

# Combinatorial QSAR Modeling of Specificity and Subtype Selectivity of Ligands Binding to Serotonin Receptors 5HT1E and 5HT1F

Xiang S. Wang,<sup>†</sup> Hao Tang,<sup>†,‡</sup> Alexander Golbraikh,<sup>†</sup> and Alexander Tropsha<sup>\*,†</sup>

Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products and Carolina Exploratory Center for Cheminformatics Research, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, and Molecular & Cellular Biophysics Program, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received November 6, 2007

The Quantitative Structure–Activity Relationship (QSAR) approach has been applied to model binding affinity and receptor subtype selectivity of human 5HT1E and 5HT1F receptor–ligands. The experimental data were obtained from the PDSP Ki Database. Several descriptor types and data-mining approaches have been used in the context of combinatorial QSAR modeling. Data mining approaches included *k* Nearest Neighbor, Automated Lazy Learning (ALL), and PLS; descriptor types included MolConnZ, MOE, DRAGON, Frequent Subgraphs (FSG), and Molecular Hologram Fingerprints (MHFs). Highly predictive QSAR models were generated for all three data sets (i.e., for ligands of both receptor subtypes and for subtype selectivity), and different individual techniques were proved best in each case. For real value activity data available for 5HT1E and 5HT1F ligand binding, models were characterized by leave-one-out cross-validated  $R^2$  ( $q^2$ ) for the training sets and predictive  $R^2$  values for the test sets. The best models for 5HT1E ligands were obtained with the *k*NN approach combined with MolConnZ descriptors ( $q^2 = 0.69$ ,  $R^2 = 0.92$ ); for 5HT1F ligands ALL QSAR method using MolConnZ descriptors gave the best results ( $R^2 = 0.92$ ). Rigorously validated classification models were also developed for the 5HT1E/5HT1F subtype selectivity data set with high correct classification accuracy for both training ( $CCR_{\text{train}} = 0.88$ ) and test ( $CCR_{\text{test}} = 1.00$ ) sets using *k*NN with MolConnZ descriptors. The external predictive power of QSAR models was further validated by virtual screening of The Scripps Research Institute (TSRI) screening library to recover 5HT1E agonists and antagonists (not present in the original PDSP data set) with high enrichment factors. The successful development of externally predictive and interpretative QSAR models affords further design and discovery of novel subtype specific GPCR agents.

## INTRODUCTION

G-Protein coupled receptors (GPCRs) represent the largest class of human proteins regulating vital biological and physiological functions. Naturally, GPCRs have been considered major targets for drug discovery. Various estimates place the percentage of modern drugs acting via GPCRs at 50–70%.<sup>1,2</sup> Still, the large proportion of GPCR ligands among current drugs by no means implies that the discovery of new pharmaceuticals acting via GPCRs is unlikely. In fact, recent advances in genomics and proteomics have led to the identification of the growing number of novel GPCRs many of which still have unknown physiological functions and are considered orphan receptors.

Rapid growth of biomolecular databases of both receptor sequences and ligands tested against panels of GPCRs such as PDSP Ki<sup>34</sup> (<http://pdsp.med.unc.edu/>) and GLIDA<sup>5</sup> (<http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/>) databases emphasize the growing need in rationalizing the wealth of information concerning the relationships between structure

and activity of GPCR ligands. The need to understand global structure–activity relationships of these ligands across the entire GPCRome is emphasized in many recent reports.<sup>6–8</sup> This challenge serves as a natural call to explore computational molecular modeling and cheminformatics approaches to accelerate the discovery of novel potent GPCR ligands.

Broadly speaking, computational drug discovery approaches include structure-based and ligands-based methods. The former require the knowledge of the three-dimensional structure of the target that can be obtained using either experimental approaches such as X-ray or NMR or computer-aided prediction relying on protein homology model building. It is well-known that even in those cases when high resolution X-ray structure of the target protein is available, accurate prediction of the bound poses of native ligands represents a formidable challenge.<sup>9,10</sup> Naturally, the use of receptor models for structure-based drug discovery studies with tools such as docking and scoring should be attempted with a great deal of caution.<sup>11</sup> Nevertheless, in the absence of experimentally characterized structures of human GPCRs (at least until recent reports on the X-ray structure of  $\beta$ -adrenergic receptor<sup>12,13</sup>), starting as early as 1992,<sup>14,15</sup> much effort has been going into 3D modeling of various GPCRs and the use of these models to search for ligands that bind to these receptors. In the meantime, ligand-based

\* Corresponding author phone: (919) 966-2955; fax: (919) 966-0204; e-mail: alex\_tropsha@unc.edu. Corresponding author address: CB #7360, Beard Hall, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360.

<sup>†</sup> School of Pharmacy.

<sup>‡</sup> School of Medicine.

methods, especially QSAR, have continued to play an indispensable role in the design of novel receptor–ligands<sup>16–18</sup> including those binding to GPCRs<sup>17,19–22</sup>

It is well-known that many ligands can bind to multiple GPCRs with high affinities.<sup>23</sup> This is especially true for amine GPCRs of Class A Rhodopsin like family, which contains major receptor targets for many psychoactive drugs. It is widely accepted that major mental illnesses are poly-genic, and many antipsychotic drugs have highly complex pharmacological profiles, with high affinity for multiple GPCRs.<sup>24,25</sup> Roth et al. proposed that antipsychotic drugs targeting a selected set of GPCRs (rather than a single receptor) could be more effective and have fewer side effects.<sup>23</sup> As part of the National Institute of Mental Health Psychoactive Drug Screening Program, the PDSP Ki Database<sup>34</sup> contains 46,275 Ki values (as of August 2007), including ca. 750 types of receptors and ca. 7000 test ligands; it is the largest database of its kind in the public domain. The majority of those receptors are GPCRs (549 types), along with various enzymes, ion channels, and transporters from a variety of species. This large data warehouse makes it possible to investigate global relationships between structure and (multiple) GPCR activities of test ligands, which is our long-term goal. One approach that we explore in this paper is to build both individual receptor-specific QSAR models as well as models that relate ligands chemical structure to the receptor subtype selectivity.

For our studies, a dual 5HT1E/5HT1F receptor system was selected from the PDSP Ki Database. The 5HT1E receptor<sup>26,27</sup> was found to be expressed predominantly in the hippocampus. Among all subtypes of 5-HT receptors, 5HT1E is the only receptor for which the biological function has not been characterized. Consequently, it has attracted significant attention in recent years. For instance, The Scripps Research Institute Molecular Screening Center (TSRI) and the NIMH Psychoactive Drug Screening Program (PDSP) have been actively screening various compound collections to identify chemical probes for the 5HT1E receptor.

The 5HT1F receptor was recently identified as a target for the acute treatment of headaches.<sup>28,29</sup> 5HT1E and 5HT1F receptors share 55% amino acid sequence identity, which is the highest among the 5HT receptor family making this receptor pair arguably the most difficult in terms of elucidating the molecular determinants of ligand specificity. Similar to many GPCRs, the 5HT1F receptor has a large number of ligands that can also bind to a broad range of other GPCRs. For example, for ketanserin, a strong binder of the 5HT1F receptor, the PDSP Ki Database also contains binding Ki values for as many as 15 other 5HT receptors, four dopamine receptors, and six other different receptors. This observation underlies a problem of the off-target binding that is likely to be observed with many novel antimigraine drugs targeting the 5HT1F receptor. The 5HT1E/5HT1F receptor pair could serve as an ideal exemplar system to study the off-target binding problem and use computational models to design highly selective 5HT1F antagonists.

Herein, we report on the development of validated and externally predictive regression QSAR models built independently for both 5HT1E and 5HT1F ligands. We further report on the development of binary classification QSAR models that were built using the selectivity index of 5HT1E/5HT1F ligands as a target property. We also explore these

models to identify specific chemical features responsible for both specificity and selectivity of receptor binding. These studies shed light on the structural determinants of 5HT1F-specific ligand binding and can help in designing potent antimigraine drugs that would bind to 5HT1F yet would not show the unwanted binding to its closest homologue, the 5HT1E receptor. To achieve the most statistically robust and predictive models we have employed the combinatorial QSAR strategy,<sup>30,31</sup> implemented within our validated QSAR modeling workflow.<sup>17</sup> This strategy enabled the identification of the best individual models of 5HT1E and 5HT1F binding and 5HT1E/5HT1F selectivity that are discussed in this paper following the methodological section.

## METHODS

**Data Sets for Model Building.** All data sets were retrieved from the PDSP Ki database (as of April 2006), incorporating the data published between 1992 and 2001 (Tables 1–3 of the Supporting Information). The 5HT1E ligand data set contains 40 compounds,<sup>32–40</sup> the 5HT1F ligand data set contains 29 compounds,<sup>38,40</sup> and the 5HT1E/5HT1F selectivity data set includes 22 compounds.<sup>38,40,41</sup> In addition to the published information, many PDSP certified data were included for model building. Though the binding data were drawn from different sources, the protocols for competitive radioligand binding were very similar. Among all the binding data for specific targets, only measurements using human cloned cell lines and <sup>3</sup>H-5HT as the radio-labeled ligands were considered. The cloning, sequencing, expression, and membrane preparation were carried out by standard techniques. The binding data were analyzed by nonlinear regression analysis, and IC<sub>50</sub> values were converted to Ki values by using the Cheng-Prusoff equation. It ensured that the biological end points from different sources are comparable and can be used in a combined data set. The range of binding affinities was 7.53 nM to 8 mM (Ki) for 5HT1E data set and 1.25 nM to 3.89 mM (Ki) for the 5HT1F data set, spanning three log units in both cases. The values of binding constants were spread relatively evenly within these ranges.

From the structural perspective, the compounds inside all three data sets are highly diverse. For example, the chemical scaffolds of ligands in the 5HT1E data set fall into at least six classes, such as tryptamine derivatives, ergoline derivatives, tertiary amines, butyrophenone derivative, yohimban derivatives, and piperazine derivatives. This diversity provides a challenge for building QSAR models. On the other hand, if successful models are built for these data sets they are likely to be robust and the applicability domain of such models is expected to be broad, i.e., they could be used prospectively to predict possible antipsychotic activity for diverse organic compounds.

**External Data Sets for Model Validation.** When our modeling studies were almost completed new binding data for 5HT1E receptor system were deposited into the PDSP Ki database.<sup>42</sup> They were employed at the later phase of modeling as the independent external validation set (vide infra). Furthermore, we also applied QSAR models developed for 5HT1E ligands for virtual screening of the chemical library examined by the TSRI screening center. TSRI is one of the ten screening centers of the NIH-funded Molecular

Libraries Screening Centers Network (MLSCN).<sup>43</sup> Specifically we focused on the active compounds of the confirmatory assays AID726 (Dose Response Cell Based Assay for Agonists of the 5HT1E) and AID749 (Dose Response Cell Based Assay for Antagonists of the 5HT1E) that are deposited in the PubChem.<sup>44</sup> We have applied QSAR models built with the PDSP data for 5HT1E ligands for virtual screening of the TSRI library of 64925 compounds. This study afforded us another opportunity to examine our models for their power to recover known experimental hits.

**MolConnZ Descriptors.** The MolConnZ software affords the computation of a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices,<sup>45–47</sup> kappa molecular shape indices,<sup>48,49</sup> topological and electrotopological state indices,<sup>50–52</sup> differential connectivity indices, graph's radius and diameter,<sup>53</sup> Wiener and Platt indices,<sup>54</sup> Shannon and Bonchev-Trinajstić information indices,<sup>55</sup> counts of different vertices, counts of paths and edges between different kinds of vertices.

We used MolConnZ version 4.05 software,<sup>56</sup> which initially produced ca. 700 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, 644 chemically relevant descriptors were initially calculated and 296 descriptors were eventually used for 5HT1E ligands data set, 282 descriptors for 5HT1F ligands data set and 273 descriptors for 5HT1E/5HT1F selectivity data set after deleting descriptors with zero value or zero-variance. MolConnZ descriptors were range-scaled prior to distance calculations since the absolute scales for MolConnZ descriptors could differ by orders of magnitude.<sup>57</sup> Accordingly, our use of range-scaling helps avoid giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional MolConnZ descriptor space.

**MOE Descriptors.** MOE descriptors include both 2D and 3D molecular descriptors. 2D descriptors include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity<sup>45–47</sup> and kappa shape indices,<sup>48,49</sup> adjacency and distance matrix descriptors,<sup>53,58–61</sup> pharmacophore feature descriptors, and PEOE partial charge descriptors.<sup>62</sup> 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors.<sup>63</sup> The conformations of each compound were generated from the SMILES string using the 3D conversion function in MOE and energy minimized. In total, 184 2D molecular descriptors and 67 3D molecular descriptors were calculated (251 total) using MOE 2006.08 software.

**DRAGON Descriptors.** DRAGON descriptors were classified into 0D, 1D, 2D, and 3D descriptors. The version 5.4 of the Dragon software<sup>64</sup> afforded 1664 descriptors total, covering a wide variety of types. For example, its 0D descriptors contain constitutional descriptors,<sup>65</sup> 1D descriptors include functional group counts and atom-centered fragments;<sup>66</sup> 2D descriptors include topological descriptors, connectivity indices, information indices and eigenvalue-based indices.<sup>67</sup> It should be pointed out that there are many novel descriptor families among 3D descriptors, such as RDF

descriptors,<sup>68</sup> 3D-MoRSE descriptors,<sup>69</sup> WHIM descriptors,<sup>70</sup> GETAWAY descriptors<sup>71</sup> and geometrical descriptors.<sup>72</sup> All descriptors were cleaned up by eliminating the constant variables and near-constant variables using the built-in function of DRAGON Professional 5.4. The pairwise correlations for all descriptors were examined and one of the two descriptors with the correlation coefficient  $R^2$  of 0.95 or higher was excluded.

**Frequent Subgraph Descriptors.** Frequent subgraph mining of chemical structures is a novel approach to generating fragment frequent subgraph (FSG) descriptors that was developed recently in our group.<sup>73</sup> The fragments are derived based on recurring substructures found in at least a subset of molecules (defined by a support value  $\sigma$ ) in the data set. These recurring substructures can implicate chemical features responsible for compounds' biological activities. First, chemical structures were converted into labeled, undirected graph representations. Fast frequent subgraph mining (FFSM) algorithm<sup>74</sup> was then used to find common frequent subgraphs for a given support value ( $\sigma$ ), which is one of the variables defined by the user. The redundant subgraphs were identified and removed leaving only so-called "closed subgraphs". A subgraph  $SG_i$  is closed in a database if there exists no supergraph  $SG_j$  such that  $SG_i \subseteq SG_j$  and  $\sigma_{SG_i} = \sigma_{SG_j}$ . However, subgraph  $SG_i$  will not be deleted if it also occurs by itself (not as part of the  $SG_j$ ) in the graph database. Removing redundant subgraphs (fragments) will reduce the number of subgraphs drastically and therefore make the subsequent calculations more efficient. The frequency of individual 'closed subgraphs' in each molecule of the data set is calculated and used as the descriptor value for each molecule.

**Molecular Hologram Fingerprints.** Molecular hologram fingerprints (MHFs),<sup>75,76</sup> or HQSAR descriptors, are built into the HQSAR module of Sybyl 7.3 (Tripos Inc.). The 2-D structures are broken into all possible linear and branched fragments, with the default size varying between 4 and 7 atoms. Each unique fragment is characterized by atom types (A), bond types (B), connectivity (C), hydrogen atoms (H), and chirality (Ch) and then assigned a specific large positive integer by the means of a cyclic redundancy check (CRC) algorithm. The fragment string could be used directly to build the integer array or to be hashed into the array bins of the fixed length  $L$  ( $L$  is generally in the range 50–500). Bin occupancies are incremented according to the fragments generated and constitute the individual values of the molecular hologram of a particular length.

In the HQSAR module, the MHFs are used as variables in the Partial Least Squares (PLS) analysis to generate QSAR models. In our studies, MHFs were used as descriptors for  $k$ NN calculations. It is expected that  $k$ NN QSAR models derived from the MHFs will be affected by a number of parameters with regard to hologram generation: fragment size, hologram length, and fragment distinction. Thus, seven combinations of the fragment distinctions, i.e., AB, ABC, ABCH, ABCHCh, ABH, ABHCh, and ABCCh were considered together with 12 default hologram length values ranging from 53 to 401 bins.<sup>77,78</sup> The MHFs were filtered to remove the invariant (identical value) descriptors and range-scaled prior to  $k$ NN QSAR calculations.

**Sphere Exclusion Algorithm.** Following our standard model development protocols, the data sets were subdivided



into multiple training/test set pairs using the Sphere Exclusion method<sup>79</sup> implemented in our laboratory. The number of compounds in the test set was varied to achieve the largest possible size of the test set, while ensuring that the training set models were still able to predict the binding affinity of the test set compounds accurately.

The procedure implemented in this study starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Let  $D_{\min}$  and  $D_{\max}$  be the minimum and maximum elements of **D**, respectively.  $N$  probe sphere radii are defined by the following formulas:  $R_{\min} = R_1 = D_{\min}$ ,  $R_{\max} = R_N = D_{\max}/4$ , and  $R_i = R_1 + (i-1)*(R_N - R_1)/(N-1)$ , where  $i = 2, \dots, N-1$ . Each probe sphere radius corresponds to one division into the training and the test set. A sphere-exclusion algorithm used in this study consisted of the following steps. (i) Select randomly a compound. (ii) Include it in the training set. (iii) Construct a probe sphere around this compound. (iv) Select compounds from this sphere and include them alternatively into the test and the training sets. (v) Exclude all compounds from within this sphere from further consideration. (vi) If no more compounds are left, then stop. Otherwise let  $m$  be the number of probe spheres constructed and  $n$  be the number of remaining compounds. Let  $d_{ij}$  ( $i = 1, \dots, m$ ;  $j = 1, \dots, n$ ) be the distances between the remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest  $d_{ij}$  value and go to step (ii).

**$k$  Nearest Neighbors ( $k$ NN) QSAR.** The  $k$ NN QSAR method<sup>57</sup> is based on the  $k$  nearest neighbors principle and the variable selection procedure. It employs the leave-one-out (LOO) cross-validation (CV) procedure and a simulated-annealing algorithm for the variable selection. The procedure starts with the random selection of a predefined number of descriptors from all descriptors. If the number of nearest neighbors  $k$  is higher than one, the estimated activities  $\hat{y}_i$  of compounds excluded by the LOO procedure are calculated using the following formula

$$\hat{y}_i = \frac{\sum_{j=1}^k y_j w_{ij}}{\sum_{j=1}^k w_{ij}} = \frac{\sum_{j=1}^k y_j w_{ij}}{k-1} \quad (1)$$

where  $y_j$  is the activity of the  $j$ th compound. Weights  $w_{ij}$  are defined as

$$w_{ij} = 1 - \frac{d_{ij}}{\sum_{j'=1}^k d_{ij'}} \quad (2)$$

and  $d_{ij}$  is Euclidean distances between compound  $i$  and its  $j$ th nearest neighbor.

For regression  $k$ NN, cross-validated  $R^2$  ( $q^2$ ) is calculated

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

where  $\bar{y}$  is the average observed activity. The summation in (3) is performed over all compounds. After each run, a predefined small number of descriptors are randomly replaced by other descriptors from the original pool, and the new value of  $q^2$  is obtained. If  $q^2$  (new) >  $q^2$  (old), then the new set of

descriptors is accepted. If  $q^2$  (new)  $\leq$   $q^2$  (old), then the new set of descriptors is accepted with probability  $p = \exp(q^2$  (new) -  $q^2$  (old))/ $T$  or rejected with probability  $(1-p)$ , where  $T$  is a simulated annealing "temperature" parameter. During this process,  $T$  is decreasing until a predefined threshold. Thus, the optimal (highest)  $q^2$  is achieved (see ref 57 for additional details). For the prediction, the final set of selected descriptors is used, and expressions 1 and 2 are applied to predict activities of compounds of the test sets.

For classification  $k$ NN, the predicted  $\hat{y}_i$  values (see expression 1) are rounded to the closest whole numbers (which are, in fact, the class numbers), and the prediction accuracy (correct classification rate,  $\text{CCR}_{\text{train}}$ ) is calculated as follows:

$$\text{CCR} = 0.5 \left( \frac{N_1^{\text{corr}}}{N_1^{\text{total}}} + \frac{N_2^{\text{corr}}}{N_2^{\text{total}}} \right) \quad (4)$$

where  $N_j^{\text{corr}}$  and  $N_j^{\text{total}}$  are the number of correctly classified and total number of compounds of class  $j$  ( $j = 1, 2$ ). Then a predefined small number of descriptors are randomly replaced by other descriptors from the original pool, and the new value of  $\text{CCR}_{\text{train}}$  is obtained. If  $\text{CCR}_{\text{train}} > \text{CCR}_{\text{train}}$ , then the new set of descriptors is accepted. If  $\text{CCR}_{\text{train}}(\text{new}) \leq \text{CCR}_{\text{train}}(\text{old})$ , then the new set of descriptors is accepted with probability  $p = \exp(\text{CCR}(\text{new}) - \text{CCR}(\text{old}))/T$  or rejected with probability  $(1-p)$ , where  $T$  is a simulated annealing "temperature" parameter. During this process,  $T$  is decreasing until a predefined threshold. Thus, the optimal (highest)  $\text{CCR}_{\text{train}}$  is achieved. For the prediction, the final set of selected descriptors is used, and expressions 1 and 2 are applied to predict activities of compounds of the test sets. Then the activities are rounded to the closest whole numbers, and the correct classification rate for the test set is calculated using formula 4.

In the case when compounds belong to two classes (e.g., active and inactive compounds), a  $2 \times 2$  confusion matrix can be defined, where  $N_{\text{act}}$  and  $N_{\text{inact}}$  are the number of active and inactive compounds in the data set, and TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives. The following classification accuracy characteristics associated with confusion matrices are widely used in QSAR studies: sensitivity ( $S = \text{TP}/N_{\text{act}}$ ), specificity ( $\text{SP} = \text{TN}/N_{\text{inact}}$ ), and enrichment  $E = \text{TP} \cdot N / [(\text{TP} + \text{FP}) \cdot N_{\text{act}}]$ . In this study, we have employed normalized confusion matrices. A normalized confusion matrix can be obtained from the non-normalized one by dividing the first column by  $N_{\text{act}}$  and the second column by  $N_{\text{inact}}$ . Normalized enrichment is defined in the same way as  $E$  but is calculated using a normalized confusion matrix:  $E_n = 2\text{TP} \cdot N_{\text{inact}} / [\text{TP} \cdot N_{\text{inact}} + \text{FP} \cdot N_{\text{act}}]$ .  $E_n$  takes values within the interval of  $[0, 2]$ .

**ALL QSAR.** For a large or diverse data set, sometimes it is difficult to establish a global linear or even nonlinear relationship between variables and the target property in the high-dimensional descriptor space. To solve this problem, the ALL QSAR method creates a series of locally weighted regression models that employ only a small fraction of compounds in the entire data set, which are chemically similar to the query compound. In the  $M$  dimensional descriptor space, the linear regression can be obtained<sup>80</sup>

$$\hat{y}(x) = \beta^T f(x) \quad (5)$$

where the  $\hat{y}(x)$  is the observed response vector corresponding to the input vector,  $x$ .  $\beta^T$  is the coefficient vector for the model, that  $\beta^T = (\beta_1, \beta_2, \dots, \beta_M)$  and  $f(x) = [f_1(x), f_2(x), \dots, f_M(x)]$ . The residual  $\epsilon$  is calculated as follows:

$$\epsilon = \sum_{k=1}^N w_k^2 (y_k - T f_k)^2 \quad (6)$$

Here  $w_k$  is the weight of each point. For a query compound in the test set, the local linear regression assigns higher weights to compounds of the training set that are closer to the test compound in the descriptor space. The weighting function, also called a kernel function, is a distance-based Gaussian function:

$$w = \exp(-d^2/2K^2) \quad (7)$$

where  $w$  is the weight of a point in the training set,  $d$  is the distance between this point and the query, and  $K$  is a smoothing parameter known as bandwidth.<sup>80</sup> The  $K$  value is optimized during the model building in search for maximal  $R^2$ , starting from 0.01 to 1.00 with the step value of 0.01. This method was developed in our group and applied successfully to several experimental chemical data sets earlier.<sup>81</sup>

**Partial Least Square (PLS) QSAR.** The PLS QSAR method was employed in the study using the QuaSAR-Model module of MOE 2006. This is arguably the most traditional and least sophisticated QSAR approach among those explored in this study. It was explored here to test if it could be possible to build reliable models for underlying data sets using the simplest approach. The number of components was set to no limit on the degree of the fit. The maximum condition number of the principal component transform of the correlation matrix **S**, the condition limit, was set to be a very large number of  $1.0 \times 10^6$ . The leave-one-out cross validation (LOO-CV) scheme was used to validate the models and the correlation coefficient ( $R^2$ ), root-mean-square error (RMSE), residual as well as Z-Score were reported.

**Applicability Domain of kNN QSAR Models.** Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, since the training set models are developed in kNN QSAR modeling by interpolating activities of the nearest neighbor compounds, a special applicability domain (i.e., similarity threshold) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules.<sup>82</sup>

In order to measure similarity, each compound could be represented by a point in the  $M$ -dimensional descriptor space (where  $M$  is the total number of descriptors in the descriptor pharmacophore) with the coordinates  $X_{i1}, X_{i2}, \dots, X_{iM}$ , where  $X_{i,s}$  are the values of individual descriptors. The molecular similarity between any two molecules is characterized by the Euclidean distance between their representative points. The Euclidean distance  $d_{ij}$  between two points  $i$  and  $j$  (which correspond to compounds  $i$  and  $j$ ) in  $M$ -dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (8)$$

Compounds with the smallest distance between them have the highest similarity. The distribution of distances (pairwise

similarities) of compounds in our training set is computed to produce an applicability domain threshold,  $D_T$ , calculated as follows:

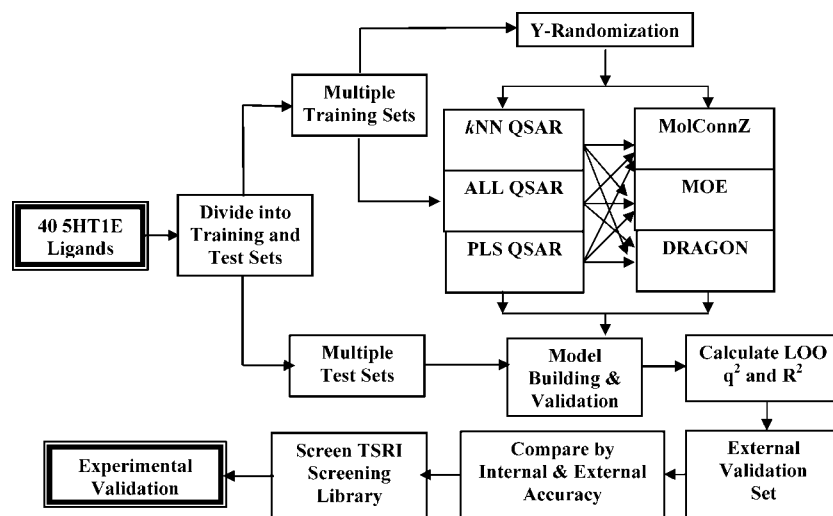
$$D_T = \bar{y} + Z\sigma \quad (9)$$

Here,  $\bar{y}$  is the average Euclidean distance of the  $k$  nearest neighbors of each compound within the training set,  $\sigma$  is the standard deviation of these Euclidean distances, and  $Z$  is an arbitrary parameter to control the significance level. Based on previous studies, we set the default value of this parameter as 0.5, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation (assuming a Boltzmann distribution of distances between each compound and its  $k$  nearest neighbors in the training set). Thus, if the distance of the external compound from at least one of its nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable.

**Y-Randomization Test.** This is a widely used validation technique to ensure the robustness of a QSAR model.<sup>83</sup> In this test, the dependent-variable vector, Y-vector, is randomly shuffled, and new QSAR models are developed using the original independent-variable matrix. This process is repeated several (typically, 10) times. It is expected that the resulting QSAR models should generally have low LOO  $q^2$  and test set  $R^2$  values. It is likely that sometimes, though infrequently, high  $q^2$  values may be obtained due to a chance correlation or structural redundancy of the training set. If all QSAR models obtained in the Y-randomization test have relatively high  $R^2$  and LOO  $q^2$ , then it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modeling method. The Y-randomization test was applied to all data sets considered in this study.

**QSAR-Based Virtual Screening.** QSAR models were used for virtual screening of the TSRI library of 64925 compounds in order to calculate the rate of recovering the experimental hits from the primary screening library (cf. Figure 1). Specifically, the active compounds of the confirmatory assays AID726 (Dose Response Cell Based Assay for Agonists of the 5HT1E) and AID749 (Dose Response Cell Based Assay for Antagonists of the 5HT1E) were used as the seed compounds. The virtual screening was conducted using 117 validated kNN models generated for the 5HT1E ligand data set (see Results), and hits were identified by consensus agreement between these models. The predictions were categorized into three classes by model coverage (MC), i.e., 90%, 70%, and 50%. The model coverage is defined as the fraction of all validated QSAR models that could make the prediction for external compounds found within a certain  $Z$  cutoff. The parameter  $Z$  to control the significance level in the definition of applicability domain for kNN models (cf. 12) was set to 0.2, which placed the similarity threshold for compounds in the external set at one-fifth of the standard deviation. As the alternative, the similarity search was also carried out in the consensus descriptor space using each compound in the modeling data set as the probe. The consensus descriptors were all the variables used in 117 validated kNN models.

In order to access the ability of QSAR models to recover the active compounds from the TSRI screening library, three criteria, i.e., hit rate, yield, and the enrichment factor (EF), were used. They were calculated using the following formulas



**Figure 1.** The workflow for QSAR model building, validation and virtual screening illustrated for 5HT1E ligands.

**Table 1.** Ten Best *k*NN QSAR Models for 5HT1E Ligands Using MolConnZ Descriptors

model ID	training set size	test set size	$k^a$	$q^2$	$R^2$	$R_0^2$	RMSE
1	33	7	1	0.69	0.92	0.90	0.22
2	33	7	1	0.64	0.92	0.84	0.28
3	33	7	1	0.64	0.82	0.82	0.26
4	33	7	2	0.63	0.81	0.76	0.36
5	33	7	1	0.65	0.80	0.75	0.30
6	33	7	2	0.60	0.78	0.74	0.38
7	31	9	1	0.61	0.75	0.72	0.33
8	33	7	2	0.62	0.75	0.74	0.33
9	33	7	2	0.60	0.75	0.75	0.30
10	31	9	1	0.67	0.74	0.73	0.42

<sup>a</sup> Number of nearest neighbors in the optimized *k*NN model (cf. eq 1).

$$\text{HitRate} = H_{\text{scr}}/D_{\text{scr}} \times 100 \quad (10)$$

$$\text{Yield} = H_{\text{scr}}/H_{\text{tot}} \times 100 \quad (11)$$

$$\text{EF} = H_{\text{scr}}/H_{\text{tot}} \times D_{\text{tot}}/D_{\text{scr}} \quad (12)$$

where  $H_{\text{scr}}$  is the number of target-specific actives recovered at a specific % level of the screening library;  $H_{\text{tot}}$  is the total number of actives for this target;  $D_{\text{scr}}$  is the number of compounds screened at a specific % level of the screening library; and  $D_{\text{tot}}$  is the total number of compounds of the screening library.

## RESULTS AND DISCUSSION

**Combinatorial QSAR Modeling of Binding Affinity for the 5HT1E/5HT1F Data Sets.** *k*NN QSAR Regression Modeling. Each original data set was divided into multiple training and test sets as described in the Methods section. The most active and most inactive compounds were always included in the training set by default. The training and test set sizes varied from 35 and 5, respectively, to 21 and 19 compounds, respectively, for the 5HT1E ligands data set and from 24 and 5, respectively, to 16 and 13 compounds, respectively, for the 5HT1F ligands data set. Generally, *k*NN models with leave-one-out cross-validated  $R^2$  ( $q^2$ ) values for the training set greater than 0.50 and linear fit predictive  $R^2$  values for the external test set greater than 0.60 were accepted.<sup>84</sup> Multiple *k*NN regression models were obtained

**Table 2.** Ten Best ALL QSAR Models for 5HT1E Ligands Using DRAGON Descriptors

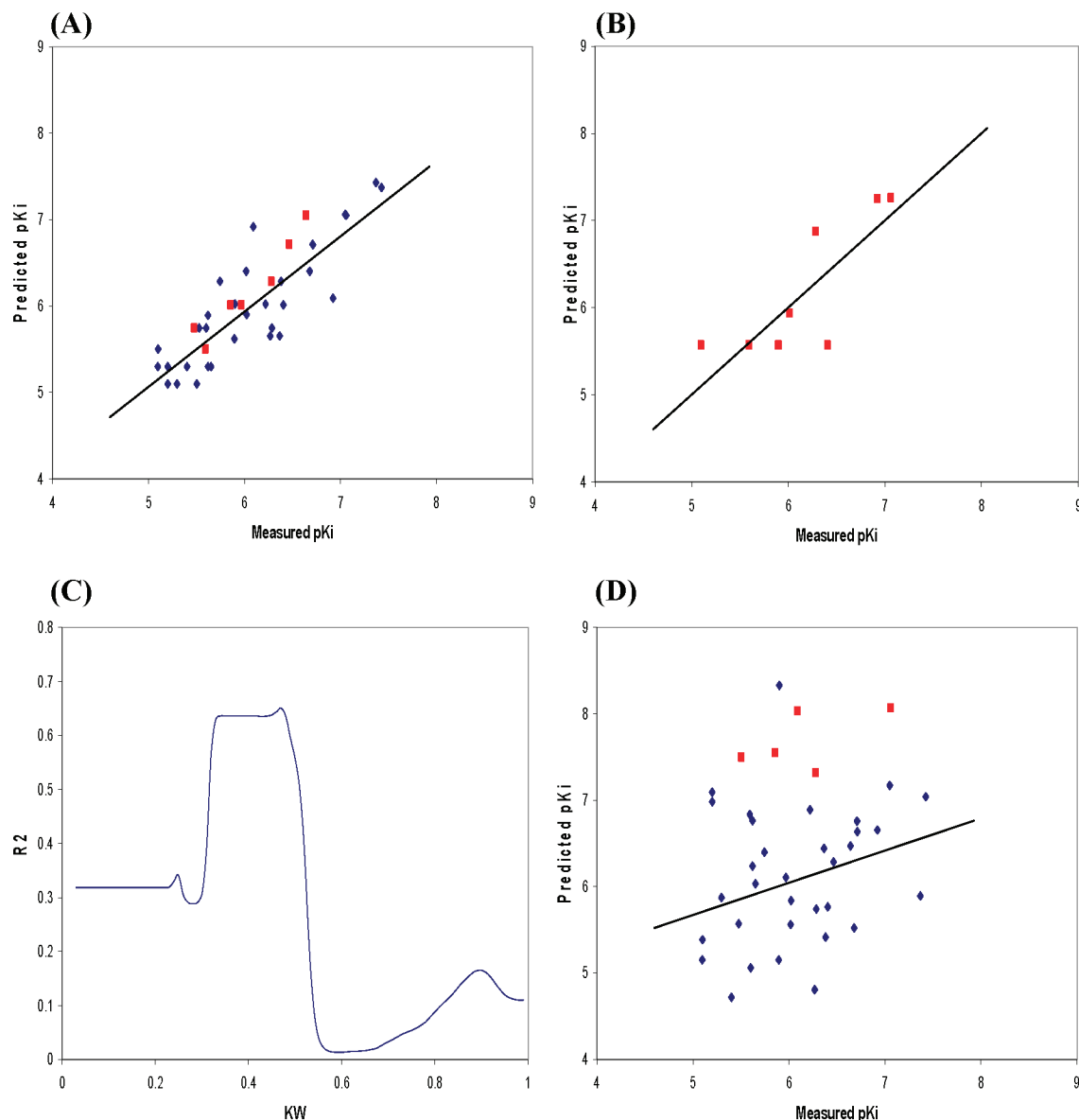
model ID	training set size	test set size	$K^a$	$R^2$	$R_0^2$	RMSE
1	35	5	0.38	0.94	0.94	0.18
2	35	5	0.37	0.94	0.94	0.18
3	35	5	0.39	0.94	0.94	0.18
4	35	5	0.36	0.94	0.94	0.18
5	35	5	0.38	0.84	0.84	0.21
6	35	5	0.37	0.84	0.84	0.21
7	35	5	0.40	0.78	0.78	0.20
8	35	5	0.39	0.74	0.74	0.22
9	32	8	0.40	0.65	0.65	0.45
10	31	9	0.47	0.65	0.65	0.43

<sup>a</sup>  $K$  is the kernel width, also known as the bandwidth (cf. eq 7).

for both 5HT1E and 5HT1F ligand data sets using MolConnZ, MOE, and DRAGON descriptors. As shown in Figure 2 and Table 1, the *k*NN QSAR method produced the best statistical models with  $q^2/R^2$  values of 0.69/0.92 for the 5HT1E ligands data set using MolConnZ descriptors. For MOE and DRAGON descriptors, the best  $q^2/R^2$  values were as high as 0.68/0.88 and 0.61/0.80, respectively (cf. Table 3). The best models for the 5HT1F ligands data set were obtained with the *k*NN method and DRAGON 5.4 descriptors ( $q^2 = 0.64$ ,  $R^2 = 0.89$ ; cf. Figure 3A). The other two approaches rendered the  $q^2/R^2$  values of 0.84/0.64 for MolConnZ descriptors and 0.61/0.68 for MOE descriptors. These results suggest that the intrinsic structure-binding affinity relationships exist for both 5HT1E and 5HT1F ligands that can be best described by *k*NN models using both independent descriptor sets.

Results of the Y-randomization test (data not shown) have confirmed that *k*NN regression models with  $q^2/R^2$  values  $\geq 0.50/0.60$  were robust. The 5HT1E/5HT1F ligand binding affinities (pKi) were randomly shuffled, and *k*NN QSAR models were generated. However, none of the models with randomized affinities of the training set compounds had  $q^2/R^2$  values  $\geq 0.50/0.60$  for either data set. These results confirmed that *k*NN models uncovered nonspurious correlations between both MOE and MolConnZ descriptors and compound binding affinity.

**Models Generated with ALL QSAR Method.** Unlike the *k*NN method, the ALL QSAR approach yielded models with



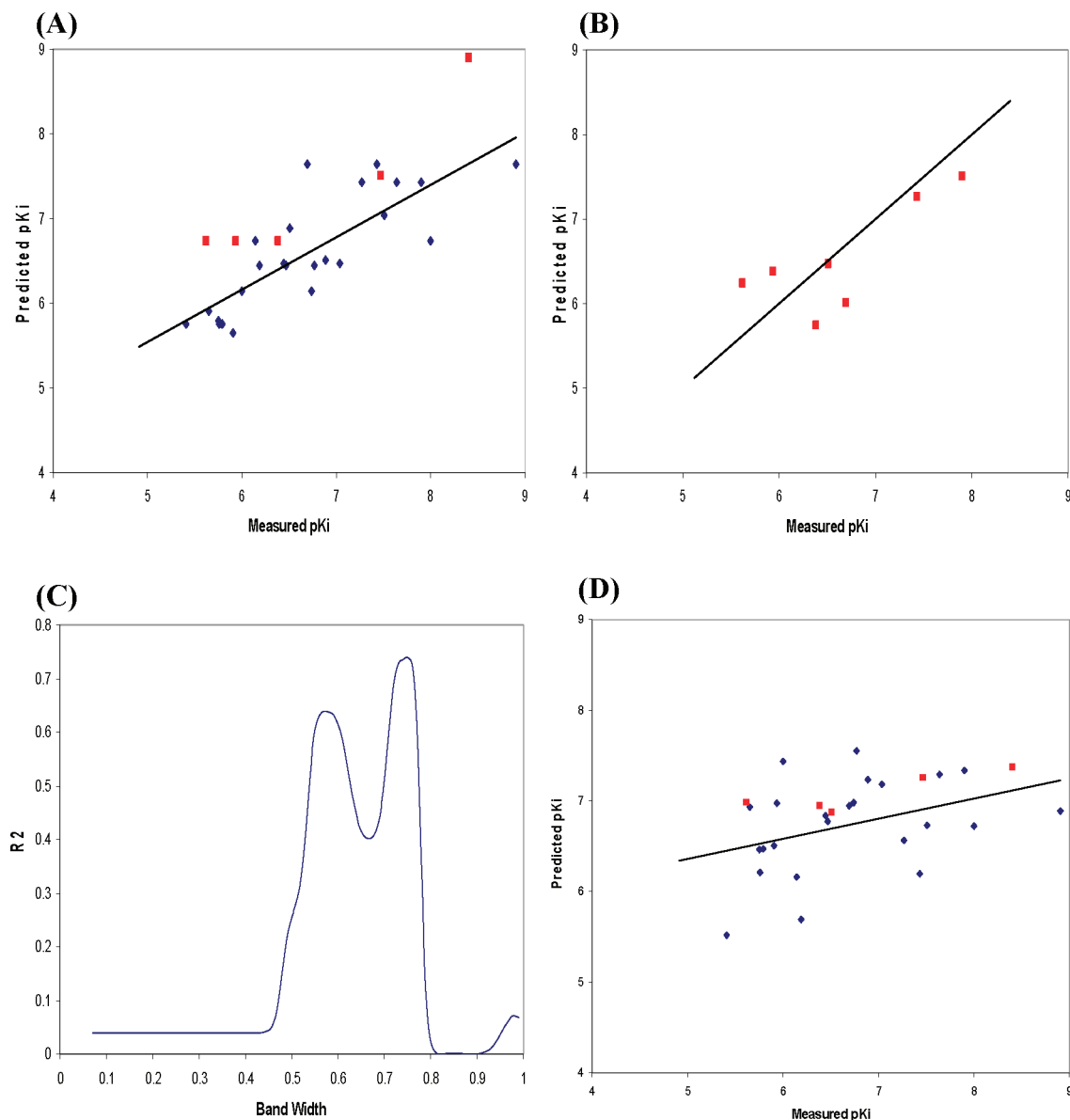
**Figure 2.** Comparison of actual vs predicted ligand binding affinity (pKi) values for the best QSAR models in each category of combi-QSAR (cf. Figure 1) generated for the 5HT1E ligands data set. **A.** Model generated using *k*NN with MolConnZ 4.05 descriptors ( $q^2 = 0.69$ ,  $R^2 = 0.92$ ). The training set contains 33 compounds (blue rhombs) and the test set contains 7 compounds (red squares). Note that the regression line is plotted using training set data only. **B.** Model generated with the ALL-QSAR and DRAGON 5.4 descriptors ( $R^2 = 0.65$ ). The test set contains 9 compounds (red squares). **C.** The changes of  $R^2$  profile vs bandwidth  $K$  of the kernel function during ALL-QSAR model optimization using DRAGON 5.4 descriptors ( $K = 0.47$ ). **D.** Model generated with the PLS method and MOE 2006 descriptors ( $q^2 = 0.09$ ,  $R^2 = 0.31$ ). The training set contains 35 compounds (blue rhombs) and the test set contains 5 compounds (red squares). Note that the regression line is plotted using training set data only.

the highest predictive power for the 5HT1E ligand data set when combined with the DRAGON descriptor ( $R^2 = 0.94$ , cf. Table 3). Table 2 summarizes the prediction power of the ten best ALL QSAR models. For illustration, the plot of actual vs predicted ligand affinity for the 5HT1E receptor is shown in Figure 2B ( $R^2 = 0.65$ ), and the trajectories of  $R^2$  optimization vs the bandwidth of kernel function during ALL QSAR model building are shown in Figure 2C. For the 5HT1F ligand data set, the best predictive model for ALL QSAR was obtained with MolConnZ descriptors ( $R^2 = 0.92$ , cf. Table 4 and Figure 3B,C). As can be observed in Figures 2C and 3, there were always several peaks of local maxima in each optimization trajectory. It should be noted that the prediction power of the models converged at the bandwidth of 0.47 (Figure 2C) or 0.75 (Figure 3C) with no further

improvement observed in other regions. These results demonstrate the dependence of the optimal bandwidth on the data set, as implied by 5. This kernel function is a distance-based Gaussian function that depends on data distribution and decays smoothly as the distance  $d$  increases. The bandwidth is therefore a very important smoothing parameter that needs to be optimized during the process of model building.

**PLS QSAR.** As shown in Tables 3 and 4, the PLS QSAR failed to render predictive models for both the 5HT1E and 5HT1F ligand data sets. The only exception is the PLS/DRAGON descriptor combination for 5HT1F, in which the  $R^2$  value is as high as 0.76 but the LOO-CV  $q^2$  is only 0.15. The poor quality of the PLS models can be further illustrated





**Figure 3.** Comparison of actual vs predicted ligand binding affinity (pKi) values for the best QSAR models in each category of combi-QSAR (cf. Figure 1) generated for the 5HT1F ligands data set. **A.** Model generated using *k*NN with DRAGON 5.4 descriptors ( $q^2 = 0.64$ ,  $R^2 = 0.89$ ). The training set contains 24 compounds (blue rhombs) and the test set contains 5 compounds (red squares). Note that the regression line is plotted using training set data only. **B.** Model generated using the ALL-QSAR method and MolConnZ 4.05 descriptors ( $R^2 = 0.74$ ). The test set contains 8 compounds (red squares). **C.** The changes of  $R^2$  profile vs bandwidth  $K$  of the kernel function during ALL-QSAR model optimization using MolConnZ 4.05 descriptors ( $K = 0.75$ ). **D.** Model generated with the PLS method and MOE 2006 descriptors ( $q^2 = 0.15$ ,  $R^2 = 0.76$ ). The training set contains 24 compounds (blue rhombs), and the test set contains 5 compounds (red squares). Note that the regression line is plotted using training set data only.

**Table 3.** Summary of Combinatorial QSAR Analysis of 5HT1E Ligands

type of descriptor	<i>k</i> NN			ALL			PLS		
	$q^{2a}$	$R^2$	RMSE <sup>b</sup>	$q^{2a}$	$R^2$	RMSE <sup>b</sup>	$q^{2a}$	$R^2$	RMSE <sup>b</sup>
MZ4.05	0.69	0.92	0.22	N/A	0.52	0.37	0.07	0.16	0.85
MOE2006.08	0.68	0.88	0.40	N/A	0.87	0.24	0.09	0.31	1.59
DRAGON5.4	0.61	0.80	0.34	N/A	0.94	0.18	0.02	0.14	0.78

<sup>a</sup>  $q^2$  is calculated by leave-one-out (LOO) cross-validated (CV) procedures for the training set. Only the best model is reported for each combination. <sup>b</sup> RMSE is the root mean square error.

by Figures 2D and 3D, that show practically no correlation between predicted and measured values.

**Comparison of QSAR Approaches for 5HT1E/5HT1F Binding Affinity Prediction.** The performance of different approaches employed as part of the combinatorial QSAR strategy for the 5HT1E and 5HT1F ligand data sets is

summarized in Tables 3 and 4. In the current study, three QSAR methods were employed, including *k*NN regression QSAR, ALL QSAR, and PLS. Each method was combined with each of the following three descriptor sets: MolConnZ, MOE, and DRAGON. Thus, nine different combinations of the QSAR methods and descriptor types were examined.

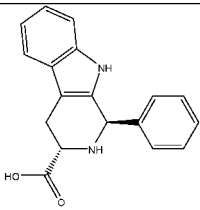
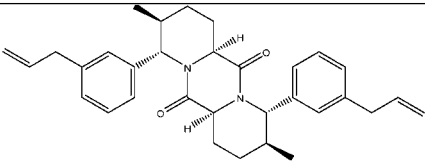
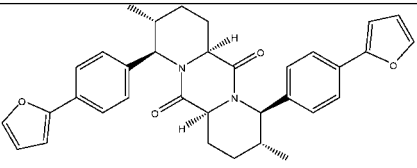


**Table 4.** Summary of Combinatorial QSAR Analysis of 5HT1F Ligands

type of descriptor	kNN			ALL			PLS		
	$q^{2a}$	$R^2$	RMSE <sup>b</sup>	$q^{2a}$	$R^2$	RMSE <sup>b</sup>	$q^{2a}$	$R^2$	RMSE <sup>b</sup>
MZ4.05	0.84	0.64	0.90	N/A	0.92	0.16	0.06	0.00	1.68
MOE2006.08	0.61	0.68	0.55	N/A	0.83	0.32	0.01	0.42	1.58
DRAGON5.4	0.64	0.89	0.68	N/A	0.62	0.82	0.15	0.76	0.83

<sup>a</sup>  $q^2$  is calculated by leave-one-out (LOO) cross-validated (CV) procedures for the training set. Only the best model is reported for each combination. <sup>b</sup> RMSE is the root mean square error.

**Table 5.** Results of Consensus Predictions of Binding Affinity (pKi) and Subtype Selectivity for Compounds Used for External Validation for the 5HT1E Ligand Data Sets

Ki ID		45772	45760	45769
				
Exp.	K <sub>i</sub> (nM)	272.50	5489.00	5680.00
	pKi	6.56	5.26	5.25
kNN	pKi <sup>d</sup>	6.25	6.12	5.73
	Standard Dev.	1.10	0.57	0.42
/MZ	Coefficient of Variation <sup>b</sup>	17.60%	9.31%	7.33%
	Classification <sup>c</sup>	0	1	1
		Selectivity	selective ligands	non-selective ligands

<sup>a</sup> The average results of consensus prediction by best kNN QSAR regression models using MolConnZ descriptors for 5HT1E ligand data set. <sup>b</sup> The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean. It allows comparison of the variation of populations that have significantly different mean and standard deviation values. <sup>c</sup> The average results of consensus prediction by 60 best kNN QSAR selectivity models with CCR<sub>train</sub> ≥ 0.75 and CCR<sub>test</sub> ≥ 0.75.

MolConnZ descriptors were found to be the best with the kNN QSAR method for the 5HT1E ligands data set, yielding the highest  $R^2$  value of 0.92. On the contrary, MolConnZ descriptors were the worst among the three descriptor sets for building 5HT1F models. The kNN/DRAGON combination gave the best predictive power for this data set ( $R^2$  = 0.89). DRAGON descriptors also afforded significant results with the ALL QSAR method for the 5HT1E ligands data set ( $R^2$  = 0.94). But for the 5HT1F ligand data set, the best combination was found to be ALL QSAR with MolConnZ descriptors ( $R^2$  = 0.92). These results reconfirm the importance of employing the combinatorial QSAR approach to reach the goal of finding the most predictive QSAR method/descriptor combination for each specific data set.

**External Validation and Prediction.** As we progressed with model development for all available data, the new binding data for the 5HT1E receptor were deposited into the PDSP Ki database.<sup>42</sup> Among those, Ki values for compounds with PDSP IDs of 45772, 45760, and 45769 were reported, so we were able to use these compounds for independent external validation of our models. Consensus predictions were carried out using the consensus kNN QSAR regression models of the 5HT1E ligands data set and the kNN QSAR classification models of the 5HT1E/5HT1F selectivity data set (Tables 1 and 6). The predicted affinities of a compound resulting from multiple models were averaged, and the mean value as well as the standard deviation were reported; we refer to this routine for affinity prediction as “consensus

prediction”. As shown in Table 5, the predicted pKi values for compounds 45772, 45760, and 45769 were close to the experimental values with prediction errors ranging from 0.31 to 0.86. Since the 5HT1F receptor was not part of the panel of the 5HT receptor array tested by Dandapani et al.,<sup>42</sup> our predictions for the subtype selectivity for compound 45772 will need to be verified experimentally.

**Virtual Screening To Identify Potential 5HT1E Ligands.** The best combination of the QSAR method and a descriptor set resulting from the exploration of combi-QSAR should be logically employed to screen an external chemical library in silico (Figure 1). If such a library contains experimentally confirmed hits that were not present in the training set, then virtual screening will both test the predictive ability of QSAR models and may help in identifying novel biologically active compounds. After the comparison of combinatorial approaches for the 5HT1E ligand data set and the external validation, the best combinations of the method/descriptor type with the highest predictive power were kNN QSAR/MolConnZ and ALL QSAR/DRAGON (Table 3). The models obtained with the former method were selected for virtual screening. The results of consensus prediction using 117 validated kNN models as applied to 27 5HT1E agonist hits (AID726) and 24 5HT1E antagonist hits (AID749) are given in Tables 4 and 5 of the Supporting Information, respectively. The results for three levels of the model coverage (MC), 90%, 70%, and 50%, are given in both tables. Furthermore, three criteria, i.e., hit rate, yield, and

**Table 6.** Ten Best *k*NN QSAR Classification Models of the 5HT1E/5HT1F Selectivity with the Highest CCR Values for All Test Sets Using MolConnZ Descriptors<sup>a</sup>

model no.	<i>k</i>	CCR <sub>train</sub>	confusion matrix parameters							statistical parameters				
			N(1)	N(0)	TP	TN	FP	FN	SE	SP	EN(1)	EN(0)	CCR <sub>test</sub>	
1	5	0.88	4	1	4	1	0	0	1.00	1.00	2.00	2.00	1.00	
2	1	0.81	6	1	6	1	0	0	1.00	1.00	2.00	2.00	1.00	
3	3	0.80	5	2	5	2	0	0	1.00	1.00	2.00	2.00	1.00	
4	1	0.79	6	2	6	2	0	0	1.00	1.00	2.00	2.00	1.00	
5	1	0.83	4	3	3	3	0	1	0.75	1.00	2.00	1.60	0.88	
6	1	0.85	7	2	5	2	0	2	0.71	1.00	2.00	1.56	0.86	
7	1	0.92	3	3	3	2	1	0	1.00	0.67	1.50	2.00	0.83	
8	1	0.92	3	3	3	2	1	0	1.00	0.67	1.50	2.00	0.83	
9	5	0.88	3	2	2	2	0	1	0.67	1.00	2.00	1.50	0.83	
10	1	1.00	3	5	3	3	2	0	1.00	0.60	1.43	2.00	0.80	

<sup>a</sup> N(1) = number of nonselective ligands, N(0) = number of selective ligands, TP = true positive (nonselective ligands predicted as nonselective ligands), FP = false positives (selective ligands predicted as nonselective ligands), FN = false negatives (nonselective ligands predicted as selective ligands), TN = true negative (selective ligands predicted as selective ligands), SE = sensitivity = TP/N(1), SP = specificity = TN/N(0), EN = the normalized enrichment, EN(1) = (2TP\*N(0))/(TP\*N(0) + FP\*N(1)), EN(0) = (2TN\*N(1))/(TN\*N(1) + FN\*N(0)), and CCR = correct classification rate.

enrichment factor (EF), were calculated (Figure 4). The *k*NN regression QSAR models did recover some active seeds from the TSRI screening library of 64925 compounds. The yield to recover the 5HT1E 27 agonist hits was 52% when the model coverage was greater than 90%. As can be seen from Figure 4A, the yield increased to 82% when the model coverage decreased to 50%. This trend is especially true for EF as its value could be as high as 124.22 (MC > 90%), demonstrating the remarkable ability of *k*NN models to identify active compounds in the screening library. Consistently, the EF value decreases to 23.17 and 10.37 when MC lowers to >70% and >50%, respectively. It should be noted that our *k*NN QSAR models of 5HT1E ligands performed better for the 5HT1E agonist data set (AID726) in comparison to the 5HT1E antagonist data set (AID749). In terms of EF, its value decreased from 124.22 for AID726 data sets to 29.95 for AID749 data sets. It increases to 79.86 if we merge the AID726 and AID749 into a single data set. This observation suggests that most of the putative 5HT1E ligands recovered by the means of virtual screening using our QSAR models are likely to be agonists.

***k*NN Classification Modeling of Subtype Selectivity for 5HT1E/5HT1F Receptors.** *Selectivity Index.* To quantify the subtype selectivity of ligands binding to both 5HT1E and 5HT1F receptors, *k*NN classification models were developed to account for 'nonselective ligands' (category 1) or 'selective ligands' (category 0) as follows. For any compound *j*, we have defined the selectivity index  $\epsilon_j$  as

$$\epsilon_j = \frac{Ki_j^{5HT1E}}{Ki_j^{5HT1F}} \quad (13)$$

The log of  $\epsilon_j$  was used to build QSAR models, where

$$\log \epsilon_j = pKi_j^{5HT1F} - pKi_j^{5HT1E} \quad (14)$$

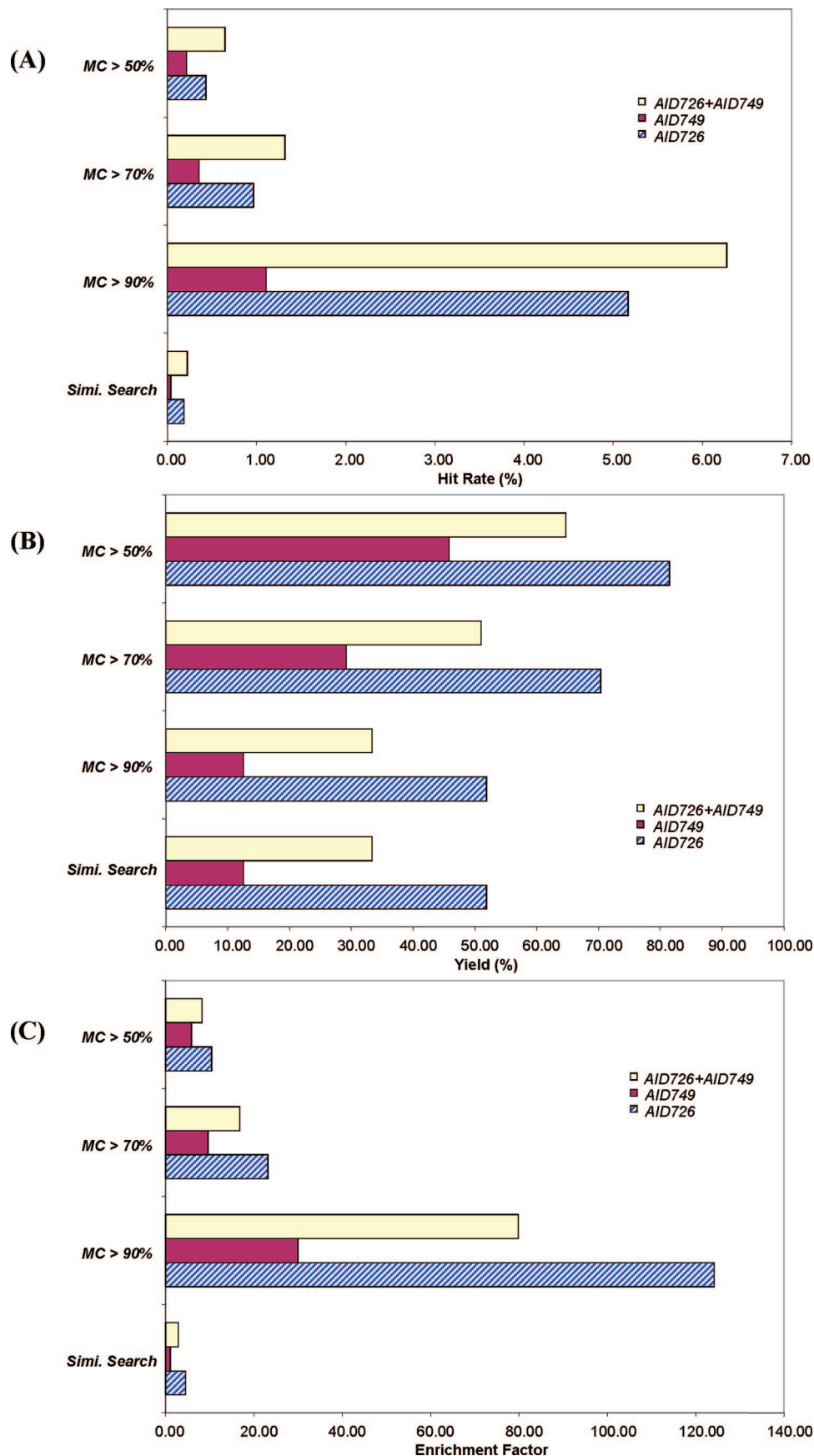
or in terms of IC50 values:

$$\log \epsilon_j = pIC50_j^{5HT1F} - pIC50_j^{5HT1E} \quad (15)$$

The 5HT1E/5HT1F selectivity data were obtained from more than 20 references in the PDSP Ki database. Inevitably, the broad source of data for different ligands is likely to be responsible for the uncertainty of Ki values used to build

the selectivity models. The errors could be due to many reasons such as the use of different radiolabeled ligands, cell lines, or assay protocols. To circumvent these errors, four thresholds were tested to build the classification models. As shown in Table 6 of the Supporting Information, when the threshold for selectivity index  $\epsilon_j$  is equal to either 0.33 or 3 [i.e., compounds are defined as selective (class "0"), when they have 3-fold or greater selectivity toward either receptor] the classification models had higher accuracy than when using the three other thresholds of  $\epsilon_j = 0.14$  or 7;  $\epsilon_j = 0.2$  or 5; and  $\epsilon_j = 1$ . Based on this investigation, compounds with the ratio of Ki values for 5HT1E vs 5HT1F binding between 1/3 and 3 were categorized as nonselective ligands (class "1"). All other compounds were then considered as selective ligands (class "0"). This definition also yielded a reasonably balanced distribution of ligands between nonselective (13 compounds) and selective (9) ligands and afforded a total of 281 acceptable models. A higher threshold, such as  $\epsilon_j = 1$  or 2 used by others,<sup>85</sup> resulted in too few nonselective ligands in our case. On the other hand, low thresholds ( $\epsilon_j = 0.14$  and 7) led to no acceptable models. Since the errors were experimentally derived and thus depended on individual data sets, the  $\epsilon_j$  should not be set to a single value arbitrarily and need to be determined on the trial basis against the target data set to find the optimal value. A similar approach was used for identifying a threshold for binary QSAR modeling in the analysis of the Maximum Recommended Therapeutic Dose for a large diverse data set of drugs.<sup>86</sup>

**Classification Models Using MolConnZ and MOE Descriptors.** For the 5HT1E/5HT1F subtype selectivity data set, the training and test set sizes varied from 17 and 5 to 12 and 10, respectively. A high fraction of variable selection *k*NN QSAR models had CCR<sub>train</sub> ≥ 0.70, and many models achieved the impressive CCR = 1.00 for the test sets. The best *k*NN classification models shown in Table 6 (built with MolConnZ descriptors) and Table 8 of the Supporting Information (MOE descriptors) include the detailed information on statistical parameters for each model for both training and test sets. Usually models with CCR<sub>test</sub> ≥ 0.70 had CCR<sub>train</sub> ≥ 0.70 as well, but the opposite was not always true. The models with high values of both CCR<sub>train</sub> and



**Figure 4.** The results of virtual screening of the 5HT1E agonists (AID: 726) and the 5HT1E antagonists (AID: 749) data sets available from PubChem using *k*NN QSAR models of 5HT1E ligands. Three thresholds of model coverage (MC), 90%, 70%, and 50%, were used. The similarity search was conducted in the consensus descriptor space using the modeling data set compounds as the probe. **A.** The hit rate of *k*NN models to recover the experimental hits from the primary screening database. **B.** The yield of *k*NN models to recover the experimental hits from the primary screening database. **C.** The enrichment factor to recover the experimental hits from the primary screening database.

**Table 7.** Statistics of the Best Classification Models for Individual Descriptors and QSAR Methods

type of descriptors	nearest neighbors no.	CCR <sub>train</sub>	confusion matrix						statistics for the models				
			N(1)	N(0)	TP	TN	FP	FN	SE	SP	EN(1)	EN(0)	CCR <sub>test</sub>
MolConnZ	5	0.88	4	1	4	1	0	0	1.00	1.00	2.00	2.00	1.00
MOE	3	0.86	4	2	4	2	0	0	1.00	1.00	2.00	2.00	1.00
FSG	1	0.73	5	2	4	2	0	1	0.80	1.00	2.00	1.67	0.90
MHF	1	0.94	4	1	3	1	0	1	0.75	1.00	2.00	1.60	0.88

<sup>a</sup> N(1) = number of nonselective ligands, N(0) = number of selective ligands, TP = true positive (nonselective ligands predicted as nonselective ligands), FP = false positives (selective ligands predicted as nonselective ligands), FN = false negatives (nonselective ligands predicted as selective ligands), TN = true negative (selective ligands predicted as selective ligands), SE = sensitivity = TP/N(1), SP = specificity = TN/N(0), EN - the normalized enrichment, EN(1) = (2TP\*N(0))/(TP\*N(0) + FP\*N(1)), EN(0) = (2TN\*N(1))/(TN\*N(1) + FN\*N(0)), and CCR = correct classification rate.

CCR<sub>test</sub> were considered acceptable. The best predictive models with MolConnZ descriptors were obtained with CCR<sub>train</sub>/CCR<sub>test</sub> = 0.88/1.00 (SE = 1.00, SP = 1.00, EN(1) = 2.00, and EN(0) = 2.00) and CCR<sub>train</sub>/CCR<sub>test</sub> = 0.86/1.00 (SE = 1.00, SP = 1.00, EN(1) = 2.00, and EN(0) = 2.00) with MOE descriptors. The 2 × 2 confusion matrices are shown for the training set (Table 7a of the Supporting Information) and the validation test set (Table 7b of the Supporting Information) of model No. 2 with MolConnZ descriptors. Remarkably, the *k*NN classification models predicted correctly all 7 nonselective ligands out of 15 compounds in the training set and all 6 nonselective ligands out of 7 compounds in the validation set.

**FSG Descriptors.** As many as 1674 frequent patterns were found for the 5HT1E/5HT1F subtype selectivity data set. After removing 754 redundant subgraphs, the number of closed subgraphs retained was 920. The size of the subgraphs is set by default, ranging from 2 nodes (i.e., atoms) to the upper limit that was dependent on the data set. The support value,  $\sigma$ , was set to 5 for the following reasons. In general, the larger value of  $\sigma$  leads to the smaller number of subgraphs identified, and the accuracies of models decrease. Certain subtle patterns that are critical to differentiate similar structures could be missed. On the other hand, as  $\sigma$  value decreases, the number of subgraphs increases exponentially. Though largely dependent on the data set, in the previous studies the best results were obtained with a  $\sigma$  value of 5–15% of the whole size for three data sets.<sup>73</sup> Where expressed by ratio or percentage, the  $\sigma$  indicates the nonselective subgraphs in at least a subset of molecules of a certain size out of the total number of molecules. However, in our case when  $\sigma$  was set to 3, a total of 99,300 closed subgraphs were generated making it impossible to use variable selection *k*NN QSAR. Thus, we set  $\sigma$  at a higher value of 5, which led to a larger number of acceptable *k*NN classification models with both CCR<sub>train</sub> ≥ 0.70 and CCR<sub>test</sub> ≥ 0.70. As shown in Table 9 of the Supporting Information, the SG descriptors afforded models comparable in their statistical characteristics to those generated with MolConnZ and MOE descriptors by *k*NN classification. The prediction accuracy was as high as CCR<sub>test</sub> = 0.90 for the test set of seven compounds, with CCR<sub>train</sub> = 0.73, SE = 0.80, SP = 1.00, EN(1) = 2.00, and EN(0) = 1.67.

**Models Using Molecular Hologram Fingerprints.** For the large combinations of parameters during MHFs generation, the fragment at the length of 353 with the distinction of ABCH produced the best results. In total, the 5HT1E/5HT1F subtype selectivity data set generated 4695 fragments. Used in *k*NN classification QSAR, these MHFs afforded many

models with both CCR<sub>train</sub> and CCR<sub>test</sub> ≥ 0.70. The ten best models are listed in Table 10 of the Supporting Information. The best *k*NN classification model was obtained with CCR<sub>train</sub> = 0.94, CCR<sub>test</sub> = 0.88, SE = 0.75, SP = 1.00, EN(1) = 2.00, and EN(0) = 1.60. It should be noted that when 4695 fragment strings were hashed into the bin arrays of fixed length of 353, each bin contained ca. 12–13 fragments. These bin arrays are still usable for model building after hashing because the majority of fragments within each bin are redundant. But in the case of a large number of fragments generated because of the chemical diversity of the data set, it is possible that certain fragments important for biological activity were assigned to the same bin, which is known as fragment collision.<sup>75,87</sup> To minimize the chance of bad fragment collision, Seel et al. recommend that unhashed fragment string should be used instead during model building.<sup>87</sup> It can be done with PLS QSAR but will be impractical for *k*NN QSAR with variable selection, especially with the large array of 4695 bins.

To ensure that these models are not spurious, the selectivity index  $\epsilon$  was randomly shuffled within the training set. It is expected that models obtained for the training set with randomized activities should have significantly lower values of CCR for the test set than the models built using the training set with real activities. Indeed, models produced with the randomized activity arrays had both CCR<sub>train</sub> and CCR<sub>test</sub> below 0.70 for all four combinations, i.e., MolConnZ descriptors, MOE descriptors, FSG descriptors, and MHFs. These results suggest that all models using actual data represent robust structure–activity correlations.

**Comparison of Combinatorial QSAR Approaches for the 5HT1E/5HT1F Subtype Selectivity.** A comparison between best models generated with five different QSAR approaches explored in this study for 5HT1E/5HT1F subtype selectivity is given in Table 7. Among four descriptor sets used in combination with the *k*NN classification method, MolConnZ and MOE descriptors afforded models with higher predictive power than those using FSG descriptors and MHFs. The accuracy of predictions for *k*NN/MolConnZ and *k*NN/MOE models was as high as CCR<sub>test</sub> = 1.00, whereas *k*NN/FSG and *k*NN/MHFs had somewhat lower accuracy. Thus, descriptors using multiple chemical connectivity indices and bonds and atomic states as well as physical property descriptors afforded statistically better models than fragment based descriptors. On the other hand, models built with FSG descriptors and MHFs offer the advantage of straightforward interpretation in terms of significant structural fragments.



**Table 8.** Major Differences between Frequent Subgraph Descriptors and Molecular Hologram Fingerprints

	FSG descriptors	MHFs
H included	no	optional
fragment size	Dependent on data set	4 to 7 <sup>a</sup>
frequent fragment	yes	no
redundant fragment	no	yes
fragment collision	no	yes
interpretable	easy	difficult

<sup>a</sup> The default range used by most studies.

**FSG Descriptors vs MHFs.** There are several major differences between FSG descriptors and MHFs, as illustrated in Table 8. For all fragment-based fingerprints or descriptors, the common problem is that the total number of fragments that can be possibly generated for a molecular data set is exceedingly large, thus making it difficult to develop robust QSAR models. To address this problem, MHFs hash the fragment strings into the array bins of the fixed length  $L$ , which is generally in the range of 50–500. The occurrences of each fragment that are hashed into the same bin of a particular length in a molecule are summed, and the resulting count constitutes the individual descriptor values of the molecular hologram. Then it is possible that several essential fragments are assigned to the same bin, causing fragment collision. It becomes worse when the total number of fragment is exceedingly bigger than the length of molecular hologram. It may not affect the statistics of QSAR models because of the nature of hash function but will make model interpretation impossible, which is an advantage of models built with fragment descriptors.

Indeed, fragment-based descriptors are particularly suitable for inverse QSAR, i.e., designing new molecules based on QSAR models. Compared to property-based descriptors or molecular indices, the information acquired from frequent fragments employed in model building can be used to reconstruct (or design) an active compound straightforwardly. Given its drawback of fragment collision, MHFs are less suitable for this purpose. In contrast with MHFs, FSG descriptors are directly mapped onto chemical fragments. Furthermore, the FSG algorithm inherently reduces the total number of descriptors for two reasons: 1. only frequent and common fragments are taken into account, and 2. redundant subgraphs are removed. The discriminative frequent substructures selected for optimized QSAR models directly implicate chemical features responsible for compounds' biological activities, and they are especially easy to interpret in cases of binary classification (binders vs nonbinders, or active vs inactive). FSG descriptors also appear more capable of discriminating between relatively similar compounds that belong to different activity classes. For instance, only 3 bins of MHFs (length 87, 145, and 268) out of 353 show different values for the two structurally similar compounds with PDSP IDs 15521 and 15523 in the 5HT1E/5HT1F selectivity data set. On the contrary, among the total of 920 closed subgraphs, as many as 449 have different numbers of occurrence for the same two compounds and consequently are sensitive to even minor structural variance.

**Model Interpretation.** We analyzed descriptors found in validated  $k$ NN QSAR models to elucidate chemical features that may be responsible for 5HT1E/5HT1F selectivity. Tables 11 and 12 of the Supporting Information list the top 25 most

frequent descriptors (MFD) found in 60 validated  $k$ NN/MolConnZ and 181  $k$ NN/MOE models with both  $CCR_{train}$  and  $CCR_{test}$  greater than or equal to 0.75. In addition, the MFD analysis was carried out on 194  $k$ NN/FSG models with the same accuracy levels for both training and test sets, and the most frequent subgraphs were identified (cf. Table 9). It is expected that the descriptors selected by most  $k$ NN models should be critical chemical determinants of respected biological activities. Among all descriptor families, the contribution of shape descriptors was found to be most significant, suggesting a high importance of steric factors in subtype selectivity. Thus, shape descriptors constituted 48% of all top 25 MFD for  $k$ NN/MolConnZ models; similarly, 68% of the top 25 MFD in  $k$ NN/MOE models were also shape-related. Most of the shape-related MOE MFD in Table 12 of the Supporting Information are associated with VDW surface areas, such as PEOE\_VSA-2, VSA\_POL, PEOE\_VSA+6, VSA\_DON, VOL, SMR\_VSA5, ASA\_H, and PEOE\_VSA\_PPOS, and two are logP-related, GCUT\_SLOGP\_0 and logS. This is in accordance with the dominance of shape descriptors found in models built with MolConnZ descriptors, which highlights the critical role of steric factors in subtype selectivity. Consistently, 32% of the top 25 most frequent subgraphs are purely aliphatic. It is interesting to note that the size of the most frequent subgraphs varies to a large extent, from simply 2 nodes (#759 and #916) to 14 nodes (#351) (cf. Table 9). Moreover, a number of the most frequent subgraphs can be mapped onto selective vs nonselective molecules from the structural prospective. Among the top-ranked 25 most frequent subgraphs, #853, #878, #698, #511, #870, #732, #916, #370, #618, #858, #632, #686, #570, and #538 were found to occur predominantly in nonselective ligands, especially in compounds 6757, 6764, 15510, and 15521. These four compounds belong to the same tryptamine family, of which five are nonselective ligands and two are selective ligands.

In fact, there is experimental evidence that steric factors do play a central role in subtype selectivity. Within the 5HT1E/5HT1F selectivity data set, methylergonovine (PDSP ID: 6751) is highly similar to methysergide (PDSP ID: 6752) structurally; these two compounds differ by one additional methyl group found in methysergide. However, this minor change raises the selectivity index  $\epsilon_j$  from 2.88 to 7.29, i.e., methysergide is much more selective toward the 5HT1F receptor. More often than not, the replacement of the primary amine in 5-methoxytryptamine (PDSP ID: 6764) by methyl groups leads to the significant increase of the selectivity (from  $\epsilon_j = 2.70$  to  $\epsilon_j = 14.20$  of DMT,5-Me). Consistently, the descriptors which are related to the methyl group had been found in highly ranked MFDs of both MolConnZ and FSG descriptors. Another functional group that was present frequently in MFDs is the hydroxyl group, suggesting its critical role in the 5HT1E/5HT1F subtype selectivity in addition to the steric factors. In summary, the results of MFDs analysis reinforce our knowledge of the subtype selectivity and may potentially provide important hints useful in the design of 5HT1E/5HT1F selective ligands.

**Table 9.** Twenty-Five Most Frequent Subgraph Fragments in *k*NN Models of the 5HT1E/5HT1F Subtype Selectivity Data Sets

Rank <sup>a</sup>	Des. Index	Freq. <sup>b</sup>	SMILES	Structure
1	759	46	CO	
2	859	44	C(CCC)C(C)C	
3	137	42	C(C)(C(CCCNC)C)NC	
4	853	38	C(CC)C(CC)CC	
5	18	38	C(C)CC	
6	878	38	C(C)CCC	
7	436	35	C(CC(CCN)=C)(CCCC)N	
8	698	34	C(C)(CCCCC(CC)-C)O	
9	511	33	C(C(CC)CC(O)CC)N	
10	870	33	C(CC)CCC	
11	732	33	C(CC(CC)C)O	
12	173	32	C(C1CCCC)NC=C1CCNCC	
13	916	32	CC	
14	762	31	CNC	
15	351	31	C(CC1)(C(CCCN(C)CC)CC1)N	
16	819	30	C(C)C(C(C)=CC)CC	
17	370	30	C(CCCO)(C(C(CC)-C)C)N	
18	618	29	C(=C(C)CCC(O)CCC)N	
19	858	29	C(CCC)C(CC)C	
20	632	29	C(=CC(C)CCCCO)N	
21	686	29	C(C)(CC(CCC)CC)O	
22	817	29	C(CC)C(C(C)=CC)CC	
23	570	29	C(=C(CCN)C(CC)CC(O)C)N	
24	538	28	C(CCC(O)CCC)N	
25	557	28	C(CC(CCN(C)CC)-C)N	

<sup>a</sup> *k*NN rank is based on the frequency of each descriptor occurred. <sup>b</sup> Frequency is the number of times each descriptor occurred in 194 models with CCR<sub>train</sub> ≥ 0.70 and CCR<sub>test</sub> ≥ 0.75.

## CONCLUSIONS

In recent years, we have advanced the combinatorial QSAR approach that explores various combinations of optimization methods and descriptor types for the analysis of experimental SAR data sets.<sup>30,31</sup> We reasoned that combi-QSAR strategy is more likely to identify highly predictive QSAR models for a particular data set than any conventional approach using only a single method and a single type of descriptors. The best QSAR models resulting from combi-QSAR can be further applied for the purpose of virtual screening. Herein, we have applied the combi-QSAR approach to three data sets taken from the PDSP Ki Database. The dual receptors system 5HT1E/5HT1F was chosen from this database not only because both types of receptors are novel targets that received significant attention lately but also because they serve as an ideal system to investigate the structural determinants of subtype selectivity.

The model building was carried out using variable selection *k*NN and ALL QSAR methods as well as PLS, combined with a variety of descriptors including MolConnZ, MOE, DRAGON, FSG, and MHFs. The best *k*NN regression models were obtained with MolConnZ and DRAGON descriptors ( $q^2$  for the training set and predictive  $R^2$  values for the test set ( $q^2/R^2$ ) of 0.69/0.92 for the 5HT1E ligands data set and 0.64/0.89 for the 5HT1F ligands data set, respectively). Using the ALL QSAR method, the best models for the 5HT1E ligands data set were obtained with DRAGON descriptors ( $R^2 = 0.94$ ; bandwidth = 0.38); the value of  $R^2$  was 0.74 for the test set including 8 compounds for the 5HT1F ligands data set when MolConnZ descriptors were used (bandwidth = 0.75). The *k*NN classification QSAR was also conducted on the 5HT1E/5HT1F subtype selectivity, and the best binary model showed  $CCR_{\text{train}} = 0.88$  and  $CCR_{\text{test}} = 1.00$  when MolConnZ descriptors were used. Our novel FSG descriptors were also successful in generating multiple validated and predictive *k*NN models. They have shown distinctive advantages in comparison with other fragment-based descriptors or molecular indices, such as MHFs, because FSG descriptors eliminate the fragment redundancy and are easy to interpret in case of binary classification models.

Notably, both regression and classification QSAR models were further validated on three recently reported 5HT1E ligands. The consensus prediction was employed, and the predicted pKi values were found to be in good agreement with the experimental values. These rigorously validated and highly predictive QSAR models have been further exploited for virtual screening<sup>20,88</sup> to recover known 5HT1E ligands from TSRI screening library. The high enrichment factor resulting from virtual screening suggests that QSAR models developed in these studies can be used to identify novel chemical probes for the 5HT1E receptor and the 5HT1F antagonist with superior potency to treat headache disorders as well as high selectivity toward one of these receptors. The approaches employed in this paper can be adopted for other types of GPCRs in the PDSP Ki Database and extended from dual receptors system to multiple receptor systems via multiclass classification QSAR modeling.

## ACKNOWLEDGMENT

We would like to thank Dr. Bryan Roth for his assistance in using the PDSP Ki database and constructive criticism

throughout this research project. We are also grateful to Drs. Weifan Zheng and Raed Khashan for the development of FSG descriptors and helpful discussions. We thank Tripos, Chemical Computing Group, and eduSoft for software grants. Finally, we acknowledge the access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The studies reported in this paper were supported in part by the NIH research grant GM066940 and the planning grant HG003898.

**Supporting Information Available:** 5HT1E, 5HT1F, and 5HT1E/5HT1F data set (Tables 1–3, respectively), consensus prediction of 5HT1E agonist and antagonist for AID 726 and 749 (Tables 4 and 5, respectively), performance of *k*NN QSAR classification models (Table 6),  $2 \times 2$  confusion matrix of the training set and the validation test set (Table 7), ten best *k*NN QSAR classification models for all test sets using MOE descriptors, frequent subgraph descriptors, and MHFs (Tables 8–10), 25 most frequent MolconnZ 4.05 and MOE descriptors (Tables 11 and 12). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) Flower, D. R. Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* **1999**, *1422*, 207–234.
- (2) Shay, J. W.; Wright, W. E. Telomerase therapeutics for cancer: challenges and new directions. *Nat. Rev. Drug Discovery* **2006**, *5*, 577–584.
- (3) Roth, B. L.; Lopez, E.; Patel, E. S.; Kroeze, W. K. The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrassment of riches. *Neuroscientist* **2006**, *6*, 252–262.
- (4) Kozikowski, A. P.; Roth, B.; Tropsha, A. Why academic drug discovery makes sense. *Science* **2006**, *313*, 1235–1236.
- (5) Okuno, Y.; Yang, J.; Taneishi, K.; Yabuuchi, H.; Tsujimoto, G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.* **2006**, *34*, D673–D677.
- (6) O'Connor, K. A.; Roth, B. L. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 1005–1014.
- (7) Armbruster, B. N.; Roth, B. L. Mining the receptorome. *J. Biol. Chem.* **2005**, *280*, 5129–5132.
- (8) Roth, B. L. Receptor systems: will mining the receptorome yield novel targets for pharmacotherapy. *Pharmacol. Ther.* **2005**, *108*, 59–64.
- (9) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
- (10) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (11) Visiers, I.; Ballesteros, J. A.; Weinstein, H. Three-dimensional representations of G protein-coupled receptor structures and mechanisms. *Methods Enzymol.* **2002**, *343*, 329–371.
- (12) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–1265.
- (13) Rasmussen, S. G.; Choi, H. J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F.; Weis, W. I.; Kobilka, B. K. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383–387.
- (14) Trumpp-Kallmeyer, S.; Hoflack, J.; Bruinvels, A.; Hibert, M. Modeling of G-protein-coupled receptors: application to dopamine, adrenaline, serotonin, acetylcholine, and mammalian opsin receptors. *J. Med. Chem.* **1992**, *35*, 3448–3462.
- (15) Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets. *Proteins* **2003**, *50*, 5–25.
- (16) Tropsha, A.; Zheng, W. F. Identification of the descriptor pharmacophores using variable selection QSAR: Applications to database mining. *Curr. Pharm. Des.* **2001**, *7*, 599–612.



- (17) Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
- (18) Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. *Curr. Pharm. Des.* **2007**, *13*, 1497–1508.
- (19) Tropsha, A.; Wang, S. X. QSAR Modeling of GPCR Ligands: Methodologies and Examples of Applications. In *GPCRs: From Deorphanization to Lead Structure Identification*; Springer-Verlag: Leipzig, 2007; pp 49–73.
- (20) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D-1 dopaminergic antagonists for database mining. *J. Med. Chem.* **2005**, *48*, 7322–7332.
- (21) Ghoneim, O. M.; Legere, J. A.; Golbraikh, A.; Tropsha, A.; Booth, R. G. Novel ligands for the human histamine H-1 receptor: Synthesis, pharmacology, and comparative molecular field analysis studies of 2-dimethylamino-5-(6-phenyl-1,2,3,4-tetrahydronaphthalenes). *Bioorg. Med. Chem.* **2006**, *14*, 6640–6658.
- (22) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217–3226.
- (23) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–359.
- (24) Harrison, P. J.; Owen, M. J. Genes for schizophrenia? Recent findings and their pathophysiological implications. *Lancet* **2003**, *361*, 417–419.
- (25) Lewis, C. M.; Levinson, D. F.; Wise, L. H.; DeLisi, L. E.; Straub, R. E.; Hovatta, I.; Williams, N. M.; Schwab, S. G.; Pulver, A. E.; Faraone, S. V.; Brzustowicz, L. M.; Kaufmann, C. A.; Garver, D. L.; Gurling, H. M.; Lindholm, E.; Coon, H.; Moises, H. W.; Byerley, W.; Shaw, S. H.; Mesen, A.; Sherrington, R.; O'Neill, F. A.; Walsh, D.; Kendler, K. S.; Ekelund, J.; Paus, T.; Lonnqvist, J.; Peltonen, L.; O'Donovan, M. C.; Owen, M. J.; Wildenauer, D. B.; Maier, W.; Nestadt, G.; Blouin, J. L.; Antonarakis, S. E.; Mowry, B. J.; Silverman, J. M.; Crowe, R. R.; Cloninger, C. R.; Tsuang, M. T.; Malaspina, D.; Harkavy-Friedman, J. M.; Svrakic, D. M.; Bassett, A. S.; Holcomb, J.; Kalsi, G.; McQuillin, A.; Brynjolfsson, J.; Sigmundsson, T.; Petursson, H.; Jazin, E.; Zoega, T.; Helgason, T. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am. J. Hum. Genet.* **2003**, *73*, 34–48.
- (26) Amlaiky, N.; Ramboz, S.; Boschert, U.; Plassat, J. L.; Hen, R. Isolation of a mouse "5HT1E-like" serotonin receptor expressed predominantly in hippocampus. *J. Biol. Chem.* **1992**, *267*, 19761–19764.
- (27) McAllister, G.; Charlesworth, A.; Snodin, C.; Beer, M. S.; Noble, A. J.; Middlemiss, D. N.; Iversen, L. L.; Whiting, P. Molecular cloning of a serotonin receptor from human brain (5HT1E): a fifth 5HT1-like subtype. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5517–5521.
- (28) Goadsby, P. J. New targets in the acute treatment of headache. *Curr. Opin. Neurol.* **2005**, *18*, 283–288.
- (29) Shephard, S.; Edvinsson, L.; Cumberbatch, M.; Williamson, D.; Mason, G.; Webb, J.; Boyce, S.; Hill, R.; Hargreaves, R. Possible antimigraine mechanisms of action of the 5HT1F receptor agonist LY334370. *Cephalalgia* **1999**, *19*, 851–858.
- (30) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W. F.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582–595.
- (31) Lima, P. D. C.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Tropsha, A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245–1254.
- (32) Bymaster, F. P.; Dreshfield-Ahmad, L. J.; Threlkeld, P. G.; Shaw, J. L.; Thompson, L.; Nelson, D. L.; Hemrick-Luecke, S. K.; Wong, D. T. Comparative affinity of duloxetine and venlafaxine for serotonin and norepinephrine transporters in vitro and in vivo, human serotonin receptor subtypes, and other neuronal receptors. *Neuropsychopharmacology* **2001**, *25*, 871–880.
- (33) Glennon, R. A.; Lee, M.; Rangisetty, J. B.; Dukat, M.; Roth, B. L.; Savage, J. E.; McBride, A.; Rauser, L.; Hufeisen, S.; Lee, D. K. 2-Substituted tryptamines: agents with selectivity for 5-HT(6) serotonin receptors. *J. Med. Chem.* **2000**, *43*, 1011–1018.
- (34) Leonhardt, S.; Herrick-Davis, K.; Titeler, M. Detection of a novel serotonin receptor subtype (5-HT1E) in human brain: interaction with a GTP-binding protein. *J. Neurochem.* **1989**, *53*, 465–471.
- (35) Lovell, P. J.; Bromidge, S. M.; Dabbs, S.; Duckworth, D. M.; Forbes, I. T.; Jennings, A. J.; King, F. D.; Middlemiss, D. N.; Rahman, S. K.; Saunders, D. V.; Collin, L. L.; Hagan, J. J.; Riley, G. J.; Thomas, D. R. A novel, potent, and selective 5-HT(7) antagonist: (R)-3-(2-(4-methylpiperidin-1-yl)ethyl)pyrrolidine-1-sulfonyl phenol (SB-269970). *J. Med. Chem.* **2000**, *43*, 342–345.
- (36) Phebus, L. A.; Johnson, K. W.; Zgombick, J. M.; Gilbert, P. J.; Van Belle, K.; Mancuso, V.; Nelson, D. L.; Calligaro, D. O.; Kiefer, A. D., Jr.; Branchek, T. A.; Flaugh, M. E. Characterization of LY5344864 as a pharmacological tool to study 5-HT1F receptors: binding affinities, brain penetration and activity in the neurogenic dural inflammation model of migraine. *Life Sci.* **1997**, *61*, 2117–2126.
- (37) Price, G. W.; Burton, M. J.; Collin, L. J.; Duckworth, M.; Gaster, L.; Gothert, M.; Jones, B. J.; Roberts, C.; Watson, J. M.; Middlemiss, D. N. SB-216641 and BRL-15572—compounds to pharmacologically discriminate h5-HT1B and h5-HT1D receptors. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **1997**, *356*, 312–320.
- (38) Schotte, A.; Janssen, P. F.; Gommeren, W.; Luyten, W. H.; Van Gompel, P.; Lesage, A. S.; De Loore, K.; Leysen, J. E. Risperidone compared with new and reference antipsychotic drugs: in vitro and in vivo receptor binding. *Psychopharmacology (Berlin)* **1996**, *124*, 57–73.
- (39) Zgombick, J. M.; Schechter, L. E.; Macchi, M.; Hartig, P. R.; Branchek, T. A.; Weinshank, R. L. Human gene S31 encodes the pharmacologically defined serotonin 5-hydroxytryptamine1E receptor. *Mol. Pharmacol.* **1992**, *42*, 180–185.
- (40) Boess, F. G.; Martin, I. L. Molecular biology of 5-HT receptors. *Neuropharmacology* **1994**, *33*, 275–317.
- (41) Adham, N.; Romanienko, P.; Hartig, P.; Weinshank, R. L.; Branchek, T. The rat 5-hydroxytryptamine1B receptor is the species homologue of the human 5-hydroxytryptamine1D beta receptor. *Mol. Pharmacol.* **1992**, *41*, 1–7.
- (42) Dandapani, S.; Lan, P.; Beeler, A. B.; Beischel, S.; Abbas, A.; Roth, B. L.; Porco, J. A., Jr.; Panek, J. S. Convergent synthesis of complex diketopiperazines derived from pipecolic acid scaffolds and parallel screening against GPCR targets. *J. Org. Chem.* **2006**, *71*, 8934–8945.
- (43) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.
- (44) PubChem. <http://pubchem.ncbi.nlm.nih.gov/>. 2007. Ref Type: Electronic Citation.
- (45) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.
- (46) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; Wiley: New York, 1986.
- (47) Randi, M. On Characterization on Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- (48) Kier, L. B. A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109–116.
- (49) Kier, L. B. Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant. Struct.-Act. Relat.* **1987**, *6*, 8–12.
- (50) Kier, L. B.; Hall, L. H. An Electrotological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801.
- (51) Kier, L. B.; Hall, L. H. An Index of Electrotological State of Atoms in Molecules. *J. Math. Chem.* **1991**, *7*, 229.
- (52) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotological State*; Academic Press: 1999.
- (53) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.
- (54) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- (55) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Wiley: Chichester, 1983.
- (56) MolconnZ. <http://www.edusoft-lc.com/molconn/>. 2006 Ref Type: Electronic Citation.
- (57) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
- (58) Balaban, A. T. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355–375.
- (59) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
- (60) Wiener, H. Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636–2638.
- (61) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- (62) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (63) Stanton, D.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
- (64) Talet srl. DRAGON for Windows (Software for Molecular Descriptor Calculations). [5.4]. 2006 Ref Type: Internet Communication.
- (65) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
- (66) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.



- (67) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517–523.
- (68) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151–164.
- (69) Schuur, J.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
- (70) Todeschini, R.; Lasagni, M.; Marengo, E. *J. Chemom.* **1994**, *8*, 263–273.
- (71) Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.
- (72) Randic, M. *Int. J. Quantum Chem. Quant. Biol. Symp.* **1988**, *15*, 201–208.
- (73) Khashan, R.; Zheng, W.; Huan, J.; Wang, W.; Tropsha, A. Development of Novel Fragment-Based Chemical Descriptors using Frequent Common Subgraph Mining Approach and Their Application in QSAR modeling. Manuscript in preparation. 2007.
- (74) Huan, J.; Prins, J.; Wang, W. Efficient Mining of Frequent Subgraph in the Presence of Isomorphism. 2003, 549–552. Ref Type: Conference Proceeding.
- (75) Heritage, T. W.; Lowis, D. R. *Molecular hologram QSAR. In Rational Drug Design: Novel Methodology and Practical Applications*; Oxford University Press: New York, 1999.
- (76) Hurst T.; Heritage T. HQSAR - A Highly Predictive QSAR Technique Based on Molecular Holograms. 1997. San Francisco, CA, 213th ACS Natl. Meeting. Ref Type: Conference Proceeding.
- (77) Honorio, K. M.; Garratt, R. C.; Andricopulo, A. D. Hologram quantitative structure-activity relationships for a series of farnesoid X receptor activators. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3119–3125.
- (78) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.
- (79) Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.
- (80) Atkeson, C. G.; Moore, A. W.; Schaal, S. Locally Weighted Learning. *Artif. Intell. Rev.* **1997**, *11*, 11–73.
- (81) Zhang, S. X.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* **2006**, *46*, 1984–1995.
- (82) 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.
- (83) Wold, S. a. E. L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; H. v. d. W., Ed.; VCH: Weinheim, 1995; pp 309–318.
- (84) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
- (85) Sutherland, J. J.; Weaver, D. F. Three-dimensional quantitative structure-activity and structure-selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309–331.
- (86) Contrera, J. F.; Matthews, E. J.; Kruhlak, N. L.; Benz, R. D. Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. *Regul. Toxicol. Pharmacol.* **2004**, *40*, 185–206.
- (87) Seel, M.; Turner, D. B.; Willett, P. Effect of parameter variations on the effectiveness of HQSAR analyses. QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS. 18[3]. 1999, 245–252. Ref Type: Abstract.
- (88) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356–2364.

CI700404C