

Published in final edited form as:

J Med Chem. 2009 July 23; 52(14): 4210–4220. doi:10.1021/jm8013772.

The Discovery of Geranylgeranyltransferase-I Inhibitors with Novel Scaffolds by the Means of Quantitative Structure-Activity Relationship Modeling, Virtual Screening, and Experimental Validation

Yuri K. Peterson^{*,#}, Xiang S. Wang^{†,#}, Patrick J. Casey^{*,‡}, and Alexander Tropsha^{†,‡}

^{*}Department of Pharmacology, Duke University Medical Center, Durham, North Carolina, 27710

[†]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, Chapel Hill, North Carolina 27599

Abstract

Geranylgeranylation is critical to the function of several proteins including Rho, Rap1, Rac, Cdc42, and G-protein gamma subunits. Geranylgeranyltransferase type I (GGTase-I) inhibitors (GGTIs) have therapeutic potential to treat inflammation, multiple sclerosis, atherosclerosis, and many other diseases. Following our standard QSAR modeling workflow, we have developed and rigorously validated Quantitative Structure Activity Relationship (QSAR) models for 48 GGTIs using variable selection *k* nearest neighbor (*k*NN), automated lazy learning (ALL), and partial least square (PLS) methods. The QSAR models were employed for virtual screening of 9.5 million commercially available chemicals yielding 47 diverse computational hits. Seven of these compounds with novel scaffolds and high predicted GGTase-I inhibitory activities were tested *in vitro*, and all were found to be *bona fide* and selective micromolar inhibitors. Notably, these novel hits could not be identified using traditional similarity search. These data demonstrate that rigorously developed QSAR models can serve as reliable virtual screening tools.

Introduction

The proper functioning of proteins often relies on post-translational modification of the polypeptide leading to changes in chemical characteristics. Found at the extreme carboxyl terminus of the protein, one post-translational “program” utilized for over 140 proteins is the so called ‘CaaX box’, where ‘C’ is a cysteine, ‘aa’ is any aliphatic dipeptide, and X is the terminal residue that directs which of two prenyl groups is added^{1,2}. The protein prenylation cascade begins with the addition of either a 15-carbon isoprene farnesyl lipid when X residues are Ser, Met, Gln, Cys, and Ala; or a 20-carbon geranylgeranyl lipid is added when the X residue is Leu³. The CaaX prenyltransferases include protein farnesyltransferase (FTase) that adds the 15-carbon farnesyl group to proteins like Ras GTPases, nuclear lamins, several protein kinases and phosphatases, as well as other regulatory proteins⁴. Protein geranylgeranyltransferase type I (GGTase-I) transfers the 20-carbon geranylgeranyl group to proteins including critical signaling molecules from many classes, e.g., the Ras superfamily

[‡]To whom correspondence should be addressed. Alexander Tropsha, CB #7360, Beard Hall, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360, Tel: 919-966-2955, Fax: 919-966-0204, E-mail: alex_tropsha@unc.edu, Patrick J. Casey, 2 Jalan Bukit Merah, Singapore 169547, Tel: (65) 6516 7251, Fax: (65) 6226 3619, E-mail: patrick.casey@gms.edu.sg.

[#]These authors contributed equally to the paper.

(including K-Ras, Rho, Rap, Cdc42 and Rac), several G-protein gamma subunits, protein kinases (rhodopsin kinase, phosphorylase kinase, and GRK7), and protein phosphatases^{5,4}.

CaaX protein lipidation is obligate for the protein to be further modified by a protease termed Rce1, which removes the three terminal 'aaX' residues. The resulting isoprenylcysteine carboxylic acid is then methylated by isoprenylcysteine carboxymethyltransferase (Icmt) to create a protein terminus with a now mature (and very hydrophobic) isoprenylcysteine carboxymethylester⁶. Protein prenylation is important in the localization, interactions, and activity of modified proteins. Many of the prenylated proteins are found at the cytoplasmic face of cell membranes, where cell signaling is concentrated. Additionally, protein prenylation is required for cellular transformation by oncogenic Ras, providing the initial evidence that prenylation-dependent localization of proteins is critical in the Ras function⁷.

The first prenyltransferase inhibitors were farnesyltransferase inhibitors (FTIs), that were rapidly developed from early CaaX peptide mimics⁸ into the small organic ligands. The first peptidomimetic protein prenyltransferase inhibitors were mixed inhibitors, but highly selective inhibitors were rapidly developed. Using the example of one of the canonical oncogenes H-Ras, rational application of FTIs have shown efficacy in leukemias, gliomas, and breast cancers, providing impetus for targeting GGTase-I in cancers driven by geranylgeranylated oncogenes^{9;10}. Moreover, some Ras-dependent tumors are resistant to FTIs. This departure from prediction is likely due to so-called cross-prenylation by GGTase-I. During FTIs treatment some proteins, most notably K-Ras, that are typically farnesylated by FTase, are found geranylgeranylated, which restores at least a portion of the activity¹¹. Dual FTase/ GGTase inhibitors have received little attention and this type of treatment would impact a large number of proteins which make result interpretations complicated.

Several GGTIs have been developed that inhibit C20 lipid modification of GGTase-I substrates. GGTIs have been primarily developed for use as cancer therapeutics, particularly in cancers that have high levels, or activating mutations of geranylgeranylated proteins^{3,5}. GGTIs are now receiving broad interest for clinical use. Besides the continuing development as anti-cancer agents, GGTIs' are now postulated to have a potential in treating a wide array of other diseases including inflammation, multiple sclerosis, atherosclerosis, viral infection (HepC/HIV), apoptosis, angiogenesis, rheumatoid arthritis, psoriasis, glaucoma, and diabetic retinopathy^{1,12}. In addition, GGTase function is prerequisite in the normal functioning of many parasites and fungi, which has led to discovery programs to develop and use non-human selective GGTIs as antifungals and antiparasitics^{13;14}.

A wide variety of GGTIs have been reported in various publications in the relatively short time (~12 years) when the enzyme has been studied. Many of these have been designed rationally based on the substrates of GGTase-I: geranylgeranyl diphosphate (GGpp) or the CaaX peptide. There are also a number of natural compounds that were identified in a screen for inhibition of GGTase-I from *Candida sp.* A comprehensive review of known GGTIs was published recently¹². Unfortunately, many of the known GGTIs' binding mode(s) were never characterized and IC₅₀ data for the same compounds are often in disagreement when measured by different laboratories. These observations make the large portion of GGTIs less than optimal for QSAR model building. However, there are two known scaffolds that have been well characterized with respect to their binding to the peptide pocket and using similar estimates for IC₅₀ values. These include a number of CaaL peptidomimetics including aminobenzoic acid derivatives such as GGTI-298 and GGTI-2154^{15,16} and benzoyleneurea-based compounds¹⁷. More recently three newer classes of GGTIs have been published including one based on a piperazin-2-one backbone¹⁸, dihydropyrole/tetrahydropyridine based small molecules¹⁹, and allenolate compounds²⁰.

At the molecular level the principal effect of GGTIs is to block interactions (either with the membrane, or decreased interaction with protein binding partners), leading to mislocalization of signaling molecules. There are ~70 protein targets for GGTase-I, however their susceptibility is not equal. Many GGTase targets are rapidly inhibited and since the modification is post-translational this suggests high turnover rates of the modified target²¹. However, some geranylgeranylated proteins ($G\gamma$ subunits in particular) appear to have very long half-lives making them resistant to GGTIs²². At the cellular level inhibition of GGTase-I leads to cell cycle arrest at G_0/G_1 at low dose^{23,24} and complete blockade typically leads to apoptosis in both normal^{25,26} and transformed cell lines²⁷.

While there are several known chemical scaffolds for selective inhibition of GGTase-I, the list of potential uses for GGTI's highlights the need for chemical diversity of inhibitors targeting the enzyme. As an example, the therapeutic targeting of glaucoma might benefit from compound characteristics that are quite different than those for treating multiple sclerosis (MS). For instance, positive characteristics for a potential topical treatment of disorders like glaucoma or psoriasis would entail limited systemic bioavailability with perhaps a short half-life, while a MS drug would need to penetrate the blood-brain barrier and would benefit from an extended half-life. Additionally, there is a major therapeutic potential for creating species-selective GGTIs for use as antipathogens. The potential to manipulate these characteristics benefits more from having the flexibility of multiple scaffolds.

The development of small molecule inhibitors for clinical use is a multi-step process with many potential dead ends. In the preclinical setting drug development typically begins with designating a target protein/enzyme whose inhibition may lead to a desired physiologic response. Generally, the next step is to use carefully designed *in vitro* assay that allows screening of small molecule libraries. The goal of this screening process is to identify active molecules as defined by the particular activity assay.

Drug discovery and development can take many forms. It is often the case that a primary aim is to increase the affinity of a drug to its target. However, in some situations it eventually becomes clear (and often quite late in the development) that the actual drug scaffold has problems, particularly with bioavailability and metabolism, which cannot be solved through traditional lead optimization. It would be of great advantage to take the knowledge gained from the drug development process to more efficiently train models and search for novel scaffolds. Novel scaffolds are also desirable means of circumventing ADME problems that are often encountered at the later stages of the drug discovery process.

Quantitative Structure Activity Relationship (QSAR) modeling has been used extensively as a major computational tool for rationalizing the experimental data on binding or inhibitory activity of chemical compounds. QSAR is typically performed in two distinct modes, frequently referred to as 2D vs. 3D QSAR. In 2D QSAR, chemical descriptors are calculated from chemical graphs and no information about three-dimensional configuration of molecules is utilized. In 3D QSAR methods such as still popular Comparative Molecular Field Analysis (CoMFA)²⁸, conformational analysis and global 3D structure alignment should take place before descriptors are calculated. 2D QSAR has an inherent advantage of being independent of drug conformation, although it has a disadvantage of being much less robust in terms of model interpretation. However, because of its compound conformation and 3D alignment independence, 2D QSAR affords much higher computational efficiency and degree of automation when models are applied for virtual screening of large external libraries²⁹.

The use of QSAR models for virtual screening has not been viewed historically as its mainstream application; on the contrary, QSAR modeling approach has been typically considered as a lead optimization technology. Nevertheless, our group has been advocating for

and advancing the use of validated, externally predictive QSAR models for virtual screening for a number of years starting as early as 2001³⁰. We have published several earlier papers demonstrating the possibility of discovering novel bioactive compounds by the means of rigorous QSAR modeling coupled with virtual screening (e.g.,³⁰⁻³⁴; see for a recent review³⁵). Critical to this approach is an extensive model validation, in which known compounds are divided into groups that are used to build the model and a separate “external” set of known compounds that is used to test if the model is capable of accurately predicting activities of external compounds³⁶. An ensemble of robust and validated models can then be used to virtually screen a chemical database for compounds with potential target receptor activity^{30;37}. The use of QSAR models as virtual screening tools and for that matter, the methodology of QSAR modeling itself remains the area of active investigation, and the choice of methodology, such as the model building algorithms and the types of chemical descriptors, can dramatically influence the success and applicability of the approach³⁸. Our current approach that we term combinatorial QSAR modeling^{38;39} relies on the concurrent use of several QSAR modeling techniques for data analysis. Our aim is to identify models (or a combination of models) that afford the highest prediction accuracy and therefore could be expected to be successful in identifying novel bioactive molecules by the means of virtual screening.

There were few computational studies on GGTIs, including QSAR, reported in the literature. The only one searchable is done by Polley *et al* in 2004⁴⁰, using bayesian regularized artificial neural network on a GGTI dataset of 446 compounds. They had one division for training and test sets, thus only one single model was generated. The statistics of the model are optimal, with R^2 of 0.893 for training set, and q^2 of 0.778 for test set. It should be pointed out that there was no cross-validation during model building, and they did not apply models to virtual screening of chemical libraries to identify novel hits.

In this study, we have employed the combinatorial QSAR modeling strategy using three different approaches (described in detail below) to develop rigorous and validated models of 44 GGTIs with two chemical scaffolds. One scaffold (the GGTI-DUx series) was identified through initial random screening with extensive iterative follow-up medicinal chemistry²². The second set (GGTI-x) was initially developed following a rational peptidomimetic approach⁴¹. The workflow of our study is shown in Figure 1. The best models were applied to virtual screening of a large collection of *ca.* 9,500,000 compounds compiled from publicly available chemical databases. These searches resulted in only 47 consensus hits⁴² (i.e., predicted active by all models), none of which were present in the original dataset or have ever been characterized as GGTIs. Seven of these hits were validated *in vitro* and all were found to be active at micromolar level. Notably, three compounds incorporated novel scaffolds that were never reported before as potential GGTIs. For comparison, the traditional fingerprint based similarity search using all training set compounds as queries was also employed and the resulting hits were found to have little overlap with 47 QSAR/VS hits. Furthermore, none of experimentally confirmed QSAR/VS hits could be identified by the similarity search. These results support the notion that the combined application of rigorous QSAR modeling and virtual screening could serve as a powerful general modeling approach towards the discovery of novel drug candidates.

Computational Methods

GGTIs Dataset

The pharmacological data for 48 GGTIs used in this study were generated as part of an iterative drug discovery program that led to GGTI-DU40²². The details of the medicinal chemistry effort that resulted in this compound will be reported separately (J.P. Strachen *et al*, in preparation). The synthesis work was conducted in Pharmaceutical Product Development, Inc. (PPD,

Research Triangle Park). All 48 GGTIs were confirmed to be of greater than 95% purity by the means of LCMS, and the detailed spectra are with the company. The structure of GGTI-DU40 can be discussed in the context of the CaaL peptide framework. There is a free amide group, a spacer domain relating to the dialiphatic motif, and critical sulfur as found in the requisite cysteine residue of GGTase-I's substrates. In even simpler terms, the structure can be described as a hydrophobic head linked to a hydrophilic tail. Four additional GGTIs included in the data set were peptidomimetics as well including GGTI-287⁴¹, GGTI-297⁴³, and GGTI-2154⁴⁴. These four compounds were developed as CaaL peptidomimetics and are reasonably similar to each other but quite dissimilar to the GGTI-DU40 series (*cf.* Chart S1 of Supporting Information). Chemical structures of all inhibitors used in QSAR modeling and their associated IC₅₀ values are given in Chart S1. The pIC₅₀ values for all compounds ranged from 3.8 to 7.6 with a near Gaussian distribution (*cf.* Figure 2). Importantly, the combination of data sets including compounds with different chemical scaffolds of the wide distribution of pairwise chemical similarities within the entire dataset (Figure S1^{of} Supporting Information), which in theory (and as we have established in this study, in practice) should have enabled the identification of chemically diverse virtual hits from virtual screening.

Generation of 2D Molecular Descriptors

All chemical structures were generated using ACD/ChemSketch software before converting them to SMILES. MolconnZ software version 4.09 (MZ4.09)⁴⁵ was used to generate the molecular topological index descriptors⁴⁶. MZ4.09 calculates more than 700 different descriptors, however, many are used for accounting purposes and several more have either zero values or zero variance. Once non-redundant descriptors were removed, a set of 274 chemically relevant descriptors remained. In order to prevent unequal weighting, descriptors were linearly normalized to fall within the range of zero to one based on the minimum and maximum values of each descriptor (i.e., range-scaled)⁴⁷. The use of range-scaling avoids giving descriptors with significantly broader ranges a disproportional weight upon distance calculations in multidimensional MZ4.09 descriptor space. We then follow our standard protocols to subdivide the whole dataset into multiple training/test set pairs using the Sphere Exclusion method⁴⁸ 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 biological activities of the test set compounds accurately.

QSAR Methods

The *k* Nearest Neighbor (*k*NN) QSAR method used in this study employs the *k*NN pattern recognition principle⁴⁹ and variable selection. In short, a subset of variables (descriptors) is selected randomly in the beginning as a Hypothetical Descriptor Pharmacophore (HDP)⁵⁰. The HDP is validated by LOO-CV, where each compound is eliminated from the training set and its GGTase-I inhibition activity is predicted as the weighted average of the activity values of the *k* most similar molecules (*k* varies from 1 to 5). The weighted molecular similarity is represented by the modified Euclidean distance between compounds in HDP multidimensional space as shown in Equation 1 and Equation 2. Essentially, the neighbor with the smaller distance from a compound is given a higher weight in calculating the predicted activity:

Supporting Information

The heatmap of self-similarity matrix for GGTIs modeling set, distributions of models for Y-randomization tests, experimental data of GGTIs screening hits FTase activities, chemical structures and pIC₅₀ values for GGTIs modeling dataset and screening hits, purity data for target compounds, and others supplementary data indicated in the text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

$$w_i = \frac{e^{-d_i}}{\sum_i e^{-d_i}} \quad (1)$$

$$\tilde{y} = \sum w_i y_i \quad (2)$$

where d_i is the Euclidean distance between the compound i and its k th nearest neighbors; w_i is the weight for the k th nearest neighbor; y_i is the experimentally measured activity value for the k th nearest neighbor; and \tilde{y} is the predicted activity value.

Simulated annealing and Metropolis-like acceptance criteria were used to optimize the selection of variables. Details of the k NN method implementation including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere⁴⁷. The statistical significance of the models were estimated by the LOO-CV q^2 in the training set, a coefficient of determination R_0^2 (Equation 3) and linear fit R^2 values for both internal and external test sets.

$$q^2(R_0^2) = 1 - \frac{\sum (\tilde{y}_k - y_k)^2}{\sum (\bar{y} - y_k)^2} \quad (3)$$

Here y_k and \tilde{y}_k are the observed and predicted activities of a compound k , respectively, and \bar{y} is the average activity of all compounds. Model acceptability cutoffs were $q^2 > 0.60$ for training set and correlation coefficient $R^2 > 0.60$ for internal test set⁵¹. Models that satisfied both criteria were applied to external validation sets.

We also employed two other methods, i.e. Automated Lazy Learning QSAR (ALL-QSAR) and Partial Least Square (PLS) QSAR, in this study. The ALL-QSAR was developed in our group and is ideal for a large or diverse dataset⁵². The PLS QSAR is arguably the most traditional and less sophisticated QSAR approach among those explored in this study. The modeling procedures were similar to those described in our previous studies^{39;52}.

Applicability Domain of 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 k NN 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. In brief, 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:

$$D_T = \bar{D} + Z\sigma \quad (4)$$

Here, \bar{D} is the average Euclidean distance of the k nearest neighbors of each compound within the training set, σ 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 the applicability domain 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. Additional details can be found in our previous publications^{36;39}.

Model Validation and Robustness

Y-randomization test is a widely used validation technique to ensure the robustness of a QSAR model⁵³. 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 , it implies that an acceptable QSAR model cannot be obtained for the given dataset by the current modeling method. Y-randomization test was applied to all QSAR methods considered in this study.

Virtual Screening of Chemical Databases

Although we have employed three different QSAR methods for model building, k NN produced the most predictive and robust models (*cf.* Table 1). It was then selected for primary use in virtual screening. The screening was performed on our Molecular Modeling Laboratory (MML) in-house collection of 9,500,000 compounds, including the ZINC7.0 database of *ca.* 6,500,000 compounds⁵⁴, the Maybridge database (2008.03) of *ca.* 56,000 compounds⁵⁵, the World Drug Index (WDI) database of *ca.* 59,000 compounds⁵⁶, the ASINEX Synergy libraries of *ca.* 11,000 compounds, the Chemizon Progenitor databases (2006 v1.1) of *ca.* 3,300 compounds⁵⁷ and several other commercial databases. None of the compounds found in the training set were present in the mining databases.

As illustrated in the workflow of Figure 1, the rigorously validated QSAR models were employed for virtual screening. A global applicability domain (calculated using all descriptors) was applied first in order to filter out compounds that differed globally in their structure from the modeling set compounds. All 48 known GGTIs were used as probes in the calculations. During the consensus prediction, the results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction and the standard deviation of estimated means across all models was small. Furthermore, we restricted ourselves to the most conservative applicability domain for each model using the cutoff (*cf.* Equation 4) $Z = 0.5$.

All the modeling and virtual screening calculations were done at a 352-processor Beowulf Linux cluster of the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The compute nodes are Intel Xeon IBM BladeCenter of Dual Intel Xeon 2.8GHz, with 2.5GB RAM on each node. The cluster runs the Red Hat Enterprise Linux 4.0 (32-bit) and the nodes communicate via a Gigabit Ethernet network. The processing speed of QSAR-based screening is relatively high, *ca.* 100K compounds per minute. As could be expected, the processing speed was found to scale linearly with the size of the screening library.

Fingerprint Based Similarity Search

The chemical similarity search was conducted with the MOE2006.08 package using the standard protocols. The MACCS structural keys were utilized with the Tanimoto Coefficient (Tc) as the similarity metric. The search was carried out independently for each of the 48 compound of the GGTIs modeling dataset. In the case that the hits from individual searches

were the same, a special Scientific Vector Language (SVL) script was employed to remove one of them based on the chemical topology.

Results and Discussions

QSAR Models and Their Robustness

The *k*NN QSAR model building employed 274 MZ4.09 descriptors derived from 48 GGTIs as the independent variables. During the calculations, the conservative value of 0.5 was used for Z_{cutoff} to define the applicability domain (*cf.* Eq. 4). In total, 6720 models were generated and only 104 models were accepted using the cutoff for both leave-one-out cross-validated q^2 values for training sets and predictive R^2 for test sets greater than 0.60. As shown in Figure 3a and Table 1, the *k*NN QSAR method afforded the best models with q^2/R^2 values as high as 0.82/0.85 for this GGTIs dataset ($R_0^2 = 0.83$). These results suggest that the intrinsic inhibition activity relationships exist for GGTIs that can be described reasonably well by *k*NN models using MZ4.09 descriptor sets.

As part of our combinatorial QSAR strategy, PLS and ALL QSAR were employed to analyze the same dataset, using the same descriptors and the same training/test set divisions generated by Sphere Exclusion. These two additional QSAR approaches were expected to increase the chances of successful modeling of GGTIs so that only best models are selected for virtual screening. Multiple predictive models by ALL QSAR method were obtained with the highest R^2 of 0.81 for 7 compounds in the test set, as can be seen in Figure 3b and Table 2. Additional model parameters for the same test set were R_0^2 of 0.91 and the RMSE of 0.21. The models produced by the PLS QSAR method were less satisfactory. As shown in Figure 3c and Table 3, all five PLS QSAR models had superior values for cross-validated q^2 of training set, ranging from 0.92 to 0.97. However, their predictive R^2 for test sets were around 0.20 ~ 0.30, except for model #2 ($R^2 = 0.87$, RMSE = 0.92). Thus, only this latter model could be used for consensus prediction.

To further evaluate the robustness of *k*NN QSAR models, the whole model building process was repeated but using randomized IC_{50} values in place of the actual measured IC_{50} values. Figure S2 of Supporting Information shows the distinctive distribution of all models for actual vs. Y-randomized data in term of q^2/R^2 values. As can be observed, the q^2 ranged from 0.40 to 0.90 for actual models while from -0.30 to 0.80 for randomized ones. It should be pointed out that no models exceeded the 0.60 cutoff for both q^2 and R^2 when the activity values were randomized. The standard one-tail hypothesis test was conducted to evaluate the statistical significance of QSAR models derived from the actual data set, in comparison to the models from the random data set. The Z score that is calculated from the q^2 value is 4.22, much higher than the tabular value of Z_c , which corresponds to the level of significance $\alpha = 0.01$. This suggests that *k*NN does not have the ability to correlate descriptors to random activities for GGTIs dataset, thus the QSAR models obtained with the real data are robust.

Comparison of Three QSAR Algorithms

Three QSAR methods were used in this study, including *k*NN regression QSAR, ALL QSAR, and PLS QSAR. Each method was combined with MZ4.09 descriptor set and applied to the same training/test sets splits of GGTIs dataset making it possible to compare the performances of different algorithms. Overall, all three QSAR methods afforded predictive models that met the statistical thresholds ($q^2, R^2 > 0.60$) though the number of acceptable models varied (104 for *k*NN QSAR, 7 for ALL QSAR, and 1 for PLS QSAR). All of these acceptable models were used for virtual screening of chemical libraries and consensus prediction downstream of our modeling workflow (*cf.* Figure 1). Among the three, the *k*NN QSAR method afforded the largest number of acceptable models and the highest statistics with q^2/R^2 values of 0.82/0.85.

The ALL QSAR yielded the best R^2 value as high as 0.91 but the number of predictive models was limited. Similarly, the PLS QSAR method generated only one good model (q^2/R^2 of 0.92/0.87) and the results were dependent on the splits of training/test sets. It should be noted that there was no external validation dataset available in addition to the test sets, because of the small size of GGTIs dataset used in this study and the limited source of literature data. Thus, the external predictive ability of all acceptable models had to be validated by the screening hits in this particular case.

Virtual Screening by Validated QSAR Models

As the first step of our QSAR-based virtual screening, the preliminary filtering of the 9.5 million compounds in our screening library yielded 79 initial hits. This was done by using the global applicability domain of all 48 GGTIs in the modeling set. After consensus predictions by 104 validated k NN models, their predicted activities (pIC_{50}) were found ranging from 4.51 to 5.96. Only 47 hits, including two pairs of stereoisomers, showed high predicted activity ($pIC_{50} > 5.50$) as well as high model coverage and were designated as the final hits. Concurrently, ALL and PLS QSAR models were employed to re-evaluate those 79 hits in order to identify the consensus hits among all three methods. In the end, seven compounds were prioritized for experimental validation based on high predicted activity, uniqueness of structure, and availability. The 2D chemical structures of the 47 compounds with predicted high inhibition activities are shown in the Chart S2 of Supporting Information. A large portion of the screening hits contained the pyridine-pyrazole-phenyl (6-5-6) ring structure which is prevalent in the training set. It was expected considering the empirical nature of QSAR modeling and the very conservative applicability domain used in the study.

Enzymatic Characterization of VS Hits

Using purified recombinant GGTase-I as an enzyme source and GGpp and Ras-CVLL as substrates, seven hit compounds were tested *in vitro* as a matter of the experimental validation. The selection was based on high predicted activity, availability and structural uniqueness. All tested compounds showed inhibition of GGTase-I with the pIC_{50} ranging from 3.63 to 5.44 (*cf.* Figure 4 and Table 4). The comparison between predicted and experimental data is shown in Table 4. Using $pIC_{50} > 4.00$ as the threshold to define the actives, it is shown that k NN QSAR predicted correctly most hit compounds as active, except for GGTI-DU.Sig3⁴². The R^2 for the prediction is 0.45 in this case. PLS QSAR also identified the same six compounds as actives, but had a large error on GGTI-DU.As1⁴² (absolute error of 2.85). ALL QSAR had the worst performance, however, predicting only 2 of 7 hits to be active. Thus, k NN based predictions were better than other methods in this case.

The unexpected result was to have several predicted actives that did not have this common ring feature in their structure. In fact, seven highly-ranked hits had no apparent relationship to any of the training set molecules. They had furan, triazole, tetrazole, and pyridine cores in their scaffolds while all non-peptidomimetic compounds of the training set were based on a pyrazole core. Therefore, the seven hit compounds without the 6-5-6 rings that were found in most non-peptidic GGTIs appear to be the structurally novel hits. Figure 4(b) list the chemical structures of three representative confirmed hits, GGTI-DU.Sig3, GGTI-DU.As2 and GGTI-DU.En2. The novel scaffolds (highlighted) among the three can be traced back to the general formulas of substructures found in the 47 screening GGTIs hits (*cf.* Figure S6 of Supporting Information). For example, GGTI-DU.Sig3 contains the novel scaffold defined by Formula IV while the new structural element in GGTI-DU.As2 belongs to Formula II. Again, these four structural formulas cannot be found in any compounds in the GGTIs training set. This observation lends additional support to the hypothesis that QSAR-based virtual screening is capable of 'scaffold hopping'.

Although these QSAR/VS derived GGTIs hits had GGTase inhibition activity, it is possible that these effects are nonspecific. The ability of the compounds to inhibit GGTase-I *in vitro* is not the only requirement for their potential as therapeutics. Another major hurdle in the development of GGTIs is their selectivity towards GGTase *versus* FTase. These two proteins share ~35% sequence identity and have been known to have cross-reacting substrates, particularly K-Ras. In fact, it was the discovery of cross-prenylation in the presence of highly selective FTIs that led to increased interest in the development of selective GGTIs. We therefore tested four representative hit compounds for inhibitory activities against this highly related FTase. Impressively, all these four compounds that inhibited GGTase showed little to no activity in the FTase assay (*cf.* the examples in Figure S4 of Supporting Information). This indicates that QSAR-based VS hits proved to be target specific.

Fingerprint Based Similarity Search

As expected, many of the 47 QSAR VS hits exhibit high degrees of similarities to the modeling set (*cf.* Chart S1). It is therefore more interesting to further analyze the 7 confirmed hits which have novel scaffolds. All 7 hits were compared to the GGTIs modeling set using the MACCS structural keys and the result is shown in Table 5. Notably, none of these confirmed hits had $T_c > 0.80$ when compared to any of the 48 GGTIs. In fact, the similarity of screening hits **89** and **92**⁴² to most similar compounds in the modeling set had $T_c < 0.70$. Thus, these 7 hits are highly dissimilar to the modeling set as measured by MACCS structural key and the associated T_c metric.

An intriguing question now emerges as to what kind of hits would a MACCS based similarity search find using compounds in the same dataset as probes and how those hits would compare to hits identified with QSAR-based VS. To create a complete picture of the differences between QSAR and fingerprint based VS, we applied MACCS based similarity search to the same virtual screening library of ~9.5 million compounds. The search generated 8,132 hits with $T_c > 0.80$, 724 hits with $T_c > 0.85$ and only 22 hits with $T_c > 0.90$; among those 22 hits there were two pairs of isomers. Notably, there were few overlaps between the hits from QSAR VS and fingerprint based similarity search. Among the 724 hits at $T_c = 0.85$ (the default similarity cutoff in MOE2006.08 package), only 20 can also be found within the 79 preliminary hits of QSAR based VS. In other words, the remaining 59 QSAR/VS hits were dissimilar to the GGTIs dataset in term of global similarity defined by MACCS structural keys. When the threshold was set as high as of $T_c = 0.90$, there was only one compound **107** (PubChem CID: 3942219) of the QSAR/VS hits that was found among the 22 hits from the similarity search. The resulting MACCS VS hits, as expected, are highly similar to the GGTIs dataset and can be divided into those that are similar to the GGTI-DUx series of pyrazoles (16 compounds) and those belonging to the GGTI-X series of peptidomimetics (6 compounds). All pyrazole-like MACCS VS hits contain the 6-5-6 ring system, a hydrophobic tail and an amine(s) linker that connects the two. The peptidomimetic MACCS VS hits were visually less similar to the modeling dataset compounds. However, close examination indicates that their entire backbones are in fact highly similar. The primary reason for the confusion is the phenyl rings found at both termini of the MACCS VS hits.

To further validate the MACCS VS hits, five of the peptidomimetic hits were tested for the GGTase assay. Notably, none of these compounds exhibited inhibition activity in the GGTase-I assay (data not shown). These results suggested that the fingerprint-based similarity search was not effective in identifying novel biological active compounds effectively.

Interpretation of Frequent Descriptors

In order to correlate the biological activities to the relevant chemical features, variable selection QSAR methods search for the optimal subsets of descriptors using different algorithms. In

current studies, both *k*NN and PLS methods identified the most relevant descriptors and many of them were found to be the same (*cf.* Figure S3 of Supporting Information). The descriptors were ranked based on their frequencies of use in models included in the consensus QSAR model. Among the frequent descriptors, the binary *nHBint10* descriptor indicates the presence of potential internal hydrogen bonds within the structure (see compounds ¹⁵ and **48** 22). The *SssO* descriptor is an integer and represents the sum of the electrotopological state indices for oxygen atoms. Its mean value is 0.935 for 25 out of 48 GGTIs in the modeling set. The *Ncarboxylicacid* is the group based descriptor, which indicates the presents of carboxylic acid functional groups. The functional groups that are encoded by frequent descriptors could be interpreted as GGTIs' pharmacophoric elements.

Conclusions

Drug discovery paradigms have been changing rapidly due to advances in high-throughput screening technologies, combinatorial chemistry and computer-aided modeling methodologies. Often, drug candidates were abandoned after a single (or groups of similar) compounds had been found to be of use-limiting bioavailabilities or toxicities. The frequent possibility that a target-specific bioactive compound could have undesired ADME/Tox properties implies that chemically diverse hits should be ideally generated in the beginning of the drug discovery cycle. The state-of-the-art QSAR methodologies that rely on variable selection and extensive model validation have become increasingly more powerful in the areas of drug lead identification and optimization. As we have demonstrated in this study, variable selection QSAR modeling followed by virtual screening could be successfully used to enable the discovery of structurally novel hits. The identification of structurally novel GGTIs will bring us closer to the goal of making a selective GGTI that could also have plausible ADME properties. Fingerprint-based similarity search is another example of a technique that is able to find "remotely-similar" compounds⁵⁸. However, typical implementations of this approach do not use variable selection (unlike many QSAR methods) to make the results more focused towards target-specific biological activity.

Despite a great interest in GGTIs only a limited number of lead scaffolds have emerged from traditional medicinal chemistry approaches. In this study, we have enabled the discovery of GGTIs with novel scaffolds by building robust QSAR models of training set compounds and then using these models for virtual screening of large chemical libraries. As we have shown in this report, using variable selection *k*NN QSAR method, we were able to generate more than a hundred of statistically robust models for a dataset including GGTIs of two types of scaffold. Alternative methods used in this study, i.e., ALL QSAR and PLS QSAR methods afforded acceptable models (values greater than 0.60 for both q^2 and R^2) but *k*NN produced more models with higher prognostic power.

Mining of the 9.5 million compound screening library for GGTIs using validated *k*NN, ALL, and PLS QSAR models, resulted in 47 hits with moderate to high predicted activity. The 7 compounds chosen for the highest predicted activity and greatest dissimilarity from the training set showed activity towards GGTase, indicating an apparent 100% success rate. None of the models afforded highly accurate quantitative prediction of the activity of experimentally confirmed hits but *k*NN models correctly predicted the order of activities. Several of these hits were also shown experimentally to be not only active but highly selective towards GGTase I. Thus, 2D-QSAR modeling was proven to be very efficient for enabling virtual screening of millions of compounds in a rapid fashion and selection of only a very small number of computational hits for the experimental validation. Notably, these novel QSAR hits cannot be obtained by traditional fingerprint based similarity search. The latter was conducted as the control but only yielded highly similar hits to the GGTIs dataset.

Most screening hits shared the 6-5-6 ring scaffold found in most of the training set GGTIs. These were expected as the QSAR/VS is designed to find chemically similar entities. However, several compounds lacking this scaffold were predicted to be GGTIs and were confirmed active experimentally. These results demonstrate that QSAR models can serve as reliable virtual screening tools capable of identifying novel biologically active scaffolds. The modeling strategy described in this report can be applied to many chemical biological systems for which experimental biological testing data for a series of chemicals is available.

Biological Methods

Materials

Farnesyl diphosphate (Fpp) and geranylgeranyl diphosphate (GGpp) were purchased from Biomol, Inc. (Plymouth Meeting, PA). ^3H -GGpp and ^3H -Fpp were purchased from PerkinElmer (Boston, MA). The FTIL-744-832 was purchased from Sigma (Saint Louis, MO). GGTI-DU40 was synthesized by the Duke Small Molecule Synthesis Facility.

Enzyme Assays

Protein GGTase-I or FTase activities were determined by following the incorporation of radiolabeled isoprenoid from ^3H -GGpp or ^3H -Fpp into Ras proteins as described previously⁵⁹. Briefly, purified mammalian GGTase-I or FTase (50 ng, expressed in Sf9 cells)⁶⁰ were used to initiate reactions containing 0.5 μM GGpp or Fpp, respectively, and 1 μM of the appropriate purified His-tagged Ras substrates (Ras-CVLL for GGTase-I; H-Ras for FTase). Final DMSO concentration was 2% for all samples. Reactions were carried out for 10 min at 30°C before precipitation and product determination. Nonspecific binding was defined by boiled enzyme and was identical to maximal inhibition by GGTI-DU40 for GGTase-I, and the well-characterized FTIL-744-832 for FTase. The data manipulation and curve fitting were performed using Prism (GraphPad, San Diego CA).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

GGTase-I, Geranylgeranyltransferase type I
GGTIs, Geranylgeranyltransferase type I inhibitors
FTase, farnesyltransferase
FTIs, farnesyltransferase inhibitors
RhoA, Ras homolog gene family member A
Cdc42, cell division cycle 42
GRK7, G-protein coupled receptor kinase 7
QSAR, Quantitative Structure Activity Relationship
*k*NN, *k* nearest neighbor
ALL, automated lazy learning
PLS, partial least square
MS, multiple sclerosis
CoMFA, Comparative Molecular Field Analysis
MZ4.09, MolconnZ software version 4.09
HDP, Hypothetical Descriptor Pharmacophore
RMSE, root mean square error
WDI, World Drug Index
Tc, Tanimoto Coefficient

MML, Molecular Modeling Laboratory
 GGTI-DU40, N-[(2S)-1-amino-1-oxo-3-phenylpropan-2-yl]-4-[2-(3,4-dichlorophenyl)-4-(2-methylsulfonyl)ethyl]-5-pyridin-3-ylpyrazol-3-yl] oxybutanamide
 GGTI-287, (2S)-2-[[4-[[[(2R)-2-azaniumyl-3-sulfanylpropyl]amino]-2-phenylbenzoyl]amino]-4-methyl]pentanoate
 GGTI-297, (2S)-2-[[4-[[[(2R)-2-azaniumyl-3-sulfanylpropyl]amino]-2-naphthalen-1-ylbenzoyl]amino]-4-methyl]pentanoate
 GGTI-298, methyl (2S)-2-[[4-[[[(2R)-2-amino-3-sulfanylpropyl]amino]-2-naphthalen-1-ylbenzoyl]amino]-4-methyl]pentanoate
 GGTI-2154, (S)-2-(5-((1H-imidazol-4-yl)methylamino)-2'-methylbiphenyl-2-yl)carboxamido)-4-methylpentanoic acid

Acknowledgments

We thank Mihir Shah, Alexander Golbraikh, Carolyn Weinbaum, and Missy Infante for technical support. We also thank Rainbo Hultman for critical evaluation of the manuscript. We acknowledge the access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. This work was supported in part by National Institutes of Health research grant F32-GM073420 (Y.K.P), GM46372 (P.J.C.), GM066940 (A.T.), the RoadMap Center planning grant P20-HG003898 (A.T.), and the UNC-CH University Research Council Research Grant A3-12988 (X.S.W.).

References

- Zhang FL, Casey PJ. Protein prenylation: molecular mechanisms and functional consequences. *Annu. Rev. Biochem* 1996;65:241–269. [PubMed: 8811180]
- Cox AD, Der CJ. Farnesyltransferase inhibitors: promises and realities. *Curr. Opin. Pharmacol* 2002;2:388–393. [PubMed: 12127871]
- Winter-Vann AM, Casey PJ. Post-prenylation-processing enzymes as new targets in oncogenesis. *Nat. Rev. Cancer* 2005;5:405–412. [PubMed: 15864282]
- Casey PJ, Seabra MC. Protein prenyltransferases. *J. Biol. Chem* 1996;271:5289–5292. [PubMed: 8621375]
- Sebti SM, Hamilton AD. Farnesyltransferase and geranylgeranyltransferase I inhibitors in cancer therapy: important mechanistic and bench to bedside issues. *Expert. Opin. Investig. Drugs* 2000;9:2767–2782.
- Reid TS, Terry KL, Casey PJ, Beese LS. Crystallographic analysis of CaaX prenyltransferases complexed with substrates defines rules of protein substrate selectivity. *J. Mol. Biol* 2004;343:417–433. [PubMed: 15451670]
- Kohl NE, Conner MW, Gibbs JB, Graham SL, Hartman GD, Oliff A. Development of inhibitors of protein farnesylation as potential chemotherapeutic agents. *J. Cell Biochem* 1995;22:145–150.
- Karp JE, Lancet JE. Farnesyltransferase inhibitors (FTIs) in myeloid malignancies. *Ann. Hematol* 2004;83:S87–S88. [PubMed: 15124688]
- Karp JE, Lancet JE. Development of farnesyltransferase inhibitors for clinical cancer therapy: focus on hematologic malignancies. *Cancer Invest* 2007;25:484–494. [PubMed: 17882662]
- Gu WZ, Joseph I, Wang YC, Frost D, Sullivan GM, Wang L, Lin NH, Cohen J, Stoll VS, Jakob CG, Muchmore SW, Harlan JE, Holzman T, Walten KA, Lador US, Anderson MG, Kroeger P, Rodriguez LE, Jarvis KP, Ferguson D, Marsh K, Ng S, Rosenberg SH, Sham HL, Zhang H. A highly potent and selective farnesyltransferase inhibitor ABT-100 in preclinical studies. *Anticancer Drugs* 2005;16:1059–1069. [PubMed: 16222147]
- Lerner EC, Zhang TT, Knowles DB, Qian Y, Hamilton AD, Sebti SM. Inhibition of the prenylation of K-Ras, but not H- or N-Ras, is highly resistant to CAAX peptidomimetics and requires both a farnesyltransferase and a geranylgeranyltransferase I inhibitor in human tumor cell lines. *Oncogene* 1997;15:1283–1288. [PubMed: 9315095]
- El Oualid F, Cohen LH, van der Marel GA, Overhand M. Inhibitors of protein: geranylgeranyl transferases. *Curr. Med. Chem* 2006;13:2385–2427. [PubMed: 16918362]

13. Chakrabarti D, Da Silva T, Barger J, Paquette S, Patel H, Patterson S, Allen CM. Protein farnesyltransferase and protein prenylation in *Plasmodium falciparum*. *J. Biol. Chem* 2002;277:42066–42073. [PubMed: 12194969]
14. Sagan SM, Rouleau Y, Leggiadro C, Supekova L, Schultz PG, Su AI, Pezacki JP. The influence of cholesterol and lipid metabolism on host cell structure and hepatitis C virus replication. *Biochem. Cell Biol* 2006;84:67–79. [PubMed: 16462891]
15. Vasudevan A, Qian Y, Vogt A, Blaskovich MA, Ohkanda J, Sebti SM, Hamilton AD. Potent, highly selective, and non-thiol inhibitors of protein geranylgeranyltransferase-I. *J. Med. Chem* 1999;42:1333–1340. [PubMed: 10212118]
16. Carrico D, Blaskovich MA, Bucher CJ, Sebti SM, Hamilton AD. Design, synthesis, and evaluation of potent and selective benzoyleneurea-based inhibitors of protein geranylgeranyltransferase-I. *Bioorg. Med. Chem* 2005;13:677–688. [PubMed: 15653335]
17. Vogt A, Sun J, Qian Y, Hamilton AD, Sebti SM. The geranylgeranyltransferase-I inhibitor GGTI-298 arrests human tumor cells in G0/G1 and induces p21(WAF1/CIP1/SDI1) in a p53-independent manner. *J. Biol. Chem* 1997;272:27224–27229. [PubMed: 9341167]
18. Peng H, Carrico D, Thai V, Blaskovich M, Bucher C, Pusateri EE, Sebti SM, Hamilton AD. Synthesis and evaluation of potent, highly-selective, 3-aryl-piperazinone inhibitors of protein geranylgeranyltransferase-I. *Org. Biomol. Chem* 2006;4:1768–1784. [PubMed: 16633570]
19. Castellano S, Fiji HD, Kinderman SS, Watanabe M, Leon P, Tamanoi F, Kwon O. Small-molecule inhibitors of protein geranylgeranyltransferase type I. *J. Am. Chem. Soc* 2007;129:5843–5845. [PubMed: 17439124]
20. Watanabe M, Fiji HD, Guo L, Chan L, Kinderman SS, Slamon DJ, Kwon O, Tamanoi F. Inhibitors of protein geranylgeranyltransferase-I and rab geranylgeranyltransferase identified from a library of allenolate derived compounds. *J. Biol. Chem.* 2008
21. Bergo MO, Gavino BJ, Hong C, Beigneux AP, McMahon M, Casey PJ, Young SG. Inactivation of IcmT inhibits transformation by oncogenic K-Ras and B-Raf. *J. Clin. Invest* 2004;113:539–550. [PubMed: 14966563]
22. Peterson YK, Kelly P, Weinbaum CA, Casey PJ. A novel protein geranylgeranyltransferase-I inhibitor with high potency, selectivity, and cellular activity. *J. Biol. Chem* 2006;281:12445–12450. [PubMed: 16517596]
23. Sun J, Qian Y, Chen Z, Marfurt J, Hamilton AD, Sebti SM. The geranylgeranyltransferase I inhibitor GGTI-298 induces hypophosphorylation of retinoblastoma and partner switching of cyclin-dependent kinase inhibitors. A potential mechanism for GGTI-298 antitumor activity. *J. Biol. Chem* 1999;274:6930–6934. [PubMed: 10066746]
24. Stark WW Jr, Blaskovich MA, Johnson BA, Qian Y, Vasudevan A, Pitt B, Hamilton AD, Sebti SM, Davies P. Inhibiting geranylgeranylation blocks growth and promotes apoptosis in pulmonary vascular smooth muscle cells. *Am. J. Physiol* 1998;275:L55–L63. [PubMed: 9688935]
25. Li X, Liu L, Tupper JC, Bannerman DD, Winn RK, Sebti SM, Hamilton AD, Harlan JM. Inhibition of protein geranylgeranylation and RhoA/RhoA kinase pathway induces apoptosis in human endothelial cells. *J. Biol. Chem* 2002;277:15309–15316. [PubMed: 11839765]
26. Morgan MA, Wegner J, Aydilek E, Ganser A, Reuter CW. Synergistic cytotoxic effects in myeloid leukemia cells upon cotreatment with farnesyltransferase and geranylgeranyl transferase-I inhibitors. *Leukemia* 2003;17:1508–1520. [PubMed: 12886237]
27. Dan HC, Jiang K, Coppola D, Hamilton A, Nicosia SV, Sebti SM, Cheng JQ. Phosphatidylinositol-3-OH kinase/AKT and survivin pathways as critical targets for geranylgeranyltransferase I inhibitor-induced apoptosis. *Oncogene* 2004;23:706–715. [PubMed: 14737105]
28. Sybyl Users Manuel. St. Louis, MO: 2002. Tripos, Inc..
29. Tropsha, A. Recent Trends in Quantitative Structure-Activity Relationships. In: Abraham, D., editor. *Burger's Medicinal Chemistry and Drug Discovery*. New York: John Wiley & Sons, Inc; 2003. p. 49-77.
30. Tropsha, A.; Cho, SJ.; Zheng, W. "New Tricks For an Old Dog": Development and application of novel QSAR methods for rational design of combinatorial chemical libraries and database mining. In: Parrill, AL.; Reddy, MR., editors. *ACS Symposium Series*. 2001. p. 719

31. Tang H, Wang XS, Huang XP, Roth BL, Butler KV, Kozikowski AP, Jung M, Tropsha A. Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation. *J. Chem. Inf. Model.* 2009
32. Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput. Aided Mol. Des* 2008;22:593–609. [PubMed: 18338225]
33. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem* 2002;45:2811–2823. [PubMed: 12061883]
34. Shen M, Beguin C, Golbraikh A, Stables JP, 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. [PubMed: 15084134]
35. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des* 2007;13:3494–3504. [PubMed: 18220786]
36. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar and Combinatorial Science* 2003;22:69–77.
37. Hoffman BT, Kopajtic T, Katz JL, Newman AH. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn Z descriptors. *J. Med. Chem* 2000;43:4151–4159. [PubMed: 11063611]
38. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng W, Wolschann P, Buchbauer G, Tropsha A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci* 2004;44:582–595. [PubMed: 15032539]
39. Wang XS, Tang H, Golbraikh A, Tropsha A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J. Chem. Inf. Model* 2008;48:997–1013. [PubMed: 18470978]
40. Polley MJ, Winkler DA, Burden FR. Broad-based quantitative structure-activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. *J. Med. Chem* 2004;47:6230–6238. [PubMed: 15566293]
41. Lerner EC, Qian Y, Hamilton AD, Sebt SM. Disruption of oncogenic K-Ras4B processing and signaling by a potent geranylgeranyltransferase I inhibitor. *J. Biol. Chem* 1995;270:26770–26773. [PubMed: 7592913]
42. Peterson Y, Wang SX, Casey P, Tropsha A. Novel Chemical Scaffolds for Protein Geranylgeranyltransferase Type I Inhibitors. USPTO. 2007
43. Qian Y, Vogt A, Vasudevan A, Sebt SM, Hamilton AD. Selective inhibition of type-I geranylgeranyltransferase in vitro and in whole cells by CAAL peptidomimetics. *Bioorg. Med. Chem* 1998;6:293–299. [PubMed: 9568283]
44. Sun J, Blaskovich MA, Knowles D, Qian Y, Ohkanda J, Bailey RD, Hamilton AD, Sebt SM. Antitumor efficacy of a novel class of non-thiol-containing peptidomimetic inhibitors of farnesyltransferase and geranylgeranyltransferase I: combination therapy with the cytotoxic agents cisplatin, Taxol, and gemcitabine. *Cancer Res* 1999;59:4919–4926. [PubMed: 10519405]
45. MolconnZ version 4.05. Edusoft, LLC; 2003.
46. Kier, LB.; Hall, LH. *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press; 1976.
47. 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. [PubMed: 10661566]
48. Golbraikh A, Shen M, Xiao ZY, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design* 2003;17:241–253. [PubMed: 13677490]
49. Sharaf, M.; Illman, D.; Kowalski, B. *Chemometrics*. New York: John Wiley & Sons; 1986.
50. Tropsha A, Zhang WF. Identification of the descriptor pharmacophores using variable selection QSAR: Applications to database mining. *Current Pharmaceutical Design* 2001;7:599–612. [PubMed: 11375770]

51. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design* 2002;16:357–369. [PubMed: 12489684]
52. Zhang S, 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. [PubMed: 16995729]
53. Wold, SaEL. Statistical Validation of QSAR Results. In: H, vdW, editor. *Chemometrics Methods in Molecular Design*. Weinheim: VCH; 1995. p. 309-318.
54. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* 2005;45:177–182. [PubMed: 15667143]
55. Maybridge Chemical Company. Maybridge: 2004.
56. Thomson Scientific. World Drug Index Database. 2007
57. Progenitor Databases. 2006. <http://www.chemizon.com>
58. Xue L, Stahura FL, Godden JW, Bajorath J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci* 2001;41:394–401. [PubMed: 11277728]
59. Zhang FL, Diehl RE, Kohl NE, Gibbs JB, Giros B, Casey PJ, Omer CA. cDNA cloning and expression of rat and human protein geranylgeranyltransferase type-I. *J. Biol. Chem* 1994;269:3175–3180. [PubMed: 8106351]
60. Zhang FL, Moomaw JF, Casey PJ. Properties and kinetic mechanism of recombinant mammalian protein geranylgeranyltransferase type I. *J. Biol. Chem* 1994;269:23465–23470. [PubMed: 8089111]

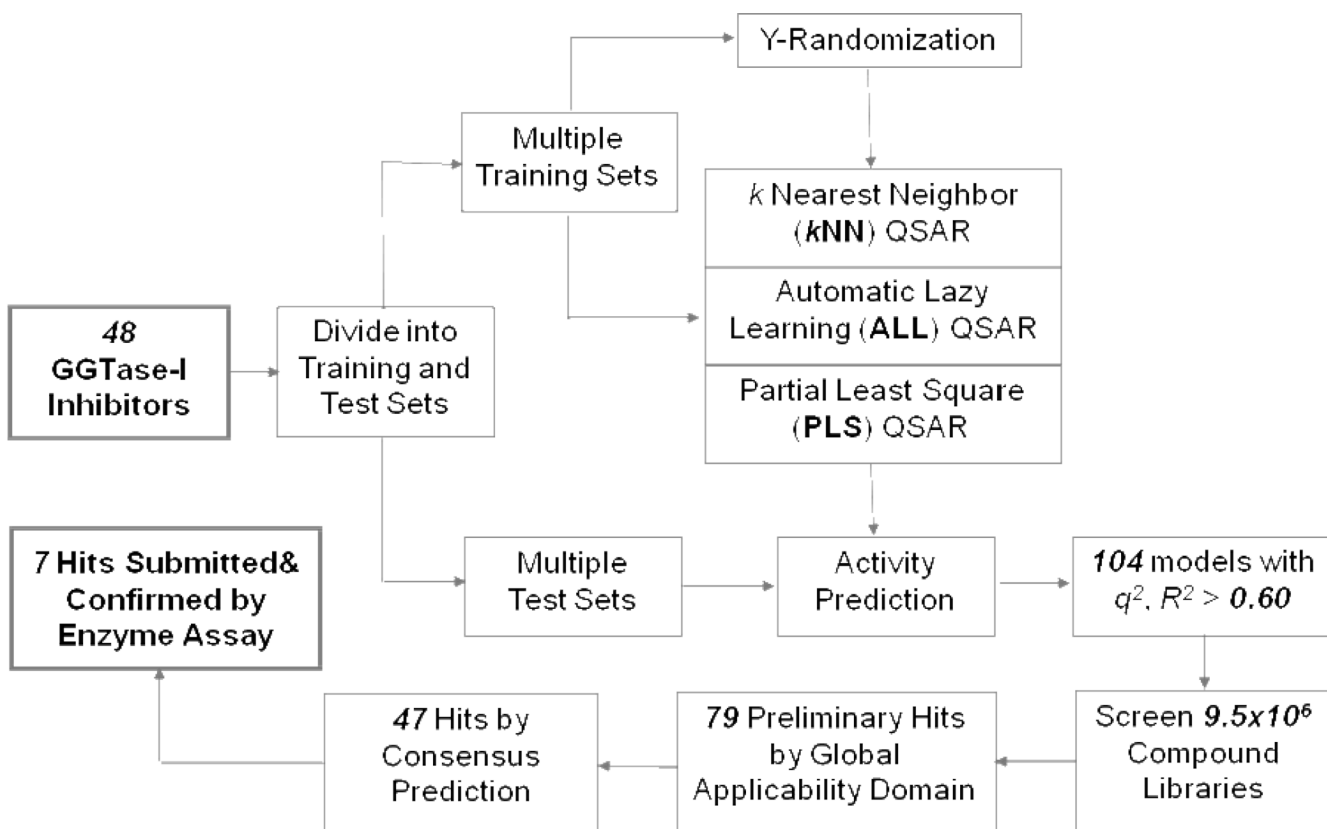


Figure 1.
The predictive QSAR modeling workflow illustrated for GGTIs.

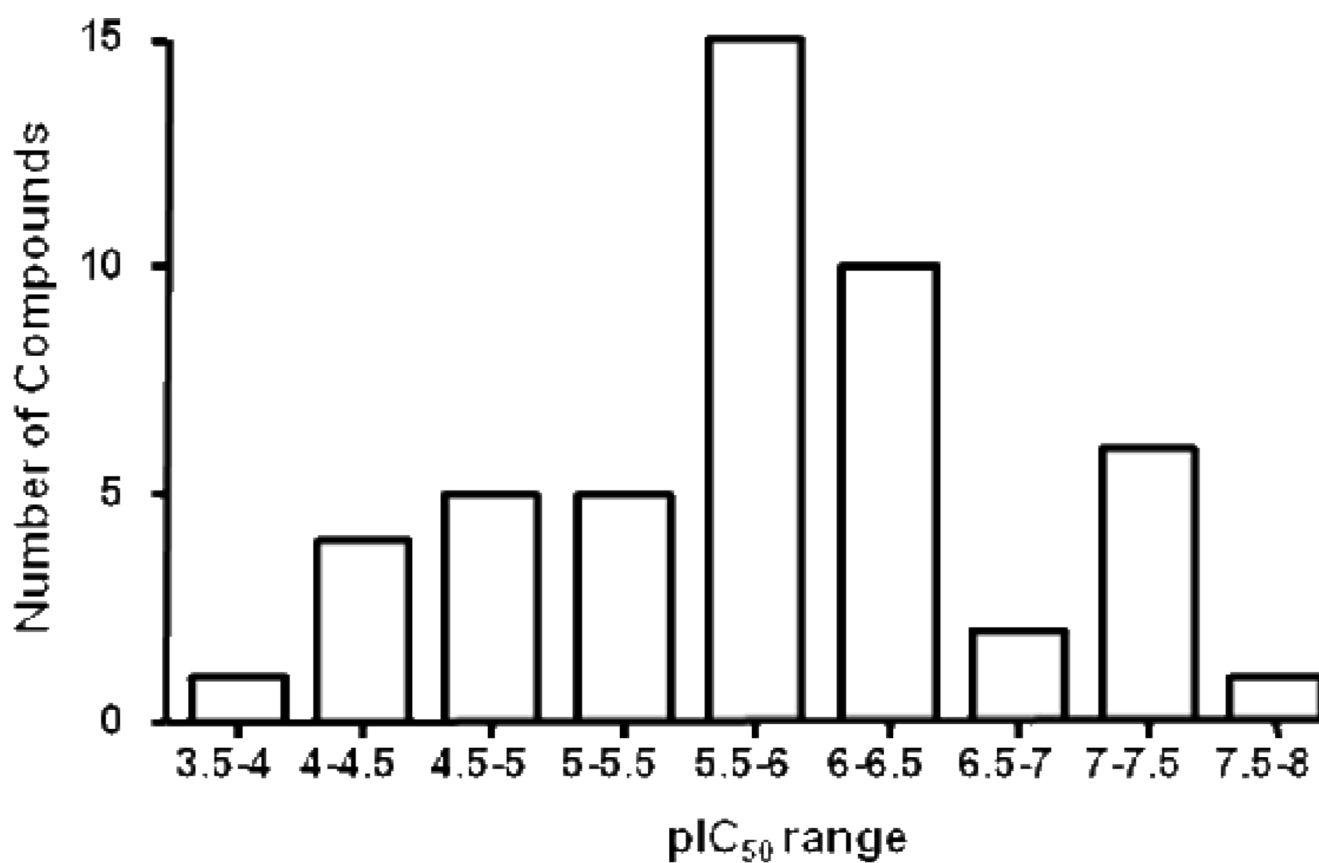


Figure 2. The activity distribution for compounds in the GGTIs dataset . 48 compounds of known *in vitro* GGTase-I inhibition activity were used for the QSAR modeling and screening studies. The IC₅₀ value was expressed in the units of molar concentration and converted to pIC₅₀ by convention.

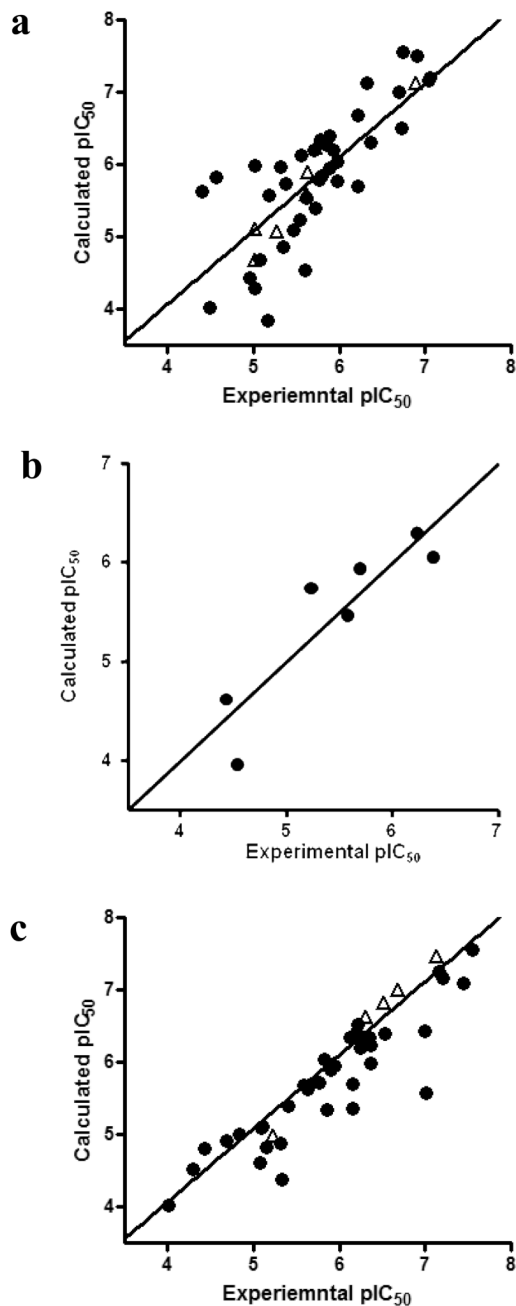
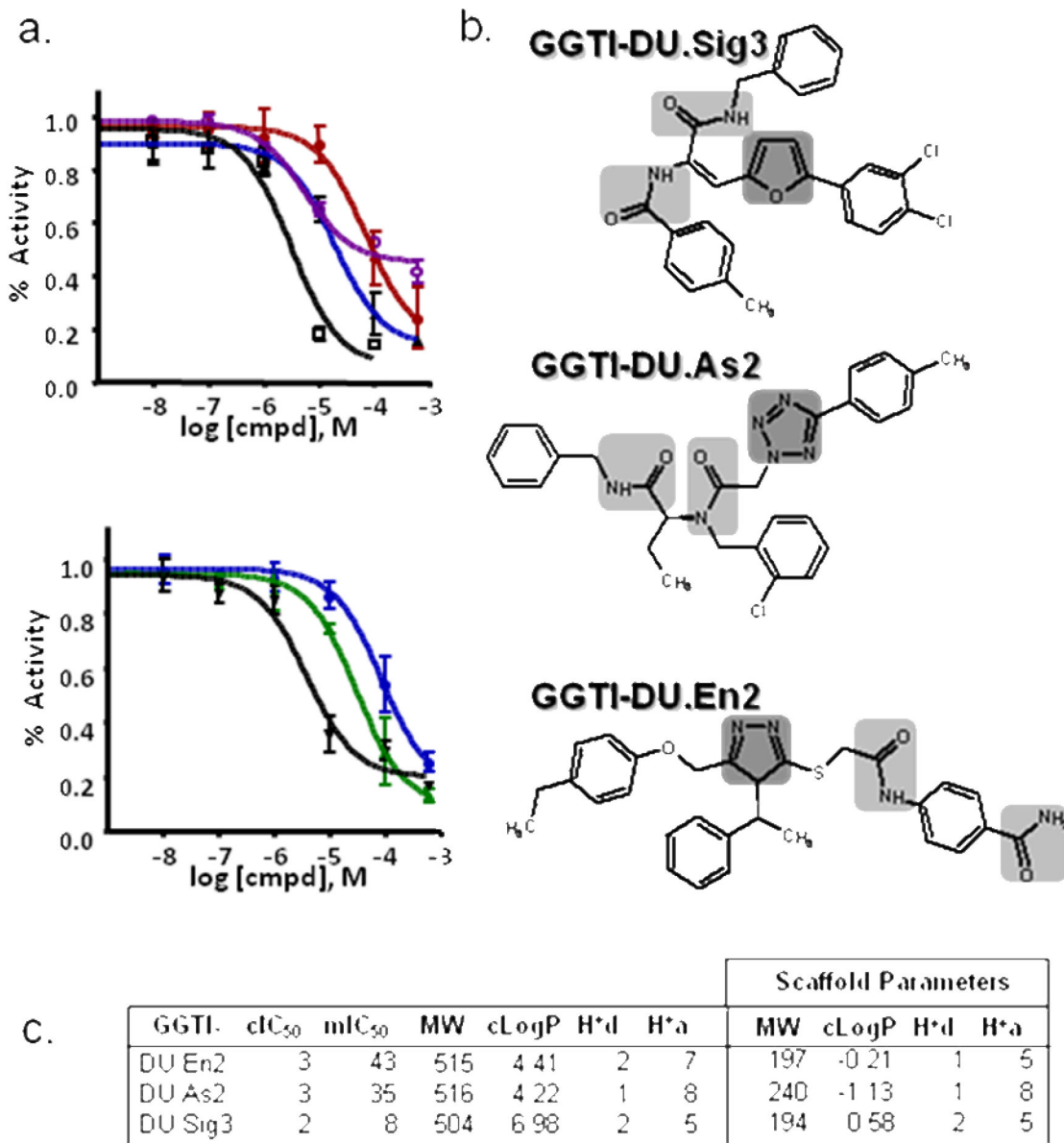


Figure 3.

Comparison of actual vs. predicted inhibition pIC_{50} values of the GGTIs dataset based on the best model developed with three methods. (a) Model generated using kNN method ($q^2 = 0.89$, $R^2 = 0.74$). The results are shown for both training set (40, solid circles) and test set compounds (8, open triangles). (b) Model generated using ALL QSAR method ($R^2 = 0.81$). The results are shown for test set compounds (7, solid circles) only. (c) Model generated using PLS method ($q^2 = 0.92$, $R^2 = 0.87$). The results are shown for both training set (43, solid circles) and test set compounds (5, open triangles).

**Figure 4.**

Experimental validations of GGTIs screening hits using GGTase-I in vitro activity assay. (a) The validation of GGTI QSAR computational hits using GGTase-I in vitro activity assay (●, GGTI-DU.As1; ▲, GGTI-DU.As2⁴²; ○, GGTI-DU.Sig1⁴²; □, GGTI-DU.Sig2⁴²; ▼, GGTI-DU.Sig3; ●, GGTI-DU.En1⁴²; ▲, GGTI-DU.En2⁴²). (b) The chemical structures of three representative confirmed hits, GGTI-DU.Sig3, GGTI-DU.As2 and GGTI-DU.En2. The novel scaffolds in the structures have been highlighted. (c) The important drug-like parameters for three representative confirmed hits. (cIC₅₀, the IC₅₀ determined by cellular assay in mM; mIC₅₀, the IC₅₀ determined by in vitro assay in μM; MW, the molecular weight; cLogP, the logP value calculated by cLogP algorithm; H^d, the number of hydrogen bond donors; H^a, the number of hydrogen bond acceptors).

Table 1
Ten best k NN QSAR models for GGTLs using MZ4.09 descriptors.

Model ID	Training Size	Test Size	k^a	Descr. Num.	q^2	R^2	R_0^2
1	40	8	1	26	0.89	0.74	0.72
2	39	9	1	24	0.89	0.63	0.61
3	39	9	1	34	0.87	0.61	0.61
4	40	8	2	28	0.85	0.66	0.66
5	39	9	1	46	0.85	0.66	0.66
6	41	7	2	24	0.85	0.61	0.61
7	40	8	2	42	0.84	0.75	0.73
8	40	8	2	32	0.84	0.68	0.55
9	40	8	2	32	0.83	0.63	0.62
10	40	8	1	44	0.82	0.85	0.83

^aThe number of nearest neighbors in the optimized k NN model.

Table 2
Ten best ALL QSAR models for GGTIs using MZ4.09 descriptors.

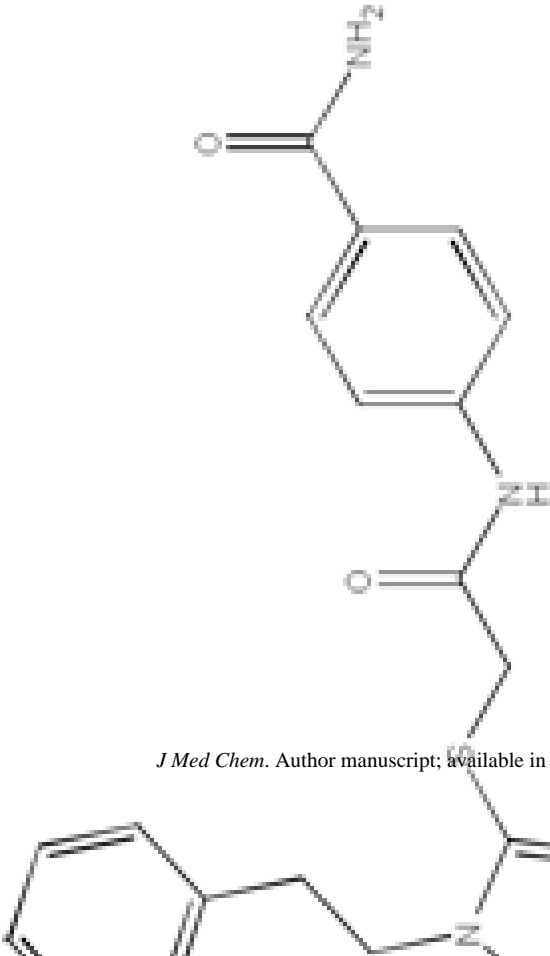
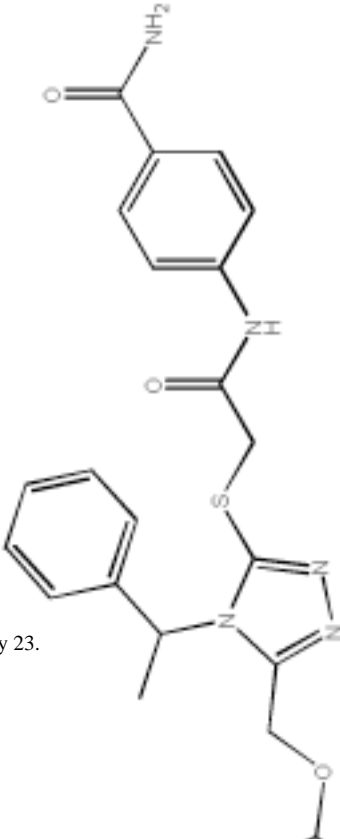
Model ID	Training Size	Test Size	KW	R^2	RMSE	R_0^2
1	40	8	0.44	0.91	0.21	0.91
2	41	7	0.42	0.81	0.34	0.81
3	42	6	0.43	0.71	0.68	0.71
4	33	15	0.64	0.69	0.58	0.69
5	35	13	0.84	0.63	0.65	0.63
6	31	17	0.75	0.63	0.61	0.63
7	42	6	0.43	0.61	0.80	0.61
8	34	14	0.79	0.59	0.66	0.59
9	33	15	0.83	0.52	0.81	0.52
10	39	9	0.91	0.49	0.85	0.49

Table 3
Five best PLS QSAR models for GGTIs using MZ4.09 descriptors.

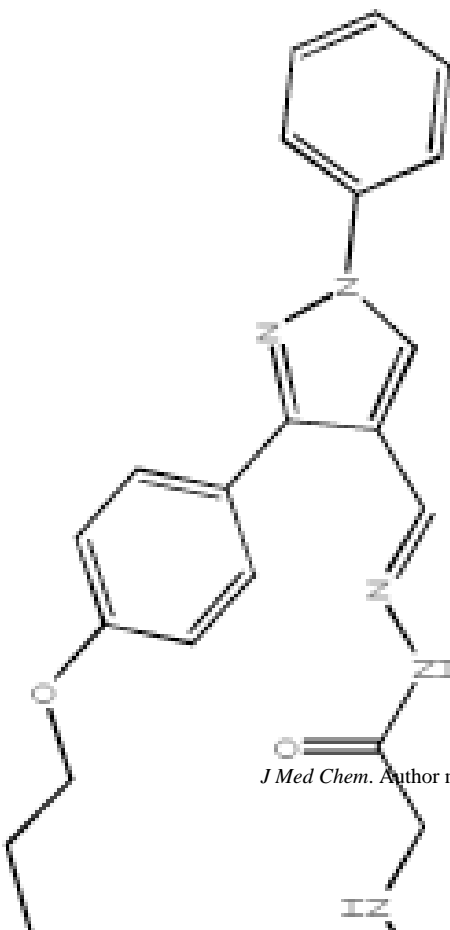
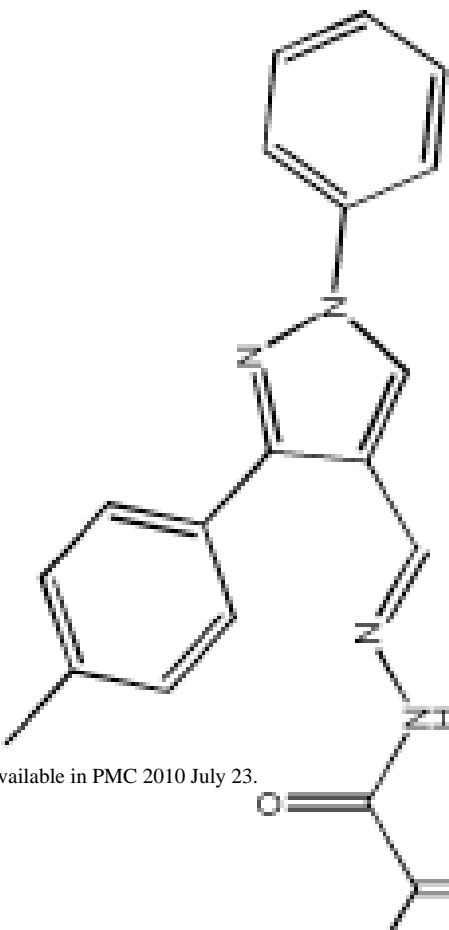
Model ID	Training Size	Test Size	q^2	R^2	RMSE
1	42	6	0.97	0.26	0.95
2	43	5	0.92	0.87	0.92
3	35	13	0.94	0.32	0.90
4	33	15	0.93	0.24	1.03
5	31	17	0.93	0.21	1.03

Table 4

inhibition efficacy (pIC₅₀) for seven confirmed screening hits using three QSAR methods. The experimentally determined pIC₅₀ values are also listed.

Structure	Compd. ID	PubChem CID	kNN QSAR	ALL QSAR	PLS QSAR	Exp.
	GGTI-DU.En1	2118978	5.31	3.37	5.56	5.56
	GGTI-DU.En2	3455185	5.57	3.69	4.60	5.56

Structure	Compd. ID	PubChem CID	kNN QSAR	ALL QSAR	PLS QSAR	Exp.
 <chem>CC(C)C1=CC=C(C=C1)N(C(=O)CCN(C(=O)CCN(C2=CC=CC=C2)C)C)C3=CN4C=NC=N43</chem>	GGTI-DU.As1	3180720	5.67	4.07	7.04	4.19
 <chem>CC(C)N1C=NC2=C1N=CN2C(=O)CCN(C(=O)CCN(C3=CC=C(C=C3)Cl)C)C4=CC=CC=C4</chem>	GGTI-DU.As2	3180738	5.55	3.37	5.14	5.56

Structure	Compd. ID	PubChem CID	kNN QSAR	ALL QSAR	PLS QSAR	Exp.
 <p>Chemical structure of GGTI-DU.Sig1: A benzimidazole core substituted with a phenyl ring at position 2, a 4-(propoxy)phenyl ring at position 5, and a 2-(2-(methylamino)acetyl)hydrazinyl group at position 6.</p>	GGTI-DU.Sig1	3311883	5.64	3.85	6.46	4.02
 <p>Chemical structure of GGTI-DU.Sig2: A benzimidazole core substituted with a phenyl ring at position 2, a 4-methylphenyl ring at position 5, and a 2-(2-(methylamino)acetyl)hydrazinyl group at position 6.</p>	GGTI-DU.Sig2	4277701	5.80	4.63	5.16	4.43

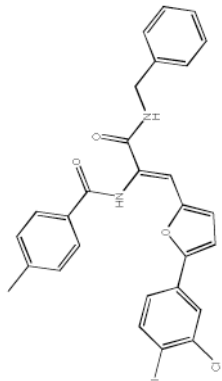
Structure	Compd. ID	PubChem CID	kNN QSAR	ALL QSAR	PLS QSAR	Exp.
	GGTI-DU.Sig3	5143450	5.65	3.78	4.75	3.84
	R^2		0.45	0.35	0.17	

Table 5

The degree of chemical similarity/dissimilarity of screening libraries and QSAR VS hits in comparison to GGTIs dataset calculated with 166 MACCS structural keys and the Tanimoto Coefficient (T_c).

Probes	T_c	Num. Hits within T_c Cutoff		
		9.5×10^6 Libraries	79 QSAR Hits	7 Confirmed Hits
48 GGTIs	0.80	8,132	30	0
	0.85	724	20	0
	0.90	22	1	0