Combinatorial Library Enumeration and Lead Hopping using Comparative Interaction Fingerprint Analysis and Classical 2D QSAR Methods for Seeking Novel GABA$_A$ α$_3$ Modulators

R. S. K. Vijayan,[†] Indrani Bera,[†] M. Prabu,[†] Sangita Saha,[‡] and Nanda Ghoshal*[,†]

Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology (a unit of CSIR), 4,
Raja S.C. Mullick Road, Kolkata-700 032, West Bengal, India and Department of Bioinformatics, West
Bengal University of Technology, BF - 142 Salt Lake, Kolkata-700 064

Received August 18, 2009

Selective modulators of GABA$_A$ α$_3$ (gamma amino butyric acid α$_3$) receptor are known to alleviate the side effects associated with nonspecific modulators. A follow up study was undertaken on a series of functionally selective phthalazines with an ideological credo of identifying more potent isofunctional chemotypes. A bioisosteric database enumerated using the combichem approach endorsed mining in a lead-like chemical space. Primary screening of the massive library was undertaken using the "Miscreen" toolkit, which uses sophisticated bayesian statistics for calculating bioactivity score. The resulting subset, thus, obtained was mined using a novel proteo-chemometric method that integrates molecular docking and QSAR formalism termed CoIFA (comparative interaction fingerprint analysis). CoIFA encodes protein−ligand interaction terms as propensity values based on a statistical inference to construct categorical QSAR models that assist in decision making during virtual screening. In the absence of an experimentally resolved structure of GABA$_A$ α$_3$ receptor, standard comparative modeling techniques were employed to construct a homology model of GABA$_A$ α$_3$ receptor. A typical docking study was then carried out on the modeled structure, and the interaction fingerprints generated based on the docked binding mode were used to derive propensity values for the interacting atom pairs that served as pseudo-energy variables to generate a CoIFA model. The classification accuracy of the CoIFA model was validated using different metrics derived from a confusion matrix. Further predictive lead mining was carried out using a consensus two-dimensional QSAR approach, which offers a better predictive protocol compared to the arbitrary choice of a single QSAR model. The predictive ability of the generated model was validated using different statistical metrics, and similarity-based coverage estimation was carried out to define applicability boundaries. Few analogs designed using the concept of bioisosterism were found to be promising and could be considered for synthesis and subsequent screening.

INTRODUCTION

Gamma amino butyric acid type A (GABA$_A$) ionotropic receptor is the major inhibitory neuronal receptor of the mammalian brain conferring fast synaptic inhibition.[1]The physiological role exerted by GABA$_A$ in regulating brain excitability and its pharmacological significance as a valuable drug target for many neuronal disorders has surged a renewed interest for a cohesive understanding of its structure and function. The heterogeneous nature of GABA$_A$ receptor and its low abundance (pmol/mg of protein), coupled with the inherent difficulties associated in isolation and purification of integral membrane proteins, have precluded structural investigations on GABA$_A$ receptors.[2,3] Purification, cloning, and sequencing of the GABA$_A$ receptor and its composite subunits have allowed the identification of 21 subunits arranged within 8 families (α1−6, β1−4, γ1−4, 1δ, 1ε, 1π, 1θ, and ρ1−3).[4] Given the heterogeneity of GABA$_A$ receptor, the pharmacological significance of identifying subtype selective modulators is increasingly being recognized.[5−7] Of the numerous structural classes of drugs that have been shown to bind to the benzodiazepine (BZ) site of GABA$_A$ receptor, relatively few display subtype selectivity and, hence, are associated with side effects such as amnesia, tolerance, dependence, and alcohol potentiation.[5−7] Recent investigations have revealed that ligands displaying functional selectivity for GABA$_A$ α$_2$ and α$_3$ receptor subtypes act as nonsedating anxiolytics in animal models.[5−7] A drug discovery project conducted by Merck research laboratories, identified a series of phthalazines that exhibit functional selectivity toward GABA$_A$ α$_3$.[8,9] Starting with the available information in hand, further exploration was undertaken in pursuit of identifying additional leads. A targeted library of privileged molecules was enumerated using the combinatorial approach. Intelligent enumeration involves the rational choice of building blocks. Bioisosteric building blocks that are ostensibly known to yield molecular entities imparting similar biological properties were considered as ideal replacements for the marked Markush fragments. Such bioisosteric replacements have proven successful particularly for the anxiolytic drug segment, where the methylated carboxamide function of diazepam when replaced by its bioequivalent

* Corresponding author. Telephone: +91-33-2499-5700, (ext. 854/836). Fax: +91-33-2473-5197. E-mail: nghoshal@iicb.res.in.

[†] Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology.

[‡] Department of Bioinformatics, West Bengal University of Technology.

10.1021/ci900309s CCC: $40.75 © 2009 American Chemical Society
Published on Web 11/05/2009

methylated triazole group lead to the identification of alprazolam.[10] In the present study, four classes of bioisosters were identified, and a virtual combinatorial library was enumerated by exploring all possible combinations on the monodentate attachment point marked on the scaffold. Thus, the principle of analog design, widely exploited in medicinal chemistry, was used to create a targeted library.

High-throughput virtual activity profiling was carried out using the molinspiration "Miscreen" toolkit,[11] which prioritizes molecules in the databases using a supervised Bayesian statistical method. Further lead mining of the combinatorial subset was carried out using a novel categorical receptor dependent three-dimensional (3D)-QSAR strategy termed as CoIFA (comparative interaction fingerprint analysis), a variant of other ligand−receptor based QSAR methods like COMBINE[12](comparative binding energy), AFMoC[13] (adaptation of fields for molecular comparison), and CoRIA[14] (comparative residue interaction analysis).

Underlying the CoIFA methodology is the generation of an interaction fingerprinting scheme that encodes the nonbonded interactions between ligands and proteins based on the putative 3D binding modes (docked solutions). Pairwise interaction fingerprints observed in the protein−ligand complex are assigned knowledge-based probabilistic contact values, which, in turn, serve as predictor (X) variables for QSAR analysis to yield categorical models that also provide mechanistic insights regarding binding modes.

In the absence of experimentally resolved structure of GABA$_A$ receptor, comparative modeling was employed to generate an atomistic model of GABA$_A$ α_3 receptor. An open-state conformation of GABA$_A$ α_3 was modeled using the standard homology modeling techniques employed by Modeler[15] based on the X-ray structure of the molluscan AchBP,[16] a homologue of GABA.

A typical docking study was carried out on the data set using the GOLD docking suite.[17] Based on the docked binding mode, probabilistic atom-pair contact values were generated using the PLIF[18,19] module of MOE,[20] and a decision tree was construed using a recursive partitioning algorithm to generate a robust categorical QSAR model.[21] The classification accuracy of the CoIFA model was validated using the standard parameters inferred from the confusion matrix. CoIFA-based QSAR models are categorical, hence, we supplemented classical QSAR-based methods for predictive lead mining.

Classical type QSAR models were developed using the two-dimensional (2D) descriptors available in MOE,[20] and fitting was carried out using two statistical methods (GFA[22] and G/PLS[23]). The developed models were subjected to rigorous validation, with an emphasis on model stability and predictivity. Activity forecasting was carried out using a consensus QSAR model, which was anticipated to provide a better model compared to the arbitrary choice of a single QSAR model.[24,25] A key component that needs to be evaluated, particularly when extrapolating QSAR models for lead mining, is to ensure that the predictions came from the domain upon which the model was calibrated. Hence, descriptor-based applicability domain estimation was carried out using an integrated approach that incorporates two pattern recognition algorithms, K-mean and K-nearest-neighbor (KNN), to define applicability boundaries based on Euclidean distance measures.[26] QSAR paradigm has been of interest in the field of medicinal chemistry generally as a retrospective tool to rationalize the lead optimization process in drug design. As an adaptive response to the changing scenario, the present study demonstrates how QSAR techniques can be used as a tool to screen millions or perhaps billions of molecules to identify potentially bioactive molecules.

MATERIALS AND METHODS

All computational studies outlined here were performed using the following software packages: Discovery studio 2.0,[27] GOLD 3.2,[28] BROOD,[29] ROCS,[29] EON[29] MOE,[20] Miscreen,[11] and TSAR 3.0[30] running on a Pentium core2 Duo workstation using Windows XP operating system. Cerius^2 4.10[31] software-based studies were carried out on a SGI fuel workstation running on IRIX 6.5 operating system.

Selection and Modeling of Data Sets. A crucial prerequisite in carrying out a successful QSAR analysis is the selection of an internally consistent data set. Hence, biological data acquired by the same group under the same assay method was pooled from literature sources to render them comparable. The skewness in the data set was removed by converting the reported K_i (nM) values[8,9] to $_pK_i$ using a simple logramethic transformation log ($1/K_i$). Data set compounds were modeled using the 3D Sketcher module of Cerius.^{2} Initial 3D structure optimization was carried out using CORINA,[32] and further charge assignment and energy optimization were carried out using the COSMIC module of TSAR.

Bioisterism Guided Virtual Combinatorial Library Enumeration. Combinatorial chemistry, which orchestrates the creation of large compound libraries, was used to enumerate a targeted library.[33] Central to the design of a targeted library is the identification of R-group that are key SAR determinants and of high-quality building blocks for intelligent enumeration.[34] The SAR report functionality of MOE was used to identify key R-group motifs[18,35] that were structurally transformed using the concept of bioisosterism.[36,37] Potential isosters for the marked Markush fragments were identified from a biosisosteric database containing fragments of synthetic tractability using the Brood software.[29] Four classes of isosters that matches the query fragment in terms of shape, atom-type, electrostatics, structure, and graph were identified. A total of 856 isosteric fragments were identified for R1 and 462 isosters for R2 after carrying out a redundancy check using a SVL script of MOE. QuaSAR-CombiGen module of MOE was then used to enumerate a virtual library of all possible products that could be combinatorially obtained from the set of fragments. The targeted library, thus, generated can be screened en masse and yields better enrichment rates than either a random or diverse library.[38] Markush fragments considered for isosteric replacement are shown in Figure 1, and the core scaffolds used for clipping the fragments are shown in Figure 2.

Bayesian Statistics-Based Activity Profiling. A targeted library consisting of 790 944 molecules was screened using "Miscreen" virtual screening toolkit,[11] which uses sophisticated Bayesian statistics for screening large compound libraries based on fragment profile comparisons. A bioactivity model was developed by generalizing the occurrence of substructural fragments among a set of active and inactive

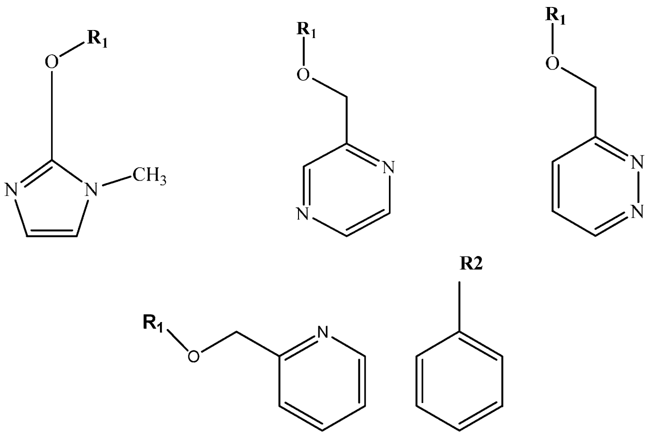

Figure 1. R-group fragments considered for isosteric replacement based on a SAR report.

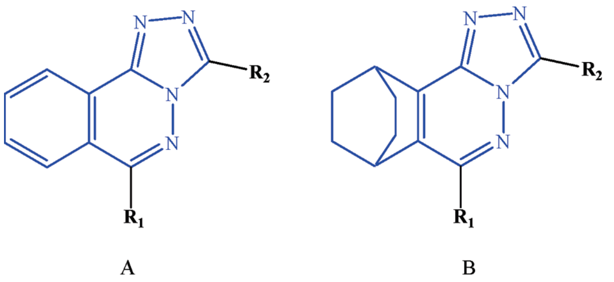

Figure 2. General scaffold for the phthalazine core-based derivatives used in the present study.

phthalazines. Construction of the Bayesian model was based on the reported $_{P}K_i$ values. Compounds with a $_{P}K_i$ value <8 were assigned to be inactive, and those with values >8 were regarded as active compounds. Such a cut-off value appears to be a reasonable starting point for hit-to-lead activity.The bioactivity model developed was validated for its enrichment ability, and the validated model was subsequently used for virtual activity profiling. Miscreen provides a "bioactivity score" for each compound that ranges between −3 to 3, which is calculated as a sum of the activity contributions of the fragments present in the molecules. Screening using the "Miscreen" toolkit is very fast (approximately 3×10^5 molecules in 30 min on a Pentium core2 processor).

CoIFA − A Novel Categorical 3D QSAR. With the ever increasing number of experimentally derived structures available across public domains, like PDB,[39] and computational methods, like homology modeling, gaining momentum, 3D QSAR methods seldom harness this valuable information. The amount of structural information incorporated in 3D QSAR methodologies has been limited to the alignment stage (docking based alignment) and to visually interpreting the results of a QSAR model.[40] Approaches like CoMFA[41] use pseudo-receptor modeling techniques to build QSAR models and are purely ligand based in nature. Variant approaches, like COMBINE,[12] CoRIA,[14] and AFMoC,[13] are termed as receptor-dependent (RD) QSAR approaches, as they explicitly use the structural information of both the ligand and the receptor to correlate the observed activity with the selected variables. COMBINE[12] and CoRIA[14] use molecular mechanics and Poission Boltzman methods for calculating residue-wise energy components, which are decomposed and translated in to a regression-based QSAR model. AFMoC[13] is a reverse approach of CoMFA[41] that works by placing a grid into a binding site that maps the pair potentials between the protein and the ligand atoms resulting in "potential fields". The pair potentials are calculated using the knowledge-based scoring function, DrugScore.

As a valuable extension to the classical 3D-QSAR paradigm, we report herein a novel proteochemometric-based QSAR formalism termed CoIFA (comparative interaction fingerprint analysis). CoIFA method integrates molecular docking and QSAR formalism for deriving a categorical QSAR model. Derivation of a CoIFA model essentially involves four steps that share analogy with other 3D QSAR approaches:

(i) Receptor-guided alignment based on docking.

(ii) Summarizing the interactions in the protein−ligand complex using knowledge-based probabilistic contact values in lieu of probes that employ simple force field based approaches to calculate energy values on a grid/surface.

(iii) Application of classification algorithms for the development of a reliable decision rule for automated classification (active/inactive).

(iv) Validation of the QSAR model for its classification accuracy using different statistical metrics.

CoIFA QSAR modeling is a fundamental shift from other RD-QSAR methods. Being categorical in nature, it has the advantage of modeling data sets with biological end points (K_i, IC_{50}, K_d) expressed as discrete and continuous data. The core strengths of CoIFA models could also be attributed to the nonlinear modeling technique (recursive partitioning method) that relies on generalizations based on weighing a set of active and inactive molecules.

Being a categorical QSAR approach that relies on a nonlinear modeling technique, the error rate would be significantly less, since no attempt is made to minimize the squared error between the model and the observed data, as in other regression-based fitting methods. Moreover, decision tree methods (recursive partitioning methods) are also known to be tolerant to noisy experimental data, and the model building process is also very fast in contrast to neural network methods that require a lengthy training phase. When one considers the merits, it is also prudent to mention the caveats of the CoIFA-based approach. Though the CoIFA approach incorporates protein−ligand cross terms based on statistical potentials that are relatively faster, the solvent and entropy terms are overlooked in the CoIFA approach. Another concern with the use of statistical potentials is that the CoIFA approach would not model the interactions correctly when the atom-pair interaction in question does not fall into the class upon which it was trained.[42] The list of the primary atom types considered to derive the statistical potential is available via ref 43. Further, it should be noted that geometry-based evaluation of nonbonded interactions are loose approximations because it is ambiguous to differentiate the interaction between two given hydrophobic groups in close proximity as attractive or repulsive without the incorporation of force field based energy terms.[44] Another obvious concern for all RD QSAR approaches is the unavailability of the 3D structure of the receptor and of also a sizable amount of active and inactive compounds. Inactive compounds are generally not available through literature sources, and in such circumstances, presumed inactives could be collected from publicly available data sets, like the DUD[45] decoy set.

In the absence of the crystal structure of GABA$_A$ α$_3$ receptor, homology modeling was carried out to model the pentameric structure of GABA$_A$ α$_3$ receptor. Amino acid sequence of human GABA$_A$ α3, β2, and γ2 were retrieved from Swiss Prot[46] (P34903, P47870, and P18507). The signaling peptide region at the N terminal was truncated, and the mature protein sequence was used for modeling. All subsequent amino acid numbering is based on the mature sequence without the signal peptide. The crystal structure of the molluscan AchBP protein data bank (PDB) (1I9B)[16] was identified as the modeling template based on a BLAST[47] search carried out against the PDB (α3 30% and β2 and γ2 27%). FUGUE-[48] based alignment that employs a sequence-structure-based alignment method, and particularly suited in the case of low-sequence identity, was employed for alignment. Standard modeling techniques were carried out using the Modeler module of the Discovery studio software suite to assign atomic coordinates to structurally conserved regions (SCR), to build intervening loops, to optimize the rotamers of amino acid side chains, and to perform an initial energy optimization of the structure. The 3D structure of GABA$_A$α3 was modeled in the presence of the crystal coordinates of HEPES (*N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulphonic acid), a buffer molecule that occupies the benzodiazepine site. This was made possible using the "copy ligand" option of Modeler. The best-ranked model, based on the probability density function (PDF) violations, was chosen and further evaluated using the routine protein structure validation tools.

To simulate the binding mode of the flexible ligands to the rigid homology modeled receptor, molecular docking was carried out using the GOLD docking program. GOLD uses an evolutionary genetic algorithm (GA)[49] for generating ligand conformations and also imparts partial receptor flexibility in an implicit manner by allowing movement around the side chain dihedrals of protein OH and NH_3+. The binding site was defined using the coordinates of HEPES (residues within 10 Å from the ligand). The standard default mode, which comprises 1×10^5 genetic operations on an initial population of 100 members divided into five subpopulations (number of islands = 5) was used. The annealing parameters for the fitness function were set at 4.0 for van der Waals and at 2.5 for hydrogen bonding. A niche size of 2 and a selection pressure of 1.1 were used, and the early termination option was turned off. The top-ranking pose was retained based on the Gold fitness function, which served as our presumed bioactive conformation for alignment.

The applications of geometry-based methods in developing knowledge-based scoring functions (PMF,[50]BLEEP,[51] and DrugScore[52]) and the success of SIFt[53] as a post-docking tool prompted us to put in place a proteochemometric-based QSAR methodology. The steric and electrostatic energy terms, commonly calculated using a probe atom on a series of grid points surrounding the aligned molecules in 3D space, has been replaced with more realistic protein−ligand atom pair contact potentials. CoIFA descriptors depict the physical interactions within the protein−ligand system based on atom pair contact propensity values. These contact propensity values are knowledge-based values estimated by statistical means from a collection of structures available in the PDB, which serves as a knowledge base.[43] Descriptor values, calculated using the PLIF module of MOE, were subjected to a categorical QSAR study using the decision tree method[21] implemented in the QuaSAR-Classify module of the MOE package. The tree classification method employs a recursive partitioning algorithm that consists of two parts: tree growing and tree pruning. The process starts with a training set consisting of preclassified records. Tree growing begins with the root node, which is the entire learning data, and the root node is set as the current node. The partitioning of the root node in to child nodes is carried out according to the rules of the form 24 < = *c* (if *x* is a continuous variable) or $x = c$ (if it is a categorical data), where *c* is the best value for splitting the node. Splitting is based on the Gini index of diversity:

$$G(t) = 1 - \sum_{I=1}^{k} p_i^2(t) \qquad (1)$$

where p_i^2 (t) is the fraction of compounds of each class I ($I = 1, ..., K$) in node t.

The goodness of a split is measured by a change in value of C termed as the change of impurity:

$$C = G(t) - P_L G(t_L) - P_R G(t_R) \qquad (2)$$

where P_L and P_R are the proportions of cases going to the left t_L and right t_R child nodes. Node splitting stops when the number of compounds in the child node is lower than a predefined threshold or when the distribution of the compounds in the child node becomes homogeneous. The misclassification rate in node t is calculated as $R(t) = 1 - n_j/n_t$, where n_j is the number of compounds of class j, and n_t is the total number of compounds in the node. The total misclassification rate is measured as

$$R(T) = N_{\text{misclassified}}/N_{\text{tot}} \qquad (3)$$

where $N_{\text{misclassified}}$ and N_{total} are the total number of misclassified compounds and the total number of compounds in the data set, respectively.

The initial tree obtained from the training data is generally too big and is prone to noise (equivalent to an over parametrized model). Hence, high correct classification rates (CCR) are obtained for the training set, and usually performs modest for the test set. Pruning solves this issue by removing smaller branches that fail to generalize, and the sub tree subsequently identified is referred to as the best sub tree. In the present study, cross validation was employed for tree pruning. A modified tree misclassification rate $R_a(T) = R(T) + aL(T)$, where $L(T)$ is the number of leaves in the tree, and $a > 0$ is a parameter that provides a solution for identifying nodes that need to be pruned. According to this equation, the size of the tree and the misclassification rate are balanced. The classification accuracy of the model was validated using different metrics derived from the confusion matrix. It should be noted that CoIFA model, like all other QSAR models, is intrinsically biased toward the scaffolds used for training, hence, extrapolations should be carried out only within the domain it was trained upon. When embarking lead hopping, it is always advisible to carry out a diversity analysis of the library using some similarity metrics to filter out potential outliers (structurally diversified compounds).

Predictive Data Mining using Consensus 2D QSAR. CoIFA models are well suited for high-throughput screening (HTS) campaigns to decide which compound should be passed to the next level of screening. CoIFA models offer a two-state prediction level (active /inactive) and cannot be a substitute to classical QSAR modeling per se. To augment our virtual screening workflow, we carried out classical QSAR studies to transform the qualitative belief inferred from CoIFA models into a quantitative measure for forecasting activity. Predictive lead mining was carried out using 2D QSAR models of proven statistical quality with extrapolations restricted to the applicability domain. Development of a predictive 2D QSAR model relies on a multitude of factors, and one such important prerequisite is the rational selection of training a test set. A nonhierarchical method called K-means clustering was performed using SPSS[54] for the rational selection of both a training and a test set.[55] Manual selection of compounds was done from each cluster so as to ensure a heterogeneous training and test sets, homogeneous with the training set in terms of activity range and "structural diversity". K-means clustering and QSAR model development were carried out using the 2D descriptors (physicochemical, structural, and topological) available in MOE. Details of the clusters are provided in Supporting Information (Table 1). The descriptors used in the study represent physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, and partial charge descriptors. Two chemometric methods, namely GA and G/PLS, implemented in Cerius^2 were employed for variable selection and fitting. Genetic algorithm-based variable selection is a stochastic process that mimics biological evolution by undertaking a survival of the fittest approach and is considered superior to models developed using the stepwise regression technique. Initial variable reduction was carried out by considering only 40% of those information-rich descriptors with highest variance in order to prevent over exploitation of independent variables (X) in fitting the dependent variable (Y). GFA,[22] which uses a combination of Holland's genetic algorithm for variable selection and Friedman's multivariate adaptive regression splines (MARS) algorithm to build regression models, was employed for model building. Population of equation, chosen at random, was evaluated using a "fitness function" or lack-of-fit (LOF) that reflects the quality of the model. This cycle is then repeated for a specified number of iterations until a final population (of descriptors) that is better suited to model the end point is obtained. The error measurement term, LOF, is determined by the following equation:

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\{1 - (c + dp)/m\}^2} \qquad (4)$$

where LSE is the least-squares error, c is the number of basis functions in the model, d is the smoothing parameter, p is the number of descriptors, and m is the number of observations in the training set. See Table 1 for results.

A variable usage graph displayed during each GFA run was used to monitor premature convergence of GA. Another statistical method termed G/PLS, which combines the best features of both GFA and PLS, was also employed. G/PLS has PLS applied to it instead of multiple linear regressions, as employed by GFA, hence, each model can have more terms in it without the danger of overfitting and is particularly valuable in cases when the number of descriptors is more than the number of samples. G/PLS[23] retains the ease of interpretation by back transforming the PLS components to the original variables. The robustness of the developed 2D QSAR were assessed for four important qualities namely goodness-of-fit, model stability, predictive ability, and domain applicability. Goodness-of-fit was adjudged using the square correlation coefficient (R^2) and explained variance ($R^2{}_{\mathrm{adj}}$). Internal predictive ability of the generated model was evaluated using the leave-one-out (LOO) based $q2$ and bootstrap R^2. The external predictive ability was validated using a reserved test set. Model stability and risk of chance correlation was evaluated using the Y-randomization procedure.

A serious concern about the use of QSAR models for activity forecasting is the reliability of the predictions. Hence, predictions were carried out within the defined applicability domain because interpolation is, in general, more reliable than that of extrapolation. Applicability domain (AD) estimation was performed using our in-house program. Our algorithm employs a three-stage approach for defining AD. In the first stage, the K-means algorithm is employed for clustering the training set compounds based on 2D descriptor space. The identified clusters represent different "structural classes". In the second stage, assignment of structural classes for the test set compound and the designed isosteric analogues were performed using the K-NN approach. In the present study, Euclidean distances were calculated for three of it is nearest-neighbors to determine the degree of structural similarity between the molecules in question and the molecules associated within the structural classes. In the final stage, an inclusion and exclusion criteria was defined for the compounds in question based on D_{T} values calculated from a Euclidean distance-based similarity matrix computed for each structural class. D_{T} values were calculated using the formula:

$$D_{\mathrm{T}} = Y + \sigma Z \qquad (5)$$

where Y is the average Euclidean distance between the compounds in question and all members belonging to that cluster, σ is the standard deviation, and Z is the arbitrary parameter to control the significance level, which was kept at 0.5, formally placing the allowed distance at half of the standard deviation.[26] D_{T} calculations were performed using an in-house program written using C programming language.

RESULTS AND DISCUSSION

Combinatorial Library Generation and Virtual Activity Profiling using Miscreen. The Bayesian bioactivity model generated using Miscreen was validated beforehand for its enrichment ability using a Perl script before undertaking an en masse virtual screening of our targeted combinatorial library consisting of 790 944 molecules. A four-fold cross-validation process was performed by randomly dividing the original data set into N subsamples. Of the N subsamples, a subset was retained as a validation set for testing the predictive ability of the model, and the remaining $N - 1$ subsamples are used as a training set. The cross-validation process was repeated four times, and a receiver operating characteristic (ROC) curve that reports the discriminatory ability of the model at various thresholds, by plotting true positive rate (TPR, y-axis) versus false positive rate (FPR,

Table 1. Structural Features with Observed and Predicted pK_i Values along with Their Substitutions (R_1 and R_2)[a]

name	actual pK_i (nM)	GPLS	GFA	C QSAR	R_1	R_2	core scaffold
1	7.39	7.29	6.81	7.05	bis-(2- methoxy-ethyl)-amine yl	4-methoxy phenyl	A
2	6.77	6.83	6.61	6.72	phenylmethoxy	4-methoxy phenyl	A
3	6.92	7.42	7.57	7.49	methoxy	phenyl	A
4	7.39	7.09	6.97	7.03	phenylmethoxy	phenyl	A
5	6.14	6.69	6.48	6.58	phenylethoxy	phenyl	A
6	6.38	6.50	6.21	6.35	phenylpropoxy	phenyl	A
7	6.59	6.82	6.62	6.72	2-chloro phenylmethoxy	phenyl	A
8	7.18	7.25	7.01	7.13	3-methoxy-phenylmethoxy	phenyl	A
9	6.77	6.61	7.08	6.84	3-cyano-phenylmethoxy	phenyl	A
10	6.52	6.79	6.73	6.76	thiophene-2-yl-methoxy	phenyl	A
11	6.62	6.90	6.84	6.87	thiophene-3-yl-methoxy	phenyl	A
12	6.74	7.10	6.85	6.98	furan-2-yl-methoxy	phenyl	A
13	7.68	7.60	7.20	7.40	furan-3-yl-methoxy	phenyl	A
14	8.11	7.91	8.20	8.05	pyridine-4-yl-methoxy	phenyl	A
15	9.15	8.58	8.64	8.61	pyrazine-2-yl-methoxy	phenyl	A
16	9.66	9.31	9.10	9.21	(1-methyl-1H-imidazol-2-yl)-methoxy	phenyl	A
17	7.43	7.23	7.63	7.43	piperidine-4-yl-methoxy	phenyl	A
18	7.77	7.75	7.94	7.85	pyridine-2-yl-ethoxy	phenyl	A
19	8.92	7.94	8.31	8.13	6-methyl-pyridine-2-yl-methoxy	phenyl	A
20	8.37	8.37	8.57	8.47	5-methyl-pyridine-2-yl-methoxy	phenyl	A
21	8.00	7.85	8.12	7.99	4-methyl-pyridine-2-yl-methoxy	phenyl	A
22	9.00	8.36	8.57	8.46	3-methyl-pyridine-2-yl-methoxy	phenyl	A
23	7.42	7.44	7.47	7.46	pyridine-2-yl-methoxy	3-pyridyl	B
24	6.82	7.00	7.15	7.08	pyridine-2-yl-methoxy	4-pyridyl	B
25	7.42	7.26	7.40	7.33	pyridine-2-yl-methoxy	3-thienyl	B
26	7.74	7.33	7.59	7.46	pyridine-2-yl-methoxy	cyclopropyl	B
27	7.28	7.50	7.36	7.43	pyridine-2-yl-methoxy	3-methoxy phenyl	B
28	7.49	8.15	8.03	8.09	pyridine-2-yl-methoxy	4-methyl phenyl	B
29	8.21	7.70	7.68	7.69	pyridine-2-yl-methoxy	2-fluorophenyl	B
30	7.24	7.29	7.28	7.28	pyridine-2-yl-methoxy	4-fluorophenyl	B
31	7.60	8.07	8.03	8.05	pyridine-3-yl-methoxy	phenyl	B
32	7.89	7.86	7.83	7.85	pyridine-4-yl-methoxy	phenyl	B
33	7.74	8.11	8.00	8.05	4-methyl1-pyridine-2yl methoxy	phenyl	B
34	7.82	8.33	8.20	8.26	5-methyl1-pyridine-2yl methoxy	phenyl	B
35	7.70	8.04	8.03	8.04	6-methyl1-pyridine-2yl methoxy	phenyl	B
36	8.62	8.06	8.30	8.18	3-propyl1-pyridine-2yl methoxy	phenyl	B
37	8.44	8.57	8.58	8.57	3,6-di methyl1-pyridine-2yl methoxy	phenyl	B
38	9.15	8.75	8.74	8.75	3,5-di methyl1-pyridine-2yl methoxy	phenyl	B
39	8.77	8.53	8.54	8.54	3,4-di methyl1-pyridine-2yl methoxy	phenyl	B
40	8.77	8.09	8.09	8.09	3,cyano 1-pyridine-2yl methoxy	phenyl	B
41	8.15	7.56	7.46	7.51	3-MeO_2C-2-pyridyl	phenyl	B
42	6.74	7.45	7.21	7.33	3-hydroxyl-1-pyridine-2yl methoxy	phenyl	B
43	8.96	8.68	8.54	8.61	3-ethoxy-1-pyridine-2yl methoxy	phenyl	B
44	8.41	9.04	8.81	8.93	3-methoxyethoxy-2-pyridyl methoxy	phenyl	B
45	8.64	8.32	8.25	8.28	3-cyclopropyl-methoxy-2-pyridinyl methoxy	phenyl	B
46	7.42	7.31	7.55	7.43	3- cyclobutyloxy-2-pyridinyl methoxy	phenyl	B
47	7.70	7.87	8.03	7.95	3-benzyloxy-2-pyridinyl methoxy	phenyl	B
48	8.07	7.75	7.87	7.81	pyridazin-3-yl methoxy	phenyl	B
49	7.48	7.56	7.55	7.56	pyrimidin-2-yl methoxy	phenyl	B
50	8.51	8.19	8.07	8.13	pyrimidine-6-yl methoxy	phenyl	B
51	8.52	7.82	7.70	7.76	pyrazine-6-yl methoxy	phenyl	B
52	8.00	8.00	8.03	8.02	6-methyl-pyridazin-3-yl methoxy	phenyl	B
53	7.11	6.82	7.12	6.97	(1,4 dihydro-quinoline-4-yl)-methoxy	phenyl	B
54	7.59	8.22	8.24	8.23	(1,2dihydro-quinoxalin-2-yl)-methoxy	phenyl	B
55	6.70	7.63	7.58	7.60	2-pyridylethoxy	phenyl	B
56	7.28	7.23	7.50	7.37	2-cyano-benzyloxy	phenyl	B
57	7.18	7.03	7.30	7.17	3-cyano-benzyloxy	phenyl	B
58	7.37	7.83	7.78	7.81	3-hydroxymethyl-benzyloxy	phenyl	B
59	7.62	7.60	7.28	7.44	3-*N*,*N*-dimethyl-aminomethyl benzyloxy	phenyl	B
60	7.51	6.87	7.17	7.02	4-cyanophenylmethoxy	phenyl	B
61	6.34	6.78	6.32	6.55	1-methyl-piperidin-3-ylmethoxy	phenyl	B
62	6.59	6.61	6.65	6.63	5-oxo-2-pyrrolidinemethoxy	phenyl	B
63	7.05	6.91	7.05	6.98	Me_2NCOCH_2O	phenyl	B
64	7.30	6.79	6.76	6.77	1-pyrrolidine-$COCH_2O$	phenyl	B
65	7.39	7.38	7.88	7.63	$BnMeNCOCH_2O$	phenyl	B
66	6.77	6.98	7.09	7.04	hydroxypropoxy	phenyl	B
67	7.72	8.06	8.07	8.06	pyrazole-2-yl methoxy	phenyl	B
68	8.64	8.55	8.39	8.47	3,5-dimethyl pyrazole 2-yl methoxy	phenyl	B
69	8.17	7.91	7.87	7.89	imidazole-2-yl methoxy	phenyl	B
70	8.11	8.16	8.04	8.10	5-methyl imidazole-2-yl methoxy	phenyl	B
71	9.00	8.74	8.70	8.72	3-ethyl imidazole-2-yl methoxy	phenyl	B
72	8.70	8.94	9.29	9.11	3-benzyl imidazole-2-yl methoxy	phenyl	B

Table 1 Continued

name	actual pK_i (nM)	GPLS	GFA	C QSAR	R_1	R_2	core scaffold
73	7.74	8.47	8.44	8.45	5-methyl-isoxazole-3-yl methoxy	phenyl	B
74	7.72	7.70	7.80	7.75	thiazol-2-yl methoxy	phenyl	B
75	7.52	7.95	7.96	7.96	4-methyl thiazol-2-yl methoxy	phenyl	B
76	7.62	8.02	7.96	7.99	5-methyl thiazol-2-yl methoxy	phenyl	B
77	7.92	7.52	7.66	7.59	5-methyl thiazol-4-yl methoxy	phenyl	B
78	6.60	6.63	6.78	6.71	benzthiazol-2-yl methoxy	phenyl	B
79	7.00	7.37	7.00	7.18	3-methyl isothiazo-5-yl methoxy	phenyl	B
80	7.29	6.97	6.99	6.98	3-methyl-1,2,4-oxadiazol-5-yl methoxy	phenyl	B
81	7.18	7.50	7.30	7.40	5-methyl-1,2,4-oxadiazol-3-yl methoxy	phenyl	B
82	7.72	7.10	7.36	7.23	1H-[1,2,4]-triazol-3-yl methoxy	phenyl	B
83	8.10	8.46	8.22	8.34	1-methyl-1,2,4-triazol-3-yl methoxy	phenyl	B
84	7.40	7.04	7.40	7.22	1-propyl-1,2,4-triazol-3-yl methoxy	phenyl	B
85	8.85	8.80	8.56	8.68	2-propyl-1,2,4-triazol-3-yl methoxy	phenyl	B
86	7.68	7.73	7.79	7.76	1-methyl-tetrazol-5-yl methoxy	phenyl	B
87[b]	6.92	7.47	7.61	7.54	2-pyridyloxymethyl	phenyl	B
88[b]	6.77	6.28	6.33	6.30	2-pyridyl-amide	phenyl	B
89[b]	7.15	7.04	7.30	7.17	4-cyano-benzyloxy	phenyl	B
90[b]	8.00	7.17	7.49	7.33	thiazol-4-yl methoxy	phenyl	B
91[b]	8.55	8.46	8.22	8.34	2-methyl-1,2,4-triazol-3-yl methoxy	phenyl	B
92[b]	8.37	8.70	8.74	8.72	3-methyl imidazole-2-yl methoxy	phenyl	B
93[b]	7.36	7.00	7.01	7.01	4-methoxy-benzyloxy	phenyl	B
94[b]	7.64	7.39	7.66	7.53	3-chloro pyridazine 6 yl methoxy	phenyl	B
95[b]	8.62	8.33	8.16	8.25	3,methoxy-1-pyridine-2yl methoxy	phenyl	B
96[b]	6.60	6.62	6.42	6.52	4-chloro phenylmethoxy	phenyl	A
97[b]	8.68	8.32	8.20	8.26	3-methyl1-pyridine-2-yl methoxy	phenyl	B
98[b]	6.72	7.26	7.01	7.13	4 -methoxy-phenylmethoxy	phenyl	A
99[b]	7.72	8.12	8.40	8.26	pyridine-3-yl-methoxy	phenyl	A
100[b]	8.89	8.23	8.44	8.34	pyrimidine-4-yl-methoxy	phenyl	A
101[b]	7.89	7.79	7.87	7.83	pyridine-2-yl-methoxy	phenyl	B
102[b]	7.28	6.96	6.79	6.88	pyridine-2-yl-methoxy	2-furyl	B
103[b]	7.54	7.91	7.76	7.84	pyridine-2-yl-methoxy	3-methoxy phenyl	B

[a] The core scaffolds A and B are shown in Figure 2. [b] Indicates test set.

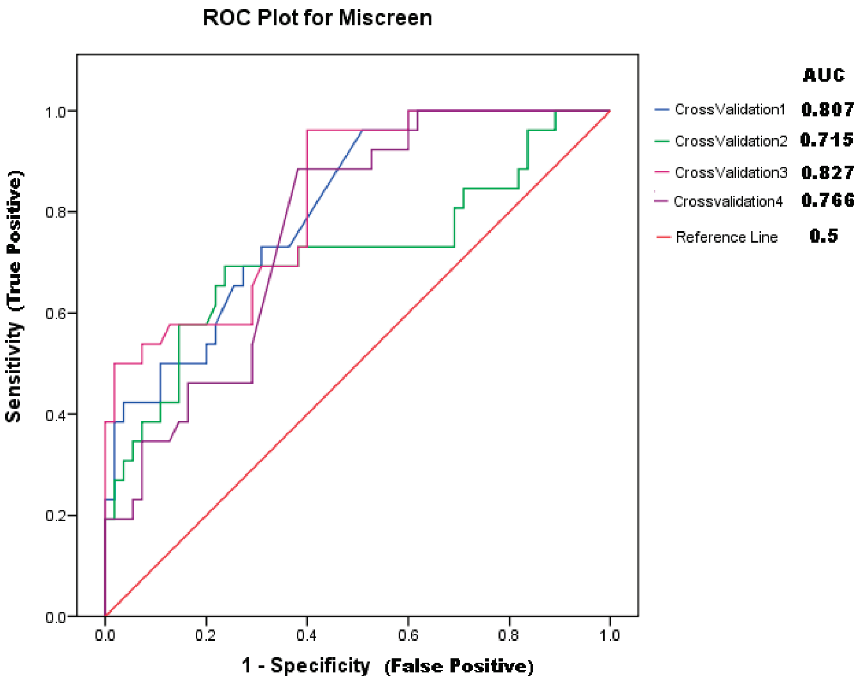

Figure 3. ROC curves showing the trade-off between sensitivity and 1-specificity calculated using SPSS. The results for the cross-validated runs are color coded and the reference line indicates random prediction.

x-axis), was carried out. The overall performance of the model was judged based on the area under the curve (AUC) values. Average AUC value of 0.778 (of four cross-validation runs) was obtained, and the enrichment ability for the developed bioactivity model is shown in Figure 3. Compounds with a bioactivity score of >0.3 were retained for further analysis. Though the current Bayesian model was applied as virtual-screening tools, it also identified the essential substructural features that discriminates active and inactive molecules. Some lists of substructural features responsible for activity are provided in Supporting Information (Figure 1) and could serve as valuable information for

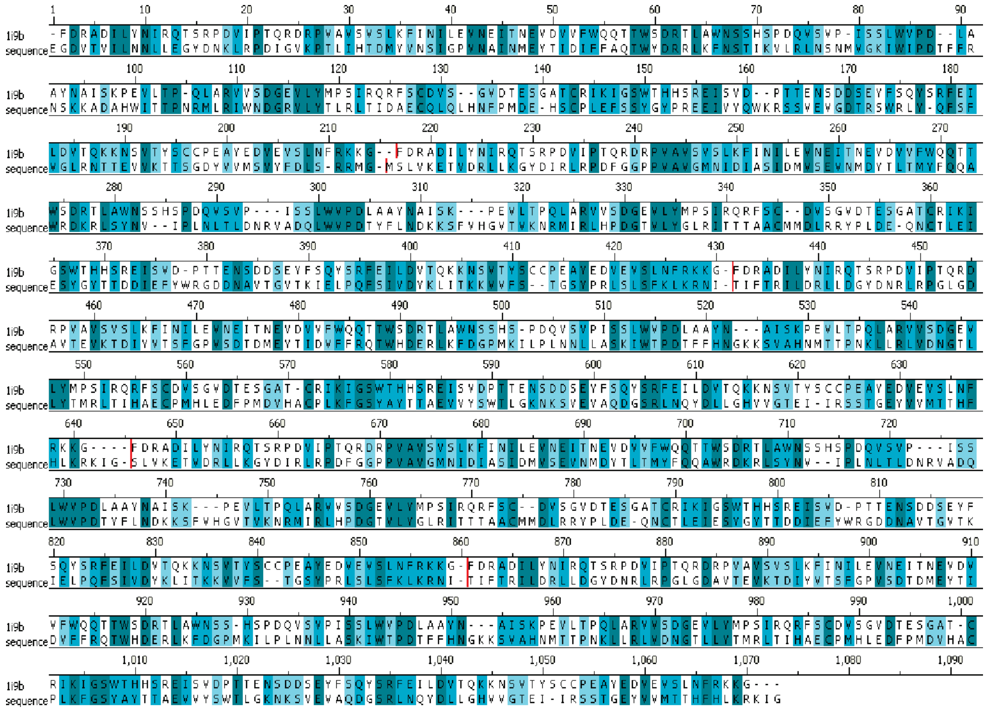

Figure 4. Aligned view of the template and the target sequence carried out using Fugue.

chemists/computational chemists during a hit-to-lead or lead optimization endeavor.

CoIFA − A Receptor-Dependent 3D QSAR Approach. To carry out CoIFA, the prerequisite receptor structure was modeled using Modeler. Fugue alignment was carried out to align the template sequence with the target sequence. The obtained alignment is shown in Figure 4, and the degree of conservity corresponds to the intensity of the color coding.

The best model was determined based on the lowest value of the Modeler objective function. The modeled structure was evaluated for stereochemical accuracy, fold reliability, and packing quality using Procheck[56] and whatif.[57] Profile-3D[58] overall self-compatibility scores for the models were much higher than the lowest possible scores conceivable for such a model. No major misfolded regions were present near the active site, as evident for a Profile-3D (Supporting Information, Figure 2) and a residue-by-residue energy assessment carried out using discrete optimized protein energy (DOPE)[59] score. Ramachandran plot[60] (Supporting Information, Figure 3) of the $\Phi-\Psi$ angle distribution showed that only 1.3% of the residues fall in the disallowed region. The backbone Cα trace of the modeled GABA$_A$ receptor model and its template AChBP overlapped well (rmsd 0.66 Å), see Figure 5 and Table 2.

The modeled structure was consistent with the overall topology as evident in ligand gated ion channels (LGIC). The modeled extracellular ligand binding domain displays a β-rich domain forming a sandwich-like architecture. The luminal (inner) and abluminal (outer) β sheets are connected by the signature disulfide bridge. The helical pore-forming domain's architecture was also consistent with other experimentally resolved LGIC. Hydrogens were added to the modeled protein using the Protonate-3D module of MOE.

Ionization states for the titratable groups were assigned, assuming a pH of 7.2. The added hydrogen's were minimized using a CHARMM force field,[61] keeping the rest of the system static to relieve steric clash and to improve hydrogen-bonding geometry. A total of 133 ligands were collected and docked to the benzodiazepine active site using GOLD. Compounds with a reported pK_i ≥ 7 were considered as active (1), and those with a pK_i < 7 were considered as inactive (0). The interaction finger prints generated using the PLIF module of MOE are shown in Figure 6.

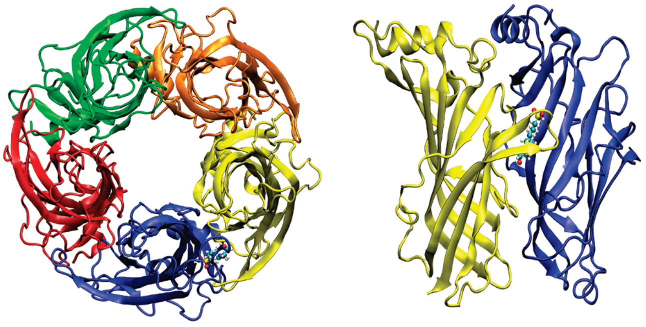

Figure 5. The pentameric, model of the GABA$_A$ α3 extra cellular domain as viewed form the extracellular side. Chain A (γ_2), blue; chain B (β_2), red; chain C (α_3), green; chain D (β_2), orange; and chain E (α_3), yellow. A close view of the α3/γ2 interface that defines the benzodiazepine binding cavity displaying a HEPES buffer molecule in the binding site rendered in ball-and-stick representation.

Table 2. Model Validation Parameters Obtained for the Homology Model

validation parameter	values
PDF total energy[a]	7493.97
$\Phi-\Psi$ violations[b]	1.3%
Profile-3D[c]	357.49
DOPE score[d]	−76 717

[a] A pseudo-energy function obtained from Modeler. [b] Percentage of residues lying in the disallowed region. [c] The minimum and maximum possible scores for a model are 207.375 and 460.833, respectively. [d] A statistical potential function for predicting errors in modeled proteins.

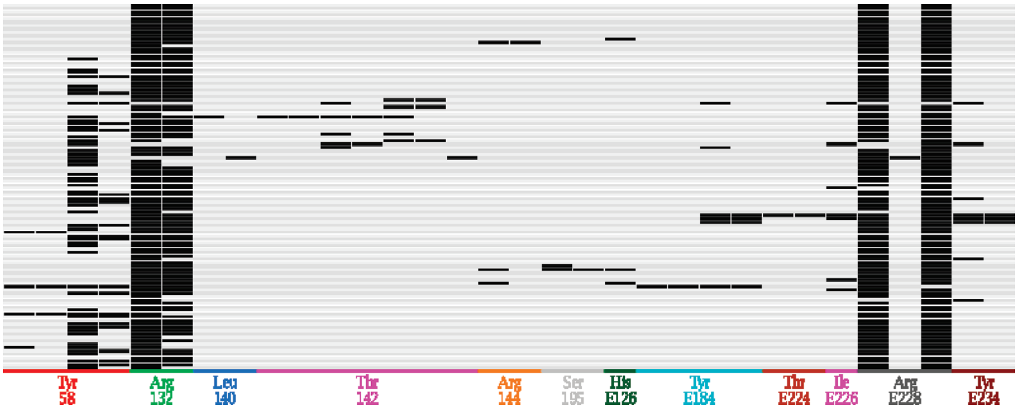

Figure 6. An interaction fingerprint generated for the docked complexes with GABA$_A$ α3. The profile shown here represents the interaction features of the ligands with the active site residues.

Table 3. Confusion Matrix Obtained for CoIFA QSAR Model

		predicted		
		active	inactive	total
observed	active ($72_{train}18_{test}$)	TP ($61_{train}16_{test}$)	FP ($8_{train}2_{test}$)	TP + FP ($69_{train}18_{test}$)
	inactive ($36_{train}7_{test}$)	FN ($11_{train}2_{test}$)	TN ($28_{train}5_{test}$)	FN + TN ($39_{train}7_{test}$)
	total ($108_{train}25_{test}$)	N_{act} = TP + FN ($72_{train}18_{test}$)	N_{inact} = FP + TN ($36_{train}7_{test}$)	$N = N_{act} + N_{inact}$($108_{train}25_{test}$)

All the interacting residues were treated equally (given equal weightage). Six types of interactions were defined, namely side chain and backbone hydrogen bonds (donors or acceptors) and ionic and surface interactions. The generated interaction fingerprints were assigned probabilistic contact values based on MOE methodology. Details of probabilistic contact value calculations are covered elsewhere and are not elaborated for reasons of brevity.[43] These values obtained using MOE served as pseudo-energy values to construct a categorical 3D QSAR model using a recursive partitioning algorithm (decision tree). Irrespective of the nature of the modeling technique employed, validation and predictivity are mandatory for any QSAR model.

The performance of the classification method was evaluated based on the obtained confusion matrix (Table 3). Different metrics, namely the classification error (E), precision (P), recall (R), values enrichment factor (EF), and the Matthews correlation coefficient (MCC) were used as a measure of its performance. The formulas used for calculating the respective metrics are stated below.

$$E = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \tag{6}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{7}$$

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{8}$$

$$\text{EF} = \frac{\text{TP}/(\text{TP} + \text{FP})}{(\text{TP} + \text{FN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \times 100\% \tag{9}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{10}$$

where the classification error (E) provides an overall error estimate, whereas recall R measures the percentage of correct predictions, and precision (P) gives the percentage of observed positives that are correctly predicted. The enrichment factor (EF) describes to what degree true positives are over-represented in the selected data set compared to that of the whole set of compounds. The obtained CoIFA model had an overall error estimate of 17.59%, a recall ability of 84.73%, a precision value of 88.4%, and an enrichement value of 131.94% for the training set. The MCC, regarded as one of the best measures to quantify the correlation coefficient between the observed and the predicted binary classification, was used to evaluate the obtained model. MCC values of 0.61 for the training set and 0.60 for the test set signify the predictive ability of the classification approach.

A modified correct classification rate (CCR) metric, as described by Weston et al.,[62] was used to prevent artifacts that occur when the numbers of compounds belonging to different classes are significantly disproportionate. Accordingly, the CCR formula suited for an unbalanced classification problem is

$$\text{CCR} = 0.5(\text{TP}/N_{act} + \text{TN}/N_{inact}) \tag{11}$$

In the above expression, N_{act} and N_{inact} signify the total number of actives and inactives, whereas TP signifies the number of compounds predicted as active, and TN signifies the number of compounds predicted as inactive. The CoIFA model obtained had a correct classification rate of 0.812 for the training set. For rigorous validation, an external test set of 25 compounds (18 actives and 7 inactives) that was not used during the model development was considered for validation. Classification accuracy of 0.80 for the test set reveals that the CoIFA model, derived based on the interaction patterns, shows good discriminative ability in distinguishing binders and non-binders. This clearly suggests the existence of an underlying classification pattern based on the interaction patterns. Such a type of study could also shed information about the hotspots in the protein active site. The encouraging results obtained from our study assures that even homology modeled structures derived from structures with reasonable homology also proves to be a valuable tool for a mining chemical database on the basis of interaction patterns and would be a value-added approach to post-process docked outputs, in light of the mediocrity of the current scoring functions. Interaction-based filtering has proven to be a successful tool for post-filtering docked outputs as evident from many studies.[53]

Classical 2D QSAR-Based Predictive Data Mining. Predictive QSAR models, developed using two statistical techniques (GFA and GPLS), were in turn used to build a consensus QSAR prediction strategy. The results obtained for the best GFA and GPLS models obtained are discussed below in detail. The GFA model was obtained on a population size of 100 with 25 000 generations. The mutation probabilities were kept at system default. The G/PLS model derived was optimal with four components, as determined by q^2 (cross-validated square correlation coefficient). The best model obtained for GFA was an explained variance of 72.1% and a predicted variance of 67.1%.

The best QSAR equation obtained for GFA model is shown in eq 12:

$$\mathrm{p}K_{\mathrm{i}} = 8.14817 - 0.06079*\text{″PEOE\_VSA\_PPOS″} + 0.016144*\text{″}S\text{log }P\text{\_VSA9″} - 0.017353*\text{″PEOE\_VSA-0″} + 0.000235*\text{″weinerPath″} + 0.010737*\text{″PEOE\_VSA\_POS″} - 0.056294*\text{″PEOE\_VSA + 3″} - 1.00127*< - 6.82572 - \text{″log }S\text{″} > -0.016909*<\text{″PEOE\_VSA + 0″} - 85.5211> \quad (12)$$

$N_{\mathrm{training}} = 86$, $N_{\mathrm{test}} = 17$, optimal number of components (ONC) = 4, $R^2 = 0.742$, adjusted R^2 (R^2_{adj}) = 0.721, Bs R^2 = 0.748, R^2_{cv} (q^2) = 0.687, randomized R^2 = 0.383, PRESS = 8.764, R^2_{pred} = 0.671, LOF = 0.234, F test = 28.171, $(R^2 - R^2_{\mathrm{O}}) / R^2 = 0$, and $k = 1$.

Regression using spline terms allows the incorporation of nonlinear modeling. The appearance of log S in the equation as a spline term with a negative regression coefficient suggests that if the value of log S is more than −6.82572, then it shows a negative contribution toward the activity. The appearance of PEOE_VSA_PPOS in the equation with coefficients of an opposite sign suggests that there may be a parabolic relationship between the activity and the parameter PEOE_VSA_PPOS. Though a bilinear modeling, incorporating squared terms could have shed more light on the effect of this variable on activity, we confine to the present model as it exhibited good predicivity and stability required to address the goal of the current study. To vindicate this speculation, the G/PLS model, as shown in eq 13, reveals the appearance of the parameter PEOE_VSA_PPOS as a spline term suggesting that a value less than 23.3534 contributes positively for activity. GFA model had three compounds (55, 44, and 19) and GPLS model had four compounds (55, 3, 51, and 41) with residual values exceeding twice the standard deviation. These compounds were considered as outliers of which 55 appeared as common outliers in both models. Removal of outliers from the QSAR models remains a contentious issue, as is often termed as a method of polishing R^2 value, unless obvious evidence supports the uniqueness of the outlier compound.[63] Activity forecasting was carried out using predictive QSAR models inclusive of outliers in light of its good statistical quality index and of the relatively low number of outliers evident from the models. The best G/PLS model obtained in terms of statistical significance is expressed in eq 13. G/PLS model obtained had an explained variance of 76.1% and a predicted variance of 67.9%. The best model was obtained on a population size of 100 with 25 000 generations, with an optimal number of components equal to three. The mutation probabilities were kept at the system default.

$$\mathrm{p}K_{\mathrm{i}} = 8.50466 + 0.230131*<\text{″Kier1″} - 20.5959 > + 0.014112*\text{″}S\text{ log }P\text{\_VSA9″} - 0.964241*< -6.87042 - \text{″log }S\text{″} > -0.01636*\text{″PEOE\_VSA-0″} - 0.136879*< 33.9457 - \text{″vsa\_pol″} > + 0.071594*<23.3534 - \text{″PEOE\_VSA\_PPOS″} > -0.046163*\text{″PEOE\_VSA + 3″} \quad (13)$$

$N_{\mathrm{training}} = 86$, $N_{\mathrm{test}} = 17$, ONC = 3, $R^2 = 0.770$, adjusted R^2 (R^2_{adj}) = 0.761, Bs R^2 = 0.754, R^2_{cv} (q^2) = 0.692, randomized R^2 = 0.362, PRESS = 8.764, R^2_{pred} = 0.679, $(R^2 - R^2_{\mathrm{O}}) / R^2 = 0$, and $k = 1$.

A short explanation of the descriptors that appeared in the final model is provided in the Supporting Information (Table 2). One important consideration in QSAR studies is the "interpretability" of the model to provide structural insights as to what modifications of the existing compounds might be promising. Since an earlier 3D QSAR study by us[55] has dealt upon those aspects and the present focus being on activity forecasting, emphasis was stressed on the predictive ability of the models. A regress diagnostic R^2 carried out to measure the goodness of interpolation of GFA and G/PLS models was statistically high (0.742 and 0.77). Since an individual QSAR model may overemphasize some descriptors and may overlook some important features completely, it seems reasonable to anticipate that a consensus QSAR model, derived from the averages of predictions from individual models, may provide a better statistical fit and predictive ability, as compared to individual models. Hence, a C-QSAR model was derived as a nonweighted average prediction of GFA and G/PLS. Apparently, the results obtained from our C-QSAR model were superior that of the individual models with an R^2 value of 0.77 and a predicted variance of 69.4%. See Figure 7.

Since the value of R^2 tends to be inflated as the number of terms in the equation increases, R^2_{adj} was also calculated for expository reasons. The near value of R^2 and R^2_{adj} for both of the models ensures the absence of over fitting. Internal predictivity of the models was validated using two resampling techniques namely LOO cross-validation and bootstrapping. Obtained R^2_{cv} ($q2$) values of 0.687 (GFA), 0.692 (G/PLS), and 0.694 (C-QSAR) ensure the predictive ability of the models. Model stability and chance correlation were evaluated by subjecting the developed models to a Y-randomization procedure that scrambles the dependent variable set and that rebuilds a new QSAR model based on the permuted response. Randomized R^2 values of 0.383 (GFA) and 0.362 (G/PLS) apparently signify the stability of the models. The above mentioned metrics are indicative of the interpolative ability of the models and do not reflect the extrapolative ability

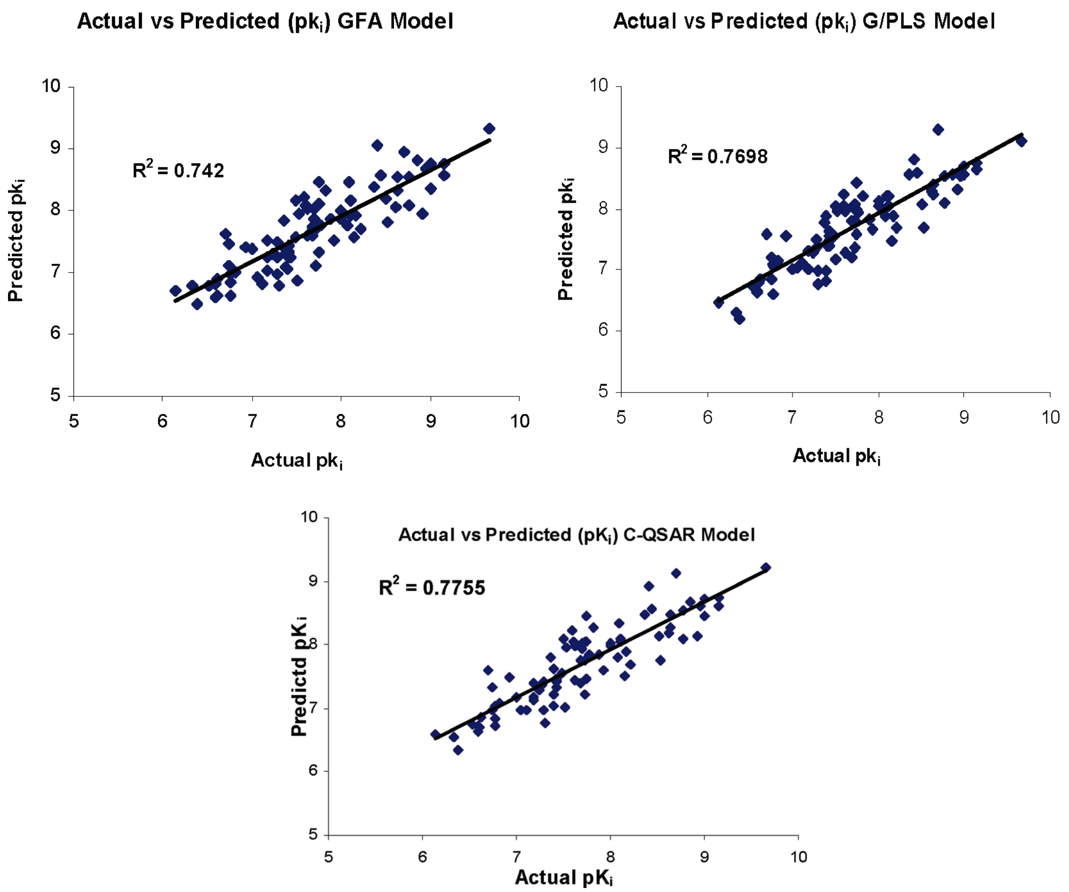

Figure 7. Plot of actual vs predicted pK_i obtained for GFA, G/PLS, and C-QSAR models.

of the models. The predictive ability and extensibility of the models were established using an external test set consisting of 17 compounds that were not considered during the model generation process. The predictive power of the models were calculated using the formula:[64]

$$R^2_{\mathrm{pred}} = \frac{\mathrm{SD} - \mathrm{PRESS}}{\mathrm{SD}} \qquad (14)$$

where SD is the sum of the square deviations between the biological activities of each molecule in the test and the mean activity of the training set molecules, and the PRESS is the sum of square deviations between the predicted and the actual activities of molecules in the test set. More rigorous validation was carried out using other parameters proposed by Tropsha et al.[64]

They recommended the use of the following formula for illustrating the predictive ability:

$$(R^2 - R_O^1)/R^2 < 0.1 \quad \text{or} \quad (R^2 - R_O^{2\prime})/R^2 < 0.1 \qquad (15)$$

and

$$k \text{ or } k' \text{close to} 1 \qquad (16)$$

where, R^2 represents the square correlation coefficient between the predicted and the observed activities. R_O^2 and $R_O^{2\prime}$ are quantities characterizing the square correlation between the predicted vs the observed and between the observed vs the predicted activity, respectively, with a y-intercept set to zero and with k, k' being their corresponding slopes. When embarking on an extrapolative adventure using QSAR models, assessing the applicability domain with the reference to the calibrated model is of immense importance to obtain predictions with confidence. A distance-based similarity method used to define the applicability domain based on a distance threshold value (D_T) was used to weed out compounds that fall outside the scope of the model domain in order to obtain reliable predictions. The overall statistical quality index, as evident from the footnotes provided in eqs 12 and 13, reveal that both the GFA and G/PLS models satisfy the acceptability criteria for a valid QSAR model ($R^2 > 0.6$, $q^2 > 0.5$, $R^2 - q^2 < 0.3$, and $R^2_{\mathrm{pred}} > 0.5$).[65]

Post Analysis of Leads. A robust insilico screening workflow of our bioisosteric library consisting of 790 944 compounds facilitated the identification of 123 consensus potential leads with predicted pK_i values greater than that of the established compounds. The reported pK_i values are based on consensuses QSAR predictions subjected to rigorous validation with extrapolation limited to the applicability domain. This ensures the highest possible accuracy in forecasting the biological activity of the compounds outside of the training set. See Figure 8.

More interestingly, it was found that some isosteric analogues, which were predicted to show good activity as per the C-QSAR model, were not predicted to be active analogs using our proteochemometric-based CoIFA approach. A close scrutiny reveals that those compounds pose an interaction pattern quite different from the other compounds. This discrepancy in the selection of leads between the classical QSAR and the proteochemometric-based CoIFA methods could be attributed to the feature selection issues

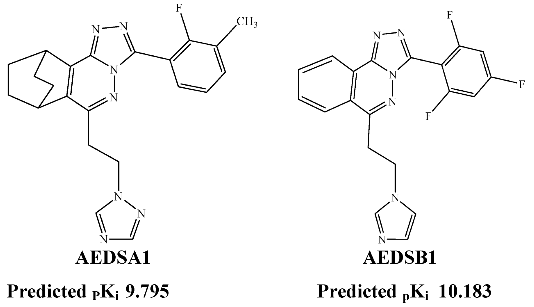

Figure 8. A 2D depiction of the top ranking hits along with their forecasted pK_i obtained, using integrated insilico screening.

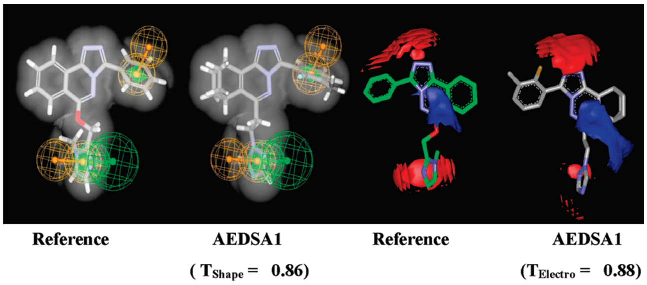

Figure 9. The lead compound AEDSA1 displaying a good amount of shape, chemistry, (pharmacophore), and electrostatic similarity with the reference compound (left).

in QSAR modeling. During the process of removing invariant variables, GA-based feature selection overlooks a small set of variant descriptors that are relatively constant to generate a statically valid model that faithfully reflects the SAR data. Hence, mining databases using QSAR models which have few variables may not be prudent, as QSAR models are mandated to faithfully correlate the variation in descriptor values with that of the target property upon which it was trained. Hence, integrated virtual screening using a proteochemometric-based QSAR method, like CoIFA which incorporates protein−ligand interaction cross terms, could be a value-added tool, since molecular recognition plays a crucial role in determining activity. As an additional measure of confidence, we used similarity metrics to establish the degree of analogy between the designed bioisosteric analogues and its progenitor molecule with regard to shape, chemistry, and electrostatics, together termed as electroform. Electroform calculations were performed using ROCS and EON. An atom-centered Gaussian shape-based overlay process optimized with respect to chemical features, as defined by a Mills and Dean force field,[66] was employed to measure the shape and the chemical similarity of the reference compound. The electrostatic potential maps of molecules were compared using EON, which employs partial charges calculated using a MMFF94 force field.[67] The obtained leads have a shape tanimoto similarity (T_{shape}) of 0.86 and an electrostatic similarity ($T_{electro}$) of 0.88, reinforcing the fact that they are true electroforms, which are more probable to be active.

A close scrutiny of the R_1 fragment of the hit coded AEDSA1, belonging to scaffold (A) with its progenitor molecules, reveals that phenyl substitutions with fluoro at position 2 and with methyl at position 3 have been explored earlier and were also shown to possess remarkable selectivity profile, but concurrent substitution of the phenyl ring with fluoro and methyl at positions 2 and 3, respectively, has not been explored. It is postulated that such a substitution could afford better affinity. The R_2 fragment shows the presence of an 1,2,4-triazol moiety, with an ethylene linker bridging the position 1 of the triazole ring and the position 6 of the 7,8,9,10-tetrahydro-7,10-ethano-1,2,4-triazolo[3,4-*a*]phthalazine core. Ethylene, an electroform to other ether linkers, like methoxy, thio ether, and amino ether, all reported earlier, is worth exploring in light of its high-predicted activity. Analysis of the potential lead coded AEDSB1 belonging to the core scaffold (B) reveals that a 2,4,6-fluorine-substituted phenyl moiety at the R_1 position is more probable to be active. Such a substitution is akin to the hydrophobic substitution (CH_3 and F) explored at various positions around the phenyl ring. Similarly, the occurrence of imidazole is not totally unexpected as some progenitor molecules have been synthesized with methyl and ethyl substitutions at various positions around the imidazole moiety, with the only variation being the attachment position of the ethylene linker. Overall, it is gratifying to see that the method employed here revealed some unexplored R-group fragments with more binding affinity that are deemed to be dissimilar based on intuition than analogs based on similarity values. See Figure 9.

CONCLUSION

Combinatorial chemistry techniques have been utilized to enumerate a targeted library of molecules by varying the Markush features using the concept of bioisosterism. High-throughput virtual profiling was carried out using Molinspiration "Miscreen" that uses Bayesian statistics. Extending the QSAR paradigm for virtual screening is an attractive approach given its general notion as a retrospective analysis tool for the ligand optimization approach. Conventional QSAR identifies relationships between some aspects of the molecular structure and the target property and seldom exploits the structural information of the receptors, deeming it to be termed a ligand-based approach. To make the best of the information in hand, we have put in practice a novel QSAR approach termed CoIFA that extends the 3D QSAR paradigm to incorporate receptor−ligand interaction profiles to classify molecules as active or inactive based on a pre-informed judgment obtained using a well validated recursive partitioning method. Such a QSAR model, when synchronized with classical QSAR models, further hones the predictive power of QSAR models as a powerful virtual-screening tool, as exemplified in the study. We believe that this extended view of QSAR modeling as a virtual screening tool presented in the paper brings a practical solution to mine novel leads using QSAR approaches rather than to confine QSAR modeling to obtain appreciable statistical values that faithfully quantify the SAR data.

ACKNOWLEDGMENT

The authors thank Council for Scientific and Industrial Research (CSIR), New Delhi, India, for providing a financial grant under the Mission Mode Program CMM 0017. Software support granted by OpenEye and Molinspiration is greatly acknowledged. Support service rendered by MOE is also acknowledged. We thank Dr. Macky Slimak of Molinspiration for his valuable support. R.S.K.V. thanks CSIR for a Senior Research Fellowship. I.B. and M.P. thank CSIR for a Project Fellowship.

Supporting Information Available: Stepwise K-means clustering of the compounds for selection of a training set and test set members, list of some important sub-structural features as evident from Bayesian modeling (Miscreen), and a profile 3D plot of the modeled GABA_A receptor. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kuffler, S. W.; Edwards, C. Mechanism of gamma aminobutyric acid (GABA) action and its relation to synaptic inhibition. *J. Neurophysiol.* **1958**, *21* (6), 589–610.

(2) Xue, H.; Chu, R.; Hang, J.; Lee, P.; Zheng, H. Fragment of GABA(A) receptor containing key ligand-binding residues overexpressed in Escherichia coli. *Protein Sci.* **1998**, *7* (1), 216–9.

(3) Haifeng, Shi; Shui, Ying, Tsang; Hui, Zheng; James, N. Sturgis; Hong, Xue. Two β-rich structural domains in GABAA receptor α1 subunit with different physical properties: Evidence for multidomain nature of the receptor. *Protein Sci.* **2002**, *11* (8), 2052–2058.

(4) Mehta, A. K.; Ticku, M. K. An update on GABAA receptors. *Brain Res. Brain Res. Rev.* **1999**, *29* (2−3), 196–217.

(5) Rudolph, U.; Crestani, F.; Benke, D.; Brunig, I.; Benson, J. A.; Fritschy, J. M.; Martin, J. R.; Bluethmann, H.; Mohler, H. Benzodiazepine actions mediated by specific gamma-aminobutyric acid(A) receptor subtypes. *Nature* **1999**, *401* (6755), 796–800.

(6) McKernan, R. M.; Rosahl, T. W.; Reynolds, D. S.; Sur, C.; Wafford, K. A.; Atack, J. R.; Farrar, S.; Myers, J.; Cook, G.; Ferris, P.; Garrett, L.; Bristow, L.; Marshall, G.; Macaulay, A.; Brown, N.; Howell, O.; Moore, K. W.; Carling, R. W.; Street, L. J.; Castro, J. L.; Ragan, C. I.; Dawson, G. R.; Whiting, P. J. Sedative but not anxiolytic properties of benzodiazepines are mediated by the GABA (A) receptor alpha1 subtype. *Nat. Neurosci.* **2000**, *3*, 587–592.

(7) Low, K.; Crestani, F.; Keist, R.; Benke, D.; Brunig, I.; Benson, J. A.; Fritschy, J. M.; Rulicke, T.; Bluethmann, H.; Mohler, H.; Rudolph, U. Molecular and neuronal substrate for the selective attenuation of anxiety. *Science* **2000**, *290* (5489), 131–4.

(8) Russell, M. G.; Carling, R. W.; Atack, J. R.; Bromidge, F. A.; Cook, S. M.; Hunt, P.; Isted, C.; Lucas, M.; McKernan, R. M.; Mitchinson, A.; Moore, K. W.; Narquizian, R.; Macaulay, A. J.; Thomas, D.; Thompson, S. A.; Wafford, K. A.; Castro, J. L. Discovery of functionally selective 7,8,9,10-tetrahydro-7,10-ethano-1,2,4-triazolo[3,4-a]phthalazines as GABA A receptor agonists at the alpha3 subunit. *J. Med. Chem.* **2005**, *48* (5), 1367–83.

(9) Carling, R. W.; Moore, K. W.; Street, L. J.; Wild, D.; Isted, C.; Leeson, P. D.; Thomas, S.; O'Connor, D.; McKernan, R. M.; Quirk, K.; Cook, S. M.; Atack, J. R.; Wafford, K. A.; Thompson, S. A.; Dawson, G. R.; Ferris, P.; Castro, J. L. 3-phenyl-6-(2-pyridyl)methyloxy-1,2,4-triazolo[3,4-a]phthalazines and analogues: high-affinity gamma-aminobutyric acid-A benzodiazepine receptor ligands with alpha 2, alpha 3, and alpha 5-subtype binding selectivity over alpha 1. *J. Med. Chem.* **2004**, *47* (7), 1807–22.

(10) Wermuth, C. G. Similarity in drugs: reflections on analogue design. *Drug Discov. Today.* **2006**, *11* (7−8), 348–54.

(11) *Molinspiration*; Molinspiration Cheminformatics: Slovensky Grob, Slovak Republic; http://www.molinspiration.com/. Accessed August 7, 2009.

(12) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38* (14), 2681–91.

(13) Gohlke, H.; Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45* (19), 4153–70.

(14) Datar, P. A.; Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *J. Comput. -Aided. Mol. Des.* **2006**, *20* (6), 343–60.

(15) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.

(16) Brejc, K.; van Dijk, W. J.; Klaassen, R. V.; Schuurmans, M.; van Der Oost, J.; Smit, A. B.; Sixma, T. K. Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature.* **2001**, *411* (6835), 269–76.

(17) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727–48.

(18) Clark, A. M.; Labute, P.; Santavy, M. 2D structure depiction. *J. Chem. Inf. Model.* **2006**, *46* (3), 1107–23.

(19) Clark, A. M.; Labute, P. 2D depiction of protein-ligand complexes. *J. Chem. Inf. Model.* **2007**, *47* (5), 1933–44.

(20) *Molecular Operating Environment (MOE)*; Chemical Computing Group: Montreal, Quebec, Canada, 2009.

(21) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, 1984.

(22) Hopfinger, A. J.; Rogers, D. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 854–866.

(23) Dunn, W. J.; Rogers, D. Genetic partial least squares in QSAR, In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London,1996.

(24) Gramatica, P.; Giani, E.; Papa, E. Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. *J. Mol. Graph. Model.* **2007**, *25* (6), 755–66.

(25) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the consensus modeling concept in genetic algorithms: application to interpretable discriminant analysis. *J. Chem. Inf. Model.* **2006**, *46* (5), 2110–24.

(26) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.

(27) *Discovery Studio*, Version 2.0; Accelrys Inc.: San Diego, CA, 2007.

(28) *GOLD*, Version 3.2; Cambridge Crystallographic Data Centre: Cambridge, U.K., 2006.

(29) *OpenEye*; OpenEye Scientific Software: Santa Fe, NM, 2006.

(30) *TSAR*, Version 3.0; Accelrys Inc.: San Diego, CA, 2007.

(31) *Cerius2*, Version 4.10; Accelrys Inc.: San Diego, CA, 2006.

(32) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comp. Method.* **1990**, *3*, 537–547.

(33) Langer, T.; Wolber, G. Virtual combinatorial chemistry and in silico screening: Efficient tools for lead structure discovery. *Pure. Appl. Chem.* **2004**, *76* (5), 991–996.

(34) John, I. M.; David, H, S. Designing Combinatorial Libraries for Efficient Screening. In *Methods in Molecular Biology Combinatorial Library*; English, L. B., Ed.; Humana Press Inc.: Totowa, NJ, 2002; pp 307−323.

(35) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47* (1), 47–58.

(36) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96* (8), 3147–3176.

(37) Pap, A.; Szommer, T.; Laszlo, B.; Gergely, G.; Ferdinandy, P.; Spadoni, C.; Ferenc, D.; Toshio, F.; Laszlo, U.; Dorman, G. Enhanced hit-to-lead process using bioanalogous lead evolution and chemogenomics: application in designing selective matrix metalloprotease inhibitors. *Expert Opin. Drug Discov.* **2007**, *2* (5), 707–723.

(38) Chen, X.; Wang, W. The use of bioisosteric groups in lead optimization. *Ann. Rep. Med. Chem.* **2003**, *38*, 338–346.

(39) http://www.rcsb.org/pdb/home/home.do (accessed Aug 7, 2009).

(40) Rebecca, C. W.; Stefan, H.; Ting, W. Using 3D protein structures to derive 3D-QSARs. *Drug Discov. Today: Technol.* **2004**, *1* (3), 241–246.

(41) Cramer, R.D., III.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–67.

(42) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovovery Des.* **2000**, *20*, 115–144.

(43) Labute, P. Probabilistic Receptor Potentials. *J. Chem. Computing Group* 2001; http://www.chemcomp.com/journal/cstat.htm (accessed Aug. 7, 2009).

(44) Shadnia, H.; Wright, J. S.; Anderson, J. M. Interaction force diagrams: new insight into ligand-receptor binding. *J Comput.- Aided. Mol. Des.* **2009**, *23* (3), 185–94.

(45) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–801.

(46) http://www.expasy.ch/sprot/(accessed Aug 7, 2009).

(47) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–10.

(48) Shi, J.; Blundell, T. L.; Mizuguchi, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **2001**, *310* (1), 243–57.

(49) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52* (4), 609–23.

(50) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791–804.

(51) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP - Potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20* (11), 1165−1176.

(52) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–56.

(53) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47* (2), 337–44.

(54) SPSS Version 15.0; SPSS Inc.: Chicago, IL, 2008.

(55) Vijayan, R. S.; Ghoshal, N. Structural basis for ligand recognition at the benzodiazepine binding site of GABAA alpha 3 receptor, and pharmacophore-based virtual screening approach. *J. Mol. Graph. Model.* **2008**, *27* (3), 286–98.

(56) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

(57) Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graphics* **1990**, *8* (1), 52−6. 29.

(58) Bowie, J. U.; Luthy, R.; Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, *253* (5016), 164–70.

(59) Shen, M. Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15* (11), 2507–24.

(60) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95–9.

(61) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545–614.

(62) Weston, J.; Perez-Cruz, F.; Bousquet, O.; Chapelle, O.; Elisseeff, A. Scholkopf. Feature Selection and Transduction for Prediction Of Molecular Bioactivity for Drug Design. *Bioinformatics* **2003**, *19*, 764–771.

(63) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSARs. *THEOCHEM* **2003**, (622), 39–52.

(64) Golbraikh, A.; Tropsha, A. Beware of q2. *J. Mol. Graph. Model.* **2002**, *20* (4), 269–76.

(65) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression based QSARs. *Environ Health Perspect.* **2003**, *111* (10), 1361–75.

(66) Mills, J. E.; Dean, P. M. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J. Comput. -Aided Mol. Des.* **1996**, 607–622.

(67) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–51.

CI900309S