Mapping Adverse Drug Reactions in Chemical Space

Josef Scheiber,*[,†,#] Jeremy L. Jenkins,[†] Sai Chetan K. Sukuru,[†] Andreas Bender,[†,▽] Dmitri Mikhailov,[†] Mariusz Milik,[†] Kamal Azzaoui,[||] Steven Whitebread,[‡] Jacques Hamon,[§] Laszlo Urban,[‡] Meir Glick,[†] and John W. Davies[†]

Lead Discovery Informatics, CPC, Novartis Institutes for Biomedical Research, 250 Massachussetts Avenue, Cambridge, Massachusetts 02139, Preclinical Safety Profiling, CPC, Novartis Institutes for Biomedical Research, 250 Massachussetts Avenue, Cambridge, Massachusetts 02139, Preclinical Safety Profiling, CPC, Novartis Pharma AG, Forum 1, 4002 Basel, Switzerland, Molecular Libraries Informatics, CPC, Novartis Pharma AG, Forum 1, 4002 Basel, Switzerland

Received December 8, 2008

We present a novel method to better investigate adverse drug reactions in chemical space. By integrating data sources about adverse drug reactions of drugs with an established cheminformatics modeling method, we generate a data set that is then visualized with a systems biology tool. Thereby new insights into undesired drug effects are gained. In this work, we present a global analysis linking chemical features to adverse drug reactions.

Introduction

Knowledge of both desired "on-target" and undesired "off target" biological activities of a potential drug is crucial in order to ensure efficacy as well as a good safety profile. Undesired effects, so-called adverse drug reactions (ADRs[a]), have gained broad public attention recently even in the mainstream media. As a consequence, over the last 10 years, 19 broadly used marketed drugs were withdrawn after exhibiting unexpected severe side effects, with Merck's Rofecoxib and Bayer's Cerivastatin being two of the most prominent cases.[1,2] Thus, it is not surprising that in addition to death and loss of quality of life for patients, ADRs are also of big economical concern for pharmaceutical companies.[3] To prevent incidents like Rofecoxib, safety issues should be detected as early as possible in development and not once the drug is marketed.[4−7] To achieve this state, it is highly desirable to create a better understanding of chemical features linked to ADRs with the aim of removing liable features during lead optimization.

In this work, we present a global analysis linking chemical features to ADRs. Starting from data sets of marketed drugs and their annotated ADRs, we extract chemical features that are highly correlated to particular effects. Figure 1 illustrates what one can learn from these analyses. Going a step further, we then compute the overlap of ADR types in chemical space to establish the degree of proximity (and, possibly, biological relatedness) between different adverse reactions. Next, links based on feature correlation information are used to generate a global map of ADR relationships. Finally, we analyze how different classes of undesired effects relate to each other on an organ level.

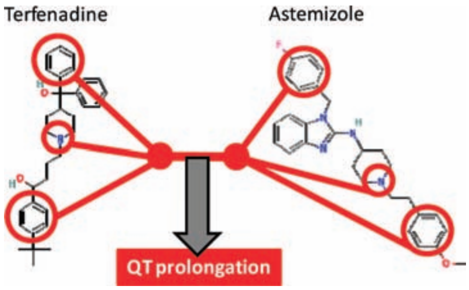

Figure 1. A well-known example that shows how chemical substructures are linked to a certain adverse drug reaction, in this case cardiac arrythmia caused by QT interval prolongation. We have performed this analysis on a large scale (thousands of ADRs).

Materials and Methods

The main source for our analyses is the PharmaPendium database from Elsevier,[8] which makes drug safety data of U.S.-approved drugs available to researchers. Compound sets were extracted that shared common ADRs or toxicities. In total we extracted 4210 different ADR terms stored in lower level Medical Dictionary for Regulatory Activities (MedDRA) terminology. MedDRA is a clinically validated international medical terminology used for ADR reporting as a standard[9] and throughout the entire regulatory process, from premarketing to postmarketing activities, for data entry, retrieval, evaluation, and presentation. MedDRA is used in the U.S., European Union, and Japan. Its use is currently mandated in Europe and Japan for safety reporting. Preclinical, clinical, and postmarketing phase information were all used as input for our analysis.

For every MedDRA term, we extracted all associated molecules from PharmaPendium. Despite the range of data set sizes between 10 and 1200 molecules associated with each MedDRA term (in total the PharmaPendium database comprises of 1842 drugs), we chose to include all of the compounds in the study to incorporate as much knowledge into the following analyses as possible.

The well-established extended connectivity fingerprints (ECFPs) with a radius of ECFP_4 were used as chemical descriptors for the molecules because it has been shown in several cases that Bayesian models built using circular fingerprints work very well in virtual screening tasks.[10−14] Multiple-category Laplacian-modified naive Bayesian classification models were built for the ADRs using components from Pipeline Pilot (Accelrys). These models assume

* To whom correspondence should be addressed. Phone: +41 (61) 324 8407. Fax: +41-61-3242163. E-Mail: mail@josef-scheiber.de.

[†] Lead Discovery Informatics, CPC, Novartis Institutes for Biomedical Research.

[‡] Preclinical Safety Profiling, CPC, Novartis Institutes for Biomedical Research.

[§] Preclinical Safety Profiling, CPC, Novartis Pharma AG.

[||] Molecular Libraries Informatics, CPC, Novartis Pharma AG.

[#] Current address: Text Mining Services, NITAS, Novartis Pharma AG, Forum 1, 4002 Basel, Switzerland.

[▽] Current address: Leiden/Amsterdam Center for Drug Research, Medicinal Chemistry Division, Einsteinweg 55, 2333 CC Leiden, The Netherlands.

[a] Abbreviations: ADR, adverse drug reaction; MedDRA, Medical Dictionary for Regulatory Activities; ECFP, extended connectivity fingerprint; SOC, system organ class.

10.1021/jm801546k CCC: $40.75
Published on Web 04/20/2009

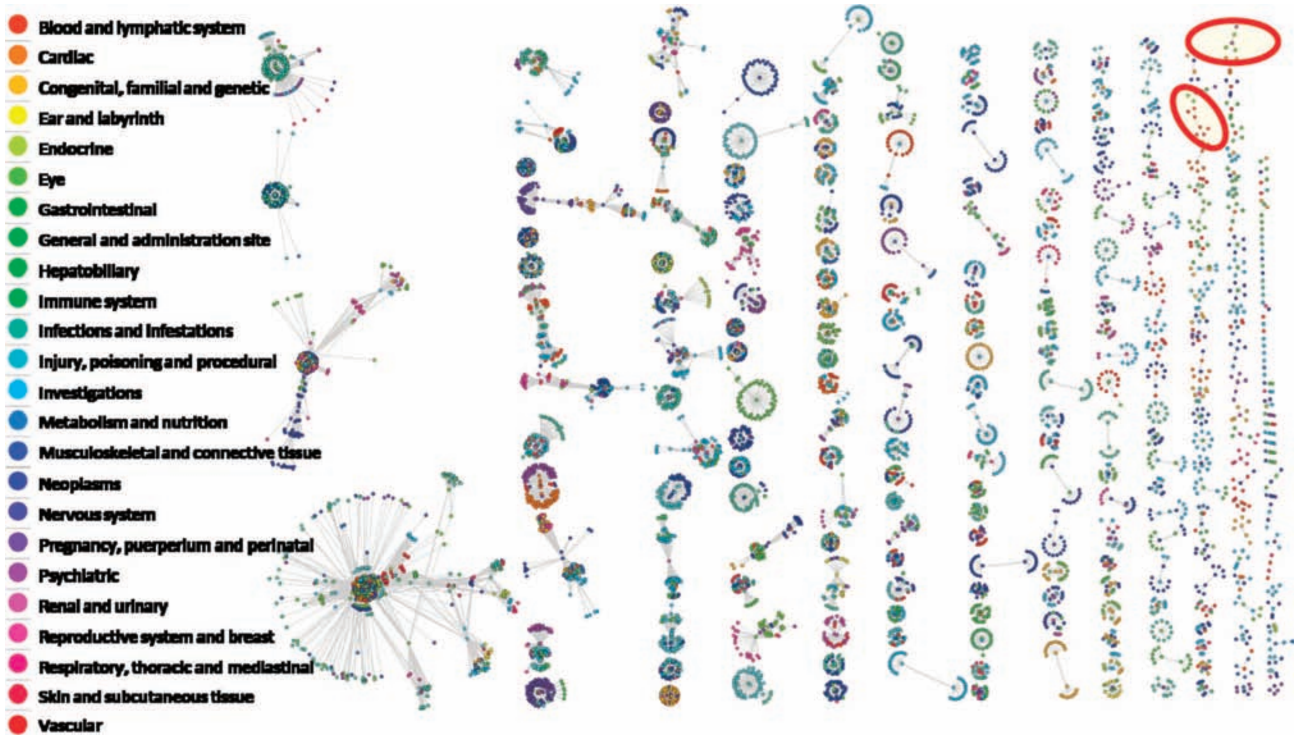
● Blood and lymphatic system
● Cardiac
● Congenital, familial and genetic
● Ear and labyrinth
● Endocrine
● Eye
● Gastrointestinal
● General and administration site
● Hepatobiliary
● Immune system
● Infections and infestations
● Injury, poisoning and procedural
● Investigations
● Metabolism and nutrition
● Musculoskeletal and connective tissue
● Neoplasms
● Nervous system
● Pregnancy, puerperium and perinatal
● Psychiatric
● Renal and urinary
● Reproductive system and breast
● Respiratory, thoracic and mediastinal
● Skin and subcutaneous tissue
● Vascular

Figure 2. The global mapping of adverse drug reactions in chemical space in the Cytoscape visualization. Each dot represents exactly one ADR. Two ADRs are connected if the chemical features they share have a Pearson correlation of 0.8 or higher; the higher the correlation, the closer the ADRs are located. The coloring represents system organ classes (as defined in MedDRA). The 2D distance in this plot is also related to the chemical similarity between those two points. The red circles mark the areas that are discussed in more detail, and they are also shown in a zoomed version in Figure 3.

that all variables are independent and use a Laplacian correction to reduce the bias caused through descriptors less prevalent in the data set. The derivation of the multiple-category Laplacian-modified Bayesian models has been described previously.[10−15] This combination has a successful track record in large-scale data analysis for various purposes. A general statistical evaluation can be found in the publication by Nidhi et al.,[14] and an in-depth evaluation of ADR models can also be found in a previous paper.[10]

Next, for any pairing of ADR models, the similarity between the two was established by computing the Pearson correlation between the normalized feature probabilities from the individual Bayesian models.[10,16] Only the 10000 most frequent features, both positively and negatively correlated, of each individual ADR model which were also present in both model sets were used. This step was found to improve the overlap of chemical substructures between the ADRs. Correlations were normalized per ADR; that is, every adverse reaction was assigned the same overall probability. In contrast to the approach of comparing targets on the basis of their overlap in small-molecule inhibitors, determining similarity via statistically correlated features, allows one to determine ADR−ADR similarity even when no exact chemical structures are in common between data sets. In other words, only important substructures of compounds need to be shared between two ADRs to find similarity. (This is important because data from pooled sources do not contain a complete experimental matrix of all drugs evaluated toward all ADRs. A similar approach has proven very successful in computing target−target similarities[16]). Recently, Campillos et al. have shown that side effect similarity between compounds can be used to predict novel targets.[17] This is in line with what Hopkins has proposed in understanding network pharmacology from the chemical point of view.[18]

In a following step, all ADR−ADR pairings that show Pearson correlation of $\rho > 0.8$ are retained as "significant" for the next step of our analysis. Therefore 1501 ADR classes remain in the analysis and were further investigated. All significant ADR pairs were loaded into Cytoscape,[19] a well-established open source bioinformatics software platform for analyzing and visualizing complex interaction networks. Within Cytoscape, a force-directed mapping was applied to visualize relationships between ADRs. Force-directed layout algorithms are a powerful and practical graph drawing heuristic that relies on an objective function that maps a particular graph layout to an energy value. Typically, such algorithms start with a random drawing of the graph and utilize standard optimization methods to minimize the energy function. The algorithms define functions in which low energies are associated with layouts where adjacent vertices are near some preferred distance from each other, and nonadjacent vertices are well-spaced. To further analyze the generated network, we used the Cytoscape-Plugin MCODE[20] to identify highly connected regions in our network. We applied the method using its standard parameters. The full outcome of the analysis and detailed results are shown in the Supporting Information (from page S306).

Results and Discussion

As a result, ADRs that show a high correlation are linked and placed close to each other. The result of this analysis is shown in Figure 2 and the raw data table is available in the Supporting Information.

Similar analyses have recently been published for networks of drugs and their targets.[21−23] It can be seen that there are many smaller clusters of ADRs, often located around one central ADR. It is worth mentioning that if the required correlation score between ADRs is lowered, more and more clusters become connected to others and the interlinking and number of nodes in the map becomes greater, making it increasingly difficult to analyze. For example, a cutoff of $\rho > 0.7$ would yield a map with 2095 nodes, which becomes ambiguous and therefore we decided to use higher correlations for our analysis. Also, it is important to emphasize that the pharmacological rationale of two ADRs being linked via chemical structures diminishes quickly when a lower cutoff is chosen. We are therefore

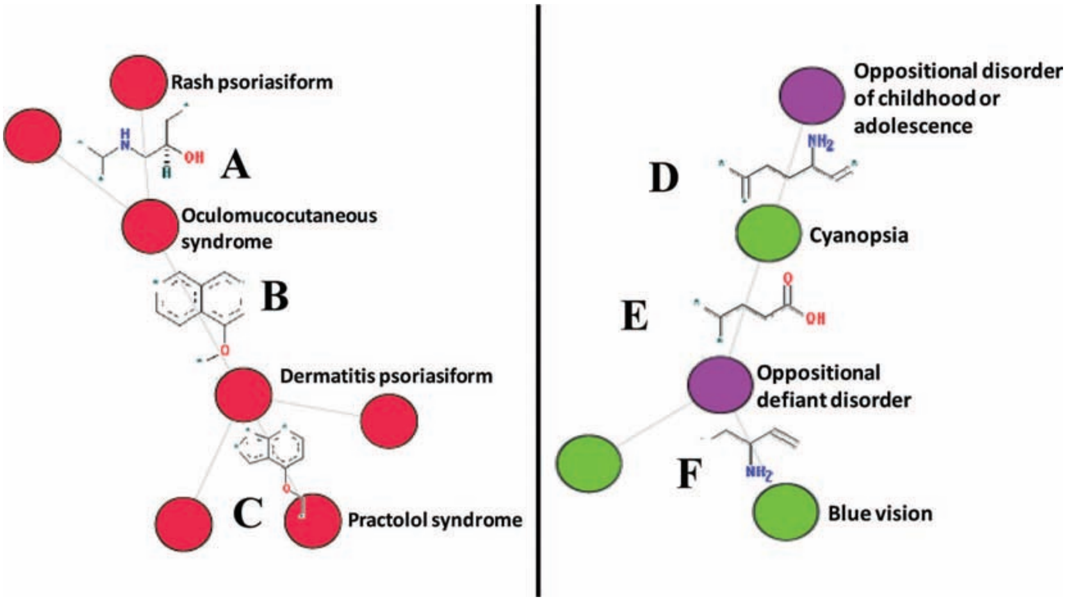
Rash psoriasiform
A
Oculomucocutaneous syndrome
B
Dermatitis psoriasiform
C
Practolol syndrome
Oppositional disorder of childhood or adolescence
D
Cyanopsia
E
Oppositional defiant disorder
F
Blue vision

Figure 3. Two specific examples for map elements; left, a homogeneous cluster of skin ADRs (marked in red), right, a cluster of eye (marked in green) and psychiatric disorders (marked in violet). The chemical features shown over the connections are always the features that have the highest score for the overlap between the two side effects.

convinced that a Pearson of 0.8 is appropriate to enable applicable follow-up analyses.

In the next step, we analyzed which ADRs are linked due to their underlying chemistry at the level of system organ class (SOC), the highest level in the MedDRA terminology. Each ADR node in the map was colored according to the SOC to which it belongs. This immediately reveals clusters where all the ADRs show the same color, suggesting the underlying chemistry leads to ADRs only in the same organ. On the other hand, there are heterogeneous clusters, indicating certain chemical features are linked to ADRs in different SOCs. Notably, distribution of compounds in the body is also known to be linked to chemical substructures.

Figure 3 shows an example for a homogeneous and a heterogeneous type of cluster, taken from the circled regions in Figure 2. On the left is a cluster of several skin side effects; for every connection shown, we extracted the chemical features with the highest Bayes score that are in common to the linked ADRs. Substructure A, linking rash psoriasiform and oculomucocutaneous syndrome, is common to many β-blockers (Prenalterol, Alprenolol, Betaxolol, Atenolol, Bisoprolol, Levobetaxolol, Acebutolol, Esmolol, Metipranolol, Metoprolol), which are used to treat hypertension, and, indeed, both of the ADRs are documented for Practolol and Propanolol in PharmaPendium. This underlines the fact that common side effects of drugs are often linked to common chemical frameworks. In the topology of our map, the latter ADR is also linked to dermatitis psoriasiform (Figure 3). In this case, another substructure B is the one with the highest correlation. It is shared by Propanolol and Duloxetine (a serotonin−norepinephrine reuptake inhibitor). The two ADRs are reported co-occurring for Practolol and Propanolol. Finally, dermatitis is linked to Practolol syndrome through another highly scoring substructure C in common to Pindolol and Propanolol, again, both terms are co-occurring for Practolol. As this example demonstrates, mapping ADRs of drugs through chemical space is possible, and it links phenotypic and mechanistic spaces through the common language of chemical structure. It is worth noting that often low-level MedDRA terms are synonyms to each other. In the example here, the ADR next to rash psoriasiform is rash psoriaform. Therefore one can get additional confidence for links between certain ADRs if synonyms are linked to the same target ADR (such as here to oculomucocutaneous syndrome).

We performed the same type of analysis for the cluster of nervous system and eye disorders shown in Figure 3 at the right. Oppositional disorder of childhood (a nervous system disorder) is linked to cyanopsia (an eye disorder) through substructure D found in Vigabatrin (an anticonvulsant). Both ADRs are reported to occur for this drug. Cyanopsia is further linked to oppositional defiant disorder (again nervous system) through substructure E found in several drugs, namely secretin, folic acid, calcitonin, chenodiol, bivalirudin, and cosyntropin. This is then linked to blue vision (again affecting the eye) through substructure F found in Vigabatrin. Interestingly, in this case, very similar ADRs from one system organ class (SOC) cluster together with very similar ADRs from another SOC, but none of them is linked to a member of the same SOC. Again, the results are supported by co-occurrence of the ADRs in marketed drugs in PharmaPendium.

In summary, both examples described here show that computing a map of ADRs in chemical space can help to better understand possible ADRs for novel compounds sharing similar substructures.

We then analyzed the connectedness of the different clusters in our network. The MCODE analysis provides us with a list of clusters that are highly interlinked, i.e., they represent the hubs within the network. The data of all these clusters is shown in the Supporting Information. Having tightly linked clusters where each member is linked to another one means basically that similar chemical structures are statistically linked to different adverse drug reactions, similar to Figure 3. The most interlinked cluster is a subset of the big cluster shown in the lower left of Figure 2. We will not describe these clusters in detail, as we believe that the description of the general concept with detailed examples along with providing the raw data will enable follow-up studies. Rather, we are performing a more global-scale analysis of our network.

To extend this analysis and to obtain a more global picture of linked ADR classes, we performed a linkage analysis for the map shown in Figure 2 by iteratively analyzing all data

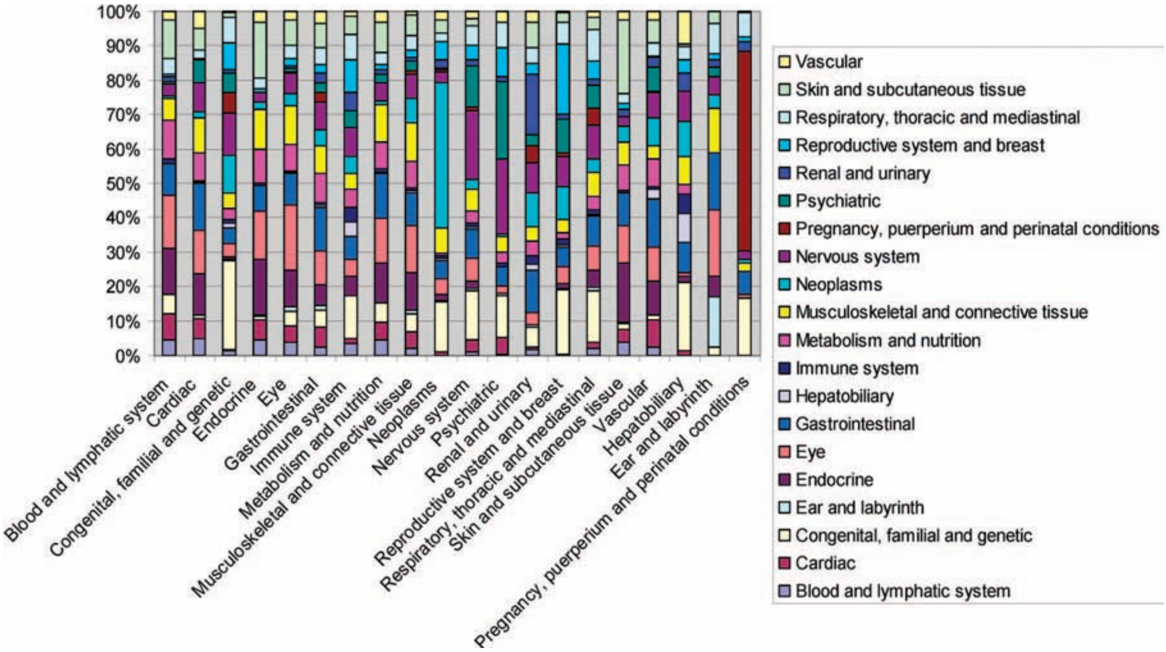

Figure 4. Analysis of ADR classes linked in chemical space. It can be seen that while ADRs classified according to MedDRA, SOCs are most significantly linked to themselves, except in cases such as ADRs related to the blood and lymphatic systems.

points assigned to a certain SOC. For example, one starts with an ADR belonging to the group of cardiac disorders and performs an analysis to find out how many other SOCs this particular ADR is linked to. This is repeated for every ADR in one SOC, and the according numbers are summed to obtain a list of frequencies of cardiac disorder linkages (e.g., links in chemical space to gastrointestinal disorders or eye disorders). After calculating relative values, one obtains a proximity measure of MedDRA SOC with the outcome visualized in Figure 4. We can see that links to ADRs within the same SOC are, as one would expect, somewhat over-represented for many SOCs. Still, there are clear exceptions like blood and lymphatic disorders or cardiac disorders where ADRs are not well-linked within an SOC. This analysis gives us an idea what the most likely other affected organ would be if one sees an ADR in a particular organ. An example would be that psychiatric disorders are very often linked to nervous system disorders or the reproductive system is often linked to congenital disorders—a finding that is also true for co-occurring ADR terms. We postulate that this information could be used to find appropriate biomarkers for certain more severe ADRs: if an ADR is linked into another SOC, one could analyze the biological parameters in the organ that has the closest connection in chemical ADR space. A simple byproduct of the present work for drug development is the application of the ADR chemical feature models to compounds at any discovery stage to get an in silico estimate of the likelihoods of the various adverse drug reactions that may occur as well as the substructural features most influencing the predicted ADRs.

Conclusions

In summary, we have developed a comprehensive data mining and analysis approach for mapping chemical features to adverse drug reactions on a large scale. We do not aim to understand the mechanistic cause of the ADR (e.g., reactivity, binding to targets, etc.) but the link between chemical structure and adverse drug reaction directly. Thereby, we are able to obtain a global picture of how different types of drug-induced adverse effects may be connected and to map these results into a single plot by exploiting network visualization tools. In addition, we performed a global linkage analysis of the map of ADRs to elucidate how adverse reactions cluster in the context of their SOCs. The presented strategy is particularly suited to early phases of drug discovery where we have previously shown[10] it can be used to annotate and potentially help weed out screening candidates prior to experimental testing. On the other hand, it might also be helpful to use desired side effects (such as an antihypertensive actions) to engineer properties into a molecule that benefit its primary mode of action or to provide ideas for repurposing drugs, i.e., identifying novel uses for already established drugs by following the same principles.

Supporting Information Available: Supporting Information containing the basis data for the generated network and the identified clusters. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Giacomini, K. M.; Krauss, R. M.; Roden, D. M.; Eichelbaum, M.; Hayden, M. R.; Nakamura, Y. When good drugs go bad. *Nature (London)* **2007**, *446* (7139), 975–977.

(2) Furberg, C.; Pitt, B. Withdrawal of cerivastatin from the world market. *Curr. Control Trials Cardiovasc. Med.* **2001**, *2* (5), 205–207.

(3) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2004**, *3* (8), 711–715.

(4) Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discovery Today* **2005**, *10* (21), 1421–1433.

(5) Classen, D. C.; Pestotnik, S. L.; Evans, R. S.; Lloyd, J. F.; Burke, J. P. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. *JAMA, J. Am. Med. Assoc.* **1997**, *277* (4), 301–306.

(6) Johnson, J. A.; Bootman, J. L. Drug-related morbidity and mortality. A cost-of-illness model. *Arch. Intern. Med.* **1995**, *155* (18), 1949–1956.

(7) Leape, L. L.; Brennan, T. A.; Laird, N.; Lawthers, A. G.; Localio, A. R.; Barnes, B. A.; Hebert, L.; Newhouse, J. P.; Weiler, P. C.; Hiatt, H. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N. Engl. J. Med.* **1991**, *324* (6), 377–384.

(8) PharmaPendium; Elsevier: New York, 2008; https://www.pharmapendium.com/(accessed Mar 2008).

(9) MedDRA: The Medical Dictionary for Regulatory Activities; Maintenance and Support Services Organization: Chantilly, VA, 2008; http://www.meddramsso.com/MSSOWeb/index.htm (accessed Mar 2008).

(10) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2* (6), 861–873.

(11) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model.* **2007**, *47* (4), 1319–1327.

(12) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9* (3), 199–204.

(13) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49* (23), 6802–6810.

(14) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46* (3), 1124–1133.

(15) Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J. Chem. Inf. Model.* **2009**, *49* (2), 308–317.

(16) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.

(17) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321* (5886), 263–266.

(18) Hopkins, A. L. Network pharmacology. *Nat. Biotechnol.* **2007**, *25* (10), 1110–1111.

(19) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13* (11), 2498–2504.

(20) Bader, G. D.; Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **2003**, *4* (Jan 13), 2.

(21) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25* (10), 1119–1126.

(22) Wishart, D. S. In silico drug exploration and discovery using DrugBank *Curr. Protoc. Bioinformatics* **2007**, Chapter 14, Unit 14.4.

(23) Lauss, M.; Kriegner, A.; Vierlinger, K.; Noehammer, C. Characterization of the drugged human genome. *Pharmacogenomics* **2007**, *8* (8), 1063–1073.

JM801546K