JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

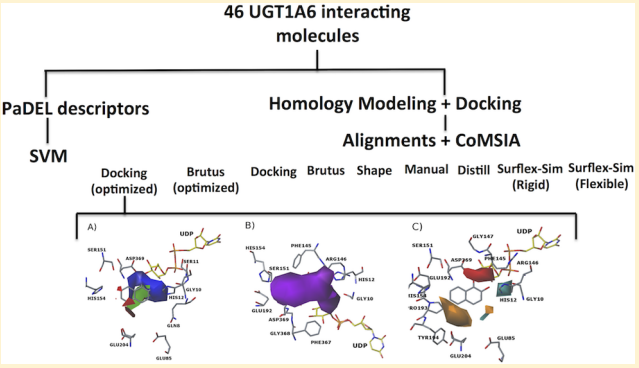
SVM Classification and CoMSIA Modeling of UGT1A6 Interacting Molecules

Leo Ghemtio,[†] Anne Soikkeli,[‡] Marjo Yliperttula,[§] Jouni Hirvonen,[‡] Moshe Finel,[†,∥] and Henri Xhaard*[,†,∥]

[†]Centre for Drug Research, [‡]Division of Pharmaceutical Technology, [§]Division of Biopharmaceutics and Pharmacokinetics, and [∥]Division of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Helsinki, 00100 Helsinki, Finland

Supporting Information

ABSTRACT: The human UDP-glucuronosyltransferase 1A6 (UGT1A6) plays important roles in elimination of many xenobiotics, including drugs. We have experimentally assessed inhibitory properties of 46 compounds toward UGT1A6 catalyzing the glucuronidation of 1-naphthol and built models for predicting compounds interactions with the enzyme. The tested compounds were divided into a training set (n = 31; evaluated by 10-fold cross-validation) and an external test set (n = 15), both of which yielded similar accuracies (80−81%) and Matthews correlation coefficients (0.61−0.63) when classified using support vector machines. Comparative molecular similarity index analysis (CoMSIA) modeling was conducted for nine methods of compound alignment. The most predictive CoMSIA model was analyzed in the light of a homology modeled UGT1A6 structure, with leave-one-out cross-validation, yielding a q^2 of 0.62 and r^2 of 0.91 on the training set and a r^2_{pred} of 0.82 on the test set. The CoMSIA contour plots highlighted the importance of H-bond donors and electrostatic field interactions, accounting for 28% and 25% contribution of the model, respectively.



INTRODUCTION

Biotransformation enzymes in the intestinal enterocytes and liver hepatocytes, such as the UDP-glucuronosyltransferases (UGTs), play important roles in protecting the body from hazardous compounds.[1] UGTs are membrane-bound enzymes of the endoplasmic reticulum that catalyze the conjugation of small and often lipophilic compounds with glucuronic acid.[2] UGT substrates include endogenous compounds, such as thyroidal hormones, neurotransmitters, and steroid hormones, as well as xenobiotics, such as drug molecules, dietary and environmental chemicals. Glucuronidation increases water solubility and stimulates elimination of the glucuronide metabolites from the cell by efflux transporters and, subsequently, from the body through bile or urine. While glucuronidation mostly neutralizes biological activity, some compounds may gain biological activity upon conjugation with glucuronic acid, for example, morphine-6-glucuronide and several acylglucuronides.[3−5]

The human genome encodes 30 different UGTs, 9 of which are inactive (pseudogenes). The active enzymes are divided into four subfamilies, named UGT1 (or UGT1A, 9 members), UGT2 (9), UGT3 (2), and UGT8 (1).[6] Developing reliable computational models for predicting individual UGTs activity is of major importance[7−14] but also difficult due to the rather broad substrate and inhibitor specificity of many of them. This study focuses on UGT1A6 that is mainly expressed in the liver, in the digestive tract, and in the airways.[15,16] The UGT1A enzymes are encoded by a single large gene that spans ∼200 kb on chromosome 2q37 and contains different (but homologous) first exons, each preceded by its own promoter, that encode the N-terminal half of the respective enzymes, and a shared set of five exons 2−5, that together encode the C-terminal half of the proteins of this subfamily.[2,6] Consequently, the N-terminal domain, about 260 amino acids long in the mature proteins (following the removal of first 24−27 amino acids "signal sequence"), is variable and assumed to harbor the binding site for the aglycone substrate, while the 246 amino acid long C-terminal domain is identical among all the UGT1As and is assumed to include the binding site for the common glucuronic acid donor, the cosubstrate UDP-glucuronic acid (UDP-GA).[2,17]

UGTs catalyze glucuronidation reactions, i.e., the transfer of the glucuronosyl group from UDPGA to aglycone substrate molecules that contain either hydroxyl, different nitrogens, or carboxylic acid functional groups, likely utilizing a serine hydrolase-like catalytic mechanism.[18,19] The resulting glucuronide is more polar and more easily excreted than the substrate molecule. Many UGTs possess the capacity to conjugate a variety of substrates, and a partial overlap in substrate specificity

Received: October 7, 2013
Published: March 3, 2014

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

across the UGT enzymes is often seen, which suits well their important role in xenobiotics detoxification.[20,21] Nevertheless, different UGTs that conjugate the same substrate molecule often do it at different rates. Substrates containing more than a single nucleophilic group may furthermore be primarily conjugated by different UGTs. For example, estradiols (17β-estradiol, the physiological estradiol, and 17α-estradiol, a synthetic analogue) have two hydroxyl groups located 9.7−12.3 Å away along the long axis of their four-membered ring system.[22] Both are glucuronidated by different UGTs with marked differences in their regioselectivity, stereoselectivity, and kinetics.[23] Most UGT enzymes have the capacity to conjugate small alcohols and phenols, and this has contributed to the view that UGTs are poorly selective. However, increasing structural complexity of substrates appears to enhance UGT enzyme selectivity due to steric, hydrophobic, and electronic interactions.[24,25]

Structurally, the human UGTs are classified as members of the GT1 family of glycosyltransferases and as such are predicted to adopt a GT-B fold.[26−30] The GT-B fold is shared among several three-dimensional (3D) structures of glycosyltransferases, including the bacterial Gtf family of enzymes that is involved in vancomycin synthesis[31−33] and plant flavonoid GTs that are involved in secondary metabolite glycosydation.[34,35] The prediction that UGTs adopt a GT-B fold was already verified for the C-terminal domain of UGT2B7 (later referred to as UGT2B7CT, residues 285−451; PDB code 2O6L).[18] The UGT2B7CT crystal structure shows a Rosman-type fold composed of a single parallel β-sheet consisting of six individual strands surrounded by seven α-helices, similarly to the fold of the C-terminal domain of other GTs.[18] UGT2B7CT is involved in UDPGA binding, and there is, despite a sequence identity of only ∼19% between the plant VvGT1 and the human UGT2B7CT, conservation of all the key amino acids that are involved in the cosubstrate UDPGA/UDP-glucose binding[18]

Much less data are available on the N-terminal domain. Nevertheless, histidine 12 of the mature protein (H37 of the premature protein, before the removal of the N-terminal "signal sequence") and aspartate 124 of the mature protein (D149 in the case of UGT1A6) in the UGT1A N-terminal domain were suggested, based on site-directed mutagenesis in several UGTs and other GTs protein combined with sequence comparisons across human UGTs, to be involved in the reaction mechanism.[2,18,36,37] They are spatially located in suitable positions to carry out the glucuronic acid transfer reaction.[38,39] Computational studies can be used to study the elusive nature of the N-terminal domain and analyze the available experimental results in terms of compounds interactions and substrate specificity of UGTs in order to be able to predict the interactions of new compounds.[40,41]

Previous work indicated that UGT1A6 is responsible for the conjugation and inactivation of popular analgesic drugs and may, therefore, have important pharmacological, toxicological, and physiological consequences.[44] UGT1A6 mainly catalyzes the glucuronidation of small and planar molecules, such as simple phenols,[20,42,43] indoles,[42] and coumarines,[43] and therefore, it is assumed to have a relatively narrow substrate-binding site in comparison to other human UGTs.[6] This quality, along with the high turnover rate that UGT1A6 often exhibits, makes it an attractive subject for trying to develop predictive in silico models. In particular, it should be possible to gather a focused data set, well balanced among active and inactive molecules, that allows studying compounds binding near the catalytic site of UGT1A6.

The interaction of a compound with UGT1A6 can be tested in vitro either directly by measuring glucuronidation rates or indirectly by detecting its effect on the glucuronidation of a probe substrate, such as 1-naphthol.[44] The indirect setup does not discriminate substrates from inhibitors, but it enables faster and broader screening. We have recently developed a fluorescence-based screening assay that is well suited for testing large sets of compounds for their effect on 1-naphthol glucuronidation by recombinant UGT1A6.[42]

In this study, we experimentally assessed the inhibitory properties of 46 compounds toward the human UGT1A6. Support vector machine (SVM) classification and 3D QSAR (CoMSIA) were then used to build predictive models of glucuronidation inhibition.

MATERIALS AND METHODS

Compounds and Biological Data. The source of the recombinant UGT1A6 is as previously described.[42] Compounds were purchased from Sigma-Aldrich Chemie, Riedelde Haën, MP Biochemicals LLC, Sekhsaria Chemicals Limited, Francis s.p.a., and TCI Europe.[42] The activity data for a set of 46 compounds were measured using an assay developed for detecting the inhibition of the 1-naphthol glucuronidation activity of recombinant human UGT1A6.[42] Some of these data were previously reported for 5,6,7,8-tetrahydro-1-napththol, 5-methylsalicylic acid, 5-bromosalicylic acid, 5-chlorosalicylic acid, 5-fluorosalicylic acid, salicylic acid, diclofenac, scopoletin, ketoprofen, and ibuprofen.[42] Briefly, compounds were first dissolved in pure DMSO at 10 mM and then diluted 1:1 with mQ-water. The assay was started by delivering 10 μL of the compound solution on the black 96-well plate. This was followed by adding 80 μL of the reaction mixture (includes recombinant UGT1A6, $MgCl_2$, and probe 1-naphthol in phosphate buffer at pH 7.4). After 10−15 min preincubation on warmed carrier at +37 °C, the reaction was started by adding 10 μL of 20 mM UDPGA solution. Immediately following UDPGA addition a starting readout was taken on Varioskan plate reader with fixed emission at 335 nm by recording the level of the excitation signal at 295−300 mM. The plate was incubated on the warmed carrier for 30 min, after which the end reading was recorded. All pipetting steps from the compound delivery step on were performed on TECAN Genesis RSP 150/8 workstation. Measurements were conducted in duplicates, each with and without UDPGA.

The biological activities are reported as % effect of probe glucuronidation rate (% Glu); the probe glucuronidation rate in the presence of 500 μM of a compound compared to the probe glucuronidation rate in the absence of the compound. As part of the CoMSIA modeling procedure, these percent effects on the probe glucuronidation rate were scaled using a procedure implemented Sybyl-X to give pGlu in negative logarithmic units.

Molecular Structures. The 3D molecular structures of all 46 compounds were generated using the compound sketch function of the Sybyl-X 1.2 software (Tripos, USA). All the structures with assigned Gasteiger−Huckel charges were first energy minimized using the standard Tripos force field (Powell method and 0.05 kcal/(mol.Å) energy gradient convergence criteria). In the case of compounds with rotable bonds (22 out of 46 compounds; 24 are rigid with one or no rotable bonds), the lowest energy conformation was picked. The data set was

Table 1. Summary of the Compounds and Respective Experimental Data Used in the Training Set of the Developed SVM and CoMSIA Models

compounds		SVM			3D-QSAR	
ID	name	% Glu_{exp}	$class_{exp}$	$class_{pred}$	$pGlu_{exp}$	COMSIA (docking) $pGlu_{pred}$
1	5-fluorosalicylic acid	45	1	1	3.21	3.25
2	5-bromosalicylic acid	23	1	1	2.77	2.67
3	5-methylsalicylic acid	49	1	1	3.28	3.31
4	5-chlorosalicylic acid	29	1	1	2.91	2.99
6	5-hydroxyisophthalic acid	47	1	1	3.24	3.24
8	2-hydroxynaphthoic acid	44	1	1	3.19	3.10
9	gentisic acid	69	2	2	3.64	3.73
11	salicylic acid	60	2	2	3.47	3.48
12	vanillic acid	52	2	1	3.33	3.47
13	3,4-dihydroxybenzoic acid	70	2	2		3.55
14	isovanillic acid	70	2	2	3.66	3.51
15	bumetanide	38	1	2	3.09	3.15
16	naphthanele-1,2-diol	1	1	1	1.30	1.49
19	1-aminonaphthalene	79	2	1	3.87	3.85
20	(+)-catechin	47	1	1	3.24	3.26
21	4-nitrophenol	64	2	2	3.55	3.51
24	3-fluorocatechol	8	1	1	2.24	2.06
27	dopamine	72	2	2	3.71	3.86
28	valproic acid	93	2	2	4.42	–
29	α-ketoglutaric acid	100	2	2	5.3	–
30	ibuprofen	84	2	2	4.02	–
31	ketoprofen	61	2	1	3.5	–
32	diclofenac	22	1	1	2.75	2.69
34	indomethacin	41	1	1	3.14	–
35	esculetin	60	2	2	3.47	3.54
36	isoscopoletin	77	2	1	3.82	3.68
37	umbelliferone	35	1	1	3.03	3.05
38	scopoletin	58	2	1	3.44	3.44
39	4-methylumbelliferone	45	1	1	3.21	3.12
41	daphnetin	2	1	1	1.61	1.51
44	2,6-dimethoxyphenol	36	1	1	3.05	3.12

Table 2. Summary of the Compounds and Respective Experimental Data of the Test Set for SVM and CoMSIA Models

compounds		SVM			3D-QSAR	
ID	name	% Glu_{exp}	$class_{exp}$	$class_{pred}$	$pGlu_{exp}$	COMSIA (docking) $pGlu_{pred}$
5	5-aminosalicylic acid	100	2	2	5.29	4.60
7	3-isopropylsalicylic acid	100	2	2	5.29	4.99
10	acetylsalicylic acid	86	2	2	4.08	–
17	5,6,7,8-tetrahydro-1-napththol	26	1	1	2.84	2.63
18	propranolol	94	2	2	4.5	5.1
22	L-thyroxin	33	1	1	2.99	3.33
23	paracetamol	86	2	2	4.08	3.67
25	4-nitrocatechol	44	1	2	3.19	2.87
26	hydrocaffeic acid	49	1	1	3.38	3.10
33	4-hydroxyindole	22	1	1	2.75	2.25
40	hesperetin	53	2	1	3.35	3.34
42	coffein	86	2	2	4.08	–
43	7-hydroxyindole	12	1	1	2.43	2.43
45	2,6-di-isopropylphenol	57	2	1	3.42	3.12
46	2-methoxy-5-nitrophenol	11	1	1	2.39	3.01

randomly divided into a training set (31 compounds, see Table 1) and a test set (15 compounds, see Table 2) to be used in both SVM modeling and CoMSIA modeling for model development and validation, respectively. The range of pGlu for the training set spans at least 3 orders of magnitude (1.30−5.29), and these values are well distributed over the entire range. Compound no. 16, naphthalene-1,2-diol, was used as a reference since it exhibited the highest probe glucuronidation inhibition.

SVM Modeling. The principle of SVM is to treat the objects belonging to different classes as points in a high-dimensional space and to find a hyperplane that best separates

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

them.[43] In this study, the 46 molecules were categorized as either Class 1, interacting (0−50% of the probe glucuronidation rate in the presence of the compounds compared to that in the absence of compounds or 50−100% inhibition), and Class 2, noninteracting compounds (50−100% remaining probe glucuronidation activity). The molecules are presented in a 856-dimensional space with the help of molecular descriptors (see below). The margin of the hyperplane was defined as the distance from the separating hyperplane to the nearest data point, and SVM finds the maximum margin separating hyperplane.

PaDEL-Descriptor,[44] an open source software, was used to calculate a total of 905 molecular descriptors (770 1D/2D and 135 3D descriptors) for each molecule. The descriptors with more than 80% zero values and small standard deviation values (<3%) were removed, and the remaining 712 descriptors were further used in the SVM model development.

A subset of descriptors was then selected to be included in the SVM model based on *F*-scores, a parameter that measures the relative importance of each descriptor in categorizing the compounds into the two classes: interacting and noninteracting; descriptors having the highest *F*-scores were gradually added to the model until the accuracy of the SVM model (see below) decreased. *F*-scores were calculated with a feature selection tool included in the LIBSVM software.[45] When descriptors shared a Pearson correlation coefficient of 0.9 or higher, those with the highest *F*-score were selected.

The SVM calculations were performed with the LIBSVM software using radial basis function (RBF) kernels. The model was trained by means of 10-fold cross-validation. RBF kernel optimization was performed with grid search algorithm, which is a learning algorithm that performs an exhaustive searching through a specified subset of the hyperparameter space to obtain a model with the best cross-validation accuracy.[45]

The prediction power of the SVM model was evaluated based on accuracy (eq 1), sensitivity (eq 2), and specificity (eq 3). We also calculated the Matthews correlation coefficient (MCC) (eq 4). The value of MCC varies between −1 (complete disagreement between prediction of classes and observation) and +1 (perfect prediction), while 0 indicates a prediction no better than random.[46]

$$\text{accuracy}(\%) = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{sensitivity}(\%) = \frac{TP}{TP + FN} \quad (2)$$

$$\text{specificity}(\%) = \frac{TN}{TN + FP} \quad (3)$$

$$\text{MCC}(\%) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (4)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

3D Modeling of UGT1A6. In order to predict the interaction of our compounds with UGT1A6 in the absence of a crystal structure, we constructed a homology model based on distantly related glycosidases. Homology modeling was performed using the Discovery Studio 3.5 platform (Accelrys Inc., San Diego, CA). Briefly, the complete amino acid sequence (532 amino acids) of UGT1A6 (access code P19224) was first downloaded from Uniprot (www.uniprot.org). A PSI-BLAST search at PDB_nr95 (nonredundant sequence database constructed from the PDB) was employed to identify the most suitable templates for homology modeling (Supporting Information S.1).[47]

The crystal structure of the C-terminal domain of UGT2B7 (PDB code 2O6L) was found, not surprisingly, to be the best template for the C-terminal domain of UGT1A6. For the N-terminal domain, four templates, PDB codes: 2PQ6, 2C1X, 3HBF, and 2ACV, with detectable sequence similarities (sequence identities in the 17−22% range) and reasonable E-value ($<4.3 \times 10^{-55}$) were selected. These four templates are all flavonoid glycosyltransferase from plants: three are from the barrel clover *Medicago truncatula*, i.e., the UGT78G1 complexed with the flavonoid myricetin (3HBF);[48] UGT71G1 (2ACV);[35] and UGT85H2 (2PQ6);[38] and one is the red grape *Vitis vinifera* enzyme UDP-glucose:flavonoid 3-*O*-glycosyltransferase VvGT1 (2C1X).[34]

A structure-based multiple sequence alignment of the templates was then constructed to align together the amino acid sequences of the five template structures (one is only the C-terminal and four are complete). Simultaneously, a sequence profile was created for the modeled sequence of UGT1A6 by collecting and aligning additional related amino sequence from Uniprot.[49] A multiple sequence alignment was then constructed between the sequence profile of UGT1A6 and the structure profile of the templates using the profile alignment method Align123.[50] This alignment, together with the five template structures, was used to build a set of 3D models using Modeler through Discovery Studio 3.5. The coordinates of the cocrystallized cosubstrate (UDP; corrected to UDP-glucose and to act as a reference of the Myricetyn ligand) were then copied from 3HBF into the model structures (superimposed on the template). Terminal residues (449 − 464) of the input sequence nonaligned with any of the templates were removed. This portion was suggested by Laakkonen et al. (2010) to be an envelope helix between the C-terminal domain and the transmembrane helix, which was not needed for our docking perspectives.[40] Five 3D models were finally evaluated using the verify protein (Profiles-3D) protocol, which assesses the compatibility of the 3D structure of a protein model with the sequence of residues it contains, and a discrete optimized protein energy (DOPE) score calculated for each model. The best model was selected for docking simulations.

Docking Simulations. In order to predict binding modes for the 46 compounds and to generate bioactive conformations usable for CoMSIA modeling, we conducted docking simulations using the GLIDE (v5.7) program from Schrödinger, Inc. (Portland, Oregon).[51] The protein was prepared using Maestro (v9.2) (Schrödinger modeling package), applying the OPLS-2005 force field. The compounds were prepared for docking using the LigPrep (v2.5) module.

For docking simulations, the standard precision (SP) parameters of GLIDE were selected with a flexible ligand and rigid receptor routine. Two distance constraints (2.5−4 Å) were added when relevant, a procedure similar to the procedure implemented by Wu et al.[52] First, a distance constraint was set between any OH group connected to an aromatic ring and any nitrogen side chain atoms from the catalytic residue His12. Eleven molecules lack any OH groups, and for these the constraint cannot be created. The second constraint was set between any two atoms of the compounds and the cosubstrate

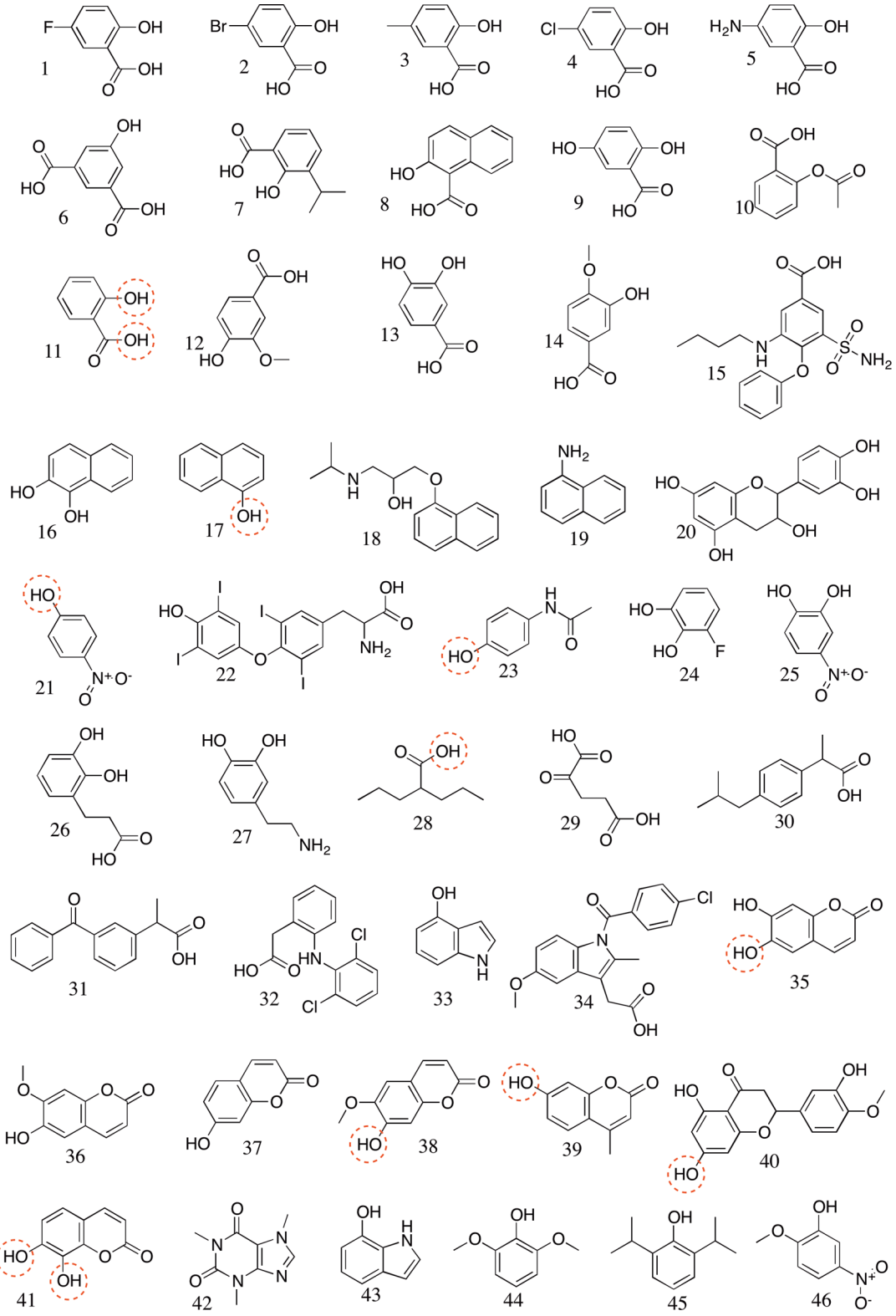


Figure 1. The set of 46 compounds experimentally tested in this study. A red circle indicates glucuronidation sites experimentally identified in the literature.

UDP-glucose. Twenty docking poses were calculated for each molecule, the docking being terminated if two consecutive solutions were within a RMSD of 0.5 Å. After docking, strains rescore was performed with Macromodel to identify ligands with too much strain; ligands with more than 4 kcal/mol energy differences between the docked and unbound conformation received penalties. Seven compounds were eliminated as part of this procedure, i.e., no satisfactory docking solution was found for them. Satisfaction of the distance constraints described above was used together with Glide score for ranking the poses and selecting the five best poses per compound. Eventually, the five best docking poses

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

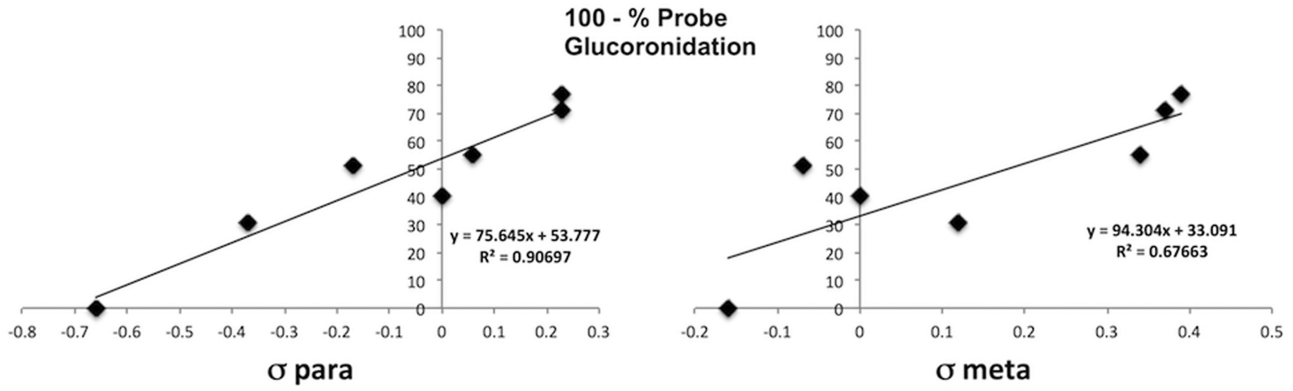


Figure 2. Plot of the Hammett σ para (left) and meta (right) substituents of seven 2-hydroxybenzoic acid core: compounds 2−6, 10, and 12.

for each of the 39 compounds were selected as bioactive conformations for CoMSIA alignment and modeling.

Compound Alignment Rules and Molecular Fields. The CoMSIA method is used to correlate and predict the biological activities of a set of aligned compounds as well as to provide a visual guide in three-dimension about the location of property fields important or detrimental for activity.[53] These property fields are based on similarity indices of aligned compound that are calculated and evaluated by a PLS analysis.

The CoMSIA models were developed using the same data set as the SVM model, 31 compounds as the training set and externally validated with 15 compounds as the test set (Tables 1 and 2). All compound alignments used in this study are provided as Supporting Information S.2 (available in .sdf format).

CoMSIA is entirely dependent on a 3D alignment of the compounds. In this study, we compared nine compound alignments constructed using automated and manual ligand based overlay methods as well as structure-based methods (docking). The seven ligand-based alignments were constructed as follows: a manual alignment that superimposed best the catecholic core of the compounds; BRUTUS, was used to generate an alignment using the rigid starting conformations that optimize shape and electrostatic/steric charge distributions;[54,55] Distill, a part of the Sybyl-X package, was used to produce a molecular scaffold (maximum common substructure) alignment; ShaEP was used to produce a shape and electrostatic potential alignment, again a rigid alignment;[56] and Surflex-Sim Rigid/Flexible, a part of the Sybyl-X package, was used to construct a surface-based morphological similarity alignment. Two docking-based alignments were also generated using Glide. In addition, the alignments using BRUTUS (five poses possibilities for each compound) and docking (five best poses generated by docking for each compound) were assessed in an optimization loop: CoMSIA models were built considering several superimposition possibilities per compounds, and later only the ones with best prediction by the model were kept.

For 10 of the compounds, compound nos. 11, 17, 21, 23, 28, 35, and 38−41 (Figure 1), the glucuronidation site has been experimentally identified as an oxygen atom of either a hydroxyl or a carboxyl group.[20,21,57−63] These glucuronidated atoms should all be located in the same region of the binding site, near the UDPGA where the catalytic mechanism takes place, interacting with the catalytic His 12.[38−40,64] Superimposition of the glucuronidated atoms in these 10 compounds was used as a posteriori validation criteria as part of the docking and compound alignment procedures.

The aligned training set of 31 molecules was positioned inside grid boxes with grid spacing value of 2.0 (default distance) in all Cartesian directions and extended beyond the molecular dimensions by 4.0 Å in all directions. Several other grid spacing values, in the range of 1−3, were also tested without any improvement in the results. CoMSIA fields were calculated using the QSAR module of the Sybyl-X with column filtering of 4.0 kcal/mol. The steric, electrostatic, hydrophobic, and H-bond donor and acceptor were evaluated using an sp^3-carbon probe atom of 1.0 Å radius. The value of the attenuation factor was set to 0.3.

CoMSIA Models: Training and Internal Validation. Sybyl-X was used to build CoMSIA models for each of the nine compound alignments mentioned above. The leave-one-out cross-validation (LOOCV) was performed to determine the optimum number of components leading to the highest cross-validated coefficient q^2 (eq 5) and to the lowest standard error of prediction, of which both are necessary but not sufficient conditions to indicate the robustness and predictive ability of models.[65] After that, non-cross-validation was performed to derive the final PLS regression models with the explained variance r^2 and F-ratio.

$$q^2 = 1 - \frac{\sum_y (Y_{\text{predict}} - Y_{\text{experimental}})^2}{\sum_y (Y_{\text{experimental}} - Y_{\text{mean}})^2} \quad (5)$$

where, Y_{predict} is the predicted scaled probe glucuronidation rate (pGlu), $Y_{\text{experimental}}$ is the experimentally determined scaled probe reaction rate (pGlu), and Y_{mean} is the mean pGlu predicted value of the training set.

Finally, CoMSIA results were graphically represented by field contour maps, where the coefficients were generated using the field type "StDev*Coeff". Favored and disfavored fields were fixed at 80% and 20%, respectively.

CoMSIA Models: External Validation. In order to test the actual predictive ability of the trained CoMSIA models, the probe glucuronidation rates (pGlu) of the external validation set (15 compounds) were predicted using the same CoMSIA calculation parameters as those used to generate the models.[65] The non-cross-validated analyses were used to display the coefficient contour maps. The actual versus predicted pGlu of the test compounds were fitted by linear regression and the r^2_{pred} (eq 6) and F ratio were determined.

$$r^2_{\text{pred}} = 1 - \frac{\sum_y (Y_{\text{predict(test)}} - Y_{\text{experimental(test)}})^2}{\sum_y (Y_{\text{experimental(test)}} - Y_{\text{mean(training)}})^2} \quad (6)$$

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

Table 3. Prediction Performance of SVM Models

	original model					accuracy of randomized models (%); training and test set random selection				
	accuracy (%)	specificity (%)	sensitivity (%)	MCC	Y = scrambling (n = 10)	1	2	3	4	5
training set (n = 31)	81	67	93	0.63	<34	77	81	85	79	80
external test set (n = 15)	80	75	86	0.61	<26	75	78	88	81	78

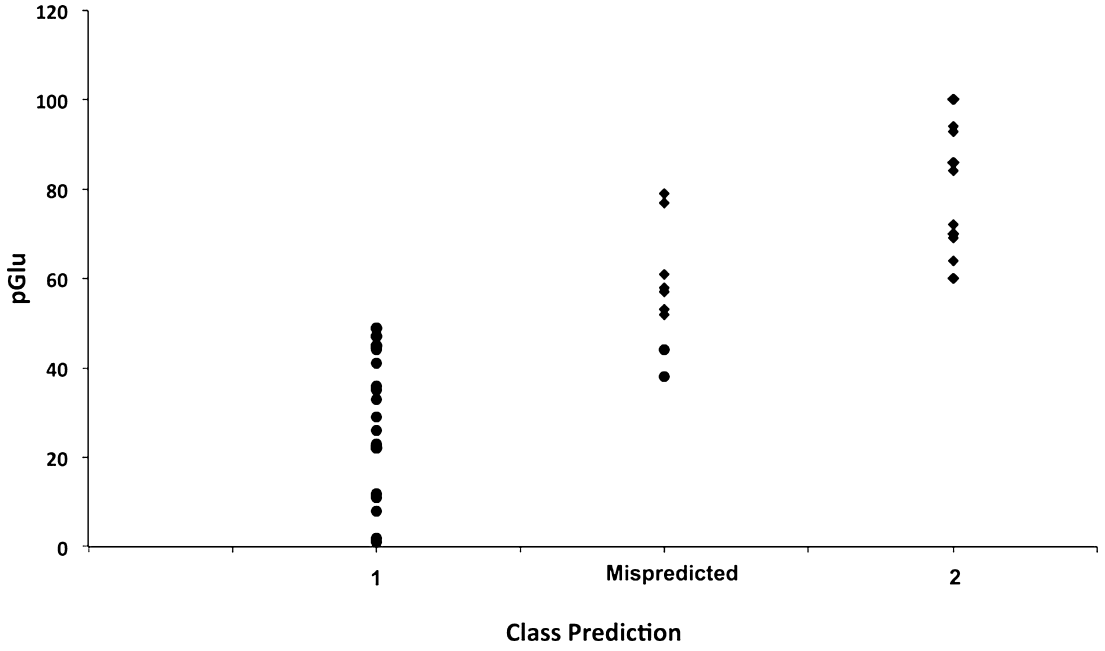


Figure 3. Prediction accuracy of the SVM models is affected by the artificial separation of continuous data into two classes. First column, compounds of class 1 (inhibitors) well predicted; second column, compounds of class 1 or class 2 predicted in the wrong class; and third column, compounds of class 2 (non inhibitors) well predicted.

where $Y_{predict(test)}$ and $Y_{experimental(test)}$ indicate the predicted and experimental probe glucuronidation rates (pGlu) of the test set compounds, respectively, and $Y_{mean(training)}$ indicates the mean probe glucuronidation rate (pGlu) of the training set.

All calculations were performed on a Linux workstation Intel x86-64 processors equipped with two Intel Core Duo CPU (3.16 GHZ) with 7 GB total RAM.

RESULTS AND DISCUSSION

The signal sequence is 26 amino acids long in the case of UGT1A6 and due to this in the "literature numbering", 26 is often added to the amino acid number in the mature sequence. In this paper, the numbering used is based on the mature sequence for UGT1A6.

Binding to UGT1A6: Simple Observations from Chemical Structures. In this manuscript we have tested a set of 46 compounds using a glucuronidation assay. Our data set covers well the known chemical space of UGT1A6, which is focused on small-size compounds. The chemical structures are shown in Figure 1, and the experimental activities are shown in Tables 1 and 2. The assay was an indirect report of binding to UGT1A6 and, therefore, provides no information as to whether the compounds that inhibit probe glucuronidation are themselves glucuronidated. Overall, 23 molecules were found to inhibit the probe glucuronidation by more than 50%, i.e., those are the active molecules characterized by low % of probe glucuronidation (% Glu <50%).

The data set is constituted of molecules that are relatively homogeneous in their chemistry: 23 molecules are built around a poly substituted phenyl ring and inhibit probe glucuronidation by 8−100% (average 57.3% ± 25.8 std). In addition, 18 molecules are built around a two-ring system, in the majority of cases two phenyl rings, with probe glucuronidation between 1 and 97% (average 44.4% ± 28.5 std), three molecules with a more complex ring system of at least three (mostly) aromatic rings separated by a linker (% Glu 41−53%) as well as two inactive molecules based on an aliphatic skeleton (% Glu 93−100%).

Our data set does not present a direct relation of either of the −COO or −OH substituents with inhibitory activity. The main substituents decorating the compounds are carboxylate, with at least one present in 23 molecules, and hydroxyl, with at least one found in 37 molecules. Sixteen compounds have both carboxylate and hydroxyl substituents. The glucuronidation activities of the benzoic acid derivatives were in the 22−100% range (average 59.4% ± 25 std), whereas the activities of molecules of same size (one or two ring systems) that do not possess a carboxylate were in the 1−94% range (average 46.8% ± 28.7 std). Similarly for −OH substituted molecules, there was no specific link with activities since they inhibited glucuronidation in the 1−100% range (average 50.1% ± 27.5 std), whereas molecules without −OH groups exhibited 38−93% glucuronidation (average 65.6% ± 26 std). Glucuronidation activities of molecules substituted by both carboxylate and −OH were in the 23−100% range (average 58.8% ± 24.6 std) and for molecules with neither carboxylate nor −OH it was in the 79−86% range (average 82.5% ± 5 std).

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

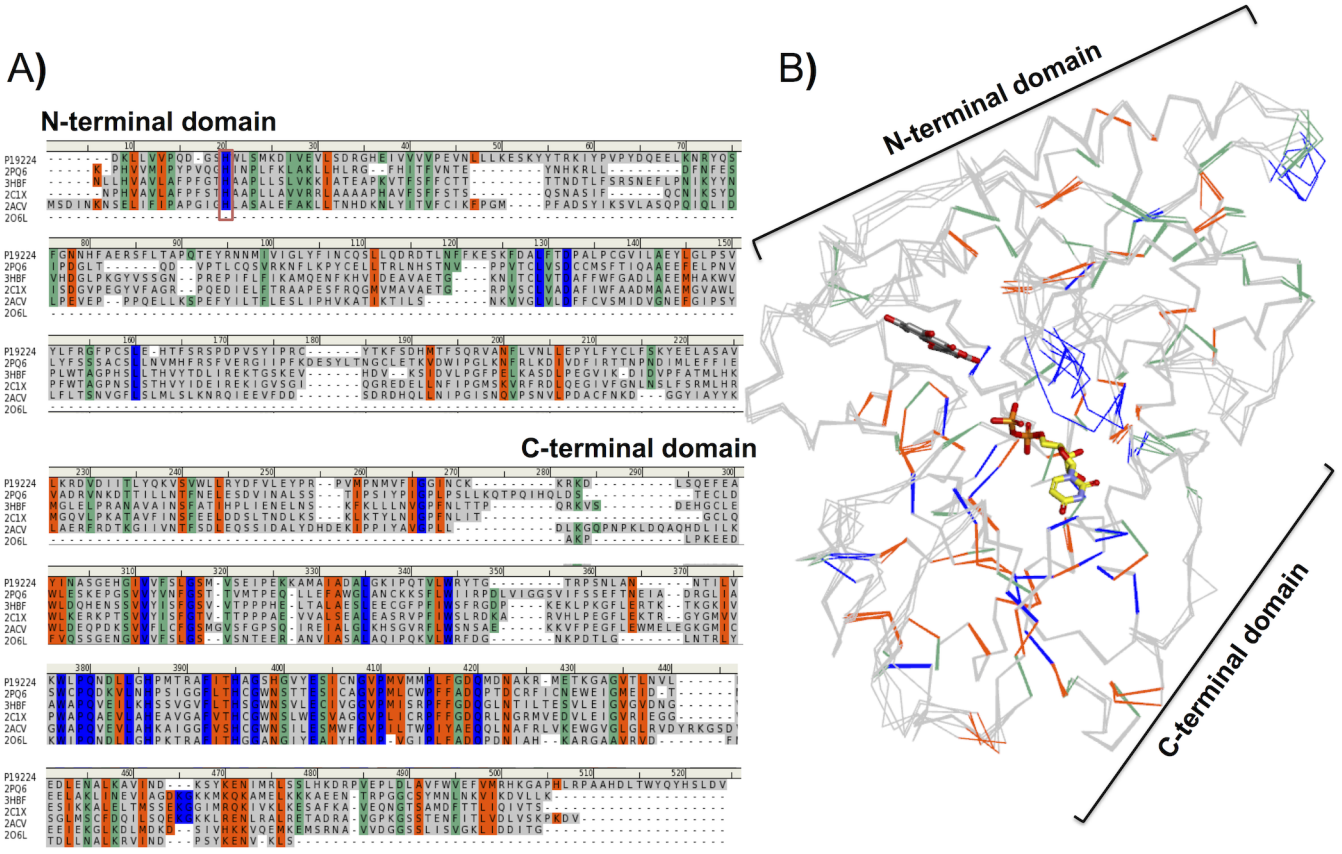


Figure 4. Homology model of UGT1A6. Conserved amino acid (blue) and similar amino acids (green and orange) indicate areas of structural conservation and therefore a higher confidence in the 3D model. They are located mostly in the C-terminal domain. (A) UGT1A6 query sequence P19224 and selected templates alignment. A red box indicates the catalytic residue His 12. (B) Superposition of four templates (PDB code: 2PQ6, 2C1X, 3HBF, and 2ACV for N-terminal) the crystal structure 2O6L of C-terminal domain of UGT2B7 for C-terminal part of the model construction. Naphthalene-1,2-diol (compound 16) is placed in the binding catalytic binding pocket near the cosubstrate UDP-glucose.

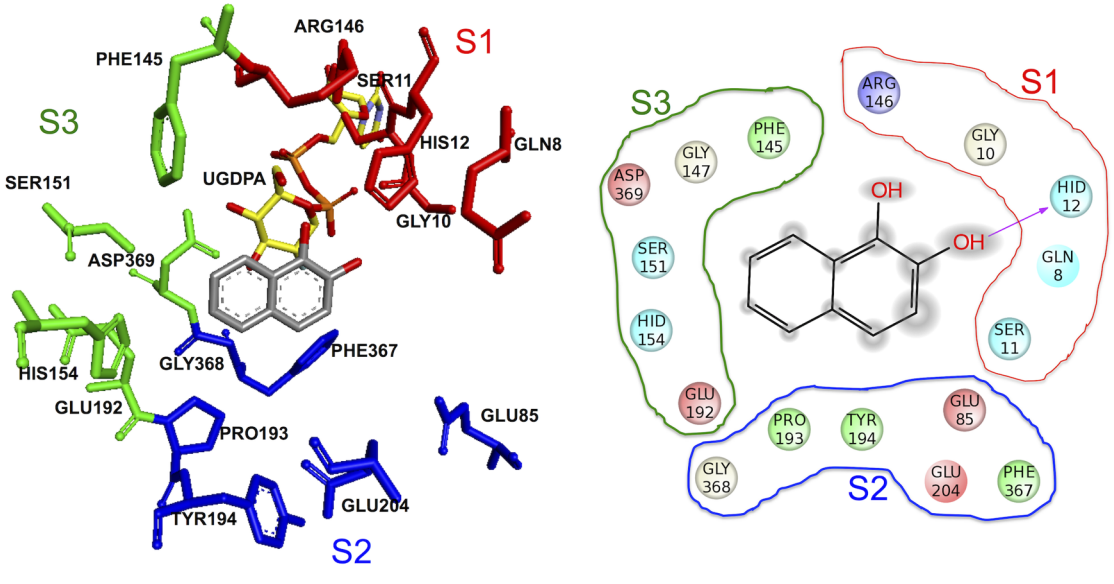


Figure 5. The catalytic binding pocket identified in the UGT1A6 homology model. (A) Amino acids located near the catalytic binding pocket. For the purpose of description three areas are defined (S1, red; S2, green; S3, blue). (B) A 2D schematic view of the substrate binding amino acids.

Multiple effects may determine inhibition, and in order to derive more accurate results, it would be better to compare a homogeneous series of compounds.[24,25] A subset of seven molecules that is based on 2-hydroxybenzoic acid core, with a substituent at the 5-meta position, could be extracted and compared independently (Figure 2). The substituents are fluorine (compounds 2), bromine (3), methyl (4), chlorine (5), amine (6), hydroxyl (10), and hydrogen (12). Ethell and co-workers previously showed that the V_{max} (but not K_m) of 10 phenolic compounds could be correlated with the substituent

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

Table 4. Comparison of CoMSIA Results with Different Methods Used to Align the Training Data Set[a]

	optimized pose selection		no selection						
	docking (1)	BRUTUS (2)	docking (3)	BRUTUS (4)	manual (5)	Distill (6)	ShaEP (7)	Surflex-Sim Rigid (8)	Sulflex-Sim Flexible (9)
q^2	0.62	0.59	0.25	0.37	0.52	0.32	0.16	0.25	0.18
r^2	0.91	0.79	0.62	0.65	0.76	0.46	0.36	0.66	0.73
NC	3	6	2	5	4	6	1	2	3
gluc	yes	yes	yes	no	yes	no	no	no	no

[a] r^2 is the q^2 without crossvalidation; NC, number of components from PLS analysis; gluc, indicates that the glucoronidated atoms (known experimentally for 10 molecules) have been verified to occupy the same region of the 3D space.

Table 5. Effect of using *Y*-Randomization and Five Different Randomly Selected Training/Test Sets on the q^2 Values of the Best CoMSIA Models

	q^2		q^2 (training and test set random selection)				
	original model	Y = scrambling, n = 10	1	2	3	4	5
CoMSIA docking (optimized) (1)	0.62	<0.23	0.58	0.64	0.71	0.59	0.58
CoMSIA BRUTUS (optimized) (2)	0.59	<0.12	0.51	0.58	0.61	0.55	0.52
CoMSIA manual (5)	0.52	<0.09	0.56	0.51	0..52	0.55	0.57

size in the case of UGT1A6 and suggested that this meant that the more bulky substrates may be accepted into the active site with higher affinity but may not be glucuronidated.[59] In our case, we did not found such correlation: correlation coefficients between volume and activity are close to 0 for both our subset of seven compounds and the complete data set of 46 molecules. Plotting the Hammett σ meta constant, taken from benzoic acid (as an approximation of 2-hydroxybenzoic acid) leads however to an interesting linear plot with a regression coefficients of 0.67. The regression coefficient of the σ para constants is even better, 0.90. Thus, inductive and resonance electronic effects may be favorable for binding to the UGT1A6. This type of result has been reported by QSAR studies for other enzymes, for example, catechol-*O*-methyl transferases, but not for UGTs.[66]

Classification of Active Compounds using SVM. We developed an SVM model predicting whether the compounds will bind to UGT1A6, which should also provide information as to whether they may be glucuronidated. This type of crude classification model is also useful to extract numerical descriptors that characterize molecular requirements for compounds to be glucuronidated or not.

The SVM models were developed using 31 compounds as a training set and an externally validated test set with 15 compounds (Tables 1 and 2). Compounds of the training and test sets were divided into two classes, interacting and noninteracting, based on their effect on the 1-naphthol glucuronidation activity of the human UGT1A6. The SVM model that was developed from the training set showed an overall statistically significant performance, with 81% accuracy and 0.63 MCC (Table 3). Its performance in terms of specificity and sensitivity was also reasonably good, 67% and 93%, respectively. These values denote the performance of the model in categorizing compounds to whether they interact with UGT1A6 or not.

In order to evaluate the robustness of our SVM model, we rebuilt five models, five times randomly distributing compounds to the training (n = 31) and test (n = 15) sets. The results show that the division of compounds into training and test sets does not affect the accuracy of the models (Table 3). Of particular importance is the finding that the pool of descriptors with the highest *F*-scores in these alternate models is the same as for the initial model. Second, the response permutation test (*Y*-randomization), which evaluates the effect of randomizing the response variable (activities), showed a significant reduction of the model performance with q^2 and r^2 to <0.26 and 0.34, respectively (Table 3).[66] This indicates that a chance correlation between response and descriptors is unlikely.

Of the 15 compounds in the test data set, the model was able to classify 12 correctly (Acc 80% and MCC 0.61) but misclassified three (Table 3). If we examine the activities of the mispredicted compounds, we see that misprediction occurred for compounds close to the borderline between the interacting and noninteracting groups, namely 50% inhibition (Figure 3). This is an artifact inherent in the artificial partition of continuous data into two classes. Using well separated training classes would certainly improve the accuracy of the prediction for the training set, but only since all difficult cases have been removed. For externally tested compounds whose activities are near the "border", there are no reasons as to why the predictive power of the model would be improved.

Even so, only one of the seven external test set compounds belonging to class 1, the most interesting class in terms of inhibitors and potential substrates for UGT1A6, was misclassified. This is highlighted by high sensitivity (true positive rate) in comparison to specificity (true negative rate) (Table 3). The specificities were higher than sensitivities in both the training and test sets, suggesting that the model is able to classify more reliably the interacting compounds in class 1 than the noninteracting compounds in the class 2.

In terms of important descriptors, five descriptors were needed to optimize the prediction performance of the model. These 2D descriptors were the χ path cluster (SPC-5 for simple path cluster, order 5 and SPC-6 for order 6) that identifies the various possible fragments of the molecules, the ATSm4 autocorrelation descriptor that is weighted by scaled atomic mass, the molecular distance edge (MDEC-11 for all primary carbons and MDEC-33 for all tertiary carbons), and the moment of inertia MOMI-YZ that calculates the moment of inertia (MOMI) values to characterize the mass distribution of a molecule along the ratios of *Y* and *Z* axes (*YZ*). These finding reveal that the constitutional, topological, and electronic properties are sufficient to characterize compounds interaction properties with the human UGT1A6.

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

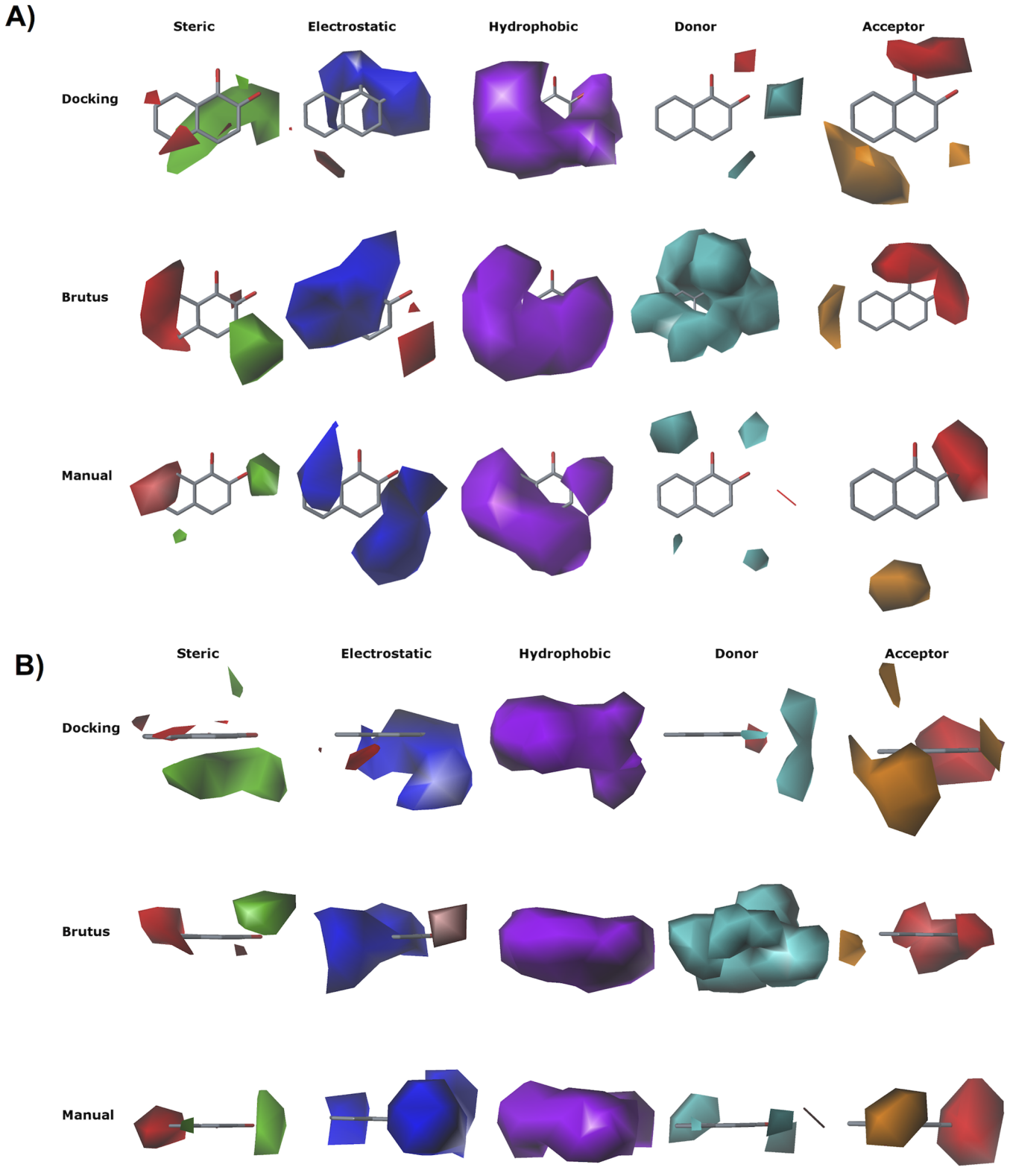


Figure 6. CoMSIA fields of the three best compounds alignment methods. Green, blue, purple, cyan, and orange indicate regions favorable for UGT1A6 activity, whereas red indicates contours not favorable for activity. Compound 16 is placed as a reference, as in Figure 5. (A) In a view above the aromatic plane. (B) View rotated by 90°.

The performance of our results is difficult to compare to that of previous studies since a different assay system, i.e., binding, was used as well as often different UGTs among the 19 known isoforms. In comparison with previous works on UGT1A6, our results largely agree with the work of Sorich et al. (2003) for predicting glucuronidation.[12] They reported that SVM predicted 81% of their test set (accuracies in the range of 78−83%) and in comparison only 66% for a linear QSAR method, PLSDA.[12] Other predictions of activity have used the K_m of glucuronidation to approximate binding, however for

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

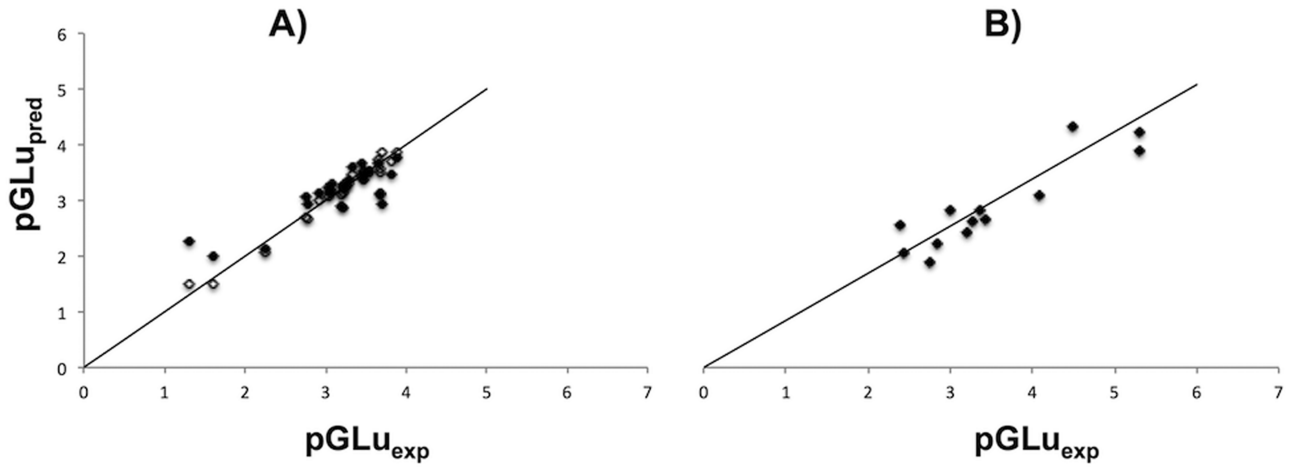


Figure 7. Scatter plots of the experimental probe glucuronidation rate values ($pGlu_{exp}$) versus predicted probe glucuronidation rate values ($pGlu_{pred}$) for the CoMSIA models. (A) Predictions of training set compounds after LOOCV (plain dots) or without (empty dots). (B) Predictions of test set compounds (plain dots).

UGT1A6 this led to a data set limited to only 10 molecules since not enough substrates could be collected.[59] Predictive models have further aimed at identifying the sites of glucuronidation based on pharmacophores or on an estimate of the likehood of a given polar atom to be glucuronidated.[9,11] Other nonlinear predictions have used for example the Volsurf descriptors to predict the K_m of UGT1A10 with good externally validated predictivity $r^2 = 0.827$ and $q^2 = 0.774$.[8]

Although SVM models may be useful to predict compounds binding to the UGT1A6, the molecular descriptors are difficult to translate into directly useful chemical information. For this reason, we used CoMSIA modeling in conjunction with homology modeling to gain more practical information about the molecular mechanism(s) of substrate interaction(s) with the UGT1A6.

Homology Model. Overall, sequence analysis combined with available structural data on related enzymes (Figure 4) showed that UGT1A6 is formed by two domains, the N-terminal domain (residues 27 to 282) that predominantly interacts with the aglycone substrate, and the C-terminal domain (residues 283 to 489) that presents most of the contact points with the cosubstrate UDP-glucose (in the related plant proteins). Each domain contains a Rossmann fold motif comprised of six parallel β-sheets separated by α-helices. A deep cleft at the interface between the two domains forms the active site. UGT1A6, like all mammalian UGTs, differs from plant and bacterial UGTs in having a hydrophobic C-terminal region, which is thought to fold as a single trans-membrane α helix that crosses the membrane, and a short C-terminal end that is exposed to the aqueous medium on the other side of the membrane.[40] The hydrophobic domain is not necessary for the function of the plant proteins, and it was not included in the present model.

In this study, we selected a single 3D UGT1A6 model for docking from a pool of five models constructed using Discovery Studio (Supporting Information S.3, all the five models are available as .mol2 files). The C-terminus fragment of UGT1A6 was modeled based on the C-terminal fraction of UGT1B9 (PDB code: 2O6L), probably leading to a good estimate due to the high (~55%) shared percentage sequence identity. The N-terminus domain of UGT1A6 was based on four plant glycosilases as templates (PDB code: 2PQ6, 2C1X, 3HBF, and 2ACV; see Methods) that share among themselves a pairwise root-mean-square deviation values from 0.5 to 2.24 Å for all the superimposed atoms. The ID percentage was much lower (~15−19%) across the UGT1A6 N-terminus and the templates, resulting in a challenging model. The key catalytic residues or residues that bind the substrate are, however, conserved in the N-terminal region; therefore the model could be useful for getting a rough estimate of the amino acids exposed to the binding region.

The selection of the model was based on DOPE assessment score and Profile-3D scores computed by the discovery studio (Supporting Information S.4 and S.5). All these five models present a similar spatial arrangement of highly conserved residues in the UGT1A6 templates (Supporting Information S.3). Moreover, superposition of the catalytic residue His12 among the selected model and 4 templates of the N-terminal domain with the presence of compounds Myricetyn and Uridine-5′-diphosphate (Supporting Information S.6) revealed a good superposition of His12 and localization of the three main components needed for UGT1A6 activity (His12, Myricetyn, and UDPGA). Overall, our final homology model of UGT1A6 shows remarkable similarity to other UGT models available in the literature (Laakkonen and Finel (1A1), 2010; Baojian Wu et al. (1A9), 2012), which is expected due to the identical or similar 3D folds of the templates used for all these models (Supporting Information S.7).

Docking Studies. The retained model was used to dock all the 46 compounds that were examined in this study, using the GLIDE program (v5.7) with default standard-precision (SP) parameters and two distances constraints. That allows interactions between the OH of the phenol ring and the catalytic residue, His12, as well as close contacts between the compounds and cosubstrate UDPGA. Only 39 compounds could be docked with a reasonably low binding energy (docking poses are presented in Supporting Information S.8) In order to check whether or not the docking program provides reasonable solutions, Myricetin was docked to UGT1A6, where it adopted a similar binding mode as in the X-ray structure 3HBF (Supporting Information S.9), supporting our docking strategy. The GLIDE scores from docked compounds, however, did not correlate ($r^2 = 0.15$) with their activities.

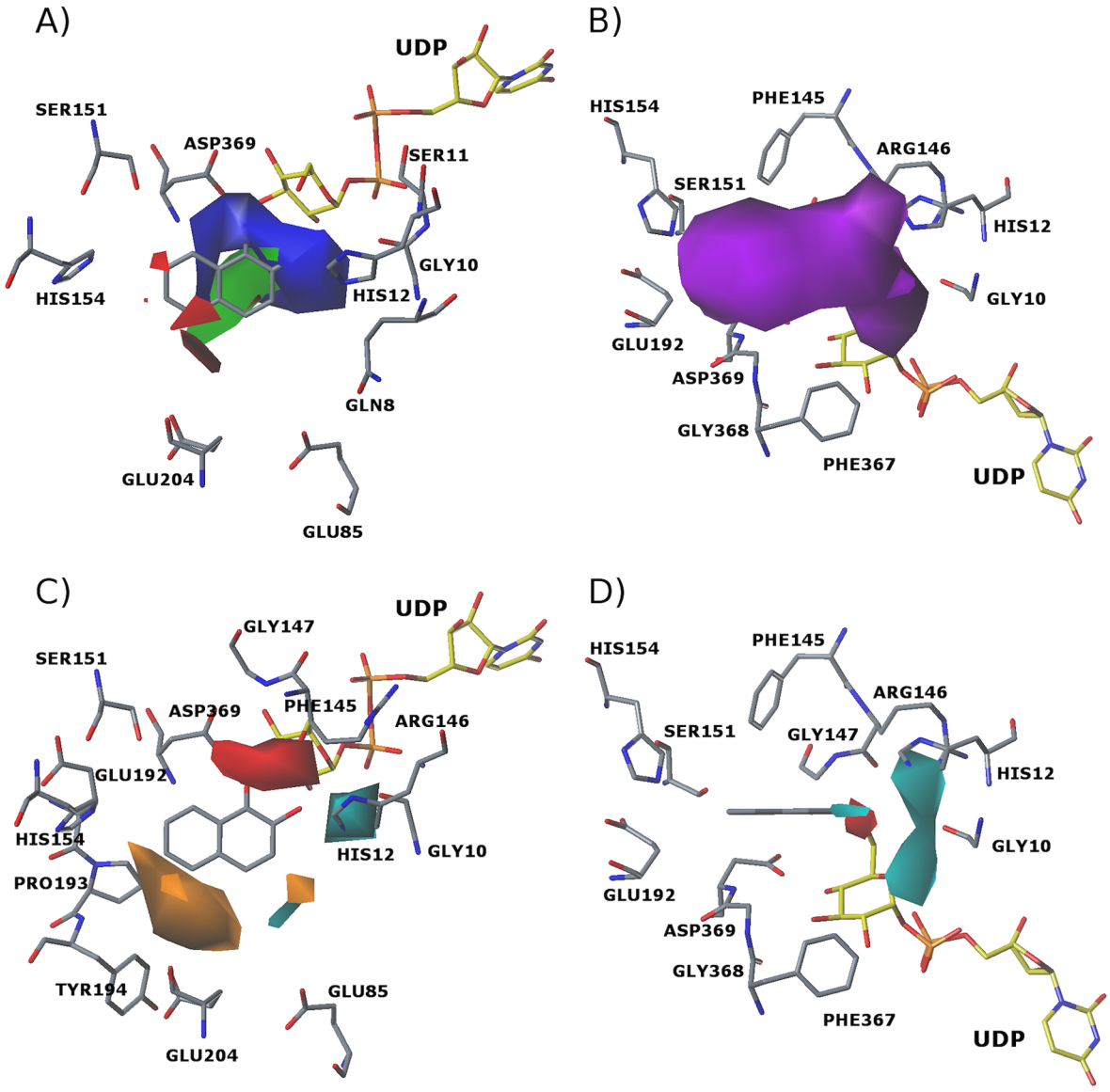


Figure 8. Contour maps of CoMSIA fields placed in the homology model of UGT1A6. Docked compound 16, naphthalene-1,2-diol is shown as a reference. Color coding as in Figure 6: red contours indicate regions not favorable the interaction. Favorable contours indicated as follows: (A) green/blue, steric/electrostatic map; (B) purple, hydrophobic map; (C,D) cyan, H-bond donor map; orange, H-bond acceptor map. In (D) the view is rotated by 90°.

The amino acid residues of the protein within 5 Å of the bound Myricetin constitute the catalytic pocket of our complex (Figure 5). The best pose of the most active compounds of the phenolic series (naphthalene-1,2-diol) were selected on the basis of GLIDE scores as shown in Figure 5. Among the listed amino acids, Phe 367, Gly 368, and Asp 369 are located in the C-terminal domain and can be placed with confidence. The compound naphthalene-1,2-diol (compounds 16) has two catecholic hydroxyl groups. In its top pose, the meta −OH group is at H-bonding distance of the side chain of His12 and of the main-chain nitrogen of Gly10. The presumably catalytic His12 was used as one of the constraint for docking. Nearby, the side chain of Gln8 and Ser11 points away. Arg 146 is located in the vicinity of His 12. The para −OH catechol hydroxyl points toward the $-CH_2OH$ of the sugar of the cosubstrate UDPGA. We will later refer to this pocket as S1 for the purpose of discussing the CoMSIA fields.

The remaining part of the substrate is mostly hydrophobic and composed of two aromatic six-membered rings that interact with the hydrophobic pocket S2, composed of Pro193, Tyr 194, Phe 367, and Gly 368. Glu 85 and Glu 204 are located further away, but according to our homology model, they face the substrate binding of UGT1A6.

Opposite the catecholic OH group along the substrate short axis is the S3 pocket. S3 faces the hydrophobic ring of the reference compound and is characterized by an intramolecular interaction between Ser 151, His 154, and Glu 192 of UGT1A6. The S3 region also includes the aromatic stacking of Phe 145 and the side chain carboxyl group of ASP 369 that makes hydrogen bonds to the 3′- and 4′- hydroxyl groups of the UDPGA sugar.

A maximum of 20 poses were generated for each 39 compounds. A large variability in poses localization is seen in the expected binding cavity (Supporting Information S.8), especially for the smallest compounds. These docking poses

dx.doi.org/10.1021/ci400577a | J. Chem. Inf. Model. 2014, 54, 1011−1026

thus suggest ambiguity in modeling. The small size of many compounds provides a larger available space and makes it difficult to find a unique pose within the active site as opposed to the relatively bulky compounds.

CoMSIA Modeling: Comparison of Alignments. We constructed and compared CoMSIA models based on nine alignments that were built using seven fully independent methods: (1,3) structure-based docking poses; (2,4) shape and electrostatic/steric charge distributions (BRUTUS), (5) manual superimposition of similar functional groups (−OH) and molecular scaffold (benzene ring); (6) substructure alignment (Distill); (7) shape and electrostatic potential (ShaEP); and (8,9) surface-based morphological similarity (Surflex-Sim Rigid and Flexible).

Some of the compound alignments led to predictive models (Table 4). The best predictive models were, not surprisingly, obtained for the alignments built in conjunction with a stochastic pose selection protocol, with a q^2/r^2 of 0.62/0.91 (docking) and 0.59/0.71 (BRUTUS). The best model without the pose optimization was the manual overlay (q^2/r^2 of 0.52/0.76). As for SVM models, we subjected the three best CoMSIA models to five-time randomization of the training and test sets followed by repeating the modeling procedure as well as the *Y*-scrambling test (Table 5). As a result, the models were robust toward division of compounds into training and test, and scrambling the activity data led to a drop in q^2, showing that the model is unlikely to be obtained by chance correlations.

The most predictive models, also not surprisingly, were those in which the spatial position of the ∼10 known sites of glucuronidation (circled in Figure 1) were controlled to occupy the spatial region in which the glucuronidation reaction takes place. Therefore, for these compounds the alignment should be reasonably realistic. It is not straightforward to compare compound alignments, however. A more detailed comparison of the BRUTUS- and docking-based alignments showed that they share 12 molecules in similar orientation to the reference molecules out of 26.

The shape of the molecular fields was then compared for the three best models: docking, BRUTUS, and Manual. Figure 6 reflects the different compound alignments. The favorable and unfavorable areas for steric field are localized in the same area of the 3D space for the three CoMSIA models. Hydrophobic fields have favorable effects in all of them and are localized in the same area, as seen in both standard and 90° rotation view of the three models. Electrostatic fields are mostly favorable, as seen in the three models in both standard and 90° rotation view. The H-bond donor fields have also mostly favorable effects and are localized in the same area for docking and manual models, but all around the compound in the BRUTUS alignment. The H-acceptor bonds have both favorable and unfavorable effects that are localized in the same area for the three alignments. The shape of the fields is somewhat similar for the three models, as seen both in standard and 90° rotation views.

Details are given herein for the best model, the docking-based model (Tables 1 and 2, Figure 7). The model has q^2 = 0.62 and r^2 = 0.91. The prediction of pGlu values for the compounds in the training set using docking-based model is shown in Table 1. The correlations between the calculated and experimental values of pGlu (from training and LOOCV) are shown in Figure 7A. This model was successfully validated by the 13 external test set compounds that were successfully docked (Table 2), which were not included in the model construction. The model was able to well describe the test set variance with a predictive r^2_{pred} of 0.82. The predicted activity values of the test set are listed in Table 2, and the correlation between the predictions and experimental values is represented in Figure 7B.

The contour plots of the CoMSIA analysis are presented in Figure 8, using catalytic pocket of UGT1A6 and naphthalene-1,2-diol as an example. The CoMSIA fields are located toward all the three S1−S3 regions described previously. Understanding the nature of the interactions with S1 in the model is complex. Electrostatic interactions are predominantly favorable toward S1 and account for 25% of the contribution to the model. The favorable interactions could be seen, for example, with compounds 1, 4, 8, 11, and 32 that have activities of 45, 29, 44, 60 and 22% Glu_{exp} inhibition that orients a COO− group and may favorably interact with His 12 or Arg 146. In addition, compounds 2, 16, 24, and 37 have activities of 23, 1, 8, and 35% Glu_{exp} and orient an −OH group to mainly favorably interactions with His_{12}, Asp_{369}, and the UDPGA sugar. The model shows favorable H-bond donors near S1, e.g., compounds 27 (72% Glu_{exp}) has a tertiary amine that should also form favorable H-bond interactions with S1. Favorable H-bond donors account for 28% of the model, while H-bond acceptors account for 19% of the model. Unfavorable H-bond acceptor fields near S1 may be explained by compounds 13, 14, and 15 (70, 70, and 38% Glu_{exp}, respectively) that have an −OH group in this area. Favorable H-bond acceptor interactions near S2 are also seen, mainly from compounds that have a H-bond acceptor in that area, e.g., carboxylate (compounds 9, 12, and 13) or nitro (compound 21) that have activities of 69, 52, 70, and 64% Glu_{exp}, respectively. Favorable hydrophobic properties represents 20% of the model and are mostly located toward S2 and S3. Compound 15 (38% Glu_{exp}) is a good example for such interactions. Interactions with S2 and S3 are typically with Phe 367, Tyr 194, and Phe 145. In contrast, the contributions of steric fields were mostly ineffective (6%).

CONCLUDING REMARKS

Herein, SVM and 3D-QSAR (CoMSIA) models were constructed for predicting compound interactions with UGT1A6, based on their effect on the probe glucuronidation rate. We have used a data set that covers a large part of the chemical space accessible to UGT1A6, and all the experimental results were obtained using the same method, the same instrument ,and in the same laboratory. This, as well as avoiding pooling data from the literature, is important for eliminating potential "noise" from the system, particularly since the previously published modeling studies often focused on substrates rather than on inhibitors or interacting molecules (both substrates and inhibitors).

In light of earlier findings with UGT1A6,[20,59,67] we have focused on small phenols, but the structures outside this core were fairly diverse, making it difficult to superimpose them in an unambiguous manner. CoMSIA, due to its smoother Gaussian function-based potentials and its being less sensitive to alignment, was able to handle this data better than comparative molecular field analysis (preliminary results not shown here).[68]

It has been previously reported that UGTs exhibit distinct but partly overlapping substrate specificity.[25,69] The two different types of models that were developed in this study, classification by SVM and 3D-QSAR, could serve as a useful

tool package together to develop tools for profiling compounds that bind to UGTs. Another perspective of this work will be to study experimentally the determinants of binding to UGT1A6 suggested by the combination of homology modeling and CoMSIA modeling. The next challenge will be in developing similar modeling systems for other human UGTs, from the development of useful activity assays to the analyses by the modeling systems that worked for the (presumably) relatively simple UGT1A6.

■ ASSOCIATED CONTENT

Supporting Information

Tables showing additional data: PSI-BLAST iterative search for UGT1A6 (S.1); scores of molecular models (S.4, S.5); and pairwise RMSDs (S.7). Figures showing superposition of five models (S.3) as well as views of specific details (S.6; S.8; S.9). All compound alignments of the training set that were used as the basis for the CoMSIA analysis are provided in the .sdf format (S.2). The homology model structures and docking poses are also provided (S.3, S.8). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: henri.xhaard@helsinki.fi. Tel: +358 9 191 59537. Fax: +359 9 191 59 725.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The University of Helsinki and the Drug Discovery and Chemical Biology, Biocenter Finland network, the Academy of Finland (project no. 12600101) and the Sigrid Juselius Foundation (to M.F.) supported this study. The CSC−IT Center for Science Ltd. is thanked for organizing computational resources. Tripos Ltd. is thanked for providing us with an Academic licence to the Sybyl-X package during the revision of this manuscript.

■ REFERENCES

(1) Suzuki, H.; Sugiyama, Y. Role of metabolic enzymes and efflux transporters in the absorption of drugs from the small intestine. *Eur. J. Pharm. Sci.* **2000**, *12*, 3−12.

(2) Mackenzie, P. I.; Gardner-Stephen, D. A.; Miners, J. O. 4.20 − UDP-glucuronosyltransferases. In *Comprehensive Toxicology*, 2nd ed.; McQueen, C. A., Ed.; Elsevier: Oxford, UK, 2010; pp 413−434.

(3) Osborne, R.; Joel, S.; Trew, D.; Slevin, M. Morphine and Metabolite Behavior after Different Routes of Morphine Administration - Demonstration of the Importance of the Active Metabolite Morphine-6-Glucuronide. *Clin. Pharmacol. Ther.* **1990**, *47*, 12−19.

(4) Tang, W. The metabolism of diclofenac enzymology and toxicology perspectives. *Curr. Drug Metab.* **2003**, *4*, 319−29.

(5) Williams, A. M.; Worrall, S.; de Jersey, J.; Dickinson, R. G. Studies on the reactivity of acyl glucuronides III. Glucuronide-derived adducts of valproic acid and plasma protein and anti-adduct antibodies in humans. *Biochem. Pharmacol.* **1992**, *43*, 745−55.

(6) Mackenzie, P. I.; Bock, K. W.; Burchell, B.; Guillemette, C.; Ikushiro, S.; Iyanagi, T.; Miners, J. O.; Owens, I. S.; Nebert, D. W. Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet. Genomics* **2005**, *15*, 677−685.

(7) Bichlmaier, I.; Finel, M.; Sippl, W.; Yli-Kauhaluoma, J. Stereochemical and steric control of the UDP-glucuronosyltransferase-catalyzed conjugation reaction: a rational approach for the design of inhibitors for the human UGT2B7. *ChemMedChem* **2007**, *2*, 1730−40.

(8) Dong, D.; Wu, B. J. In silico modeling of UDP-glucuronosyltransferase 1A10 substrates using the volsurf approach. *J. Pharm. Sci.* **2012**, *101*, 3531−3539.

(9) Sorich, M. J.; McKinnon, R. A.; Miners, J. O.; Smith, P. A. The importance of local chemical structure for chemical metabolism by human uridine 5′-diphosphate-glucuronosyltransferase. *J. Chem. Inf. Model.* **2006**, *46*, 2692−7.

(10) Sorich, M. J.; McKinnon, R. A.; Miners, J. O.; Winkler, D. A.; Smith, P. A. Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J. Med. Chem.* **2004**, *47*, 5311−7.

(11) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Smith, P. A. Multiple pharmacophores for the investigation of human UDP-glucuronosyltransferase isoform substrate selectivity. *Mol. Pharmacol.* **2004**, *65*, 301−8.

(12) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019−24.

(13) Sorich, M. J.; Smith, P. A.; McKinnon, R. A.; Miners, J. O. Pharmacophore and quantitative structure activity relationship modelling of UDP-glucuronosyltransferase 1A1 (UGT1A1) substrates. *Pharmacogenetics* **2002**, *12*, 635−45.

(14) Wu, B.; Morrow, J. K.; Singh, R.; Zhang, S.; Hu, M. Three-dimensional quantitative structure-activity relationship studies on UGT1A9-mediated 3-O-glucuronidation of natural flavonols using a pharmacophore-based comparative molecular field analysis model. *J. Pharmacol. Exp. Ther.* **2011**, *336*, 403−13.

(15) Ohno, S.; Nakajin, S. Determination of mRNA Expression of Human UDP-Glucuronosyltransferases and Application for Localization in Various Human Tissues by Real-Time Reverse Transcriptase-Polymerase Chain Reaction. *Drug Metab. Dispos.* **2009**, *37*, 32−40.

(16) Court, M. H.; Zhang, X.; Ding, X.; Yee, K. K.; Hesse, L. M.; Finel, M. Quantitative distribution of mRNAs encoding the 19 human UDP-glucuronosyltransferase enzymes in 26 adult and 3 fetal tissues. *Xenobiotica* **2012**, *42*, 266−77.

(17) Gong, Q. H.; Cho, J. W.; Huang, T.; Potter, C.; Gholami, N.; Basu, N. K.; Kubota, S.; Carvalho, S.; Pennington, M. W.; Owens, I. S.; Popescu, N. C. Thirteen UDPglucuronosyltransferase genes are encoded at the human UGT1 gene complex locus. *Pharmacogenetics* **2001**, *11*, 357−68.

(18) Miley, M. J.; Zielinska, A. K.; Keenan, J. E.; Bratton, S. M.; Radominska-Pandya, A.; Redinbo, M. R. Crystal structure of the cofactor-binding domain of the human phase II drug-metabolism enzyme UDP-glucuronosyltransferase 2B7. *J. Mol. Biol.* **2007**, *369*, 498−511.

(19) King, C. D.; Rios, G. R.; Green, M. D.; Tephly, T. R. UDP-glucuronosyltransferases. *Curr. Drug Metab.* **2000**, *1*, 143−61.

(20) Ebner, T.; Burchell, B. Substrate specificities of two stably expressed human liver UDP-glucuronosyltransferases of the UGT1 gene family. *Drug Metab. Dispos.* **1993**, *21*, 50−5.

(21) Luukkanen, L.; Taskinen, J.; Kurkela, M.; Kostiainen, R.; Hirvonen, J.; Finel, M. Kinetic characterization of the 1A subfamily of recombinant human UDP-glucuronosyltransferases. *Drug Metab. Dispos.* **2005**, *33*, 1017−1026.

(22) Zhurova, E. A.; Zhurov, V. V.; Chopra, D.; Stash, A. I.; Pinkerton, A. A. 17Alpha-estradiol x 1/2 H2O: super-structural ordering, electronic properties, chemical bonding, and biological activity in comparison with other estrogens. *J. Am. Chem. Soc.* **2009**, *131*, 17260−9.

(23) Itaaho, K.; Mackenzie, P. I.; Ikushiro, S.; Miners, J. O.; Finel, M. The configuration of the 17-hydroxy group variably influences the glucuronidation of beta-estradiol and epiestradiol by human UDP-glucuronosyltransferases. *Drug Metab. Dispos.* **2008**, *36*, 2307−15.

(24) Bowalgaha, K.; Elliot, D. J.; Mackenzie, P. I.; Knights, K. M.; Miners, J. O. The glucuronidation of Delta4−3-Keto C19- and C21-hydroxysteroids by human liver microsomal and recombinant UDP-glucuronosyltransferases (UGTs): 6alpha- and 21-hydroxyprogesterone are selective substrates for UGT2B7. *Drug Metab. Dispos.* **2007**, *35*, 363−70.
(25) Miners, J. O.; Knights, K. M.; Houston, J. B.; Mackenzie, P. I. In vitro-in vivo correlation for drugs and other compounds eliminated by glucuronidation in humans: pitfalls and promises. *Biochem. Pharmacol.* **2006**, *71*, 1531−9.
(26) Breton, C.; Snajdrova, L.; Jeanneau, C.; Koca, J.; Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **2006**, *16*, 29R−37R.
(27) Campbell, J. A.; Davies, G. J.; Bulone, V.; Henrissat, B. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem. J.* **1997**, *326*, 929−939.
(28) Coutinho, P. M.; Deleury, E.; Davies, G. J.; Henrissat, B. An evolving hierarchical family classification for glycosyltransferases. *J. Mol. Biol.* **2003**, *328*, 307−317.
(29) Hu, Y. N.; Walker, S. Remarkable structural similarities between diverse glycosyltransferases. *Chem. Biol.* **2002**, *9*, 1287−1296.
(30) Morera, S.; Lariviere, L.; Kurzeck, J.; Aschke-Sonnenborn, U.; Freemont, P. S.; Janin, J.; Ruger, W. High resolution crystal structures of T4 phage beta-glucosyltransferase: Induced fit and effect of substrate and metal binding. *J. Mol. Biol.* **2001**, *311*, 569−577.
(31) Mulichak, A. M.; Losey, H. C.; Lu, W.; Wawrzak, Z.; Walsh, C. T.; Garavito, R. M. Structure of the TDP-epi-vancosaminyltransferase GtfA from the chloroeremomycin biosynthetic pathway. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9238−43.
(32) Mulichak, A. M.; Losey, H. C.; Walsh, C. T.; Garavito, R. M. Structure of the UDP-Glucosyltransferase GtfB that modifies the heptapeptide aglycone in the biosynthesis of vancomycin group antibiotics. *Structure* **2001**, *9*, 547−557.
(33) Mulichak, A. M.; Lu, W.; Losey, H. C.; Walsh, C. T.; Garavito, R. M. Crystal structure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: Interactions with acceptor and nucleotide ligands. *Biochemistry* **2004**, *43*, 5170−5180.
(34) Offen, W.; Martinez-Fleites, C.; Yang, M.; Kiat-Lim, E.; Davis, B. G.; Tarling, C. A.; Ford, C. M.; Bowles, D. J.; Davies, G. J. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO J.* **2006**, *25*, 1396−405.
(35) Shao, H.; He, X.; Achnine, L.; Blount, J. W.; Dixon, R. A.; Wang, X. Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from Medicago truncatula. *Plant Cell* **2005**, *17*, 3141−54.
(36) Breton, C.; Fournel-Gigleux, S.; Palcic, M. M. Recent structures, evolution and mechanisms of glycosyltransferases. *Curr. Opin. Struct. Biol.* **2012**, *22*, 540−549.
(37) Patana, A. S.; Kurkela, M.; Finel, M.; Goldman, A. Mutation analysis in UGT1A9 suggests a relationship between substrate and catalytic residues in UDP-glucuronosyltransferases. *Protein Eng., Des. Sel.* **2008**, *21*, 537−43.
(38) Li, L.; Modolo, L. V.; Escamilla-Trevino, L. L.; Achnine, L.; Dixon, R. A.; Wang, X. Crystal structure of Medicago truncatula UGT85H2–insights into the structural basis of a multifunctional (iso)flavonoid glycosyltransferase. *J. Mol. Biol.* **2007**, *370*, 951−63.
(39) Li, Y.; Baldauf, S.; Lim, E. K.; Bowles, D. J. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of Arabidopsis thaliana. *J. Biol. Chem.* **2001**, *276*, 4338−43.
(40) Laakkonen, L.; Finel, M. A Molecular Model of the Human UDP-Glucuronosyltransferase 1A1, Its Membrane Orientation, and the Interactions between Different Parts of the Enzyme. *Mol. Pharmacol.* **2010**, *77*, 931−939.
(41) Lewis, B. C.; Mackenzie, P. I.; Miners, J. O. Homodimerization of UDP-glucuronosyltransferase 2B7 (UGT2B7) and identification of a putative dimerization domain by protein homology modeling. *Biochem. Pharmacol.* **2011**, *82*, 2016−2023.
(42) Soikkeli, A.; Kurkela, M.; Hirvonen, J.; Yliperttula, M.; Finel, M. Fluorescence-based high-throughput screening assay for drug interactions with UGT1A6. *Assay Drug Dev. Technol.* **2011**, *9*, 496−502.
(43) Vapnik, V. N. An overview of statistical learning theory. *Neural Networks, IEEE Trans.* **1999**, *10*, 988−999.
(44) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−74.
(45) Chang, C. C.; Lin, C. J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, 2.
(46) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−24.
(47) Schaffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V.; Altschul, S. F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **2001**, *29*, 2994−3005.
(48) Modolo, L. V.; Li, L.; Pan, H.; Blount, J. W.; Dixon, R. A.; Wang, X. Crystal structures of glycosyltransferase UGT78G1 reveal the molecular basis for glycosylation and deglycosylation of (iso)-flavonoids. *J. Mol. Biol.* **2009**, *392*, 1292−302.
(49) The Universal Protein Resource (UniProt) **2009**. *Nucleic Acids Res. 2009*, 37, D169-74.
(50) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.
(51) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(52) Wu, B.; Wang, X.; Zhang, S.; Hu, M. Accurate prediction of glucuronidation of structurally diverse phenolics by human UGT1A9 using combined experimental and in silico approaches. *Pharm. Res.* **2012**, *29*, 1544−61.
(53) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−46.
(54) Ronkko, T.; Tervo, A. J.; Parkkinen, J.; Poso, A. BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 227−36.
(55) Tervo, A. J.; Ronkko, T.; Nyronen, T. H.; Poso, A. BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. *J. Med. Chem.* **2005**, *48*, 4076−86.
(56) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inform. Model.* **2009**, *49*, 492−502.
(57) Brand, W.; Boersma, M. G.; Bik, H.; Hoek-van den Hil, E. F.; Vervoort, J.; Barron, D.; Meinl, W.; Glatt, H.; Williamson, G.; van Bladeren, P. J.; Rietjens, I. M. Phase II metabolism of hesperetin by individual UDP-glucuronosyltransferases and sulfotransferases and rat and human tissue samples. *Drug Metab. Dispos.* **2010**, *38*, 617−25.
(58) Court, M. H.; Duan, S. X.; von Moltke, L. L.; Greenblatt, D. J.; Patten, C. J.; Miners, J. O.; Mackenzie, P. I. Interindividual variability in acetaminophen glucuronidation by human liver microsomes: identification of relevant acetaminophen UDP-glucuronosyltransferase isoforms. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 998−1006.
(59) Ethell, B. T.; Ekins, S.; Wang, J.; Burchell, B. Quantitative structure activity relationships for the glucuronidation of simple phenols by expressed human UGT1A6 and UGT1A9. *Drug Metab. Dispos.* **2002**, *30*, 734−8.
(60) Ge, G. B.; Liang, S. C.; Liu, H. X.; Zhang, Y. Y.; Yang, L. Characterization of human UDP-Glucuronosyltransferase isoforms responsible for the in vitro glucuronidation of esculetin. *Drug Metab. Rev.* **2009**, *41*, 141−142.

(61) Kuehl, G. E.; Bigler, J.; Potter, J. D.; Lampe, J. W. Glucuronidation of the aspirin metabolite salicylic acid by expressed UDP-glucuronosyltransferases and human liver microsomes. *Drug Metab. Dispos.* **2006**, *34*, 199−202.
(62) Kuehl, G. E.; Lampe, J. W.; Potter, J. D.; Bigler, J. Glucuronidation of nonsteroidal anti-inflammatory drugs: identifying the enzymes responsible in human liver microsomes. *Drug Metab. Dispos.* **2005**, *33*, 1027−35.
(63) Liang, S. C.; Ge, G. B.; Liu, H. X.; Zhang, Y. Y.; Wang, L. M.; Zhang, J. W.; Yin, L.; Li, W.; Fang, Z. Z.; Wu, J. J.; Li, G. H.; Yang, L. Identification and Characterization of Human UDP-Glucuronosyltransferases Responsible for the In Vitro Glucuronidation of Daphnetin. *Drug Metab. Dispos.* **2010**, *38*, 973−980.
(64) Patana, A. S.; Kurkela, M.; Goldman, A.; Finel, M. The human UDP-glucuronosyltransferase: identification of key residues within the nucleotide-sugar binding site. *Mol. Pharmacol.* **2007**, *72*, 604−11.
(65) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−76.
(66) Lotta, T.; Taskinen, J.; Backstrom, R.; Nissinen, E. PLS modelling of structure-activity relationships of catechol O-methyltransferase inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 253−72.
(67) Taskinen, J.; Ethell, B. T.; Pihlavisto, P.; Hood, A. M.; Burchell, B.; Coughtrie, M. W. Conjugation of catechols by recombinant human sulfotransferases, UDP-glucuronosyltransferases, and soluble catechol O-methyltransferase: structure-conjugation relationships and predictive models. *Drug Metab. Dispos.* **2003**, *31*, 1187−97.
(68) Buolamwini, J. K.; Assefa, H. Overview of novel anticancer drug targets. *Methods Mol. Med.* **2003**, *85*, 3−28.
(69) Miners, J. O.; Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Mackenzie, P. I. Predicting human drug glucuronidation parameters: application of in vitro and in silico modeling approaches. *Ann. Rev. Pharm. Toxicol.* **2004**, *44*, 1−25.