Journal of
Chemical Information
and Modeling

Article

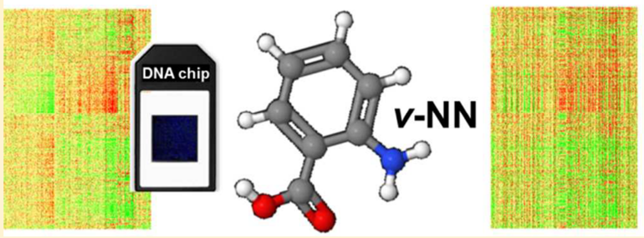
pubs.acs.org/jcim

Molecular Structure-Based Large-Scale Prediction of Chemical-Induced Gene Expression Changes

Ruifeng Liu,* Mohamed Diwan M. AbdulHameed, and Anders Wallqvist*

Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland 21702, United States

Ⓢ *Supporting Information*



ABSTRACT: The quantitative structure−activity relationship (QSAR) approach has been used to model a wide range of chemical-induced biological responses. However, it had not been utilized to model chemical-induced genomewide gene expression changes until very recently, owing to the complexity of training and evaluating a very large number of models. To address this issue, we examined the performance of a variable nearest neighbor (ν-NN) method that uses information on near neighbors conforming to the principle that similar structures have similar activities. Using a data set of gene expression signatures of 13 150 compounds derived from cell-based measurements in the NIH Library of Integrated Network-based Cellular Signatures program, we were able to make predictions for 62% of the compounds in a 10-fold cross validation test, with a correlation coefficient of 0.61 between the predicted and experimentally derived signatures—a reproducibility rivaling that of high-throughput gene expression measurements. To evaluate the utility of the predicted gene expression signatures, we compared the predicted and experimentally derived signatures in their ability to identify drugs known to cause specific liver, kidney, and heart injuries. Overall, the predicted and experimentally derived signatures had similar receiver operating characteristics, whose areas under the curve ranged from 0.71 to 0.77 and 0.70 to 0.73, respectively, across the three organ injury models. However, detailed analyses of enrichment curves indicate that signatures predicted from multiple near neighbors outperformed those derived from experiments, suggesting that averaging information from near neighbors may help improve the signal from gene expression measurements. Our results demonstrate that the ν-NN method can serve as a practical approach for modeling large-scale, genomewide, chemical-induced, gene expression changes.

■ INTRODUCTION

Quantitative structure−activity relationship (QSAR) modeling has been routinely applied to predict a broad range of chemical-induced biological responses.[1] However, its ability to predict chemical-induced gene expression changes in cultured cells has only recently been investigated.[2] A major obstacle to using this approach in predicting transcriptional changes is the complexity of the problem, given that the human genome consists of over 20 000 coding and noncoding genes. Even though the expression levels of a subset of independent genes may be sufficient to infer those for the remaining genes, because some genes tend to be coregulated, recent studies indicate that ∼1000 independent genes are still required to properly represent the expression patterns of the human genome.[3] This means that a conventional QSAR approach needs to generate at least ∼1000 models, one for each gene, to predict genomewide changes in gene expression. A further caveat in modeling gene expression changes is that data derived from high-throughput measurements contain significant experimental variability. This can obscure the fundamental assumption of the QSAR approach—that similar structures have similar activities—and undermine any modeling attempts. For instance, Chen et al., in a comprehensive analysis of the correlation between the similarity of molecular structures among 11 000 compounds and the similarity of gene expression patterns induced by these compounds, found that the likelihood of two compounds with a Tanimoto similarity of at least 0.85 showing similar gene expression profiles was only 20%.[4] Nevertheless, they did show a correlation, albeit weaker than expected,[5] thereby providing a basis for further exploring QSAR modeling of chemical-induced changes in gene expression.

In a recent study, Hall et al. used the QSAR approach to model chemical-induced gene expression by deploying multiple machine learning methods in combination with over 30 000 molecular descriptors.[2] They used the expression data of ∼1000 genes—derived from the A673 cell line exposed to 175 compounds at 10 μM for 6 h—to build ∼600 000 QSAR models. They selected ∼20 000 of the best-performing models, 20 for each gene, as their final QSAR models. This brute force approach, although impressive, is impractical because the processes of building, storing, retrieving, and evaluating 600 000 static models require substantial computational resources.

Received: May 18, 2017
Published: August 10, 2017

DOI: 10.1021/acs.jcim.7b00281
J. Chem. Inf. Model. 2017, 57, 2194−2202

We recently developed a variable nearest neighbor (v-NN) method that only uses information from qualified near neighbors to make predictions.[6] This method does not require the creation or storage of any static models, and has the advantage that it can use up-to-date experimental data to make predictions without retraining. The v-NN method is thus potentially well suited for predicting genome-wide gene expression changes.

In this study, we evaluated the performance of the v-NN method in predicting expression changes of 987 genes in cultured cancer cells, induced by exposure to 13 150 chemicals. We derived a single parameter set that allowed our v-NN method to predict gene expression signatures for 8154 of the 13 150 chemicals, in 10-fold cross validation, with a correlation coefficient of 0.61 between the predicted and experimentally derived gene expression values. We further compared the ability of experimentally derived and v-NN predicted gene expression signatures to identify drugs with the potential to cause human organ injuries and found that the predicted signatures performed at least as well as the experimental signatures. Our work shows that a QSAR algorithm of low computational complexity can serve as a practical approach to predicting chemical-induced genomewide changes in gene expression.

■ MATERIALS AND METHODS

Gene Expression Data. To develop useful QSAR models, data from a large number of compounds are required to train and evaluate the models.[7] To date, the largest gene expression data set derived from a single experimental platform with a consistent protocol is the one generated by the U.S. National Institutes of Health Library of Integrated Network-based Cellular Signatures (LINCS) program (http://www.lincsproject.org/). This data set consists of changes in gene expression levels derived from chemical treatments of different human cell lines across a range of concentrations and treatment durations, typically 5−10 μM for either 6 or 24 h. To reduce costs, the LINCS program only monitored the expression levels of 987 carefully selected genes, using the L1000 bead-based technology. These genes are called “landmark genes”, as the LINCS team uses their expression levels to infer the expression levels of all other genes in the human genome. Thus, far, more than 20 000 compounds have been tested, using multiple replicates in different cell lines. The results of each chemical treatment are expressed as z-values of the landmark genes calculated from their expression levels across all chemical treatments. The set of z-values of the landmark genes derived from a chemical treatment is called the gene expression signature of that treatment. Chen et al. analyzed 475 251 signatures of 11 000 compounds downloaded from the LINCS Web site in September 2013 and found that the likelihood of two compounds with a Tanimoto similarity of at least 0.85 having similar gene expression profiles was 20%.[4] However, the LINCS team has indicated that the signatures in the LINCS data set are not all of the same quality; whereas roughly 50% of the signatures are of high quality (termed the “gold” set), the others are less reproducible.

Because every chemical treatment of a cell line results in a gene expression signature, a single compound may have multiple signatures in the LINCS data set. However, for the purpose of modeling chemical-induced gene expression, each compound should preferably have a single consensus signature. Brueggeman and collaborators recently derived such consensus gene expression signatures from the LINCS data and made them publicly available at https://figshare.com/articles/L1000_Drug_level_Consensus_Expression_Profiles/1476293/2. To derive the consensus signatures, they developed a method in consultation with the L1000 team. We briefly summarize this method below:

(1) Use only the higher-quality gold data and, thereby, remove ∼50% of the gene expression signatures that are nonreproducible or indistinct from the LINCS data set.

(2) For a compound, calculate the Spearman correlation coefficients between the signature of a treatment and the signatures of all other treatments (replicates, doses, treatment durations, and cell types).

(3) Calculate the mean correlation coefficient for each signature and scale the mean correlation coefficients so that they sum to 1. The scaled correlation coefficients are the weights, w_i, for calculating consensus profiles.

(4) To derive a consensus signature for a compound, the z-value of a gene is calculated as a weighted average of the z-values of all k treatments as

$$z \sim \frac{\sum_{i=1}^{k} w_i z_i}{\sqrt{\sum_{i=1}^{k} w_i^{2}}} \quad (1)$$

This is a robust approach for generating consensus signatures, because the contribution of an invidual signature to the consensus signature is scaled by the correlation between it and all other signatures. As a result, the signature most dissimilar to the other signatures contributes the least. Using this method, Brueggeman and colleagues created consensus gene expression signatures for more than 14 000 compounds. We downloaded these consensus signatures and mapped them to well-defined molecular structures for 13 150 compounds. In this study, we used the consensus signatures of these compounds as the “experimental” gene expression data.

Machine Learning Method for Predicting Chemical-Induced Gene Expression. There are multiple mechanisms, most of which are unknown, by which a chemical can perturb the biological state of a living cell and induce changes in gene expression. Without knowing the molecular mechanisms, it is impossible to develop mechanism-based prediction models. The only principle we can rely on in a QSAR study is that similar structures should plausibly elicit similar activities. Thus, we base our prediction approach on information on structurally similar compounds. From this perspective, the k-nearest neighbor (k-NN) method is a reasonable choice, because it always uses information on the k nearest neighbors to make predictions.[8] Mathematically, a k-NN prediction for the value y of a test subject is calculated as

$$y = \frac{\sum_{i=1}^{k} y_i e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{k} e^{-\left(\frac{d_i}{h}\right)^2}} \quad (2)$$

In this equation, y_i is the value of the ith nearest neighbor of the test subject, d_i is the distance between the ith nearest neighbor to the test subject, k is a constant number of nearest neighbors whose information is used in the predictions, and h is a smoothing factor that modulates the distance penalty, i.e., the contributions of distant neighbors to the predictions.

With the k-NN method, once a distance metric is chosen, the only model parameters that need to be determined with training data are k and h. A shortcoming of the method, as applied to chemical activities, is that it always gives a prediction for a compound based on information from k nearest neighbors, irrespective whether the nearest neighbors are structurally similar enough to ensure that their activities are also similar. As an example, consider a training set consisting of congeneric compounds, i.e., molecules belonging to the same structural class (and therefore highly similar). A k-NN prediction for a test compound of the same structural class would be reasonably satisfactory, because there would very likely be near neighbors in the training set with sufficient structural similarity to the test compound to ensure that their activities were also similar. However, if a test compound does not belong to the same structural class and is structurally very different from the training set compounds, the k-NN method would be expected to perform poorly. In this scenario, therefore, no prediction is better than a misleading prediction. However, by design, the k-NN method always gives a prediction no matter how structurally different the nearest neighbors are from the test compound. To correct for this shortcoming, we propose a modification to the method.[6] Instead of using training data to determine k and h, we propose to determine h and a distance threshold d_0. The assumption here is that when the distance between a training compound and a test compound is greater than d_0, the structural similarity is insufficient to ensure that the two compounds have similar activities; therefore, information on the training compound should not be used in making predictions. Thus, instead of using eq 2, we use the following equation to make predictions:

$$y = \frac{\sum_{i=1}^{\nu} y_i e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{\nu} e^{-\left(\frac{d_i}{h}\right)^2}} \qquad (3)$$

where ν is the count of all nearest neighbors that satisfy the condition $d_i \leq d_0$ in the training set. This inevitably leads to a variable number of qualified nearest neighbors for different test compounds. Hence, we denote this construction as the variable nearest neighbor (ν-NN) method. When there is no qualified nearest neighbor to a test compound, the method makes no prediction. Thus, the distance threshold naturally defines an intuitive applicability domain.

We have assessed the performance of the ν-NN method for a range of assay end points with satisfactory results.[6,9−11] In these studies, we examined different distance metrics and found that the Tanimoto distance, derived from the extended connectivity fingerprint with a diameter of 4 chemical bonds (ECFP_4),[12] gives the best results. Therefore, we used the same ν-NN approach to develop prediction models for the chemical-induced gene expression signatures presented here.

■ RESULTS AND DISCUSSION

Dependence of Prediction Performance on d_0 and h. To determine an optimal Tanimoto distance threshold d_0 and smoothing factor h for the ν-NN method, we need to select a performance metric. Owing to the nature of high-throughput measurements, the resulting expression data contain significant experimental variability (noise). It is not uncommon for the results of repeated measurements of a specific gene under the same condition to vary substantially.[13] Hence, the closeness between the predicted and measured values of individual genes is not an ideal performance measure. Instead of the agreement between predicted and consensus values of individual genes, we used the correlation coefficient between the predicted and the consensus z-values of all landmark genes in all compounds as a performance measure. To determine the optimal d_0 and h values, we performed 10-fold cross validation calculations by (1) splitting the consensus signature data set randomly into ten equal-sized groups, (2) making ν-NN predictions for one group of compounds based on the consensus signatures of the other nine groups, (3) repeating the process nine times so that every group was predicted once, and (4) calculating the correlation coefficient between the predicted and experimental consensus signatures of all compounds.

We repeated the 10-fold cross validation calculations 25 times by systematically varying both d_0 and h from 0.1 to 0.5 in 0.1-step increments. Figure 1 shows the resulting correlation coefficients, together with the percentages of compounds for which we could make a prediction (i.e., coverage) as a function of the Tanimoto distance threshold. The best performance was achieved with a Tanimoto distance threshold of 0.1, irrespective of the smoothing factor. However, at this distance threshold, the ν-NN method had the lowest coverage (56%). Although coverage increased with the Tanimoto distance threshold, prediction performance, as measured by the correlation between the predicted and consensus signatures, deteriorated. In selecting optimal values for d_0 and h, we need to balance both the desired performance and the tolerated coverage.

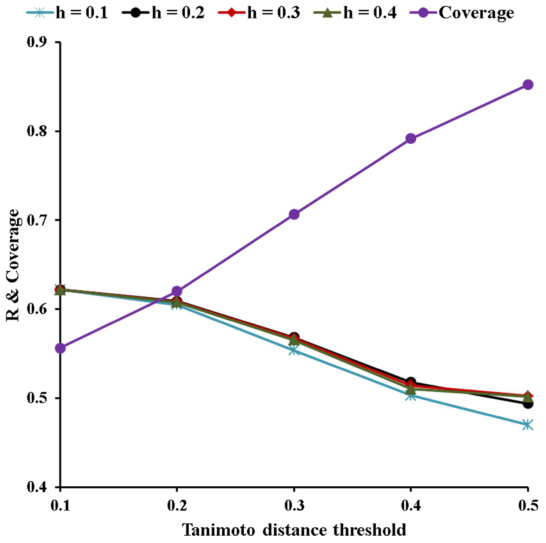


Figure 1. Performance of the ν-NN method, as measured by the correlation (R) between the predicted and consensus signatures of 13 150 compounds calculated with the smoothing factor fixed (0.1, 0.2, 0.3, or 0.4), plotted as a function of the Tanimoto distance threshold. Coverage refers to the percentage of compounds for which ν-NN predictions can be made at a specific Tanimoto distance threshold.

With a Tanimoto distance threshold of 0.1, the highest correlation coefficient between the predicted and consensus signatures achieved by ν-NN was 0.62. To place this value in perspective, we examined the high-throughput data used to generate the consensus signatures and the reproducibility of the underlying experimental data used to create the ν -NN model.

Correlation Between Consensus Signatures Derived from Different Replicate Measurements. We assessed the reproducibility of the consensus signatures by examining compounds with a large number of replicate measurements in the LINCS data set. Subsets of these measurements can be used to create multiple consensus signatures for each of these

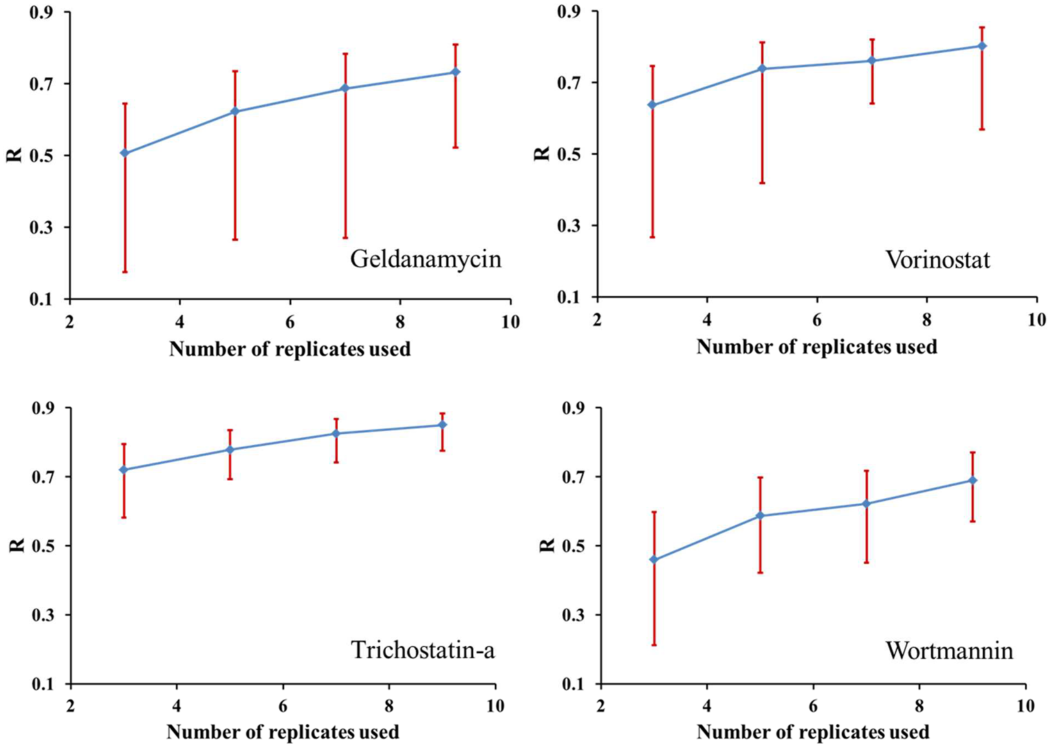


Figure 2. Correlation coefficients between consensus signatures of the same compound but derived from different numbers of replicate measurements following chemical treatment of PC3 cells at 10 μM for 24 h. Only LINCS gold data were used to calculate the correlation coefficients. Blue dots denote mean correlation coefficients, and red vertical bars show the range of correlation coefficients derived from sampling.

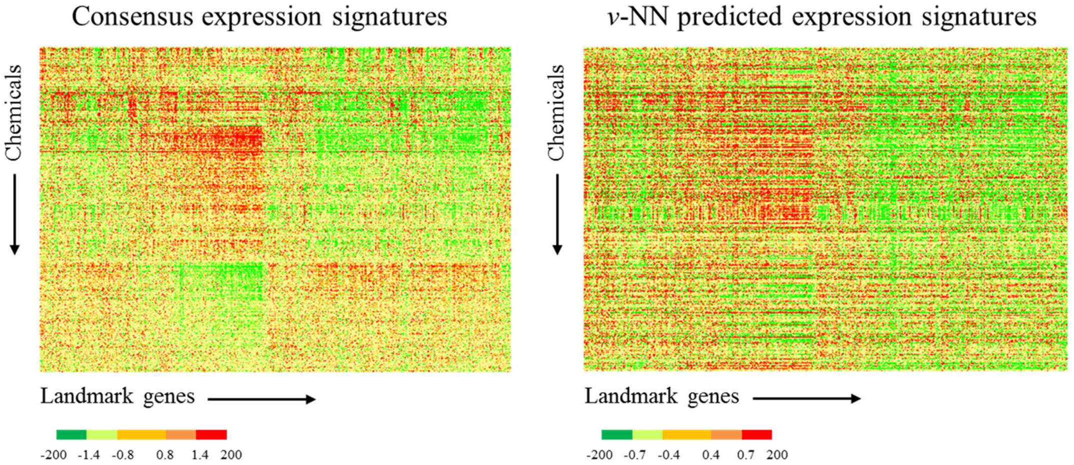


Figure 3. Heat maps of consensus and ν-NN predicted gene expression signatures of 8154 compounds, which show coherence between the consensus and predicted signatures. For purposes of visual comparison, the same hierarchical clustering order for chemicals and signatures derived from the experimental signatures (left) was used to generate the corresponding heat map based on the ν-NN predicated gene signatures (right).

compounds and calculate correlation coefficients between the signatures. This is not feasible for most compounds in the LINCS data set, because under a specific condition (cell line, compound concentration, and treatment duration), the standard protocol calls for three replicates for a given sample. The results of replicate measurements might also have been discarded if they did not satisfy quality control criteria. Thus, most compounds in the LINCS data set have only a small number of replicate measurements. However, the L1000 team selected a small number of compounds whose gene expression is highly reproducible and used these compounds as positive controls. As a result, the gold data set contains a large number of replicates of these compounds. For example, among the treatments applied to PC3 cells at a chemical concentration of 10 μM for 24 h, there are 151 replicates of geldanamycin (LINCS ID: BRD-A19500257), 138 of vorinostat (BRD-K81418486), 135 of trichostatin-a (BRD-A19037878), and 105 of wortmannin (BRD-A75409952). These measurements afford us the opportunity to assess the reproducibility of consensus signatures created from a subset of replicate treatments.

We created consensus signatures for these compounds with an increasing number of randomly selected replicate measurements. We then calculated the mean and range of the pairwise correlation coefficients of the consensus signatures created from a fixed number of replicates. For all four compounds, the mean correlation coefficients between consensus signatures

Table 1. Examples of Consensus and Predicted Signatures (Only the First Five Landmark Genes) Illustrating Different Numerical Values of the Consensus and Predicted Signatures

	consensus signature-derived LINCS gold data					predicted signature based on molecular structure				
drug name	GNPDA1	CDH3	HDAC6	PARP2	MAMLD1	GNPDA1	CDH3	HDAC6	PARP2	MAMLD1
geldanamycin[a]	−6.7	16.1	6.3	−40.0	4.4	2.1	0.1	−5.3	−6.2	−0.2
doxorubicin[b]	−6.5	11.3	5.9	−45.3	4.7	2.9	7.6	8.0	−32.2	8.6
altanserin[c]	2.5	−1.2	0.9	0.4	1.1	−1.8	−0.2	−2.2	−2.0	−1.1
ketanserin[d]	−1.8	−0.2	−2.2	−2.0	−1.1	2.5	−1.2	0.9	0.4	1.1
daunorubicin[e]	−2.0	3.0	2.2	−5.0	−0.5	3.1	7.4	6.2	−29.9	6.6

[a] LINCS sample ID: BRD-A19500257. [b] LINCS sample ID: BRD-A52530684. [c] LINCS sample ID: BRD-K00610438. [d] LINCS sample ID: BRD-K49671696. [e] LINCS sample ID: BRD-A37630846.

created from three replicates ranged from 0.5 to 0.7 and were associated with the largest variability (Figure 2). As the number of replicates increased, the mean correlation coefficient increased with a concomitant reduction in variability. This highlights the importance of replicate measurements for generating statistically significant results. However, most compounds were tested with a limited number of replicates. Even with seven replicates, the correlation coefficient between consensus signatures still ranged between 0.6 and 0.8.

Because these compounds were chosen as positive controls owing to their good signal reproducibility and all replicate measurements were performed using a single compound concentration and the same treatment duration in a single cell line, the signature reproducibility should be higher than that for other compounds in the LINCS data set. Given these considerations, a correlation coefficient of 0.6 between v-NN predicted and consensus signatures of 13 150 compounds (Figure 1), achieved with d_0 and h both set to 0.20, is comparable to the expected experimental reproducibility of the consensus signatures. Thus, as the final v-NN parameters for the study, we set d_0 to 0.20 and h to 0.20. With these parameters, the v-NN method had a coverage of 62% for the 13 150-compound data set and achieved a correlation coefficient of 0.61 between the predicted and consensus signatures (Figure 1).

Figure 3 gives a visual overview of the coherence between the consensus and v-NN predicted gene signatures of 8154 compounds, in the form of clustered heat maps of the experimental values and the corresponding predicted values based on the same clustering order. Gross features of common color clustering patterns were visible, although the correspondence was not perfect, as might be expected from an overall correlation of 0.61.

Performance of Predicted Signatures in Practical Applications. Although the overall correlation between the predicted and consensus signatures was comparable to that between consensus signatures of the same compound derived from repeated experimental measurements, the z-values of individual genes in the predicted and consensus signatures may differ considerably. Table 1 highlights a few extreme examples in which the z-values of specific genes in the predicted versus the consensus signatures markedly differed.

Given the difference in the predicted and the consensus z-values of individual genes, an alternative way of gauging the utility of the predicted signatures is to compare the performance of predicted and consensus signatures in applications that use gene signatures as input to infer a biological effect. We recently evaluated the feasibility of using gene expression signatures to identify chemicals with the potential to cause distinct and specific human organ injuries.[14] Briefly, we implemented the following procedure:

(1) We calculated the median z-values for LINCS signatures derived from chemical treatment of VCaP cells as the consensus signatures of the chemicals.

(2) For each compound, we selected the 50 landmark genes with the highest median z-values as the genes upregulated by the chemical and the 50 landmark genes with the lowest median z-values as the genes downregulated by the chemical. We considered the remaining landmark genes to be unperturbed by exposure to the compound.

(3) We identified drugs in the data set that were reported to cause specific human liver, heart, and kidney injuries upon chronic use.

(4) We used the drugs causing human organ injuries as positives and the remaining chemicals in the LINCS VCaP data set as baseline samples and developed Bayesian models to score the chemicals in terms of their potential to cause the same organ injuries. Extensive cross-validation analyses indicated that the models developed from the gene expression signatures performed satisfactorily.

To evaluate the predicted gene expression signatures, we built drug-induced human organ injury models based on the consensus signatures, and compared the performance of the predicted and consensus signatures in identifying drugs known to cause these injuries. Details of the evaluation and results are as follows:

(1) Among the consensus signatures, we identified 58 drugs with a moderate to high risk of inducing long QT syndrome from peer-reviewed publications,[15−17] 116 drugs that potentially cause liver cholestasis according to the SIDER[18] and OFFSIDE[19] databases, and 109 drugs with the potential to cause interstitial nephritis from a peer-reviewed publication.[20]

(2) We made v-NN prediction of gene expression signatures for these drugs based on the consensus signatures, with the model parameters d_0 and h each set to 0.20. Because some of the drugs did not have qualified nearest neighbors, we could only make predictions for 23 of the 58 drugs causing long QT syndrome, 43 of the 116 drugs causing cholestasis, and 39 of the 109 drugs causing nephritis.

(3) We converted both the predicted and consensus signatures into up- and down-regulated genes for each compound. The 50 genes with the highest z-values were considered as up-regulated, and the 50 genes with the

lowest z-values were considered as down-regulated. We considered the remaining genes as unchanged.

(4) To build a drug-induced organ injury model, we separated the drugs causing an organ injury into two groups: a group without predicted signatures to be used as training set positives and a group with both predicted and consensus signatures to be used as test set positives for comparing signature performance. We then randomly split the remaining compounds in the consensus signature data set, with 60% of the compounds to be used as training set baseline samples and 40% as test set baseline samples. We used consensus signatures as descriptors and developed Bayesian models for predicting compounds with the potential to cause the organ injuries. We then used the models to examine the enrichment of positive drugs in the test set based on their consensus and predicted signatures separately. We also used the models and signatures to calculate areas under the receiver operating characteristic curves (AUCs). Figure 4 shows how we split the data set for model building and signature performance evaluation. We refer the reader to our recent publication for additional details on using chemical genomic signatures for Bayesian modeling.[14]

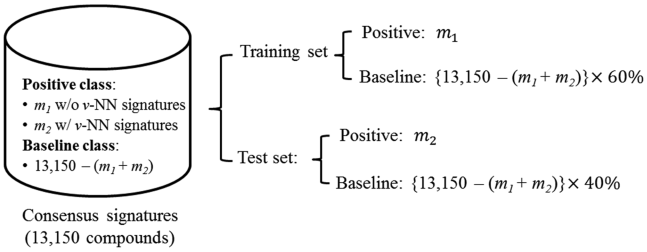

Positive class:
• m_1 w/o ν-NN signatures
• m_2 w/ ν-NN signatures
Baseline class:
• 13,150 − (m_1 + m_2)
Consensus signatures
(13,150 compounds)
Training set
Positive: m_1
Baseline: {13,150 − (m_1 + m_2)} × 60%
Test set:
Positive: m_2
Baseline: {13,150 − (m_1 + m_2)} × 40%

Figure 4. Schematic showing the segregation of the consensus signature data set of 13 150 compounds into a training set for developing a Bayesian model of drug-induced organ injury and a test set for comparing the performance of ν-NN predicted signatures to that of consensus signatures.

Figure 5 presents enrichment curves of the predicted and consensus signatures for drugs inducing long QT syndrome, cholestasis, and nephritis, as well as the AUCs calculated from the consensus and predicted signatures. The diagonal blue lines represent the performance of completely random selection of samples. The black and red curves represent the performance of the models using the consensus and predicted signatures, respectively. The greater the AUC values, the better the signatures are at identifying drugs with potential to cause the relevant organ injuries. The results indicated that the predicted signatures performed at least as well as the consensus signatures derived from experimental measurements.

A closer examination revealed that the predicted signatures appeared to outperform slightly the consensus signatures in two of the three cases. This was a counterintuitive result, given that predicted data normally provide results that are less accurate than the experimental data on which they are based. This prompted us to examine ν-NN predicted signatures in greater detail.

Of the 23 long QT-inducing drugs with ν-NN predicted signatures, 15 had one qualified nearest neighbor in the consensus signature data set and 8 had more than one near neighbor. Similarly, of the 43 cholestasis-inducing drugs with ν-NN predicted signatures, 26 were predicted with one qualified near neighbor and 17 with more than one qualified neighbor. Finally, of the 39 nephritis-inducing drugs with ν-NN predicted signatures, 26 were predicted with one qualified neighbor and 13 with more than one qualified neighbor.

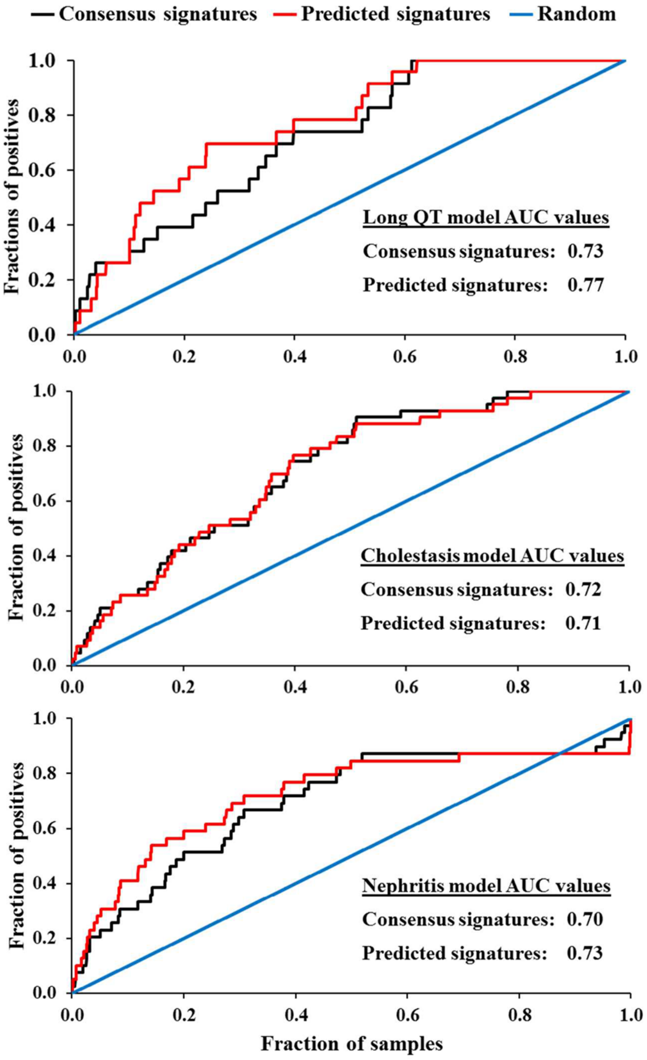


Figure 5. Enrichment curves for models of drug-induced long QT syndrome (top), liver cholestasis (middle), and interstitial nephritis (bottom) calculated with the consensus signatures (black curve) and ν-NN predicted signatures (red curve). The graphs show that the performance obtained with the predicted signatures is similar to that obtained with the consensus signatures derived from experimental measurements. Area under the curve (AUC).

Because ν-NN predictions are based on information from qualified near neighbors, the predicted results are likely more reliable with a greater number of qualified near neighbors. To test this hypothesis, we separated the drugs with a single qualified near neighbor from those with more than one qualified near neighbor. We then recalculated the enrichment curves and AUCs for the drugs using the ν-NN predicted and consensus signatures.

Figures 6, 7, and 8 show the resulting enrichment curves and AUC values for the models of drugs inducing long QT syndrome, liver cholestasis, and interstitial nephritis, respectively. In identifying drugs with potential to induce long QT syndrome, ν-NN signatures predicted from one qualified near neighbor performed comparably to the consensus signatures

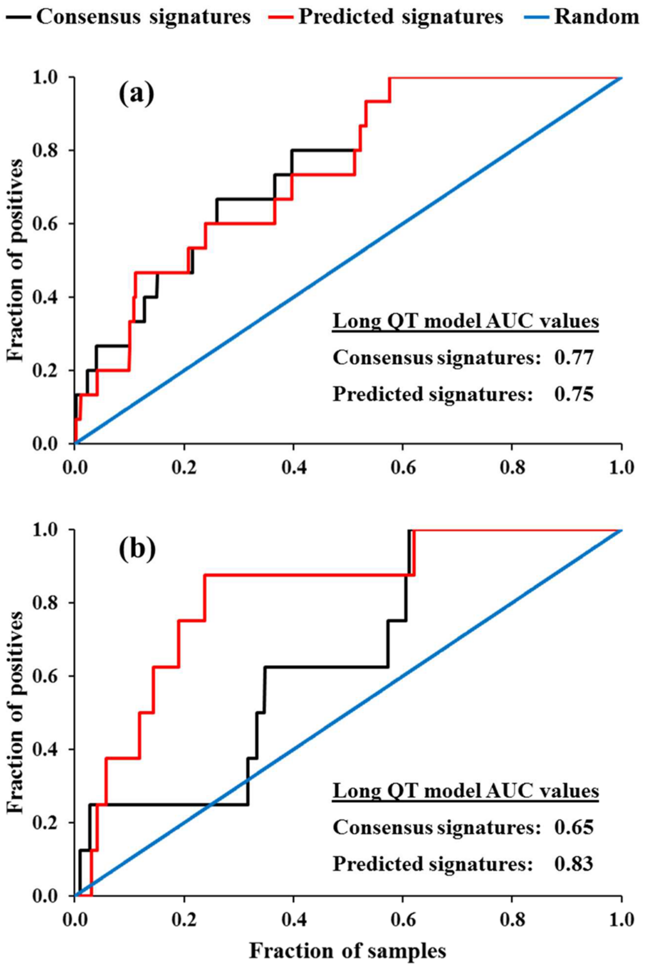


Figure 6. (a) Enrichment curves and areas under the receiver operating characteristic curve (AUC values) of the drug-induced long QT model calculated with consensus signatures (black curve) and *ν*-NN signatures predicted with one qualified near neighbor (red curve). (b) Enrichment curves and AUC values of the drug-induced long QT model calculated with consensus signatures (black curve) and *ν*-NN signatures predicted with more than one qualified near neighbor (red curve).

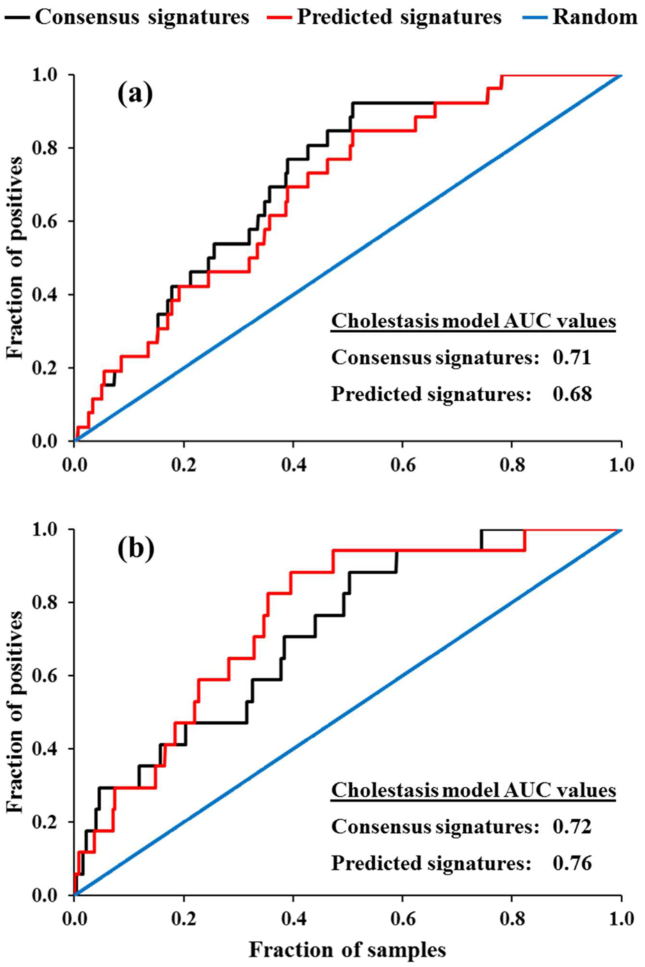


Figure 7. (a) Enrichment curves and areas under the receiver operating characteristic curve (AUC values) of the drug-induced liver cholestasis model calculated with consensus signatures (black curve) and *ν*-NN signatures predicted with one qualified near neighbor (red curve). (b) Enrichment curves and AUC values of the drug-induced liver cholestasis model calculated with consensus signatures (black curve) and *ν*-NN signatures predicted with more than one qualified near neighbor (red curve).

(Figure 6a). In contrast, *ν*-NN signatures predicted from more than one qualified near neighbor performed significantly better than the consensus signatures (Figure 6b). In identifying drugs with the potential to induce liver cholestasis, *ν*-NN signatures predicted from one qualified near neighbor performed slightly worse than the consensus signatures (Figure 7a), whereas those predicted from more than one qualified near neighbor performed slightly better than the consensus signatures (Figure 7b). Finally, in identifying drugs with potential to induce interstitial nephritis, *ν*-NN signatures predicted from one qualified near neighbor performed nearly identically to the consensus signatures (Figure 8a), and those predicted from more than one qualified near neighbor performed slightly better than the consensus signatures (Figure 8b).

In all three examples, the *ν*-NN signatures predicted from one qualified near neighbor were comparable in quality to consensus signatures derived from experimental measurements, whereas those predicted from more than one qualified near neighbor showed better performance than the experimentally derived consensus signatures. Overall, the quality of *ν*-NN predicted signatures was at least as good, if not better, than that of the consensus signatures.

It may appear counterintuitive that *ν*-NN signatures predicted with information from more than one qualified near neighbor are better than consensus signatures derived from experimental measurements. As shown in Figure 2, owing to variability in the high-throughput experimental data, a fair number of replicates are needed to derive statistically reliable results. If we assume that each consensus signature was generated from three replicates on average and a *ν*-NN signature was predicted using three qualified near neighbors, then the *ν*-NN signature would have been based on information from nine replicates. Thus, increasing the number of replicates improves not only the statistical significance of the consensus signatures, but also the quality of *ν*-NN signatures predicted with an increasing number of qualified near neighbors.

Efficiency of *ν*-NN Method in Modeling Chemical-Induced Gene Expression Changes. Compared to most other machine learning methods, the distance-weighted *ν*-NN method is highly efficient. Using a single Intel Xeon E5-2665 processor E5-2665 (2.40 GHz), the method required 39 min to

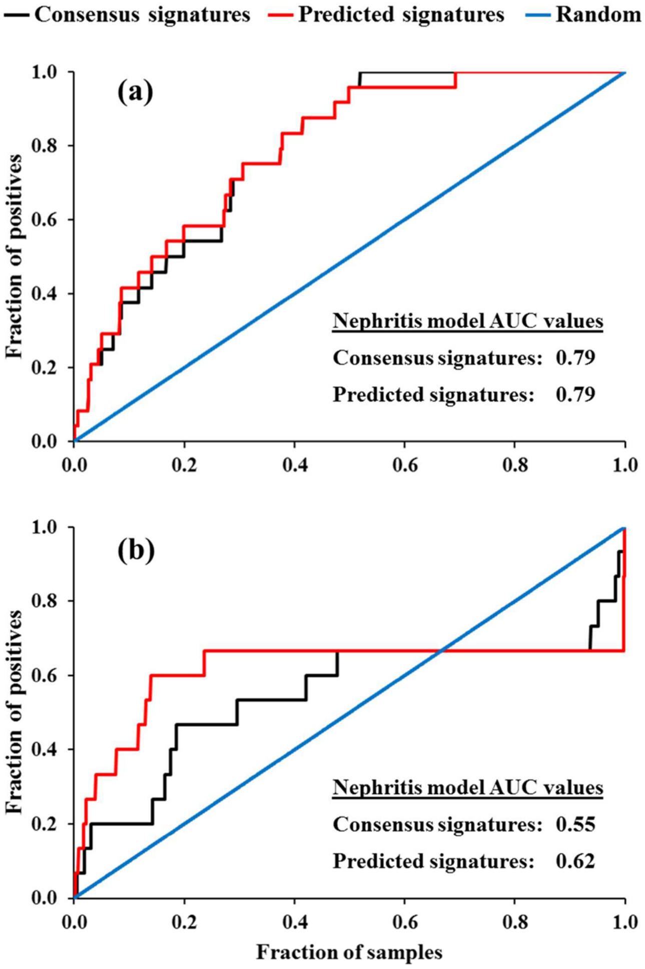


Figure 8. (a) Enrichment curves and areas under the receiver operating characteristic curve (AUC values) of the drug-induced interstitial nephritis model calculated with consensus signatures (black curve) and *v*-NN signatures predicted with one qualified near neighbor (red curve). (b) Enrichment curves and AUC values of the drug-induced interstitial nephritis model calculated with consensus signatures (black curve) and *v*-NN signatures predicted with more than one qualified near neighbor (red curve).

generate *v*-NN predictions for 13,150 compounds. The total computing time, using the same single processor, for running 25 10-fold cross validation steps with d_0 and h systematically varied from 0.1 to 0.5 in 0.1-step increments was 23.7 h. This calculation generated the data for Figure 1 and was the basis of our selection of model parameters (d_0 = 0.2 and h = 0.2). Unlike most other machine learning methods, the *v*-NN method does not build any static models. Furthermore, it uses whatever data are available to make predictions. Thus, as new data are generated from experimental measurements and made available in a repository, *v*-NN predictions can use them without retraining any model. This is a significant advantage, especially for modeling many properties simultaneously, because of the effort involved in training and retraining a large number of models. For example, in modeling the same 978 landmark genes based on a data set of 175 compounds, Hall et al. created ~600 000 static models and evaluated their performance to select the final ~20 000 models.[2] Although the cost of digital data storage has declined steadily over the years, storing and retrieving ~600 000 static models may still be a challenge for most computers. Retraining ~600 000 models periodically with increasing amounts of data from an increasing number of compounds would be even more challenging.

Applicability of *v*-NN to Drug-Centered Biomedical Research. The ability of the *v*-NN method to predict gene expression signatures depends on the existence of qualified near neighbors with experimentally derived gene expression signatures. To assess the applicability of the method to drug-like space, we downloaded the molecular structures of all approved drugs (including those withdrawn from the market) and investigational drugs in DrugBank on June 29, 2017. We standardized the structures by retaining the largest disconnected fragments (thus removing counterions in salts), protonating acids, deprotonating bases, and then removing all replicate structures. We collected a total of 2491 structurally unique molecules. Using a Tanimoto distance threshold of 0.20, the *v*-NN prediction model made predictions for 1266 of them—representing a coverage of roughly 51% of the drug-like space. This coverage would be higher if approved inorganic compounds, such as oxygen (O_2), nitrogen (N_2), carbon monoxide (CO), carbon dioxide (CO_2), and potassium chloride (KCl), were excluded. In addition, the LINCS data used in this study were all obtained experimentally a few years ago. As additional LINCS data are released, the coverage will naturally increase. In our previous studies, we have demonstrated that not only coverage of *v*-NN predictions but also prediction performance improves with increasing data set size.[10]

SUMMARY

In this study, we evaluated the feasibility of using a distance-weighted nearest neighbor method to predict chemical-induced genomewide changes in gene expression. Building on the principle that similar structures have similar activities, the method uses information on only qualified neighbors to make predictions, and in so doing defines a natural and intuitive applicability domain. Computationally, the method is highly efficient because it employs only two adjustable model parameters that can be optimized based on the original training data set. The method can take advantage of up-to-date experimental data, as soon as they are made available and incorporated into the data set used by the method, without the need to retrain any model. This makes it different from most other machine learning methods that require static models to be created, stored, and periodically retrained to take advantage of data derived from new experimental measurements.

Using the consensus gene expression signatures of over 13 000 compounds derived from the LINCS program, we demonstrated that the signatures predicted from one qualified near neighbor are similar in quality to experimentally derived consensus signatures, whereas those predicted from more than one qualified near neighbor are generally better than the consensus signatures. Hence, the distance-weighted variable nearest neighbor method provides a practical approach to predicting chemical-induced gene expression, in terms of both the reliability of the results and the computational resources required.

Our results also suggest that the conclusion of Chen et al., that the likelihood of two compounds with a Tanimoto similarity of at least 0.85 showing similar gene expression profiles is only 20%,[4] is at least partly compromised by the variability of data in the full LINCS data set, which contains both reproducible (high-quality) and nonreproducible (low-quality) data. A comparison of the performance of *v*-NN signatures and that of consensus signatures derived from the gold set of LINCS data indicates a much higher probability of two compounds having similar gene expression profiles when their Tanimoto similarity is at least 0.80.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00281.

> List of drugs reported to cause long QT syndrome, liver cholestasis, and interstitial nephritis. We compared the performance of the predicted and consensus gene expression signatures of these drugs in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: rliu@bhsai.org (R.L.).
*E-mail: sven.a.wallqvist.civ@mail.mil (A.W.).

ORCID

Ruifeng Liu: 0000-0001-7582-9217

Funding

The research was supported by the U.S. Army Medical Research and Materiel Command (Ft. Detrick, MD) as part of the U.S. Army's Network Science Initiative, and by the Defense Threat Reduction Agency (Grant No. CBCall14-CBS-05-2-0007).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Dr. Tatsuya Oyama in editing the manuscript. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

■ ABBREVIATIONS

QSAR = quantitative structure activity relationship
ν-NN = variable nearest neighbor method
LINCS = library of integrated network-based cellular signatures
AUC = area under receiver operating characteristic curve

■ REFERENCES

(1) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. Qsar Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(2) Hall, M. L.; Calkins, D.; Sherman, W. Automated Protocol for Large-Scale Modeling of Gene Expression Data. *J. Chem. Inf. Model.* **2016**, *56*, 2216−2224.

(3) Duan, Q.; Flynn, C.; Niepel, M.; Hafner, M.; Muhlich, J. L.; Fernandez, N. F.; Rouillard, A. D.; Tan, C. M.; Chen, E. Y.; Golub, T. R.; et al. A LINCS Canvas Browser: Interactive Web App to Query, Browse and Interrogate Lincs L1000 Gene Expression Signatures. *Nucleic Acids Res.* **2014**, *42*, W449−460.

(4) Chen, B.; Greenside, P.; Paik, H.; Sirota, M.; Hadley, D.; Butte, A. J. Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data across 11,000 Compounds. *CPT: Pharmacometrics Syst. Pharmacol.* **2015**, *4*, 576−584.

(5) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(6) Liu, R.; Tawa, G.; Wallqvist, A. Locally Weighted Learning Methods for Predicting Dose-Dependent Toxicity with Application to the Human Maximum Recommended Daily Dose. *Chem. Res. Toxicol.* **2012**, *25*, 2216−2226.

(7) Tropsha, A. Best Practices for Qsar Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(8) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175−185.

(9) Liu, R.; Wallqvist, A. Merging Applicability Domains for in Silico Assessment of Chemical Mutagenicity. *J. Chem. Inf. Model.* **2014**, *54*, 793−800.

(10) Liu, R.; Schyman, P.; Wallqvist, A. Critically Assessing the Predictive Power of Qsar Models for Human Liver Microsomal Stability. *J. Chem. Inf. Model.* **2015**, *55*, 1566−1575.

(11) Schyman, P.; Liu, R.; Wallqvist, A. Using the Variable Nearest Neighbor Method to Identify P-Glycoprotein Substrates and Inhibitors. *ACS Omega* **2016**, *1*, 923−929.

(12) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(13) Draghici, S.; Khatri, P.; Eklund, A. C.; Szallasi, Z. Reliability and Reproducibility Issues in DNA Microarray Measurements. *Trends Genet.* **2006**, *22*, 101−109.

(14) Liu, R.; Yu, X.; Wallqvist, A. Using Chemical-Induced Gene Expression in Cultured Human Cells to Predict Chemical Toxicity. *Chem. Res. Toxicol.* **2016**, *29*, 1883−1893.

(15) Barnes, B. J.; Hollands, J. M. Drug-Induced Arrhythmias. *Crit. Care Med.* **2010**, *38*, S188−197.

(16) Behr, E. R.; Roden, D. Drug-Induced Arrhythmia: Pharmacogenomic Prescribing? *Eur. Heart J.* **2013**, *34*, 89−95.

(17) Viskin, S.; Justo, D.; Halkin, A.; Zeltser, D. Long Qt Syndrome Caused by Noncardiac Drugs. *Prog. Cardiovasc. Dis.* **2003**, *45*, 415−427.

(18) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The Sider Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44*, D1075−1079.

(19) Tatonetti, N. P.; Ye, P. P.; Daneshjou, R.; Altman, R. B. Data-Driven Prediction of Drug Effects and Interactions. *Sci. Transl. Med.* **2012**, *4*, 125ra31.

(20) Naughton, C. A. Drug-Induced Nephrotoxicity. *Am. Fam. Physician* **2008**, *78*, 743−750.