophore hypotheses prompting us to use the equation that gave the best bioactivity estimate for prediction purposes, that is, eq. (10).

The fact that the predicted Log ($1/K_i$) values exceeded the upper and lower bioactivity limits of the training compounds prompted us to use bioactivity predictions merely to rank the corresponding hits to minimize the impact of any possible extrapolatory prediction errors on decisions regarding hits that merit subsequent *in vitro* testing.[67]

The highest ranking 11 anti-β-D-galactosidase NCI hits were requested for experimental validation. One agrochemical (oxadixyl) was purchased and tested against β-D-galactosidase. Figure 10 shows the chemical structures of tested hits, whereas Table 4 lists the corresponding experimental bioactivities and fit values of the tested hits against corresponding pharmacophore models.

The tested hits were evaluated by measuring the percentage of enzyme inhibition at 7, 10, 40, and 100 μM and by comparing the enzyme activity in the presence and absence of the particular hit.

Hit **42** (IC_{50} = 2.4 μM, Fig. 10) was the most promising. The fact that **42** combined excellent anti-β-D-galactosidase potency and tight fit values against all β-D-galactosidase pharmacophores (fit values ranging 8.9–15.8, Table 4) underlines the significance of our QSAR-selected pharmacophore models, particularly when combined together.

Figures 3–7 show how hit **42** fits the different β-D-galactosidase pharmacophores. Table 4 shows the fit values of different NCI and DAC hits against the different pharmacophores, QSAR estimated K_i values and the *in vitro* anti β-D-galactosidase.

Conclusions

Our results suggest that pharmacophore modeling combined with QSAR analysis can be a useful tool for the discovery of new scaffold of β-D-galactosidase inhibitors. The exploration of the pharmacophoric space of different β-D-galactosidase inhibitors was performed using CATALYST-HYPOGEN to identify high quality binding model(s). Consequently, QSAR analysis was used to obtain a model that explains bioactivity variation. Five successful anti–β-D-galactosidase pharmacophores emerged from five independent equations suggesting the existence of multiple binding modes accessible to ligands within β-D-galactosidase pocket. The QSAR equations and the associated pharmacophoric models were validated by ROC curve analysis and experimental identification of several β-D-galactosidase inhibitors retrieved from NCI database and our in house built structural database of established drug and agrochemicals. NCI hit **42,** which maps all the successful β-D-galactosidase pharmacophores, was found to possess potent anti-β-D-galactosidase bioactivity (IC_{50} = 2.4 μM).

References

1. Scofield, A. M.; Witham, P.; Nash, R. J.; Kite, G. C.; Fellows, L. E. Comp Biochem Phys A 1995, 112, 187.
2. Scofield, A. M.; Witham, P.; Nash, R. J.; Kite, G. C.; Fellows, L. E. Comp Biochem Phys A 1995, 112, 197.
3. Gerber-Lemairer, S.; Juillerat-Jeanneret, L. Mini-Rev Med Chem 2006, 6, 1043.
4. Lillelund, V. H.; Jensen, H. H.; Liang, X.; Bols, M. Chem Rev 2002, 102, 515.
5. Markad, S. D.; Karanjule, N. S.; Sharma, T.; Sabharwal, S. G.; Dhavale, D. D. Bioorg Med Chem 2006, 14, 5535.
6. Merrer, Y. L.; Gauzy, L.; Gravier-Pelletier, C.; Depezay, J. C. Bioorg Med Chem 2000, 8, 307.
7. Robina, I.; Vogel, P. Synthesis 2005, 5, 675.
8. Shitara, E.; Nishimura, Y.; Kojima, F.; Takeuchi, T. Bioorg Med Chem 1999, 7, 1241.
9. Asano, A.; Nash, R. G.; Molyneuxc, R. J.; Fleet, G .W. G. Tetrahedron-Asymmetr 2000, 11, 1645.
10. Asano, N. Glycobiology 2003, 13, 93.
11. Berecibar, A.; Grandjean, C.; Siriwardena, A. Chem Rev 1999, 99, 779.
12. Kim, J. H.; Ryu, Y. B.; Kang, N. S.; Lee, B. W.; Heo, J. S.; Jeong, I. Y.; Park. K. H. Biol Pharm Bull 2006, 29, 302.
13. Li, H.; Schütz, C.; Favre, S.; Zhang, Y.; Vogel, P.; Sinay, P.; Bl'eriot, Y. Org Biomol Chem 2006, 4, 1653.
14. Pandey, G.; Dumbre, S. G.; Khan, M. I.; Shabab, M. J Org Chem 2006, 71, 8481.
15. Coutinho, P. M.; Henrissat, B. In Recent Advances in Carbohydrate Bioengineering; Royal Society of Chemistry: Cambridge, UK, 1999; pp. 3–12.
16. Davies, G. J.; Ducros, V. M.-A.; Varrot, A.; Zechel, D. L. Biochem Soc Trans 2003, 31, 523.
17. Dubost, E.; Tschamber, T.; Streith, J. Tetrahedron Lett 2003, 44, 3667.
18. Dubost, E.; Nouën, D. L.; Streith, J.; Tarnus, C.; Tschamber, T. Eur J Org Chem 2006, 2006, 610.
19. Frankowski, A.; Deredas, D.; Dubost, E.; Gessier, F.; Jankowski, S.; Neuburger, M.; Seliga, C.; Tschamber, T.; Weinberg, K. Tetrahedron 2003, 59, 6503.
20. Gessier, F.; Tschamber, T.; Tarnus, C.; Neuburger, M.; Huber, W.; Streith, J. Eur J Org Chem 2001, 2001, 4111.
21. Tschamber, T.; Gessier, F.; Dubost, E.; Newsome, J.; Tarnus, C.; Kohler, J.; Neuburger, M.; Streith, J. Bioorg Med Chem 2003, 11, 3559.
22. Heightman, T. D.; Vasella, A. T. Angew Chem 1999, 38, 750.
23. Vasella, A.; Davies, G. J.; Bohm, M. Curr Opin Chem Biol 2002, 6, 619.
24. Schramm, V. Acc Chem Res 2003, 36, 588.
25. Schramm, V. Curr Opin Struct Biol 2005, 15, 604.
26. Amyes, T.; Richard, J. ACS Chem Biol 2007, 2, 711.
27. Sutherland, J.; O'Brien, L.; Weaver, D. J Med Chem 2004, 47, 3777.
28. Taha, M. O.; Bustanji, Y.; Al-Ghussein, M.; Mohammad, M.; Zalloum, H.; Al-Masri, I. M.; Atallah, N. J Med Chem 2008, 51, 2062.
29. Al-masri, I. M.; Mohammad, K.; Taha, M. O. ChemMedChem 2008, 3, 1763.
30. Taha, M. O.; Dahabiyeh, L. A.; Bustanji, Y.; Zalloum, H.; Saleh, S. J Med Chem 2008, 51, 6478.
31. Taha, M. O.; Atallah, N.; Al-Bakri, A. G.; Paradis-Bleau, C.; Zalloum, H.; Younis, K. S.; Levesque, R. C. Bioorg Med Chem 2008, 16, 1218.
32. Taha, M. O.; Bustanji, Y.; Al-Bakri, A. G.; Yousef, M.; Zalloum, W. A.; Al-Masri, I. M.; Atallah, N. J Mol Graphics Model 2007, 25, 870.
33. Abu Hammad, A. M.; Taha, M. O. J Chem Inf Model 2009, 49, 978.
34. Abu Khalaf, R.; Abu Sheikha, G.; Bustanji, Y.; Taha, M. O. Eur J Med Chem 2010, 45, 1598.
35. Accelrys Software, Inc. Catalyst 4.11 User Guide; Accelrys Software, Inc.: San Diego, CA, 2005.
36. Sprague, P. W.; Hoffmann, R. Curr Tools Med Chem (VHCA: Basel) 1997, 223.
37. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. J Chem Inf Comput Sci 1996, 36, 563.
38. Smellie, A.; Teig, S.; Towbin, P. J Comput Chem 1995, 16, 171.
39. Li, H.; Sutter, J.; Hoffmann, R. Pharmacophore Perception, Development, and Use in Drug Design; International University Line: California, 2000; pp. 173–189.
40. Sutter, J.; Güner, O. F.; Hoffmann, R.; Li, H.; Waldman, M. Pharmacophore Perception, Development, and Use in Drug Design; International University Line: California, 2000; pp. 501–511.
41. Kurogi, Y.; Güner, O. Curr Med Chem 2001, 8, 1035.
42. Bersuker, I. B.; Bahçeci, S.; Boggs, J. E. Pharmacophore Perception, Development and Use in Drug Design; International University Line: California, 2000; pp. 457–473.
43. Poptodorov, K.; Luu, T.; Langer, T.; Hoffmann, R.Methods and principles in Medicinal Chemistry, Pharmacophores and Pharmacophores Searches, Vol.2; WILEY-VCH: Weinheim, 2006; pp. 17–47.
44. Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Bioorg Med Chem Lett 2003, 13, 4355.
45. Taha, M. O.; Qandil, A. M.; Zaki, D. D.; AlDamen, M. A. Eur J Med Chem 2005, 40, 701.
46. Keller, P. A.; Bowman, M.; Dang, K. H.; Garner, J.; Leach, S. P.; Smith, R.; McCluskey, A. J. J Med Chem 1999, 42, 2351.
47. Karki, R. G.; Kulkarni, V. M. Eur J Med Chem 2001, 36, 147.
48. Taha, M. O.; Al-Bakri, A. G.; Zalloum, W. A. Bioorg Med Chem Lett 2006, 16, 5902.
49. Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. J Chem Inf Model 2008, 48, 719.
50. Fisher, R. The Principle of Experimentation Illustrated by a Psycho-Physical ExpeHafner Publishing Co., 8th ed.; Hafner Publishing: New York, 1966.

51. Krovat, E. M.; Langer, T. J Med Chem 2003, 46, 716.
52. Accelrys, Inc. CERIUS2, version 4.10 QSAR Users' Manual; Accelrys, Inc.: San Diego, CA, 2005; pp. 221–235.
53. Hahn, M. J Chem Inf Comput Sci 1997, 37, 80.
54. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Adv Drug Deliv Rev 2001, 46, 3.
55. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. J Med Chem 2002, 45, 2615.
56. Sheridan, R. P.; Kearsley, S. K. Drug Discov Today 2002, 7, 903.
57. Ramsey, L. F.; Schafer, W. D. The Statistical Sleuth, 1st ed.; Wadesworth Publishing Company: USA, 1997.
58. Tropsha, A.; Gramatica, P.; Gombar, V. K. QSAR Comb Sci 2003, 22, 69.
59. Sivakumar, P. M.; Babu, S. K. G.; Doble, M. Chem Biol Drug Des 2008, 71, 447.
60. Verdonk, M. L.; Marcel, L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. J Chem Inf Comput Sci 2004, 44, 793.
61. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. J Comput Aided Mol Des 2008, 22, 213.
62. Irwin, J. J.; Shoichet, B. K. J Chem Inf Comput Sci 2005, 45, 177.
63. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. J Med Chem 2005, 48, 2534.
64. Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostroem, H.; Norinder, U. J Med Chem 2003, 46, 5781.
65. Gao, H.; Williams, C.; Labute, P.; Bajorath, J. J Chem Inf Comput Sci 1999, 39, 164.
66. Gloster, T. M.; Roberts, S.; Perugino, G.; Rossi, M.; Moracci, M.; Panday, N.; Terinek, M.; Vasella, A.; Davies, G. J. Biochemistry 2006, 45, 11879.
67. Cronin, M. T. D.; Schultz, T. W. J Mol Struct 2003, 622, 39.