Human Nephrotoxicity Prediction Models for Three Types of Kidney Injury Based on Data Sets of Pharmacological Compounds and Their Metabolites

ARTICLE *in* CHEMICAL RESEARCH IN TOXICOLOGY · OCTOBER 2013

Impact Factor: 3.53 · DOI: 10.1021/tx400249t · Source: PubMed

6 AUTHORS, INCLUDING:

Sehan Lee
EPA
7 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Young-Mook Kang
Bioinformatics and Molecular Design Researc…
5 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Kyoung Tai No
Yonsei University
141 PUBLICATIONS **1,752** CITATIONS
SEE PROFILE

Chemical Research in Toxicology

Article

pubs.acs.org/crt

Human Nephrotoxicity Prediction Models for Three Types of Kidney Injury Based on Data Sets of Pharmacological Compounds and Their Metabolites

Sehan Lee,[†] Young-Mook Kang,[‡] Hyejin Park,[†] Mi-Sook Dong,[§] Jae-Min Shin,[†] and Kyoung Tai No*[,†,‡]

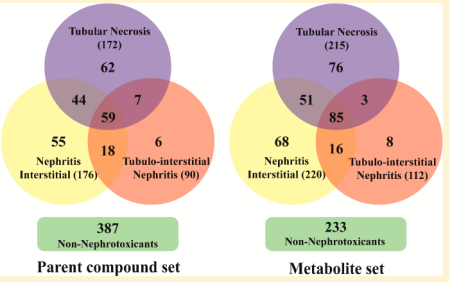
[†]Bioinformatics and Molecular Design Research Center, Seoul 120-749, Korea

[‡]Department of Biotechnology, Yonsei University, Seoul 120-749, Korea

[§]School of Life Sciences and Biotechnology, Korea University, Seoul 136-701, Korea

Supporting Information

ABSTRACT: The kidney is the most important organ for the excretion of pharmaceuticals and their metabolites. Among the complex structures of the kidney, the proximal tubule and renal interstitium are major targets of nephrotoxins. Despite its importance, there are only a few in silico models for predicting human nephrotoxicity for drug candidates. Here, we present quantitative structure−activity relationship (QSAR) models for three common patterns of drug-induced kidney injury, i.e., tubular necrosis, interstitial nephritis, and tubulo-interstitial nephritis. A support vector machine (SVM) was used to build the binary classification models of nephrotoxin versus non-nephrotoxin with eight fingerprint descriptors. To build the models, we constructed two types of data sets, i.e., parent compounds of pharmaceuticals (251 nephrotoxins and 387 non-nephrotoxins) and their major urinary metabolites (307 nephrotoxins and 233 non-nephrotoxins). Information on the nephrotoxicity of the pharmaceuticals was taken from clinical trial and postmarketing safety data. Though the mechanisms of nephrotoxicity are very complex, by using the metabolite information, the predictive accuracies of the best models for each type of kidney injury were better than 83% for external validation sets. Software to predict nephrotoxicity is freely available from our Web site at http://bmdrc.org/DemoDownload.



INTRODUCTION

Nephrotoxicity is a poisonous effect of a compound and its metabolites on the kidney, with 19−25% of all cases of acute kidney injury caused by drug exposure.[1] Therefore, the accurate measurement or estimation of human nephrotoxicity is crucial for drug discovery.[2] Unfortunately, kidney histopathology from *in vivo* animal studies, the gold standard for preclinical testing of compounds, is time-consuming and expensive and raises a number of ethical issues. Cell-based *in vitro* assays have advantages over *in vivo* assays as they can provide an early indication of the toxicity characteristics of the drug candidates.[3] There have been numerous research reports on the use of cellular assay systems to detect nephrotoxicity.[4−6] However, the scope of these studies has been restricted by either the types of cells employed or the drugs applied to the cells.[7]

In the past decades, several *in silico* toxicity prediction models have been developed.[2,8,9] Some toxicity prediction algorithms rely on the fact that toxic behavior is often associated with chemical structural motifs, like DEREK,[10] HazardExpert,[11] and OncoLogic.[12] Another approach is to use quantitative structure−activity relationship (QSAR) methods, such as CASE,[13] TOPKAT,[14] and PreADMET.[15] Compared with other toxicological end points, there are very few *in silico* prediction models for nephrotoxicity due to the diversity and biological complexity of the nephrotoxic end points as well as the paucity of data suitable for model development.[2,8,16] Jolivette et al.[17] developed a linear QSAR model using the lowest unoccupied molecular orbital energy (E_{LUMO}) to predict glutathione dependent haloalkene bioactivation that leads to nephrotoxicity. Since they focused only on haloalkenes, the resulting model has a limited applicability domain. Matthews et al.[18] recently developed QSAR models for predicting six types of drug-induced urinary tract injury in humans using approximately 1600 pharmaceuticals based upon observations in pharmaceutical clinical trials and postmarket surveillance by the FDA. The best QSAR models had high specificity (average 89%) but very low sensitivity (average 35%).

The mechanisms of nephrotoxicity may differ among various drugs or drug classes, and they are generally categorized based on the histological component of the kidney that is affected.[19] Renal tubular cells, particularly proximal tubule cells, are vulnerable to the toxic effects of drugs because their role in concentrating and reabsorbing the glomerular filtrate exposes them to high levels of circulating toxins. Drugs that cause tubular cell toxicity do so by impairing mitochondrial function, generating endoplasmic reticulum (ER) stress, interfering with tubular transport, increasing oxidative stress, or forming free radicals.[19−25]

Received: July 8, 2013
Published: October 18, 2013

Drugs can cause inflammatory changes in the glomerulus, renal tubular cells, and the surrounding interstitium, leading to fibrosis and renal scarring. Acute interstitial nephritis, which can result from an allergic response to a suspected drug, develops in an idiosyncratic, nondose-dependent fashion.[26]

Generally, drugs cannot be excreted in urine via the kidney until the functional groups that make drugs more soluble in water are introduced. Drug metabolism usually consists of two phases. Phase I reactions involve the formation of a new or modified functional group or cleavage (e.g., oxidation, reduction, hydrolysis, decyclization). Phase II reactions involve conjugation with an endogenous substance (e.g., glucuronic acid, sulfate, and glycine). These metabolisms play a crucial role in the bioactivation mechanism of nephrotoxins by forming proximate toxic metabolites or stable reactive intermediates. For example, the nephrotoxic effects of ifosfamide are attributed to its hepatic metabolite chloroacetaldehyde,[27] and the nephrotoxicity of methoxyflurane seems to result from O-demethylation, which forms both fluoride and dichloroacetic acid.[28]

In this study, we developed binary classification models for three common patterns of drug-induced kidney injury, i.e., tubular necrosis (TN), interstitial nephritis (IN), and tubulo-interstitial nephritis (TIN), with data sets of 638 pharmacological compounds. To reflect the metabolism effect on the nephrotoxicity of the pharmacological compounds, we also developed the models with data sets of 540 major urinary metabolites of the compounds. Our approach relied on the fact that biological activity is often associated with structural motifs.[29,30] We assumed that substructures that were frequently found in toxic compounds but only rarely found in nontoxic ones were likely to cause the toxic effect. Models with reasonably high predictive accuracy were built using a support vector machine (SVM) with fingerprints.

MATERIALS AND METHODS

Data Set Construction. Two types of data sets were constructed: a parent compound set and a metabolite set for each end point (TN, NI, and TIN). Parent compound sets consisted of nephrotoxic and non-nephrotoxic pharmacological compounds, and metabolite sets consisted of their major urinary metabolites. The compositions of the data sets are summarized in Figure 1.

The entire clinical information on the nephrotoxicity of the pharmacological compounds was obtained from the PharmaPendium database from Elsevier,[31] which integrates clinical trial data and postmarketing surveillance reports by the FDA. Entries containing inorganic compounds, proteins, large peptides, mixtures, and duplicated compounds were excluded during data set construction. Salts were converted into the corresponding acids or bases. To secure the reliability of the data, nephrotoxins that had more than five clinical trial and postmarketing reports were selected. Non-nephrotoxins with no evidence of nephrotoxicity were obtained from the PharmaPendium toxicity/side effect database by eliminating the following compounds: (i) compounds causing any renal or urinary disorders and (ii) compounds that had less than 10 clinical trial and postmarketing safety reports based on the assumption that those compounds had not been sufficiently verified. By applying the above criteria, 251 nephrotoxic and 387 non-nephrotoxic pharmacological compounds were obtained.

Information on the major urinary metabolites of the pharmacological compounds in the parent compound sets was collected from clinical or animal studies. The majority of a metabolite in urine was identified by urinary concentration of the metabolite (usually more than 10% of the total radioactivity in urine). The metabolite sets comprised 307 metabolites that came from 172 nephrotoxic compounds and 233 metabolites that came from 123 non-nephrotoxic compounds. Out of the 172 nephrotoxins and 123 non-nephrotoxins, 110 and 49, respectively, were excreted in urine as a major metabolite.

Training and External Validation Sets. The parent compound sets are imbalanced with less than 30% of nephrotoxic compounds. Most classification algorithms trained on the imbalanced data tend to classify test samples as the majority class and, therefore, have high specificity but low sensitivity.[32] Among the strategies developed to effectively learn from the imbalanced data sets, undersampling where only a portion of the majority class is selected to construct the training set exhibited the capability of improving the performance of predictive models. The drawback of this technique is information loss, which can hinder classification performance in some cases. To cope with data imbalance in this work, the training set and corresponding external validation set to develop and evaluate the models were selected from each data set as follows: (1) 80% of randomly selected compounds from the metabolite sets and nephrotoxins from the parent compound sets were used as a training set and (2) non-nephrotoxins (majority) of the parent compound sets were partitioned into training and external validation sets in a ratio of 35 to 65 to rebalance class distributions. Although TIN sets were still imbalanced (nephrotoxin/non-nephrotoxin ≈2) due to the limited amount of data, additional undersampling was not performed to minimize information loss. The final compositions of training and external validation sets for each model are summarized in Table 1.

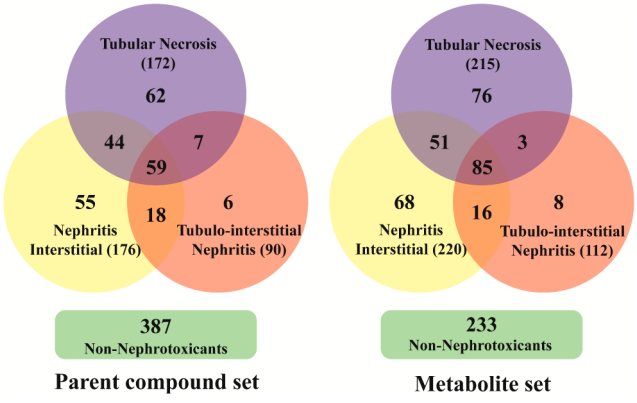


Figure 1. Composition of the data sets summarized by Venn diagrams. Some nephrotoxic compounds cause more than one type of kidney injury. Metabolite sets consist of major urinary metabolites of pharmaceuticals in parent compound sets whose metabolite information is available. The number in parentheses is the number of compounds in each class.

Table 1. Detailed Statistical Description of Chemicals Used in the Training Set and Test Set[a]

Parent Compound Set			
	data sets		
nephrotoxicity	training	external validation	total
TN	138	34	172
NI	141	35	176
TIN	72	18	90
non-nephrotoxin	136	251	387
Metabolite Set			
	data sets		
nephrotoxicity	training	external validation	total
TN	172	43	215
NI	176	44	220
TIN	90	22	112
non-nephrotoxin	186	47	233

[a]TN, tubular necrosis; NI, nephritis interstitial; TIN, tubulo-interstitial nephritis.

Evaluation of Models Performance. The quality of the models was evaluated in terms of classification accuracy (CA), sensitivity (SE), specificity (SP), and Matthews correlation coefficient (MCC).

$$\mathrm{CA} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}) \quad (1)$$

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (2)$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (3)$$

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

where TP, TN, FN, and FP are the number of true positives, true negatives, false negatives, and false positives, respectively. MCC takes into account true and false positives and negatives; thus, the uneven distribution of the objects in the two classes is less likely to result in biased models.

Descriptors. Eight kinds of topological fingerprints implemented in PaDEL-Descriptor[33] were introduced: the CDK fingerprint (FP), CDK extended fingerprint (ExtFP), Estate fingerprint (EstateFP), Graph only fingerprint (GraphFP), MACCS fingerprint (MACCSFP), PubChem fingerprint (PubChemFP), Substructure fingerprint (SubFP), and the Klekota-Roth fingerprint (KRFP). Detailed descriptions of these fingerprints can be found in the original literature.[33,34]

Descriptor Selection. In order to reduce the computation time and the noise originated from undesirable descriptors during the descriptor selection process, only 20 descriptors with the highest information gain (IG) were considered. The IG provides a measure of the reduction of the uncertainty (entropy) associated with the presence of a variable in the data set. The IG of a particular descriptor X for a data set T of samples classified into two subsets v (i.e., v = nephrotoxin and non-nephrotoxin) is given by the following relationship:

$$IG(T, X) = H(T) - \sum_{g} \frac{|T_g|}{|T|} H(T_g) \quad (5)$$

where g is a possible value of descriptor X (i.e., g = presence and absence), and T_g is a set of samples, which have a descriptor value g. $|T|$ and $|T_g|$ are the number of samples in sets T and T_g, respectively. $H(T)$ and $H(T_g)$ are the information entropy of sets T and T_g, respectively. $H(T)$ is given by the following equation:

$$H(T) = -\sum_{v} \frac{|T_v|}{|T|} \log_2 \frac{|T_v|}{|T|} \quad (6)$$

where $|T_v|$ is the number of samples in subset v.

The typical combinatorial method of the SVM with genetic algorithm (GA-SVM) was used to capture the most informative descriptors, and a step-by-step selecting process was used. For each step, the population size was set to 100, and the current step was terminated after reproducing 10 generations. In each generation, MCCs from a 10-fold cross-validation for each chromosome were used as the fitness value of each chromosome. MCC was used as the score since it is a balanced measure of robustness.

Modeling Methods. SVM was performed by the RapidMiner5.2 open-source data mining tool (http://rapid-i.com).[35] The SVM algorithm, originally developed by Vapnik[36] for pattern recognition, aims at minimizing the structural risk under the frame of the Vapnik–Chervonenkis (VC) theory. Many studies have demonstrated that SVM is one of the best methods for classification modeling.[37,38]

The ith sample in a data set is defined as $D_i = (x_i, y_i)$, where x_i is an N-dimensional real vector, and y_i is the corresponding class label (1 = nephrotoxic class and −1 = non-nephrotoxic class). The separating hyperplane $f(x)$ is defined as follows:

$$f(x) = w \cdot x_i + b \quad (7)$$

where w is a vector normal to the hyperplane, and b is a scalar quantity. The SVM attempts to find an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\max_{w,b} \frac{2}{\|w\|} \text{subject to } y_i(w \cdot x_i + b) - 1 \geq 0,\ i = 1, \ldots, N \quad (8)$$

where $(2/\|w\|)$ is the margin. When perfect separation is not possible, slack variables ξ_i are introduced for sample vectors that are within the margin, and the optimization problem can be reformulated:

$$\max_{w,b} \frac{2}{\|w\|} + C \sum_{i=1}^{N} \xi_i \text{ subject to } y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0;\ \xi_i \geq 0 \quad (9)$$

where C is the penalty parameter, which should be predetermined by the user.

The nonlinear (non)separable cases could be easily transferred to linear cases by projecting the input variable into a high-dimensional feature space by using a kernel function $K(x_i, x_j)$. In this work, we use the Gaussian radial basis function kernel:

$$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2) \quad (10)$$

The penalty parameter C and different kernel parameter γ were tuned based on the training set using a grid search strategy by 10-fold cross-validation in order to obtain a SVM model with optimal performance.

Internal Validation. The Y-randomization test is a tool that is widely used to establish the robustness of a model.[36] The test consists of rebuilding models using randomly shuffled activities of the training set and subsequently assessing the model statistics. This process is repeated several times. It is expected that models obtained for the training set with randomized activities should have significantly lower predictivity for the test set with real activities than the models built using the training set with real activities. If this condition is not satisfied, real models built for this training set are not reliable and should be discarded.

The Y-randomization test was performed for the best performing models. The calculations were repeated five times based on the original descriptor pool and the original model building procedure. The robustness of the models was examined by comparing these models to those derived from data sets with randomized activity using Z-score statistics. The Z-score is calculated as follows:

$$Z = (\mathrm{MCC}_{\mathrm{ori}} - \mathrm{MCC}_{\mathrm{rand}}^{\mathrm{mean}})/\sigma \quad (11)$$

$\mathrm{MCC}_{\mathrm{ori}}$ is the MCC of the original data set with actual activity values, $\mathrm{MCC}_{\mathrm{rand}}^{\mathrm{mean}}$ represents the mean MCC values of the data sets with randomized activity values, and σ is the standard deviation from $\mathrm{MCC}_{\mathrm{rand}}^{\mathrm{mean}}$ of the distribution of MCC values of the random models. The Z-score serves as a measure of the uniqueness of the models built with actual data as opposed to those generated with the randomized activity data. Models with Z-scores exceeding 3 are regarded as statistically significant.

Analysis of Privileged Substructures. The privileged substructures for nephrotoxic compounds were identified using information gain and substructure fragment analysis.[39,40] The term privileged structure was used by Evans et al. in 1998 to represent substructures that confer activity to two or more different receptors.[41] The implication was that the privileged structure is able to provide ligands for diverse receptors. If a substructure was more frequently presented in nephrotoxic chemical class, this substructure was called a privileged substructure involved in drug induced kidney injury. The frequency of a fragment in nephrotoxins was defined as follows:

$$\text{frequency of a fragment} = \frac{(N_{\text{fragment.class}} \times N_{\text{total}})}{(N_{\text{fragment.total}} \times N_{\text{class}})} \quad (12)$$

where $N_{\text{frag\_class}}$ is the number of compounds containing the fragment in nephrotoxic class; N_{total} is the total number of compounds; $N_{\text{frag\_total}}$ is the total number of compounds containing the fragment; and N_{class} is the number of nephrotoxins.

RESULTS AND DISCUSSION

Data Set Analysis. The quality and diversity of data, in terms of capturing the wide variety of adverse effects and underlying

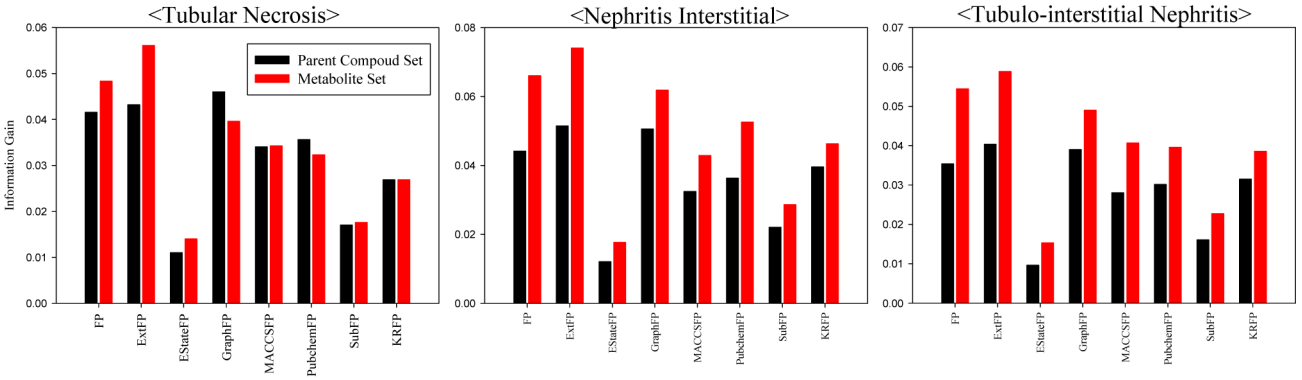


Figure 2. Average values of the 20 highest information gains. Eight different fingerprints were computed for parent compound sets (black bars) and metabolite sets (red bars). The average values were higher for FP, ExtFP, and GraphFP than the other fingerprints and higher for metabolite sets than parent compound sets.

mechanisms of action, is an important issue for QSAR and other *in silico* models for toxicological end points. In the past decades, most of the models have been developed based on specific chemical classes, and consequently, they have limited applicability domains.[9] We developed here data sets of heterogeneous pharmacological compounds for nephrotoxicity end points. The parent compound sets consisted of pharmacological compounds of diverse drug classes (analgesics, antivirals, antibiotics antineoplastics, diuretics, and so on) targeting more than 140 proteins.

The chemical spaces of the data sets were investigated by calculating the Ghose-Crippen LogKow (ALogP) and molecular weight (MW) of the training and external validation sets. The distribution scatter diagrams were presented in Figures S1−S3 of the Supporting Information. As shown in Figures S1−S3 (Supporting Information), the chemical spaces of each external validation set were within the scope of the corresponding training set. Since the data sets are composed of pharmaceuticals, most compounds have a MW less than 500 Da and ALogP not greater than 5.

The clinical information on the nephrotoxicity of the compounds came from clinical trial and postmarketing safety data. As a way to secure the reliability of data, nephrotoxic and non-nephrotoxic compounds that had more than 5 and 10 reports, respectively, were used. Information on the major urinary metabolites of the pharmacological compounds came mainly from clinical studies, i.e., 167 and 107 out of 172 nephrotoxic and 123 non-nephrotoxic compounds, respectively.

Information Gain (*IG*) Analysis. The information content of the individual substructure for a data set was analyzed using the *IG*. Higher *IG* is related to lower information entropy of the subsets defined by the presence and absence of a particular substructure (eq 5). This effectively describes better separation between nephrotoxic and non-nephrotoxic structures by that particular substructure. Supporting Information, Figure S4 shows the distributions of the *IG*s of each fingerprint computed for the parent compound sets and the metabolite sets. For all fingerprints, most substructures had low *IG* regardless of data sets, which means that the discriminatory power of those substructures is very low. However, when comparing the 20 most informative substructures used in the descriptor selection process, all fingerprints calculated for NI and TIN sets showed much more informative distributions in the metabolite sets than in the parent compound sets. In the case of TN, there were inconsistent distributions of *IG*s between the fingerprints. While FP and ExtFP showed more informative distributions for the metabolite set, GraphFP showed more informative distribution for the parent compound set, and the other fingerprints were comparable. Among the eight fingerprints, FP, ExtFP, and GraphFP showed the most informative distributions for all data sets (Figure 2).

Performance of the Models. In this work, eight different fingerprints (FP, ExtFP, EstateFP, GraphFP, MACCSFP, PubChemFP, SubFP, and KRFP) and two different types of data sets (parent compound sets and metabolite sets) were used to develop highly predictive models for three types of kidney injury (TN, NI, and TIN). The performances of the models based on the parent compound sets and the metabolite sets are summarized in Tables 2 and 3, respectively.

Parent Compound Set Based Models. Comparing the eight different fingerprints, the TN model with ExtFP and the NI and TIN models with KRFP were better than the other seven in terms of CA in external validation: CAs for TN, NI, and TIN models were 0.72, 0.79, and 0.89, respectively. However, the TIN model had a high SP but a very low SE (SP = 0.94 and SE = 0.17), and therefore, MCC was very low (MCC = 0.11). In terms of balanced accuracy, the TIN model with GraphFP yielded the best performance (CA = 0.80 and MCC = 0.32).

Metabolite Set Based Models. The best TN, NI, and TIN models, in terms of CA and MCC in external validation, employed MACCSFP, ExtFP, and FP, respectively. CAs were 0.84, 0.85, and 0.83, and MCCs were 0.69, 0.69, and 0.62 for the TN, NI, and TIN models, respectively. The performances of the models with FP and ExtFP were always within third best. This result coincides with the results shown in Figures S4 and 2 (Supporting Information), in which FP and ExtFP show the most informative *IG* distributions for the metabolite sets.

The prediction abilities of the models based on the parent compound set and the metabolite set were compared. The results indicated that when the only difference between the models was the type of data set, the models based on the metabolite sets always outperformed the models based on the parent compound sets in terms of CA and MCC in external validation. In the case of TN and NI, the average CA and MCC of the models based on the metabolite sets were 0.77 and 0.55, respectively, whereas those of the models based on the parent compound sets were 0.68 and 0.23, respectively. In the case of TIN, both types of models yielded the same average CAs of 0.80 but much different average MCCs, 0.51 and 0.25 for the models based on the metabolite set and the parent compound set, respectively.

The nephrotoxicity of a pharmacological compound is induced by the parent compound as well as its metabolites. However, because of the paucity of information on which compounds are

Table 2. Performance of Classification Models for the Training and Validation Using Different Fingerprints and Parent Compound Sets[a]

	Tubular Necrosis							
	10-fold cross-validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.76	0.74	0.75	0.50	0.65	0.69	0.69	0.23
Ext	0.75	0.78	0.77	0.53	0.65	0.73	0.72	0.27
Estate	0.75	0.70	0.72	0.45	0.65	0.67	0.67	0.22
Graph	0.75	0.76	0.76	0.52	0.68	0.69	0.69	0.25
MACCS	0.64	0.79	0.72	0.44	0.76	0.67	0.68	0.29
Pubchem	0.71	0.79	0.75	0.51	0.56	0.71	0.69	0.19
Sub	0.92	0.49	0.71	0.46	0.56	0.70	0.68	0.18
KR	0.86	0.60	0.73	0.47	0.76	0.51	0.54	0.18
	Interstitial Nephritis							
	10-fold cross-validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.80	0.71	0.76	0.52	0.80	0.66	0.68	0.31
Ext	0.86	0.66	0.76	0.53	0.89	0.61	0.64	0.32
Estate	0.66	0.74	0.70	0.40	0.54	0.73	0.71	0.19
Graph	0.72	0.75	0.73	0.47	0.51	0.71	0.69	0.16
MACCS	0.69	0.74	0.71	0.42	0.60	0.69	0.68	0.20
Pubchem	0.60	0.79	0.69	0.40	0.54	0.73	0.71	0.19
Sub	0.68	0.76	0.72	0.44	0.69	0.70	0.70	0.27
KR	0.61	0.80	0.70	0.42	0.54	0.83	0.79	0.30
	Tubulo-Interstitial Nephritis							
	10-fold cross-validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.78	0.83	0.81	0.60	0.61	0.82	0.81	0.27
Ext	0.85	0.82	0.83	0.64	0.78	0.71	0.71	0.26
Estate	0.63	0.86	0.78	0.50	0.61	0.82	0.81	0.27
Graph	0.65	0.91	0.82	0.60	0.72	0.81	0.80	0.32
MACCS	0.53	0.90	0.77	0.47	0.61	0.83	0.81	0.27
Pubchem	0.56	0.88	0.76	0.46	0.56	0.80	0.79	0.22
Sub	0.68	0.84	0.78	0.52	0.61	0.80	0.79	0.25
KR	0.43	0.92	0.75	0.41	0.17	0.94	0.89	0.11

[a]SE: sensitivity. SP: specificity. CA: classification accuracy. MCC: Matthews Correlation Coefficient. FP: CDK fingerprint. ExtFP: CDK extended fingerprint. EstateFP: Estate fingerprint. GraphFP: CDK graph only fingerprint. MACCSFP: MACCS fingerprint. PubChemFP: PubChem fingerprint. SubFP: Substructure fingerprint. KRFP: Klekota-Roth fingerprint.

responsible for the observed nephrotoxicity (i.e., TN, NI, TIN, and nontoxic), the data sets were developed based on the assumption that the nephrotoxicity of parent compounds and metabolites correspond with that of pharmacological compounds observed in clinical trials and/or postmarket surveillance. From the standpoint of QSAR, for the parent compound sets, it is clear that the assumption is correct only if parent compounds contain structural features representing corresponding nephrotoxicity. However, according to the metabolite information from the metabolite sets (see section on data set construction), most pharmacological compounds are excreted via the kidney after extensive metabolism, and thus, the assumption can be a source of model error. For the metabolite sets, the assumption would be reasonable for non-nephrotoxins because the metabolites do not represent any notable nephrotoxicity during excretion via the kidney. However, in the case of nephrotoxins, it is unclear which ones are nephrotoxic because metabolism can result in both bioactivation and detoxification. Because of the obscurity of the nephrotoxicity of the compounds, especially in the nephrotoxic class of the metabolite sets, along with the imbalance of TIN training sets, there is a tendency for SP to be relatively higher than SE for most of the models.

Validation by Y-Randomization. The reliability of the models was assessed using Y-randomization to exclude the possibility of chance correlations. Y-Randomization was performed for each model wherein training set activities were randomized. The statistical data on MCCs and Z-scores for five runs are summarized in Table 4. Usually, MCCs were close to zero (i.e., no better than random prediction), which proves that the predictive ability of the actual models is real. The Z-scores for TN, NI, and TIN models were 3.46, 14.01, and 3.99 for the parent compound sets and 5.28, 7.19, and 4.84 for the metabolite sets, respectively, indicating the high statistical significance of models built with actual data.

Privileged Substructure for Drug Induced Kidney Injury. To further explore the structural features of nephrotoxins, information gain and substructure fragment analysis were performed on each data set (combining the training set and external validation set) using KRFP. The privileged substructure fragments, information gain, and frequency of fragment enrichment factor were identified in Supporting Information, Schemes S1−S3. Selected substructures were usually common to both types of data sets (parent compound sets and metabolite sets) for all types of kidney injuries (TN, NI, and TIN).

Table 3. Performance of Classification Models for the Training and Validation Using Different Fingerprints and Metabolite Sets[a]

	Tubular Necrosis							
	10-fold cross validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.76	0.84	0.80	0.60	0.74	0.85	0.80	0.60
Ext	0.78	0.87	0.83	0.65	0.84	0.81	0.82	0.65
Estate	0.65	0.75	0.70	0.40	0.60	0.74	0.68	0.35
Graph	0.64	0.87	0.76	0.53	0.67	0.83	0.76	0.51
MACCS	0.78	0.76	0.77	0.55	0.88	0.81	0.84	0.69
PubChem	0.81	0.70	0.75	0.51	0.79	0.72	0.76	0.51
Sub	0.65	0.75	0.70	0.40	0.70	0.77	0.73	0.47
KR	0.72	0.71	0.71	0.42	0.67	0.74	0.71	0.42
	Interstitial Nephritis							
	10-fold cross validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.85	0.81	0.83	0.66	0.84	0.83	0.84	0.67
Ext	0.85	0.83	0.84	0.68	0.86	0.83	0.85	0.69
Estate	0.64	0.82	0.73	0.47	0.52	0.91	0.73	0.48
Graph	0.76	0.82	0.79	0.58	0.75	0.79	0.77	0.54
MACCS	0.80	0.76	0.78	0.56	0.73	0.83	0.78	0.56
PubChem	0.70	0.83	0.77	0.53	0.68	0.81	0.75	0.50
Sub	0.60	0.85	0.73	0.47	0.64	0.85	0.75	0.50
KR	0.53	0.90	0.72	0.47	0.48	0.96	0.73	0.50
	Tubulo-Interstitial Nephritis							
	10-fold cross-validation				external validation			
fingerprint	SE	SP	CA	MCC	SE	SP	CA	MCC
FP	0.80	0.85	0.84	0.64	0.82	0.83	0.83	0.62
Ext	0.71	0.88	0.82	0.59	0.68	0.87	0.81	0.56
Estate	0.48	0.94	0.79	0.49	0.55	0.94	0.81	0.54
Graph	0.66	0.93	0.84	0.63	0.59	0.89	0.80	0.51
MACCS	0.59	0.95	0.83	0.61	0.36	0.89	0.72	0.31
PubChem	0.66	0.89	0.82	0.57	0.77	0.81	0.80	0.56
Sub	0.53	0.91	0.79	0.50	0.36	1.00	0.80	0.53
KR	0.62	0.92	0.83	0.59	0.50	0.91	0.78	0.47

[a]SE: sensitivity. SP: specificity. CA: classification accuracy. MCC: Matthews Correlation Coefficient. FP: CDK fingerprint. ExtFP: CDK extended fingerprint. EstateFP: Estate fingerprint. GraphFP: CDK graph only fingerprint. MACCSFP: MACCS fingerprint. PubChemFP: PubChem fingerprint. SubFP: Substructure fingerprint. KRFP: Klekota-Roth fingerprint.

Table 4. Y-Randomization Data for the Best Models[a]

	parent compound set			metabolite set		
model	TN	NI	TIN	TN	NI	TIN
original[b]	0.27	0.30	0.32	0.67	0.69	0.62
Y-randomization[c]	0.00 ± 0.08	−0.03 ± 0.02	0.01 ± 0.08	0.01 ± 0.08	−0.08 ± 0.11	−0.0 ± 0.14
Z-score	3.46	14.01	3.99	5.28	7.19	4.84

[a]TN: Tubular Necrosis. NI: Nephritis Interstitial. TIN: Tubulo-Interstitial Nephritis. [b]The MCCs of the best models developed in this work. [c]The Y-randomization values represent the mean ± standard deviation of MCCs from 5 independent runs.

Interestingly, based on the comparison of the numbers of nephrotoxins and non-nephrotoxins with a specified fragment, N_T and N_{NT}, in the metabolite set to those in the parent compound set, N_Ts were comparable or increased, and N_{NT}s were decreased dramatically (i.e., bioactivation of nephrotoxins and detoxification of non-nephrotoxins). For instance, N_T and N_{NT} with [!#1]S[!#1] were 29 and 53 in the parent compound set, and 37 and 14 in the metabolite set, respectively, for TN (Supporting Information, Scheme S1). Therefore, the proper reflection of metabolism effect on a chemical structure is very important for accurate prediction of nephrotoxicity of the compound.

CONCLUSIONS

In this study, we developed binary classification models for three common patterns of drug-induced kidney injury with high predictive accuracy using 2D fingerprint-based SVM. Since kidney injury is caused by not only pharmacological compounds but also their metabolites, each model was built based on the sets of pharmacological compounds and their metabolites. The results of 10-fold cross-validation and external validation showed a high accuracy of our models, and they thus indicate the importance of the inclusion of metabolism information. The results also demonstrate that fingerprints, as attributes for classification models, are powerful tools and that it is possible to

predict nephrotoxicity from a 2D structure. All of the tools used in this study for model development are free of charge and easily accessible. Software for nephrotoxicity prediction is freely available for downloading at http://bmdrc.org/DemoDownload.

■ ASSOCIATED CONTENT

Supporting Information

Chemical spaces, distributions of information gains, and privileged substructures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 82-2-393-9551. Fax: 82-2-393-9554. E-mail: ktno@yonsei.ac.kr.

Funding

This study was supported by grants from the Korea Healthcare Technology R&D Project, Ministry for Health, Welfare & Family Affairs, Korea [A100096], and from KRICT and Ministry of Knowledge, Korea [SI-1304].

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

QSAR, quantitative structure−activity relationship; SVM, support vector machine; TN, tubular necrosis; IN, interstitial nephritis; TIN, tubulo-interstitial nephritis; CA, classification accuracy; SE, sensitivity; SP, specificity; MCC, Matthews correlation coefficient; IG, information gain

■ REFERENCES

(1) Bonventre, J. V., Vaidya, V. S., Schmouder, R., Feig, P., and Dieterle, F. (2010) Next-generation biomarkers for detecting kidney toxicity. *Nat. Biotechnol. 28*, 436−440.

(2) Lapenna, S., Fuart-Gatnik, M., and Worth, A. (2010) Review of QSAR Models and Software Tools for predicting Acute and Chronic Systemic Toxicity, EUR 24639 EN.

(3) Zang, R., Li, D., Tang, I.-C., Wang, J., and Yang, S.-T. (2012) Cell-based assays in high-throughput screening for drug discovery. *Int. J. Biotechnol. Well. Ind. 1*, 31−51.

(4) White, D. J., and Seaman, C. (1995) LCC-RK1 cell screening test for nephrotoxicity. *Methods Mol. Biol. 43*, 11−16.

(5) Pabla, N., and Dong, Z. (2008) Cisplatin nephrotoxicity: mechanisms and renoprotective strategies. *Kidney Int. 73*, 994−1007.

(6) Rankin, G. O., Racine, C., Sweeney, A., Kraynie, A., Anestis, D. K., and Barnett, J. B. (2008) In vitro nephrotoxicity induced by propanil. *Environ. Toxicol. 4*, 435−442.

(7) Pfalleer, W., and Gstraunthaler, G. (1998) Nephrotoxicity testing in vitro - what we know and what we need to know. *Environ. Health Perspect. 106*, 559−569.

(8) van de Waterbeemd, H., and Gifford, E. (2003) ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov. 2*, 192−204.

(9) Dearden, J. C. (2003) In silico prediction of drug toxicity. *J. Comput.-Aided Mol. Des. 17*, 119−127.

(10) Sanderson, D. M., and Earnshaw, C. G. (1991) Computer prediction of possible toxicity action from chemical structure. *Hum. Exp. Toxicol. 10*, 261−273.

(11) Smithing, M. P., and Darvas, F. (1992) HazardExpert: An Expert System for Predicting Chemical Toxicity, *Food Safety Assessment*, pp 191−200, ACS Symposium Series, American Chemical Society, Washington, DC.

(12) Woo, Y., Lai, D., Argus, M., and Arcos, J. (1995) Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicol. Lett. 79*, 219−228.

(13) Klopman, G. (1984) Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc. 106*, 7315−7321.

(14) Enslein, K., Gombar, V. K., and Blake, B. W. (1994) International commission for protection against environmental mutagens and carcinogens. Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat. Res. 305*, 47−61.

(15) http://preadmet.bmdrc.org/ (accessed Oct 2012).

(16) Perazella, M. A. (2009) Renal vulnerability to drug toxicity. *Clin. J. Am. Soc. Nephrol. 4*, 1275−1283.

(17) Jolivette, L. J., and Anders, M. W. (2002) Structure−activity relationship for the biotransformation of haloalkenes by rat liver microsomal glutathione transferase 1. *Chem. Res. Toxicol. 15*, 1036−1041.

(18) Matthews, E. J., Ursem, C. J., Kruhlak, N. L., Benz, R. D., Sabaté, D. A., Yang, C., Klopman, G., and Contrera, J. F. (2009) Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Requl. Toxicol. Pharmacol. 54*, 23−42.

(19) Naughton, C. A. (2008) Drug-induced nephrotoxicity. *Am. Fam. Physician 78*, 743−750.

(20) Loeffler, M., and Kroemer, G. (2000) The mitochondrion in cell death control: certainties and incognita. *Exp. Cell. Res. 256*, 19−26.

(21) Ferri, K. F., and Kroemer, G. (2001) Organelle-specific initiation of cell death pathways. *Nat. Cell Biol. 3*, E255−263.

(22) Hacki, J., Egger, L., Monney, L., Conus, S., Rosse, T., Fellay, I., and Borner, C. (2000) Apoptotic crosstalk between the endoplasmic reticulum and mitochondria controlled by Bcl-2. *Oncogene 19*, 2286−2295.

(23) Nakagawa, T., Zhu, H., Morishima, N., Li, E., Xu, J., Yankner, B. A., and Yuan, J. (2000) Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-beta. *Nature 403*, 98−103.

(24) Cihlar, T., Ho, E. S., Lin, D. C., and Mulato, A. S. (2001) Human renal organic anion transporter 1 (hOAT1) and its role in the nephrotoxicity of antiviral nucleotide analogs. *Nucleosides Nucleotides Nucleic Acids 20*, 641−648.

(25) Karahan, İ., Ateşşahin, A., Yılmaz, S., Çeribaşı, A. O., and Sakin, F. (2005) Protective effect of lycopene on gentamicin-induced oxidative stress and nephrotoxicity in rats. *Toxicology 215*, 198−204.

(26) Eddy, A. A. (1996) Molecular insights into renal interstitial fibrosis. *J. Am. Soc. Nephrol. 7*, 2495−2508.

(27) Dubourg, L., Michoudet, C., Cochat, P., and Baverel, G. (2001) Human kidney tubules detoxify chloroacetaldehyde, a presumed nephrotoxic metabolite of ifosfamide. *J. Am. Soc. Nephrol. 12*, 1615−1623.

(28) Kharasch, E. D., Schroeder, J. L., Liggitt, H. D., Ensign, D., and Whittington, D. (2006) New insights into the mechanism of methoxyflurane nephrotoxicity and implications for anesthetic development (Part 2): Identification of nephrotoxic metabolites. *Anesthesiology 105*, 737−745.

(29) Horton, D. A., Bourne, G. T., and Smythe, M. L. (2002) Exploring privileged structures: the combinatorial synthesis of cyclic peptides. *J. Comput. Aided. Mol. Des. 5−6*, 415−430.

(30) DeSimone, R. W., Currie, K. S., Mitchell, S. A., Darrow, J. W., and Pippin, D. A. (2004) Privileged structures: applications in drug discovery. *Comb. Chem. High Throughput Screen. 7*, 473−493.

(31) (2008) Parmapendium, Elsevier, New York, https://www.pharmapendium.com (accessed May, 2012).

(32) He, H., and Garcia, E. A. (2009) Learning from imbalanced data. *IEEE Trans. Knowledge Data Eng. 21*, 1263−1284.

(33) Yap, C. W. (2011) PaDel-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem. 32*, 1466−1474.

(34) Kletota, J., and Roth, F. P. (2008) Chemical substructures that enrich for biological activity. *Bioinformatics 24*, 2518−2525.

(35) Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD: New York, pp 935−940.
(36) Corinna, C., Vladimir, V., Cortes, C., and Vapnik, V. (1995) Support-vector networks. *Mach. Learn. 20*, 273−297.
(37) Byvatov, E., and Schneider, G. (2003) Support vector machine applications in bioinformatics. *Appl. Bioinformatics 2*, 67−77.
(38) Yang, Z. R. (2004) Biological applications of support vector machines. *Briefings Bioinf. 5*, 328−338.
(39) Jensen, B. F., Vind, C., Padkjaer, S. B., Brockhoff, P. B., and Refsgaard, H. H. (2007) In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem. 50*, 501−511.
(40) Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., and Tang, Y. (2012) In silico prediction of chemical Ames mutagenicity. *J. Chem. Inf. Model. 52*, 2840−2847.
(41) Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., Lundell, G. F., Veber, D. F., Anderson, P. S., Chang, R. S., et al. (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem. 31*, 2235−2246.