Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

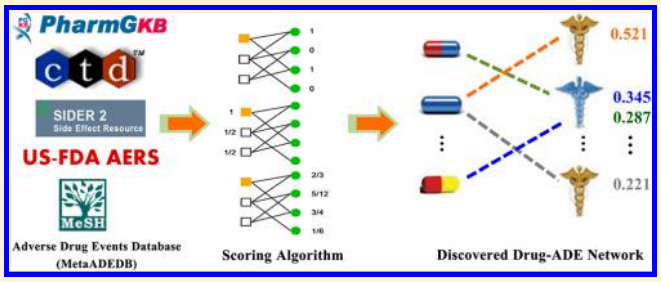
Adverse Drug Events: Database Construction and in Silico Prediction

Feixiong Cheng,[†] Weihua Li,[†] Xichuan Wang,[‡] Yadi Zhou,[†] Zengrui Wu,[†] Jie Shen,[†] and Yun Tang*,[†]

[†]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

[‡]Department of Surgery, Shanghai MCC Hospital, 456 Chunlei Road, Shanghai 200941, China

S Supporting Information



ABSTRACT: Adverse drug events (ADEs) are the harms associated with uses of given medications at normal dosages, which are crucial for a drug to be approved in clinical use or continue to stay on the market. Many ADEs are not identified in trials until the drug is approved for clinical use, which results in adverse morbidity and mortality. To date, millions of ADEs have been reported around the world. Methods to avoid or reduce ADEs are an important issue for drug discovery and development. Here, we reported a comprehensive database of adverse drug events (namely MetaADEDB), which included more than 520 000 drug−ADE associations among 3059 unique compounds (including 1330 drugs) and 13 200 ADE items by data integration and text mining. All compounds and ADEs were annotated with the most commonly used concepts defined in Medical Subject Headings (MeSH). Meanwhile, a computational method, namely the phenotypic network inference model (PNIM), was developed for prediction of potential ADEs based on the database. The area under the receive operating characteristic curve (AUC) is more than 0.9 by 10-fold cross validation, while the AUC value was 0.912 for an external validation set extracted from the US-FDA Adverse Events Reporting System, which indicated that the prediction capability of the method was reliable. MetaADEDB is accessible free of charge at http://www.lmmd.org/online_services/metaadedb/. The database and the method provide us a useful tool to search for known side effects or predict potential side effects for a given drug or compound.

■ INTRODUCTION

Adverse drug events (ADEs, also known as drug side effects) are important human phenotypic resources, which measure the harms associated with uses of given medications at normal dosages. Every year, thousands of people were reported to die from serious ADEs around the world. For example, there are about two million serious ADEs reported per year in the United States, resulting in about 100 000 deaths.[1] Therefore, ADEs have resulted in serious social and economic problems, and it is urgent to detect or determine ADEs before medication is used.

In order to avoid potential ADEs from taking place, it is necessary to collect high quality reported ADEs first. To date, several ADE-associated databases have been constructed.[2−4] Kuhn et al. developed a computer-readable side effect resource (SIDER) that connects 888 drugs to 1450 side effect items.[3] Davis et al. developed the comparative toxicogenomics database (CTD) which is a public resource of expanded chemical−gene−disease associations data.[4] Tatonetti et al. presented a comprehensive database of drug side effects (OFFSIDES) and a database of drug−drug interaction side effects (TWOSIDES).[2] Though SIDER, CTD, and OFFSIDES provide useful sources for obtaining human phenotypic information, the same drug or side effect is often expressed in different forms from each other. For example, "noninsulin dependent diabetes", "type 2 diabetes", and "NIDDM" mean the same disease, but these are expressed differently in the three databases.[5,6]

Besides known ADEs, it is also very important to identify potential ADEs for an approved drug in drug discovery and postmarketing surveillance. A variety of experimental approaches have been developed to detect ADEs. For example, Ingelman-Sundberg reviewed the recent progress of pharmacogenomic biomarkers in detection of severe adverse drug reactions.[7] However, experimental detection of ADEs is costly and laborious. In particular, many ADEs are not detectable in clinical trials until the drug was approved for clinic usage.[2] It is necessary to develop new alternative or supplementary methods, such as computational methods for prediction of drug side effects at the stages of the drug development process.

In recent years several computational methods have been developed for ADE prediction, in order to reduce drug-related morbidity and mortality.[8−21] Cami et al. developed an approach called predictive pharmacosafety networks (PPNs).[22] They first collected 809 drugs and 852 related adverse events on a widely used drug safety database and constructed a drug−ADE network to build the logistic regression predictive model. The model achieved an area under the receiver operating characteristic curve (AUC) of 0.87, with a sensitivity of 0.42 and a specificity of 0.95 when predicting new drug−ADE associations appeared in the external validation set. Liu et al. proposed a

Received: January 4, 2013
Published: March 22, 2013

dx.doi.org/10.1021/ci4000079 | J. Chem. Inf. Model. 2013, 53, 744−752

machine learning-based method for prediction of ADEs by integrating the phenotypic characteristics of a drug, drug's chemical and biological properties, protein targets, and pathways information.[17] Yamanishi et al. built several extensions of a kernel regression model by integrating the chemical space of drug chemical structures and biological space of drug target protein for side effect prediction.[23] Tatonetti et al. developed a signal detection algorithm for identifying hidden drug−drug interactions and related side effects.[24] Several similar signal detection methods were also reported for prediction of ADEs.[25,26] Therefore, computational methods and models demonstrated the potential value and application in side effect prediction.

In this study, a comprehensive computer-readable drug adverse events database, named MetaADEDB, was constructed on the basis of three databases SIDER, CTD, and OFFSIDES, utilizing Medical Subject Headings (MeSH) to annotate all compounds and diseases systematically. To facilitate linking to other databases and reuse for academic research, all drugs and their Anatomical Therapeutic Chemical Classification System (ATC) codes were mapped with DrugBank[27] and Therapeutic Target Database (TTD).[28] Furthermore, a phenotypic network inference model (PNIM) was developed to predict potential ADEs with high accuracy (Figure 1). This work provided a computational framework and useful tool for side effect prediction in drug discovery.

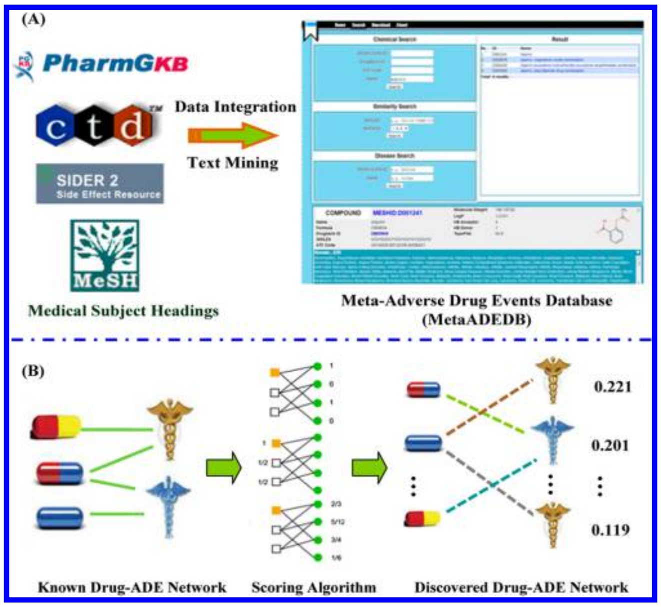


Figure 1. Workflow of construction of meta adverse drug events database, named MetaADEDB (**A**), development and application of the phenotypic network inference model for new drug side effect prediction (**B**).

■ METHODS AND MATERIALS

Data Source. The original data were downloaded from the Web sites of three ADE associated databases: CTD,[4] SIDER[3] (version 2), and OFFSIDES[2,29] (http://www.pharmgkb.org/downloads.jsp). Only data points with experimental evidence were collected. All compounds and ADE items were annotated with the most commonly used MeSH vocabularies (2012 release, xml format) downloaded from the Web site of National Library of Medicine (http://www.nlm.nih.gov/mesh/gcm.html). All compounds in MetaADEDB were mapped with those in DrugBank[27] and TTD[28] in order to facilitate linking to other databases and reuse for academic research. At last, the duplicated compound−ADE associations in MetaADEDB were excluded.

Data Management and Implementation. *Physicochemical Property Filtering.* In MetaADEDB, five classic physicochemical properties, namely numbers of hydrogen bond acceptor and donor, Log P, TopoPSA, and molecular weight, were calculated for each drug using OpenBabel v2.3.1.[30]

Similarity Search. The structural similarity search is assessed by the Tanimoto coefficient using the MACCS keys implemented with OpenBabel v2.3.1.[30]

Visualization Features. The two-dimensional (2D) chemical structures were generated using Marvin v5.8.0 (http://www.chemaxon.com/). The database browser was organized using a Python script and cascading style sheet (CSS).

Implementation. The MetaADEDB system was built with Django v1.4.0 on Apache v2.2.20 with mod_wsgi v3.3 and MySQL v5.1.61, installed on Ubuntu-Server v11.10. A detailed description can be found in our previous work.[31]

Construction of Drug−ADE Association Network. A bipartite network was built to represent the associations of drugs and ADEs. Denoting the drug set as $D = \{d_1, d_2, ..., d_n\}$, the ADE set as $P = \{p_1, p_2, ..., p_l\}$, the drug−ADE binary pairs were then described as a bipartite drug−ADE graph G (D, P, E), where $E = \{e_{ij}: d_i \in D, p_j \in P\}$. An edge was drawn between a drug and an ADE if there was a reported side effect annotated in MetaADEDB. The bipartite network of drug−ADE associations could be represented as an $n \times m$ adjacent matrix $\{a_{ij}\}$, where $a_{ij} = 1$ represents a reported side effect; otherwise, $a_{ij} = 0$.

Measurement of Network Topological Feature. Three classical topological features of complex network, namely connectivity (k), clustering coefficient (C), and betweenness (B) were systemically investigated using the toolbox of NetworkX version 1.6 (http://networkx.lanl.gov/).

Connectivity. The connectivity (k), or degree, measures how many edges of a node connecting to other nodes in the network. In the bipartite networks of CTD, SIDER, OFFSIDES, and MetaADEDB, each network can be represented by an adjacency matrix **A**, where $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ if there does not exist an edge between nodes i and j. The connectivity (k_i) of node i is calculated as follows:

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{1}$$

The connectivity distribution, $P(k)$ denotes the fraction that a given node has exactly k degree.[32]

Clustering Coefficient. The clustering coefficient denotes the degree of interconnectivity in the neighborhood of a node. The average cluster coefficient $\langle C \rangle$ represents the overall tendency of nodes to form clusters. The bipartite clustering coefficient is a measure of local density of links defined as follows:[33]

$$C_u = \frac{\sum_{v \in N(N(v))} C_{uv}}{|N(N(u))|} \tag{2}$$

Where $N(N(u))$ are the second-order neighbors of u in N excluding u, and C_{uv} is the pairwise clustering coefficient between node u and v.

Betweenness. The betweenness (B) is a measurement of a node's centrality in a network. It denotes the ratio of all of the shortest paths between all nodes in a network that pass through a given node. For a node x, B is computed by taking the sum of the number of shortest paths between pairs of nodes that pass

Table 1. Statistics of Adverse Drug Event Association Pairs in Three Different Published Databases and Our Curated Database[a]

database	N_{Drug}	N_{ADE}	N_{A}
CTD	2406	2789	52987
SIDER	945	4026	96561
OFFSIDES	1180	9905	405493
MetaADEDB	3059	13255	527216

[a] N_{Drug}: the number of drugs. N_{ADE}: the number of adverse drug events. N_{A}: the number of drug−ADE association pairs.

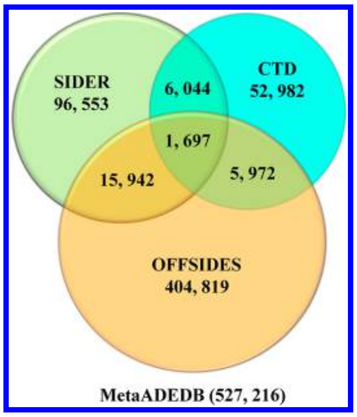


Figure 2. Detailed coverage of chemical and ADE terms in three databases, namely SIDER, CTD, and OFFSIDES, and our MetaADEDB.

Table 2. Average Topological Properties of Drugs versus All Adverse Drug Events (ADEs) in Four Different Drug−ADE Association Bipartite Networks

networks	topological metrics	mean (drugs)	mean (ADEs)
CTD	connectivity	22.021	19.004
	clustering	0.055	0.056
	betweenness	5.19×10^{-4}	4.21×10^{-4}
SIDER	connectivity	102.173	24.006
	clustering	0.108	0.053
	betweenness	1.56×10^{-3}	2.13×10^{-4}
OFFSIDES	connectivity	343.067	40.899
	clustering	0.068	0.037
	betweenness	1.26×10^{-3}	7.01×10^{-5}
MetaADEDB	connectivity	172.293	39.772
	clustering	0.037	0.034
	betweenness	4.78×10^{-4}	6.33×10^{-5}

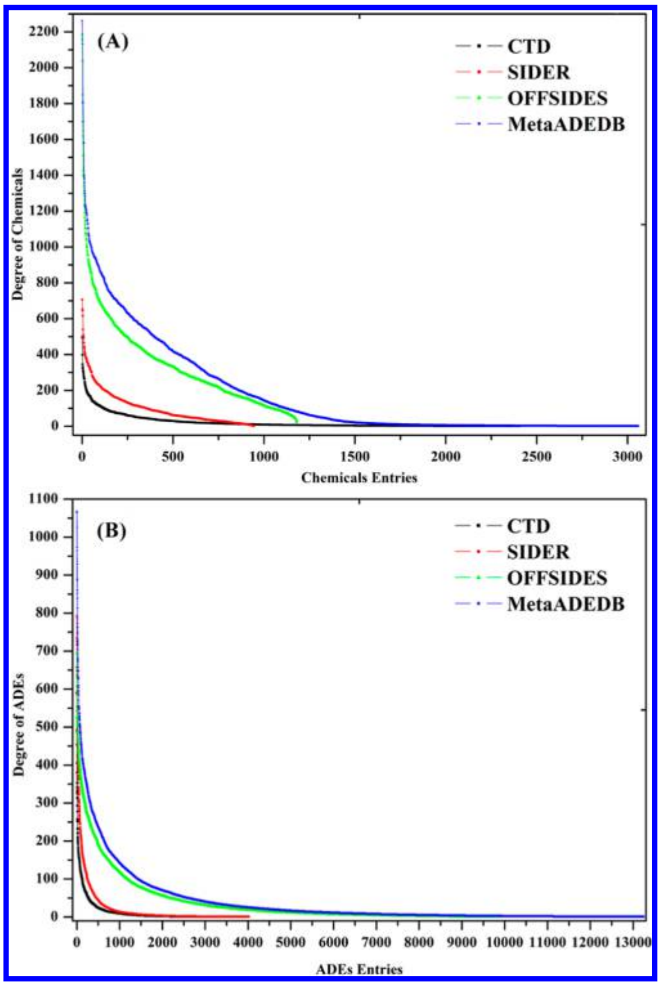


Figure 3. Degree distribution of four different drug−ADE association networks. (**A**) Degree distribution of drug nodes in four different drug−ADE association networks. (**B**) Similar to A, the degree distribution of ADE nodes.

through node x divided by the total number of shortest paths between pairs of nodes.

Development of Phenotypic Network Inference Model (PNIM). On the basis of the above bipartite drug−ADE network of $G(D, P, E)$, the PNIM was built to predict new drug−ADE associations using our previously developed NBI algorithm.[34,35] As shown in Figure 1, for a given drug node d_i in $G(D, P, E)$, supposing that a kind of resource is initially located in the ADEs interacted with d_i, the resource will diffuse to all ADEs in the drug−ADE network after network-based resource allocation process. Each ADE node averagely distributes its resource to all neighboring drug nodes and then each drug node redistributes the received resource to all neighboring ADE nodes. The final resource in ADE nodes that are not connected with the drug node c_i could be considered as the ranking score of each ADE, and the ADEs with higher ranking score are more likely to associate with drug d_i.

Denoting $B_{0n\times m}$ as the initial resource and $B_{0ij} = a_{ij}$, $R_{n\times n}$ as the total resource (degree) of each drug node, a_{ij} as the initial resource between d_i and p_j and $\mathbf{R} = \mathrm{diag}(\sum_{j=1}^{m} a_{1j}, \sum_{j=1}^{m} a_{2j}, ..., \sum_{j=1}^{m} a_{nj})$, $H_{m\times m}$ as the total resource (degree) of each ADE node, and $\mathbf{H} = \mathrm{diag}(\sum_{i=1}^{n} a_{i1}, \sum_{i=1}^{n} a_{i2}, ..., \sum_{i=1}^{n} a_{im})$, the resource matrix will be obtained as $B_{1n\times m}$, and $B_1 = B_0 W_{m\times m}$ or $B_1^{\mathrm{T}} = B_0^{\mathrm{T}} W_{n\times n}$, where transfer matrix $W_{m\times m} = (B_0\mathbf{H}^{-1})^{\mathrm{T}}(\mathbf{R}^{-1}B_0)$ or $W_{n\times n} = (\mathbf{R}^{-1}B_0)(B_0\mathbf{H}^{-1})^{\mathrm{T}}$.

Performance Assessment. *Cross-validation.* To test the performance of the method, a 10-fold cross-validation technique was used and each result was repeated by 10 independent simulated times. For each data set, all the drug−ADE binary pairs were randomly divided into 10 parts with equal size, respectively. Each part was taken in turn as test set, while the remaining nine parts were served as the training set.

External Validation. In order to test the performance of our method on the external validation set, a new data set was collected from US-FDA Adverse Events Reporting System (AERS). The third quarter of 2011 AERS reports (file name: aers_ascii_2011q3.zip) was downloaded as the external validation set from the Web site: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveill ance/AdverseDrugEffects/ucm083765.htm. In the original file, every ADE report was given an ISR ID, and the detailed aspects of the reports were stored in several files separately. The drugs labeled as primary suspect drug (PS) and secondary

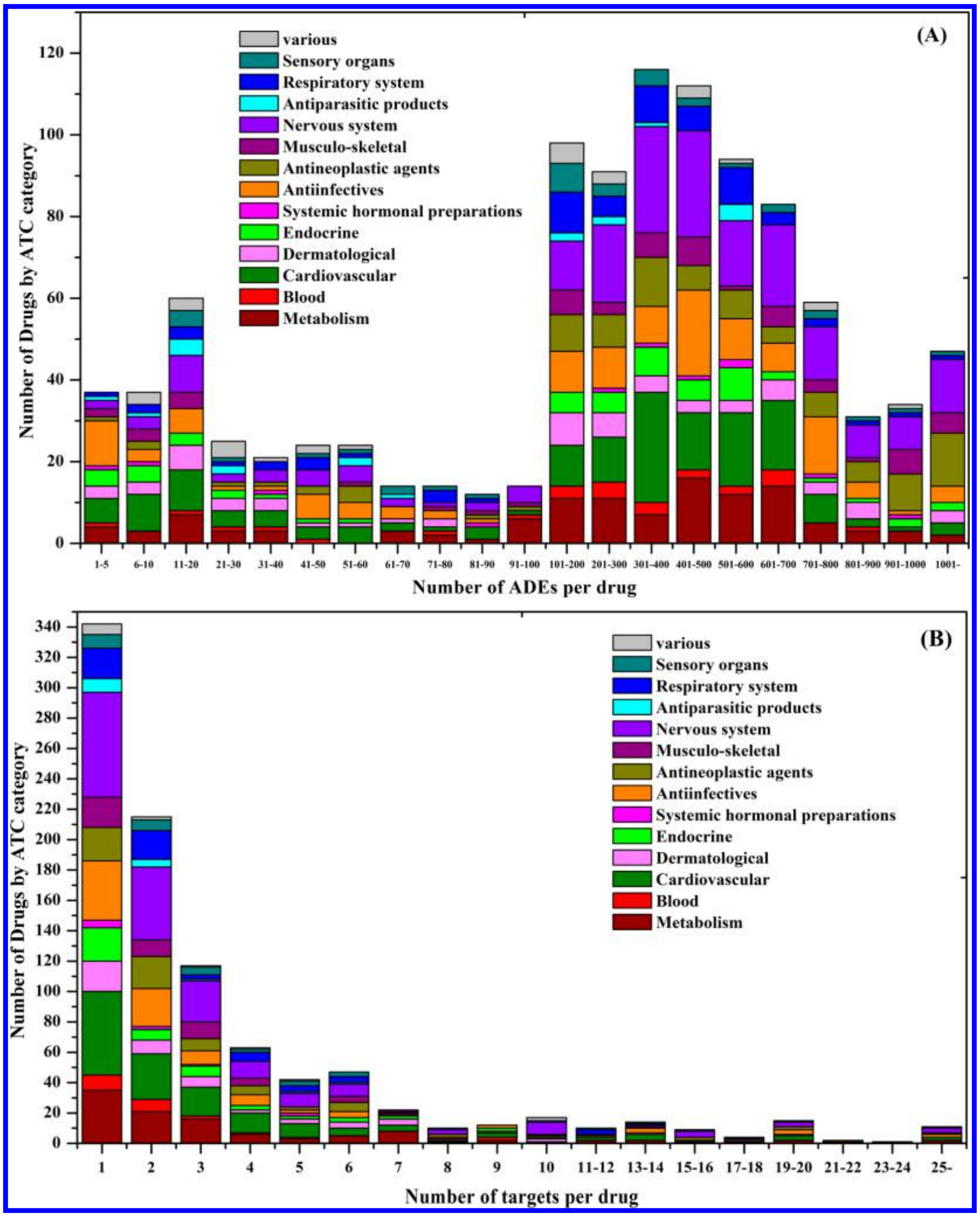


Figure 4. Distribution of the number of ADEs per approved drug (**A**) in MetaADEDB and the number of target proteins per approved drugs (**B**) in DrugBank and TTD. The drugs were colored by the first class of ATC Classification System.

suspect drug (SS) were used based on Takarabe's work.[36] All drug and ADE terms were annotated using MeSH vocabularies as describing above, and duplicated compound–ADE associations between AERS and MetaADEDB were removed.

Several criteria were used to assess the performance of the method. At first AUC was calculated.[35] Meanwhile, the recall (R) and precision (P) were calculated, and the precision–recall (PR) curve was plotted. The detailed descriptions of P, R, and AUC calculation were given in our previous work.[34,35]

■ RESULTS

Development of MetaADEDB. A comprehensive ADE-related database, namely MetaADEDB, was constructed by integrating CTD, SIDER, and OFFSIDES together. As shown in Table 1 and Figure 2, CTD[4] contained 52 987 drug–ADE associations among 2406 chemical items and 2789 ADEs; SIDER[3] (version 2) included 96 561 drug–ADE associations among 945 drugs and 4026 ADEs; while OFFSIDES[2] consisted of 405 493 drug–ADE associations among 1180 drugs and 9905 ADEs. The resulting MetaADEDB contained 527 216 drug–ADE associations connecting 3059 chemical items (including more than 1300 approved, withdrawn, and experimental drugs) and 13 255 ADEs, after all duplicates were removed. All chemicals and ADE items were annotated with the most commonly used medical/biological items of Unified Medical Subject Headings (MeSH) or Medical Language System (UMLS) vocabularies[37] to improve the quality of our database.

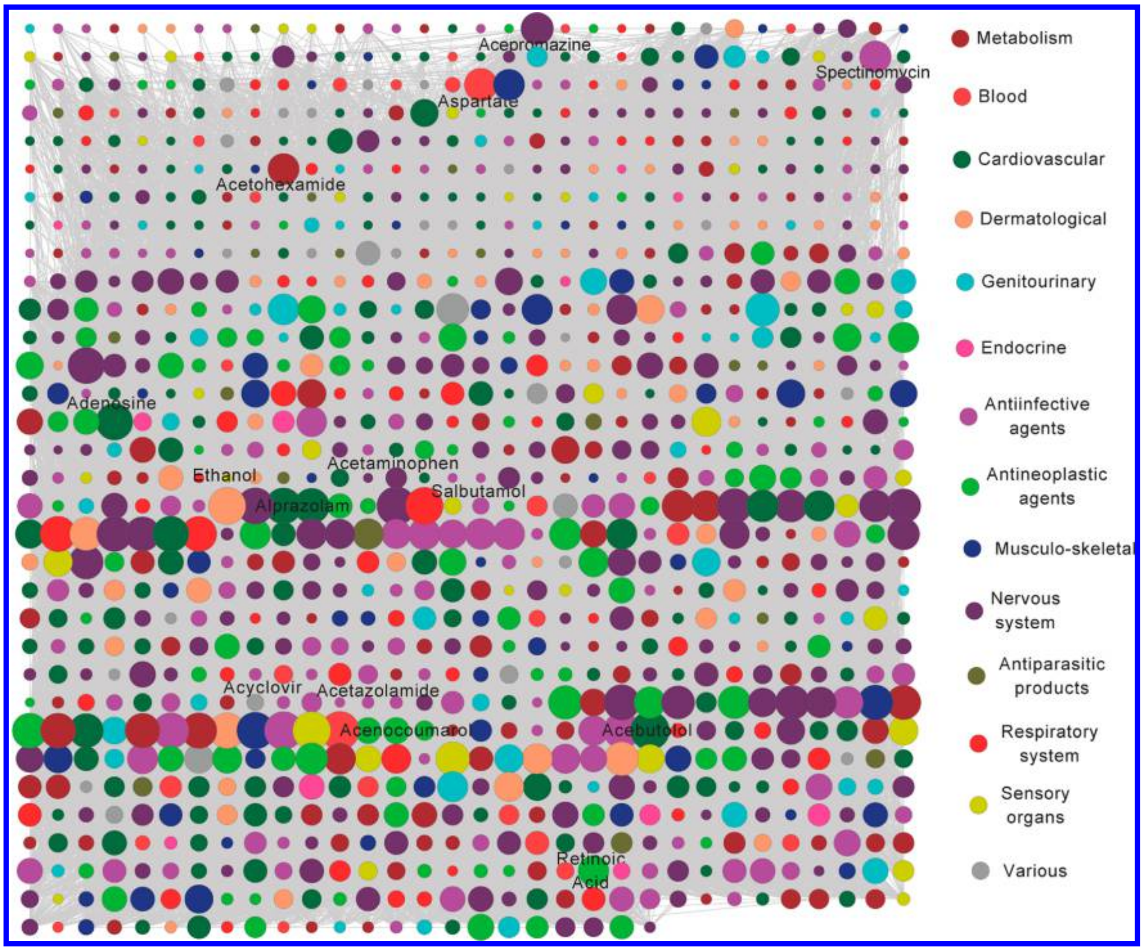


Figure 5. ADE-centered drug−drug network. Two drugs are connected if they are associated with the same ADE. The network consisted of 51 972 pairs among 1045 drugs (Supporting Information Table S2). The nodes represent the interacting drugs, and the edges denote the associations. The size of the node is the fraction of the number of drugs that the drug linked. Drug nodes are colored according to the first class of their ATC Classification. This figure was prepared with Cytoscape (http://www.cytoscape.org/).

The new MetaADEDB was compared with the three old ones, CTD, SIDER, and OFFSIDES, by measuring three classical network topological features, namely connectivity (k), clustering coefficient (C), and betweenness (B). As given in Table 2, the mean connectivity of drugs in MetaADEDB was 172, which was higher than those in CTD and SIDER. The mean clustering and betweenness was 0.037 and 4.78×10^{-4}, respectively, in MetaADEDB, which was lower than those in CTD, SIDER, and OFFSIDES. As shown in Figure 3, the connectivity distribution $P(k)$ of chemical and ADE terms in MetaADEDB (blue line) was overall higher than those in the other three databases. These data demonstrated that MetaADEDB is better than the others and provides a comprehensive human phenotypic network source for prediction of ADEs.

MetaADEDB is available at http://www.lmmd.org/online_services/metaadedb/ for free use. A user-friendly web-based query tool enables the database to be queried by drug or disease name, MeSH/ULMS ID, DrugBank ID, ATC code, SMILES, and structural similarity search. The structure similarity search is assessed by the Tanimoto coefficient using the MACCS keys implemented with OpenBabel v2.3.1.

Drug−ADE Network. The 3059 chemical items in MetaADEDB were mapped with those in DrugBank[27] and TTD,[28] and 1331 drugs were overlapped (Supporting Information Table S1). Then, the 1331 drugs were grouped based on ATC code, to investigate how specific the ADEs are to drug ATC classes. In total, 1047 drugs with known ATC code extracted from DrugBank were grouped. As shown in Figure 4A, most ADEs can occur in more than one drug class. The overlap of ADEs between drugs from different classes indicated that those drugs might have common mechanism-of-action (MOA) or one drug may have multiple anatomical effects. Drugs used in metabolism, cardiovascular, or nervous systems, and anti-infective, antineoplastic agents often link with more ADEs. For example, the norepinephrine-dopamine reuptake inhibitor bupropion (DB01156) and atypical antipsychotic drug clozapine (DB00363) are linked with 1811 and 1359 ADEs, respectively. The estrogen receptor antagonist tamoxifen (DB00675) and proteasome inhibitor bortezomib (DB00188) belonging to antineoplastic agents are linked with 1646 and 1585 ADEs in MetaADEDB, respectively.

Drug−Target Interaction Network. To further investigate the MOA of drug involved in multiple ADEs, the targets annotated in DrugBank[27] and TTD[28] were extracted for 953 FDA approved drugs with known ATC codes in MetaADEDB. A drug−target interaction network was constructed among 953 approved drugs and 1200 targets. As shown in Figure 4B, there were obvious polypharmacological features for approved drugs. Only 35.9% (342) drugs linked with one target. Drugs involved in metabolism, cardiovascular, and nervous systems often link with multitarget genes. For example, the drug pyridoxal phosphate (DB00114) links with 72 targets in DrugBank, which is involved in 246 ADEs in MetaADEDB. The atypical antipsychotic drug clozapine (DB00363) is associated with 1359 ADEs links and 26 targets. The verapamil (DB00661) links with 16 targets, which is involved in 550 ADEs in MetaADEDB. The data indicated that the promiscuity of approved drugs was consistent with the multi-ADEs found in clinical use. The high promiscuous drugs have high risk of side effects to patient safety.

Drug−Drug Network. An ADE-centered drug−drug network was constructed, in which two drugs are connected if they are associated with the same ADE. The drug−drug network was represented graphically in Figure 5 using Cytoscape, including 51 971 drug−drug pairs among 1044 drugs (Supporting Information Table S2). The nodes represent the interaction drugs and the edges denote the associations of two drugs with the same ADE. It is a nonhomogeneous network. The average degree of drugs in drug−drug network is 49.7. Only 14 out of 1044 drugs have a degree of one. And 632 of 1044 approved drugs are linked with more than 50 drugs in the ADE-centered drug−drug network. The top ten drugs with the highest degree were shown in Figure 5. They are acebutolol, acetaminophen, acenocoumarol, acyclovir, acetazolamide, salbutamol, ethanol, allopurinol, adenosine, and alprazolam.

Prediction of New ADEs for Known Drugs. Prediction of potential ADEs before a drug enters into clinical trials is a tantalizing goal to reduce drug-related morbidity and mortality. In this study, a method named PNIM was built to predict new drug−ADE associations using our previously developed network-based inference (NBI) algorithm.[34,35] For PNIM, it could prioritize new ADEs for a given drug or prioritize new drugs for a given ADE. As shown in Figure 6, by 10 times simulation of 10-fold cross-validation using PNIM, the mean AUC values were 0.912 ± 0.002, 0.936 ± 0.001, 0.881 ± 0.001, and 0.902 ± 0.001 for the CTD, SIDER, OFFSIDES, and MetaADEDB data sets, respectively. The detailed results of AUC value were given in Supporting Information Table S3. The PR curve is more informative than the ROC curve when the number of positive examples is much lower than that of negative examples.[23,38] Figure 7A shows the PR curves for the four different PNIMs built in the CTD, SIDER, OFFSIDES, and MetaADEDB data sets respectively by 10-fold cross-validation. The results of PR curves indicate similar tendencies as those observed in the AUC value in Figure 6, which demonstrated that the high performance was yielded for our predictive PNIMs when evaluated by cross-validation. Since only cross-validation was used to validate the model, it is unclear if it is possible to generalize the prediction to the molecules significantly different from those in the original training set. Therefore, the two-dimensional structural similarity of 1331 drugs was calculated via the Tanimoto similarity metrics using MACCS keys, freely available from OpenBabel v2.3.1. As shown in Figure S1 of the Supporting Information, the mean Tanimoto similarity of 1331 drugs is 0.312, which indicated the high diversity of our training set.

In order to further test the performance of our method on external validation set, a new data set was collected from the US-FDA AERS for external validation, which included 25 259 new drug−ADE association pairs connected 913 unique drugs and 1440 unique ADEs (Supporting Information Table S4). The resulted AUC value was as high as 0.912 for this new data set. The PR curve was also plotted in Figure 7B, which indicated the high predictive performance for the external validation set. In the drug−ADE bipartite network of MetaADEDB, there are some ADE nodes which are associated with almost all drugs (e.g., nausea, dizziness, vomitting, and rash), which may influence the performance of the PNIM. In order to test the influence of ADE with the high nodes (those ADEs connected with almost all drugs), we removed the top 50 kinds of ADE nodes which were associated with the most drugs in the MetaADEDB data set and built the new PNIM model using the rest of the data set. As shown in the Table S5 of the Supporting Information, a high AUC value of 0.897 ± 0.001 was yielded by 100 independent simulated tests of 10-fold cross-validation. In addition, we also removed the top 50 kinds of drug nodes which were associated with the most ADEs in the MetaADEDB data set and built the new PNIM model. A high AUC value of 0.902 ± 0.001 was yielded by 100 independent simulated tests of 10-fold cross-validation (Table S5). The data indicated that the performance of our model was not influenced by ADE nodes which are associated with almost all drugs.

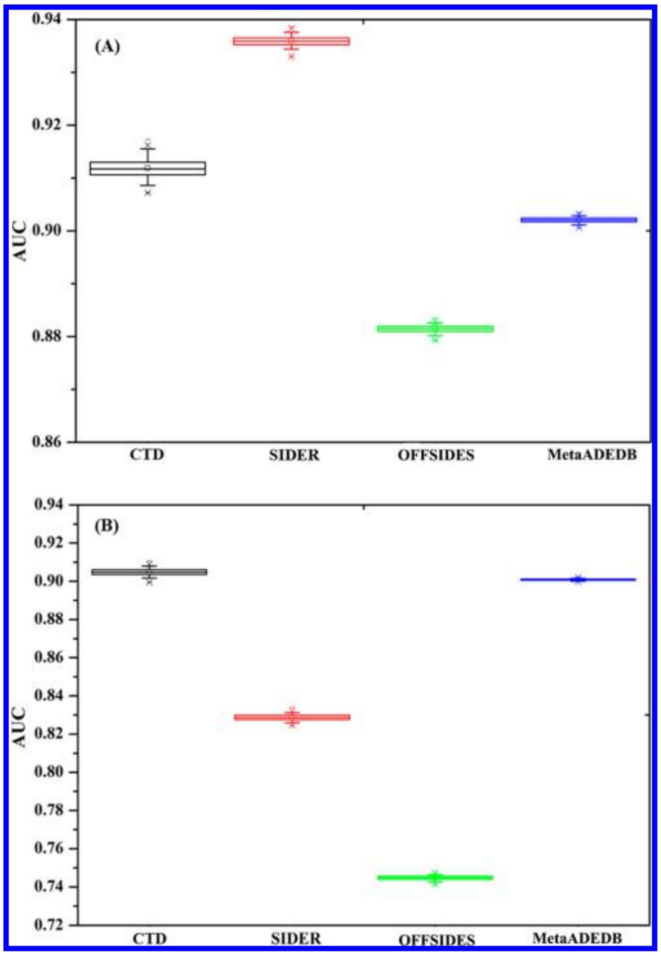


Figure 6. Box plots of showing the minimum, lower quartile, median, upper quartile, and maximum AUC values using PNIM to prioritize potential ADEs for a given drug (**A**) or prioritize new drugs for a given ADE (**B**) for four different drug−ADE association networks (CTD, SIDER, OFFSIDES, and MetaADEDB).

Moreover, to assess the feasibility of the method, the top 20 potential drug−ADE associations with highest ranking scores in CTD, SIDER, OFFSIDES, and MetaADEDB were predicted via PNIM. All predicted lists were also deposited into the database MetaADEDB.

■ DISCUSSION

In this study, we reported a human phenotypic network inference method to predict potential side effects of drugs. Due to the importance of high quality data in method development, a

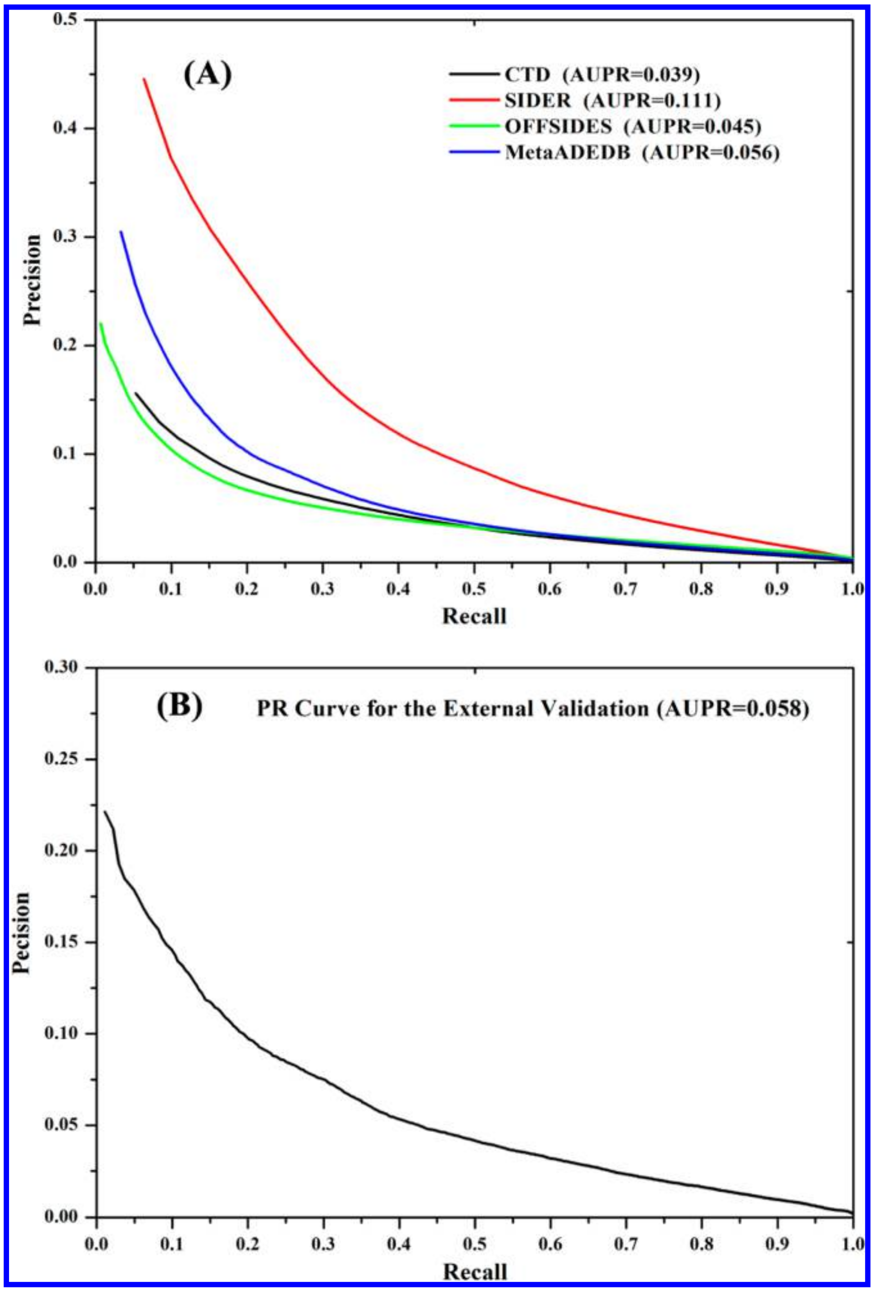


Figure 7. Precision−recall (PR) curves based on the 10 times independent 10-fold cross-validation (**A**) for four phenotypic network inference models (PNIMs) built on the CTD, SIDER, OFFSIDES, and MetaADEDB data sets, respectively, and the external validation set (**B**) by PNIM built on MetaADEDB data set when prioritizing new ADEs for a given drug. The PR curve is the plot of precision as a function of recall, where the detailed calculation about precision and recall were given in previous work.[34] The AUPR represents the area under the PR curve.

database related to drug side effects, namely MetaADEDB, was constructed at first, which contained 527 216 drug−ADE associations connecting 3059 compounds (including more than 1300 drugs) and 13 200 ADE items, the largest database on drug side effects so far. With the PNIM method, the AUC value was 0.912 for the external validation set, which indicated that the method was highly reliable for predicting potential ADEs of drugs in postmarketing surveillance.

Compared to other published methods, ours showed great advantages in ADE prediction. In Liu's work,[17] an inherent problem was that the negative drug−ADE association pairs were randomly constructed, which could easily generate noise in the process of model building. In Cami's work,[22] they calculated the network covariates on the topological structure of the observed drug−ADE association network and then built the logistic classifiers using the calculated network covariates as input features for prediction of unknown ADEs. In that work, the same inherent problem as Liu's work existed. Our method only used simple drug−ADE network topological similarity and yielded high predictive performance (Figures 6 and 7). Therefore, our method takes full advantage of the labeled and unlabeled information encoded in the full drug−ADE association network, thereby simultaneously exploiting both topological and functional modularity of drug phenotypic network. As described in our previous work,[34,35] since PNIM only utilized known drug−ADE bipartite network information, for a new drug without known ADE information in the training set, the method could not predict ADEs for this new drug. This is the weakness of our method. Recently, Yamanishi et al. developed several extensions of kernel regression model by integration of chemical space of drug chemical structures and biological space of drug target protein for side effect prediction.[23] The models in Yamanishi's work were built using the chemical substructure and target protein profiles, which could predict potential side effects for novel drug molecules. In terms of this, we will further modify our method so that it could predict side effects of novel molecules.

Compared with CTD,[4] SIDER,[3] and OFFSIDES,[2] the data source and the network topological properties of MetaADEDB were obviously larger than theirs. In addition, a set of new drug−ADE associations predicted by our PNIM were stored in MetaADEDB. The practitioner can follow up the high scoring candidates through clinical investigation. Although the comprehensive data sources were collected and available in MetaADEDB, two limitations need to be pointed out. First, the side effect frequencies of ADEs and placebo administration were not recruited. In particular, the data about side effect frequencies at the population level is a valuable resource to determine the correlation between ADE incidence, plasma concentrations, drug targets, dose level, and variability in the ethnic differences. Second, a small fraction of the published drug−ADE associations may be false positive due to various reasons, such as the placebo effect in healthy volunteers.[39]

In past decade, the best and most detailed data on human disease and drug discovery were proprietary in pharmaceutical companies. Recently, GlaxoSmithKline announced its intention to release "patient-level" raw data from clinical trials of approved drugs and failed investigational compounds, which could catalyze a growth in the understanding of disease and help avoid repeating mistakes made in failed trials and the early process of drug discovery.[40] Because the unlocked data from pharmaceutical companies will help interrogate the data to try to understand whether the failure was due to the mechanism of action of the drug, the wrong end points being measured, or the drug not working on the disease. To ensure the usefulness of MetaADEDB, the database will be updated monthly with additional data, such as the data points of the side effect frequencies of placebo administration at population level, pharmacogenetic biomarkers on the FDA Web site (http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm 083378.htm), FDA package inserts data on DailyMed (http://dailymed.nlm.nih.gov/dailymed/about.cfm), and further unlocked clinical data from pharmaceutical companies.

■ CONCLUSION

Here, we developed a useful human phenotypic source MetaADEDB with the largest drug side effect data so far, connecting more than 13 000 unique adverse drug events and 3000 unique chemicals including more than 1300 drugs. On the basis of the database, a method named PNIM was developed to predict unknown side effects. High AUC value of 0.912 was yielded for the method on the external validation set. This work provided useful tools for drug side effect prediction and drug discovery.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1 and Tables S1−S5. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*To whom correspondence should be addressed. Tel.: +86-21-6425-1052. Fax: +86-21-6425-3651. E-mail: ytang234@ecust.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities (WY1113007), the 863 Project (Grant 2012AA020308), the National S&T Major Project of China (Grant 2011ZX09307-002-03), and the Shanghai Committee of Science and Technology (11DZ2260600).

■ REFERENCES

(1) Lazarou, J.; Pomeranz, B. H.; Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA, J. Am. Med. Assoc.* **1998**, *279*, 1200−1205.

(2) Tatonetti, N. P.; Ye, P. P.; Daneshjou, R.; Altman, R. B. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **2012**, *4*, 125ra131.

(3) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **2010**, *6*, 343.

(4) Davis, A. P.; King, B. L.; Mockus, S.; Murphy, C. G.; Saraceni-Richards, C.; Rosenstein, M.; Wiegers, T.; Mattingly, C. J. The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.* **2011**, *39*, D1067−1072.

(5) Jensen, L. J.; Saric, J.; Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **2006**, *7*, 119−129.

(6) Amberger, J.; Bocchini, C. A.; Scott, A. F.; Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **2009**, *37*, D793−796.

(7) Ingelman-Sundberg, M. Pharmacogenomic biomarkers for prediction of severe adverse drug reactions. *N. Engl. J. Med.* **2008**, *358*, 637−639.

(8) Harpaz, R.; Dumouchel, W.; Shah, N. H.; Madigan, D.; Ryan, P.; Friedman, C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clin. Pharmacol. Ther.* **2012**, *91*, 1010−1021.

(9) Pouliot, Y.; Chiang, A. P.; Butte, A. J. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin. Pharmacol. Ther.* **2011**, *90*, 90−99.

(10) Ball, R.; Botsis, T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? *Clin. Pharmacol. Ther.* **2011**, *90*, 271−278.

(11) Huang, L. C.; Wu, X.; Chen, J. Y. Predicting adverse side effects of drugs. *BMC Genom.* **2011**, *12* (Suppl 5), S11.

(12) Harpaz, R.; Perez, H.; Chase, H. S.; Rabadan, R.; Hripcsak, G.; Friedman, C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clin. Pharmacol. Ther.* **2010**, *89*, 243−250.

(13) Tatonetti, N. P.; Liu, T.; Altman, R. B. Predicting drug side-effects by chemical systems biology. *Genome Biol.* **2009**, *10*, 238.

(14) Harpaz, R.; Perez, H.; Chase, H. S.; Rabadan, R.; Hripcsak, G.; Friedman, C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clin. Pharmacol. Ther.* **2011**, *89*, 243−250.

(15) Adkins, D. E.; Aberg, K.; McClay, J. L.; Bukszar, J.; Zhao, Z.; Jia, P.; Stroup, T. S.; Perkins, D.; McEvoy, J. P.; Lieberman, J. A.; Sullivan, P. F.; van den Oord, E. J. Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. *Mol. Psychiat.* **2011**, *16*, 321−332.

(16) Chiang, A. P.; Butte, A. J. Data-driven methods to discover molecular determinants of serious adverse drug events. *Clin. Pharmacol. Ther.* **2009**, *85*, 259−268.

(17) Liu, M.; Wu, Y.; Chen, Y.; Sun, J.; Zhao, Z.; Chen, X. W.; Matheny, M. E.; Xu, H. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J. Am. Med. Inf. Assoc.* **2012**, *19*, e28−e35.

(18) Abernethy, D. R.; Woodcock, J.; Lesko, L. J. Pharmacological mechanism-based drug safety assessment and prediction. *Clin. Pharmacol. Ther.* **2011**, *89*, 793−797.

(19) Yang, L.; Wang, K.; Chen, J.; Jegga, A. G.; Luo, H.; Shi, L.; Wan, C.; Guo, X.; Qin, S.; He, G.; Feng, G.; He, L. Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome–clozapine-induced agranulocytosis as a case study. *PLoS Comput. Biol.* **2011**, *7*, e1002016.

(20) Xie, L.; Wang, J.; Bourne, P. E. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3*, e217 DOI: 10.1371/journal.pcbi.0030217.

(21) Xie, L.; Li, J.; Bourne, P. E. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.* **2009**, *5*, e1000387.

(22) Cami, A.; Arnold, A.; Manzi, S.; Reis, B. Predicting adverse drug events using pharmacological network models. *Sci. Transl. Med.* **2011**, *3*, 114ra127.

(23) Yamanishi, Y.; Pauwels, E.; Saigo, H.; Stoven, V. Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions. *J. Chem. Inf. Model.* **2012**, *52*, 3284−3292.

(24) Tatonetti, N. P.; Fernald, G. H.; Altman, R. B. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J. Am. Med. Inf. Assoc.* **2012**, *19*, 79−85.

(25) Bate, A.; Evans, S. J. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidem. Dr. S.* **2009**, *18*, 427−436.

(26) Szarfman, A.; Machado, S. G.; O'Neill, R. T. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety* **2002**, *25*, 381−392.

(27) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−1041.

(28) Zhu, F.; Shi, Z.; Qin, C.; Tao, L.; Liu, X.; Xu, F.; Zhang, L.; Song, Y.; Zhang, J.; Han, B.; Zhang, P.; Chen, Y. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1128−1136.

(29) Hernandez-Boussard, T.; Whirl-Carrillo, M.; Hebert, J. M.; Gong, L.; Owen, R.; Gong, M.; Gor, W.; Liu, F.; Truong, C.; Whaley, R.; Woon, M.; Zhou, T.; Altman, R. B.; Klein, T. E. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **2008**, *36*, D913−918.

(30) Boyle, N. M; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminf.* [Online] **2011**, *3*, Article 33; http://www.jcheminf.com/content/3/1/33 (accessed Nov 19, 2011).

(31) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099−3105.

(32) Seebacher, J.; Gavin, A. C. SnapShot: Protein-protein interaction networks. *Cell* **2011**, *144*, 1000−1000.e1.

(33) Matthieu, L.; Mahnien, C.; Vecchio, N. D. Basic notions for the analysis of large two-mode networks. *Social Networks* **2008**, *30*, 31−48.

(34) Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503.

(35) Cheng, F.; Zhou, Y.; Li, W.; Liu, G.; Tang, Y. Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One* **2012**, *7*, e41064.

(36) Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics* **2012**, *28*, i611−i618.

(37) Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267−270.

(38) Davis, J.; Goadrich, M. The Relationship Between Precision−Recall and ROC Curves. In *Proceedings of the Twenty Third International Conference on Machine Learning*; Pittsburgh, PA, June 25−29, 2006; Cohen, W. W. A. M., Ed.; ACM Press: PA, 2006; pp 233−240.

(39) Rosenzweig, P.; Brohier, S.; Zipfel, A. The placebo effect in healthy volunteers: influence of experimental conditions on the adverse events profile during phase I studies. *Clin. Pharmacol. Ther.* **1993**, *54*, 578−583.

(40) Harrison, C. GlaxoSmithKline opens the door on clinical data sharing. *Nat. Rev. Drug Discovery* **2012**, *11*, 891−892.