Analysis of Chemical and Biological Features Yields Mechanistic Insights into Drug Side Effects

Miquel Duran-Frigola[1] and Patrick Aloy[1,2,*]
[1]Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), c/ Baldiri Reixac 10-12, 08028 Barcelona, Spain
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain
*Correspondence: patrick.aloy@irbbarcelona.org

http://dx.doi.org/10.1016/j.chembiol.2013.03.017

SUMMARY

Side effects (SEs) are the unintended consequence of therapeutic treatments, but they can also be seen as valuable readouts of drug effects, resulting from the perturbation of biological systems by chemical compounds. Unfortunately, biology and chemistry are often considered separately, leading to incomplete models unable to provide a unified view of SEs. Here, we investigate the molecular bases of over 1,600 SEs by navigating both chemical and biological spaces. We identified characteristic molecular traits for 1,162 SEs, 38% of which can be explained using solely biological arguments, and only 6% are exclusively associated with the chemistry of the compounds, implying that the drug action is somewhat unspecific. Overall, we provide mechanistic insights for most SEs and emphasize the need to blend biology and chemistry to surpass intricate phenomena not captured in the molecular biology view.

INTRODUCTION

Side effects (SEs) are additional, usually undesirable, consequences of therapeutic treatments. Safety issues may arise at any developmental stage of a drug and even after its marketing and testing on the population at large (Giacomini et al., 2007). Apart from fatal postmarketing consequences, which are relatively rare, unwanted SEs can often lead to harmful or unpleasant reactions. The current picture in Western countries is that adverse drug reactions rank closely behind cancer and heart diseases as the major cause of mortality and morbidity (Wu et al., 2010).

Although unintended, drug SEs constitute a valuable readout of drug effects in humans, and accordingly, substantial efforts have been recently invested to catalog the drug-adverse events scattered in clinical reports (Tatonetti et al., 2012) and drug package inserts (Kuhn et al., 2010). When appropriately delineated in databases, SEs may be used as phenotypic profiles that can be linked with precise molecular data. However, this is not straightforward because SEs are the result of complex relationships emerging from the exposure of a biological system to a chemical entity and its degradation by-products, in the same way that are the therapeutic treatments. Unfortunately, biological and chemical entities are often considered separately, which leads to incomplete models unable to provide a unified view of SEs, and concerted strategies able to merge and quantify their individual contributions to the global behavior of the system are highly desirable.

The molecular biology view of drug action is that medicines exert their desired and undesired effects by interacting with molecular targets. This rationale inspired a systematic analysis to identify novel drug-target interactions using SE profiles (Campillos et al., 2008) and has been further reformulated to capture more complex biological entities, such as pathways (Wallach et al., 2010) or cellular processes (Lee et al., 2011). From this perspective, a SE could be seen as a consequence of the modulation of a known therapeutic target, for instance, because of its presence in different body tissues. This is the case of morphine, which achieves analgesic effect by modulating γ-opioid receptors in the brain and causes constipation when targeting the same receptors in the gut (Holzer, 2009). Alternatively, SEs can be explained by the binding of the drug molecule to nonintended proteins (i.e., off-targets). For example, Rescriptor, which is an HIV-1 reverse-transcriptase inhibitor, causes severe rashes by also interacting with the histamine H4 receptor (Keiser et al., 2009). However, unfortunately, only a few relationships between targets and SEs are well documented (Vedani et al., 2012), and despite remarkable recent progress (Lounkine et al., 2012), the mechanistic details for most undesired events remain unknown (Bauer-Mehren et al., 2012).

The traditional pharmacological view of drug action and discovery is somewhat different, as it was based on documenting the effects of drugs on animal models, rather than exploring and characterizing all biological mechanisms. Currently, there is a growing awareness that bioactivity data, and information on the behavior of a compound in a biological context, is mostly encoded within its chemical structure. This assumption has permitted to relate targets showing similar pharmacological profiles (Garcia-Serna et al., 2010; Keiser et al., 2007; Mestres et al., 2006) into clusters of seemingly disparate proteins, providing new insights into the function of the targets (Hillenmeyer et al., 2010).

It is known that some complex or unspecific biological phenomena can be triggered by very well-defined chemical features, such as the toxicity initiated by electrophilic compounds, formed either as a drug by-product (e.g., a quinoneimine from paracetamol), or as a result of enhanced cellular production of reactive oxygen or nitrogen species (Williams and Park, 2003). Recently, such structural alerts were mined in a collection of SEs (Bender et al., 2007) and used in a follow-up study to build a chemical map of adverse events that showed a systemic organization (Scheiber et al., 2009), similar to the pharmacological organization that targets show when described from the perspective of their ligands. However, the general applicability of this strategy is uncertain, and there are no estimates as to the number, and types, of chemical moieties that can trigger SEs independently of their protein targets. Moreover, it is currently unfeasible to decode structure-activity relationships (SARs) at the organism level for every drug molecule, because the data are scarce and they do not explore a significant region of the chemical space.

Some SEs can be understood by analyzing the biology of the modulated targets, whereas others are better explained by considering only the chemical properties of the drug compound, suggesting that no definitive methodology exists to approach drug SEs departing solely from molecular concepts. Indeed, the best event predictors are trained mostly from observational clinical reports that relate drug intakes to SEs (Cami et al., 2011; Tatonetti et al., 2012), with little attention paid to either biological or chemical details. Other attempts, alternatively, included all information available on drugs (Gottlieb et al., 2011) at the expense of interpretability. Yet, knowledge about the specific biological or chemical features able to trigger an adverse reaction is highly desirable as these molecular details can aid the drug discovery process by informing decisions on a binding assay choice or a moiety to avoid, for example, and, at the same time, by building a mechanistic explanation.

Universal solutions are rare in biology, and the faculty to explain a phenotypic event with a few molecular concepts is often a matter of choosing the right focus. Accordingly, we analyze here a comprehensive collection of drug SEs, and for each of them, we assess whether a biological or a chemical perspective, or a combination of both, is the most adequate to reveal its molecular bases. That is, we rationally blend two disparate perspectives to quantify their contribution to SEs, which should ultimately lead to a unified view of drug effects in humans.

RESULTS AND DISCUSSION

Our analysis consisted in a top-down approach, where drugs were classified at the phenotypic level, and associations with molecular characteristics were agnostically proposed therefrom. We gathered and organized molecular details for each compound in a way that represents the current biological and chemical mindsets (Figure 1).

The first layer of biological details included the intended targets of a drug molecule. Then we complemented therapeutic targets with proteins linked to the drug molecule in a chemical-protein interactome, in order to include possible off-target effects. Additionally, we considered the possibility that the modulation of distinct proteins participating in the same cellular process, or devoted to a particular function, might have similar phenotypic implications (Lee et al., 2011; Wallach et al., 2010). With these levels of biological organization, we expect to gain power in detecting simple biological descriptions that are related to SEs, although not necessarily through the interaction with the same proteins.

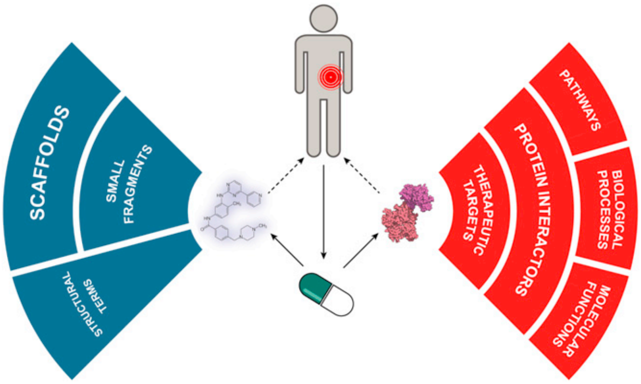

Figure 1. A Top-Down Approach into Drug Side Effects
Chemical and biological details of drugs are deployed in order to link detailed molecular features to phenotypic events.

Regarding the chemical perspective, we used in the first place small 2D fragments, which have proven useful to build interpretable SARs (Pauwels et al., 2011) and have the virtue of being recognizable by the organic and the medicinal chemist, despite not including properties like rigidity or lipophilicity (Oprea and Gottfries, 2001). Similarly, we built a network of molecular scaffolds (Varin et al., 2011) following Murcko's rules (Bemis and Murcko, 1996). These rules imply a recursive pruning of molecules and have been proposed to guide the navigation of chemical libraries in a manner that is biologically relevant (Wetzel et al., 2009). Finally, we also explored an organic chemistry vocabulary that organizes the structural features of compounds (de Matos et al., 2010).

Summary of the Methodology

A schema of the methodological strategy is presented in Figure 2. To identify overrepresented biological or chemical features that might be indicative of a given effect, we analyzed a collection of 1,626 clinical events extracted from 992 drug package inserts (Kuhn et al., 2010). We then classified the drugs into different SEs, where a single drug might occur in multiple SEs depending on the number of reported events, and we tested the overrepresentation of molecular features among drugs using appropriate statistical approaches (see the Experimental Procedures). The outcome of the enrichment analysis is, for each SE, a collection of different overrepresented traits, which could potentially play a role in the development of the SE. Our top-down approach generates simple associations at different depths of biological and chemical details (from therapeutic targets to pathways, and from small fragments to large scaffolds), thus favoring individual, well-defined, and testable findings (Figure 1).

Finally, we used the enriched molecular features to propose simple drug classifiers, which now allow for features to be combined, and evaluated their performance by means of the F_1-score (Equation 1), a common measure to balance precision and recall. Here, precision is calculated as the number of drugs correctly identified as causing a given SE (i.e., true positives) over the total number of drugs predicted to cause it (i.e., true positives + false positives), and the recall is given as fraction of true positives over the total number of drugs causing the SE

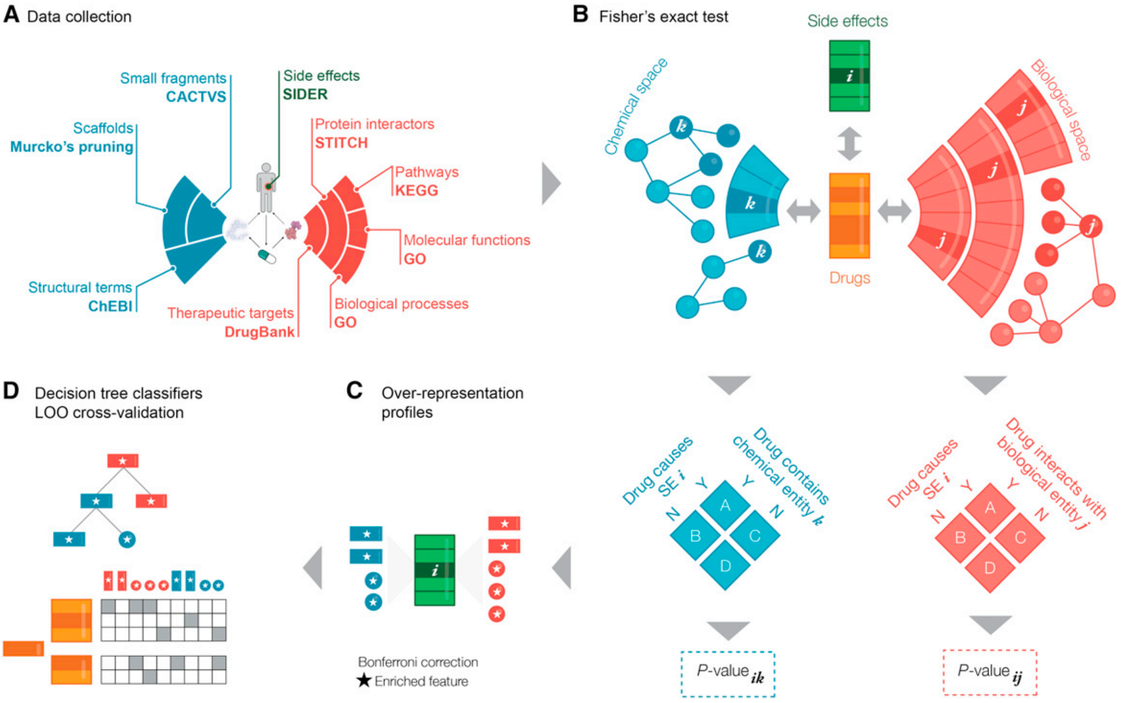

Figure 2. Scheme of the Methodology

(A) Drugs that cause each of the SEs are collected, and the biological and chemical profiles of these drugs are extracted from several resources.

(B) An enrichment analysis is performed for each SE feature pair, based on contingency tables counted on the number of drugs that cause the event and the number of drugs related to the feature.

(C) After a multiple testing correction, an overrepresentation profile is proposed for the different types of features.

(D) A cross-validated classification exercise helps to address the comprehensiveness of the overrepresentation profile and the possible advantages of combining observations of different types.

(i.e., true positives + false negatives). We would like to stress that these classifiers should not be considered as the endpoint of our analysis. As explained above, for an exhaustive collection of SEs, we perform a search for individual molecular traits that appear strongly associated to the events. In this context, an F_1-score measures if such molecular details are specific to the drugs that cause the event (precision) and account for most of the drugs (recall). Therefore, high F_1-scores correspond to those SEs for which our findings are sufficient to identify many of the related drugs, with strong certainty that safe compounds (i.e., those not causing the SE) do not share the proposed molecular profiles. On the contrary, low F_1-scores correspond to cases for which our findings only belong to a fraction of the SE-causing drugs, possibly without a specific and complete description because safe drugs exist with the traits associated with the event. Note that scarce drug safety records, and the fact that drug target profiles are not complete (Mestres et al., 2008), contribute to low F_1-scores. Thus, overall, we argue that classification performance can be seen as a measure of comprehensiveness of our molecular discoveries, i.e., the proportion of drugs to which molecular models are applicable and the certainty that drugs fulfilling the molecular requirement will cause the given SEs. Accordingly, F_1 is used to prioritize our SE models in this global analysis but should not be taken as an assessment of the true association between individual molecular traits and SEs, which we controlled in the statistical analysis. For discussion in this paper, we considered that combinations of biological and/or chemical features with F_1-scores > 0.5 provide descriptions that are fairly comprehensive.

Flagging Biological and Chemical Details

Figures 3 and 4A illustrate the features most commonly associated to SEs and three representative portions of the molecular models built around each SE, respectively. The complete results are provided in Tables S1 and S2 (available online). Concretely, for each SE, we indicate which biological and/or chemical features are overrepresented and assess their use to build classifiers with the F_1-score. We also propose an optimal classifier that may benefit from the integration of different feature types, specifying its precision and recall.

We could find overrepresented biological or chemical features in 1,162 (71%) SEs; all statistics hereafter refer to this subset of events. Out of these 1,162 SEs, we obtained a model with an F_1-score > 0.5 for 164 (14%) by using enriched features as parameters. These are cases for which we have identified combinations of biological and/or chemical traits whose occurrence is almost invariably associated to a SE (average precision of 0.74 within the [0.49–1] range) and that are present in most drugs known to cause this effect (average recall of 0.59 within the [0.36–1] range). On the other hand, in 432 (37%) of the SEs, our features were not sufficient to alert the event for any of the drugs in a cross-validation. In between, almost half of the SEs

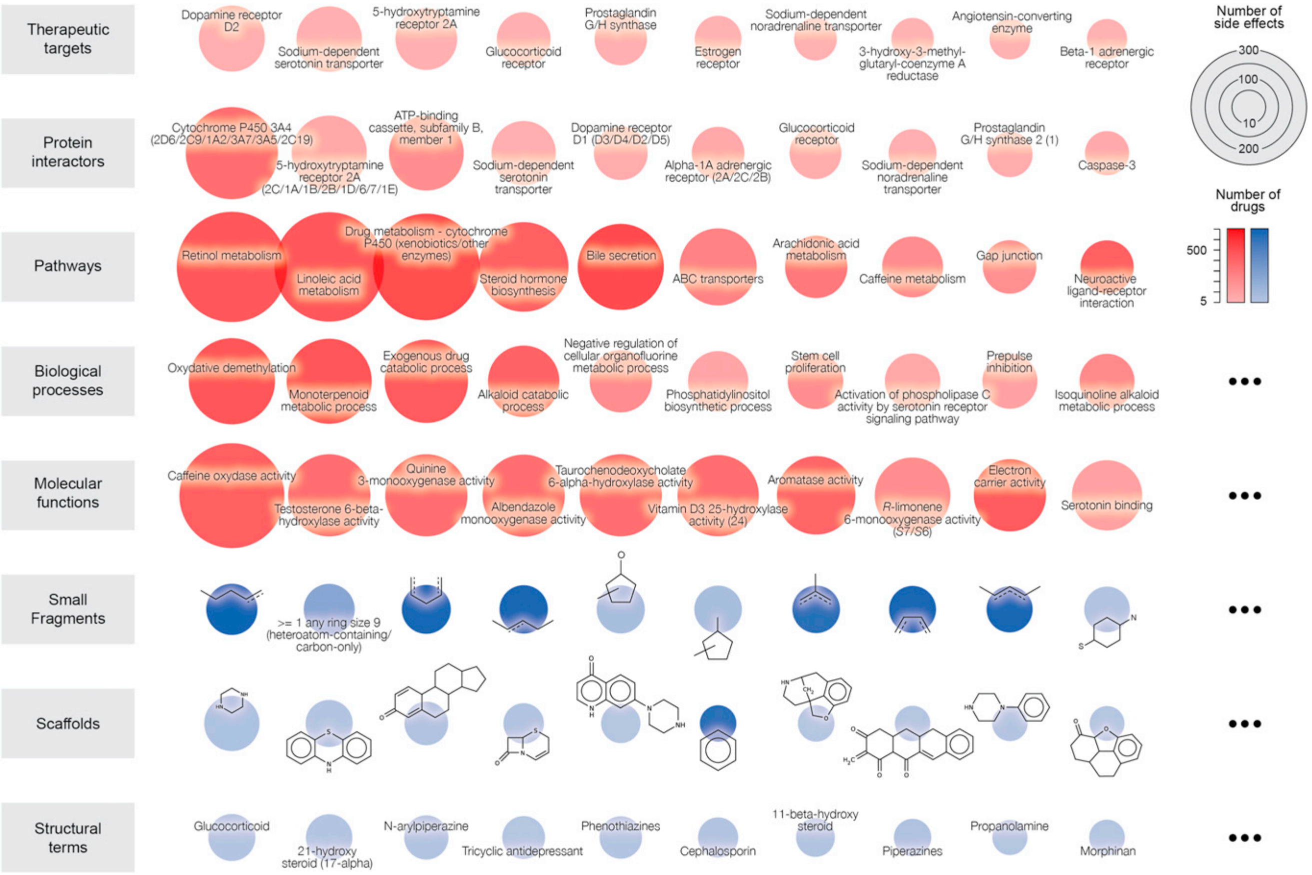

Figure 3. Features Most Commonly Associated to SEs

Each row represents a feature type, and features are ranked by decreasing number of associated SEs; only the top ten features are shown. For each feature, the size of the bubble is proportional to the number of associated SEs, and the opacity corresponds to the frequency of the feature among drugs in the data set. Some of the bubbles refer to more than one feature; in such cases, the properties of the bubble are those of the top-ranking feature, and the remaining features are specified within parentheses. A detailed description of all of the traits associated to SEs is provided in Table S2.

had an overrepresentation profile that was moderately comprehensive ($0 < F_1 \leq 0.5$), with an average precision and recall of 0.62 and 0.21, respectively.

Abundant, yet Insufficient, Biological Knowledge

We associated biological features to a majority (94%) of SEs. As expected, for most SEs we captured entities of different types (biological processes, pathways, proteins, etc.) because all these types refer to the same initial drug-protein interactome (Figure 4C). To assess these findings, we first checked whether our systematic approach was able to flag proteins that are commonly alerted by toxicologists, related to endocrine and metabolic disruption, cardiogenicity, and cardiotoxicity (Vedani et al., 2012). Thirteen out of the 16 proteins associated to these adverse events were present in our drug-protein interaction data. We found 10 of them overrepresented in at least one SE, and remarkably, 9 were enriched in a number of SEs above the median (Table S4). In addition, we conducted a bibliometric study to assess how many of the known associations we are able to recapitulate as well as how many have not been reported before. Our analyses suggest that the majority of the relationships identified may be novel, with the estimation that our biological explanation had already been reported for only 16% of the SEs (Table S3; Supplemental Experimental Procedures).

Interestingly, in 674 (58%) of the cases, the adverse and the therapeutic mechanisms correlated. For example, all of the 10 drugs intended to target the angiotensin-converting enzyme (ACE) cause pulmonary eosinophilia (Figure 4B), which agrees well with what is known about ACE inhibitors, their mechanism of action, and possible off-target effects (Trifilieff et al., 1993; Vasquez-Pinto et al., 2010).

A potential pitfall when considering only intended therapeutic targets is that we can erroneously infer noncausal associations between these and SEs, when the actual effectors are off-targets shared by molecules of the same therapeutic class. However, the real causal relationships should emerge if we include off-target proteins in the analyses. In doing so, we proposed liable proteins in 77% of the SEs. On average, six proteins were overrepresented per SE (Table 1). In principle, these findings include causal drug-protein interactions and can thus hint mechanistic insights. For instance, our analyses show that four out of the seven drugs that cause respiratory paralysis interact with the KCNQ1 and SCN10A voltage-gated channels, thereby interfering with the neuromuscular apparatus (Figure 4B), in

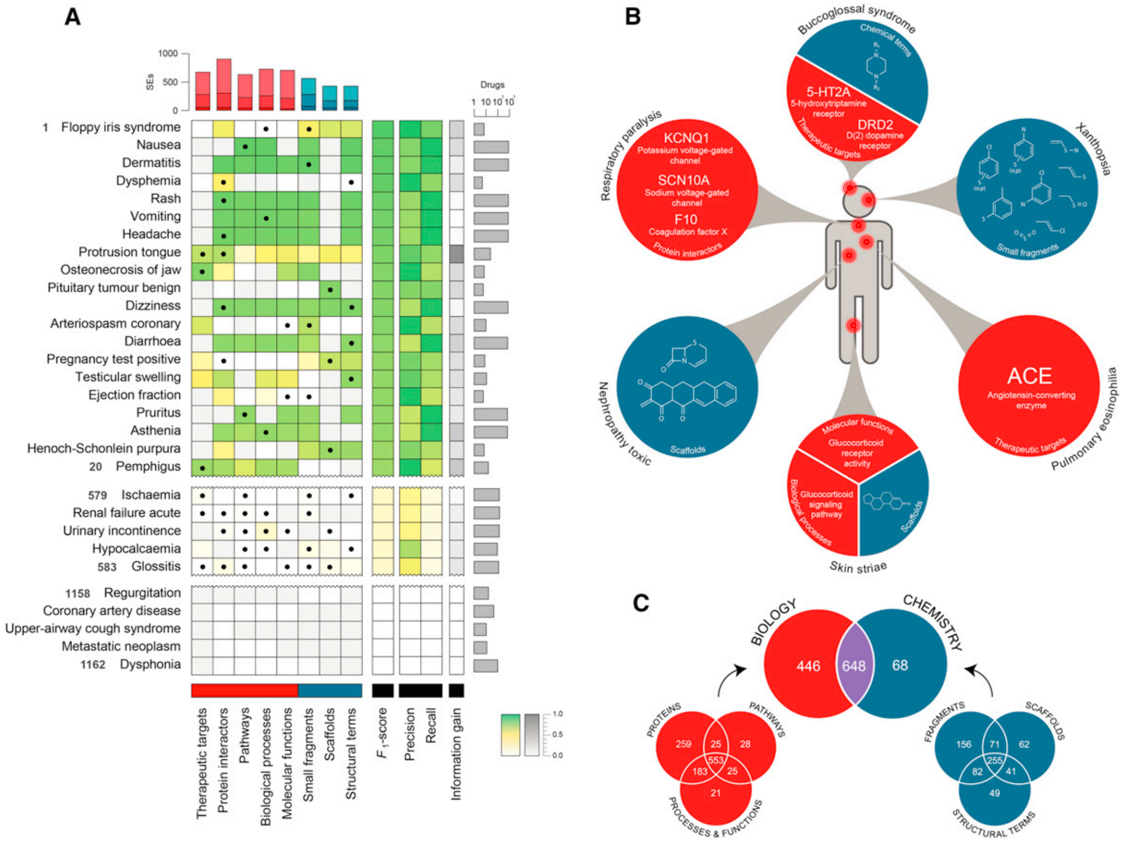

Figure 4. Overrepresented Features to Explain Drug SEs

(A) This panel shows three representative portions of the results SE-wise (see also Table S1). Filled cells in the main matrix refer to SEs for which we detected enrichment signals. Cells in the main matrix are color-graded according to the F_1-score. Columns on the right refer to the optimal classifiers, built by combining the spotted molecular views. In addition, a normalized measure of the information gain upon classification is given in gray-scale. The bar plot on the right represents the number of drugs that were reported to cause each SE. Finally, the bar plot on the top adds up the SEs with overrepresented features. The proportion of these that led to classifiers and contributed to an optimal model is progressively dark shaded.

(B) Illustrative examples linking adverse events to detailed molecular features.

(C) The underlying biology and chemistry of SEs. Venn diagrams join SE sets for which enrichment signals were found. For simplicity, the "proteins" circle groups the "therapeutic targets" and the "protein interactors" categories, and "processes and functions" joins "biological processes" and "molecular functions". See also Figures S1 and S2.

agreement with some work (Senanayake and Roman, 1992). The other three drugs, namely, dextrose, magnesium ion, and pyridostignin, are pharmacologically distant, but they are all reported to bind the coagulation factor X. How this translates into respiratory paralysis is more difficult to comment on, as to the best of our knowledge, mechanistic links between factor X and respiratory paralysis has not been established.

Beyond rationalizing and quantifying known associations between drug targets and SEs, our method is able to hypothesize relationships. An effective strategy to extract compelling mechanistic insights among these putative SE-protein associations is to look for overrepresented proteins that interact with disease-related genes (see the Supplemental Experimental Procedures for further details). For example, 35 drugs in our data set may induce ileus paralytic, which refers to intestinal pseudo-obstruction due to severe abnormality of gastrointestinal motility. Dopamine receptors (DRs) D2 and D3 are overrepresented as targets among these 35 drugs and have not been previously related to this SE. Also, drugs that target DRD2 and/or DRD3 do not interact with any of the ileus-related gene products (Davis et al., 2013), meaning that no alternative and trivial mechanism of action can be proposed. Interestingly, DRD2 and DRD3 physically interact with filamin A (FLNA); duplication of *FLNA* gene has been associated with intestinal pseudo-obstruction phenotypes (Clayton-Smith et al., 2009). Also, FLNA is known to play an important and selective role in the localization of DRD2/3 (Lin et al., 2001). Thus, we hypothesize that blockade of DRD2 and/or DRD3 may result in aberrant behavior of this mechanism, possibly acting on the same molecular processes that are altered upon the disease-related duplication of the *FLNA* gene.

Another interesting example regards the pathologic deposition of calcium salts in tissues, termed calcinosis. We found that three of the seven drugs that cause calcinosis interact with matrix metallopeptidase 1 (MMP1). This enzyme shows a certain specificity to cleave the monocyte chomoattractant protein 2 (MCP2) (McQuibban et al., 2002), and the integrity of the MMP1 catalytic domain has been proposed to be required for

Table 1. Summary of Results

	SEs	SE-Features	Distinct Features	$F_1 > 0.5$
Therapeutic targets	674	1,457	79	24
Protein interactors	905	5,466	323	40
Pathways	631	4,281	154	31
Biological processes	727	9,727	835	38
Molecular functions	705	5,766	310	43
Small fragments	564	6,630	452	44
Scaffolds	429	586	74	19
Structural terms	427	694	105	26
Biology	1,904	2.7×10^4	1,701	73
Chemistry	716	7.9×10^3	631	59

"SEs" refers to the number of SEs for which enriched features were found. "SE-Features" is the total number of SE-feature relationships; "Distinct Features" counts the unique features that were overrepresented in at least one SE. Finally, "$F_1 > 0.5$" refers to the number of comprehensive models.
See also Tables S3 and S4.

complex formation with its natural inhibitor, tissue inhibitor of metallopreoteinases-1 (TIMP1) (Vallon et al., 1997). Drugs that interact with MMP1 can thus interfere with its ability to engage MCP2 and TIMP1. Interestingly, these two genes were detected in a recent study to rationalize therapeutic effects in aortic valve calcification (Shuvy et al., 2011), strongly suggesting that drug interference with MMP1 could trigger calcinosis.

Whereas some SEs can be rationalized as shown in the previous examples, others are more related to drug metabolism issues. In this regard, it is worth noting that cytochromes P450 were recurrently overrepresented, e.g., CYP3A4 was associated with 214 SEs (Figure 3). This observation is in good agreement with the recent estimate that targets of drug metabolites may be involved in up to 40% of the adverse events (Bauer-Mehren et al., 2012). We decided not to include drug metabolites in our analysis, because drug by-products are only identified upon administration in body systems or predicted, most notably, by means of cytochrome catabolism experiments (Fura, 2006). Our approach illustrates the tendency to overlook the identity of metabolites at the drug discovery stage (Fura, 2006). In line with this, and aware of the lack of completeness in drug-protein interaction data (Mestres et al., 2008), we propose that our biological insights may be read as a quantification of the usefulness of our current knowledge on the molecular biology of medicines to explain their adverse events.

Remarkably, we could only propose a model with $F_1 > 0.5$ for 73 (6%) SEs, even when we had 2.7×10^4 SE associations with proteins, pathways, biological processes, and molecular functions (Table 1). This means that, although there is already abundant biological information, it is not yet enough to explain the complexity of effects observed in clinical trials. Moreover, the predictive power of proteins was similar to that of pathways, biological processes, or molecular functions, suggesting that, in the majority of cases, these annotations alone are not enough to scale-up to a phenotype. In order to gain predictive power, our identification of potentially harmful proteins should be integrated with other types of biological data (i.e., presence of paralogs that can compensate for each other's function, backup circuits, expression in different tissues) into systems biology models able to weight their contribution and establish the functional causes of SEs. For this purpose, our findings, which constitute individual biological alerts, can be particularly useful as seeds in the elaboration of complex network biology models (Soler-López et al., 2011).

Predictive, yet Limited, Chemical Details

We detected chemical traits for 716 (61%) SEs. On average, we spotted 11.8 small fragments, 1.4 scaffolds, and 1.6 structural terms for each SE. The several small fragments related to each SE may be used in SAR models, whereas the more informative scaffolds and structural terms may be proposed as liable features to avoid when designing safer drugs. It should be noted that molecules in this study correspond to optimized chemotypes, and thus our findings are enriched with structural alerts that might not be currently known to have SEs. Interestingly, 19% of the associated small fragments could be found among disease-related environmental chemicals (Table S4).

Overall, we collected 7.9×10^3 associations between SEs and chemical traits, three times less than in the case of biological features. Still, we could generate 59 comprehensive classifiers solely based on chemical details (Table S1). Of these, 27 lacked a plausible biological model, suggesting that the chemistry of the compounds alone may be responsible for the SEs. Overall, we observe that, although the number of chemical features associated to a SE is scarcer than biological traits, their predictive power is on average superior. An example is xanthopsia, which refers to the predominance of yellow in vision. In drugs causing it, we observed an overrepresentation of the small fragments depicted in Figure 4B. The biological mechanism underlying xanthopsia is not yet clear, despite the interest aroused by the suspicion that this condition influenced Vincent van Gogh's "yellow period" (Arnold and Loftus, 1991). Nonetheless, the small fragments we flagged, combined to form e.g., thiazides or sulfonamides, recalled with full precision 7 out of the 13 drugs that cause this SE. Similarly, by considering only the cephalosporin and the tetracycline scaffolds, we predicted nephropathotoxicity with 47% recall and 85% precision. Several antibiotics are mounted on these two scaffolds long known to be toxic in the kidney. In fact, the first attempts to elucidate the mechanism of cephalosporin toxicity were purely chemocentric (Tune and Fravert, 1980); we recovered this knowledge in an unbiased systematic fashion.

One can suspect that, if no biological signal is found, then the drug action is somewhat unspecific. We found that as low as 6% of the 1,162 SEs can exclusively be associated with chemistry (Figure 4C), implying that only a small proportion of SEs appear unrelated to drug targets. However, this may well be an underestimate given the abundance of overtargeted cytochromes in our findings: drug-metabolizing enzymes can certainly produce reactive by-products, thereby leading to nonspecific alterations.

Chemistry to Complement Biology

For 648 (56%) SEs, we detected both biological and chemical features. In some cases, the two views seemed complementary. For instance, uncontrolled movement of the body, termed buccoglossal syndrome, is caused by six drugs in our data set. Buccoglossal syndrome is a common event among antipsychotics (Blanchet et al., 2012). Accordingly, we found an

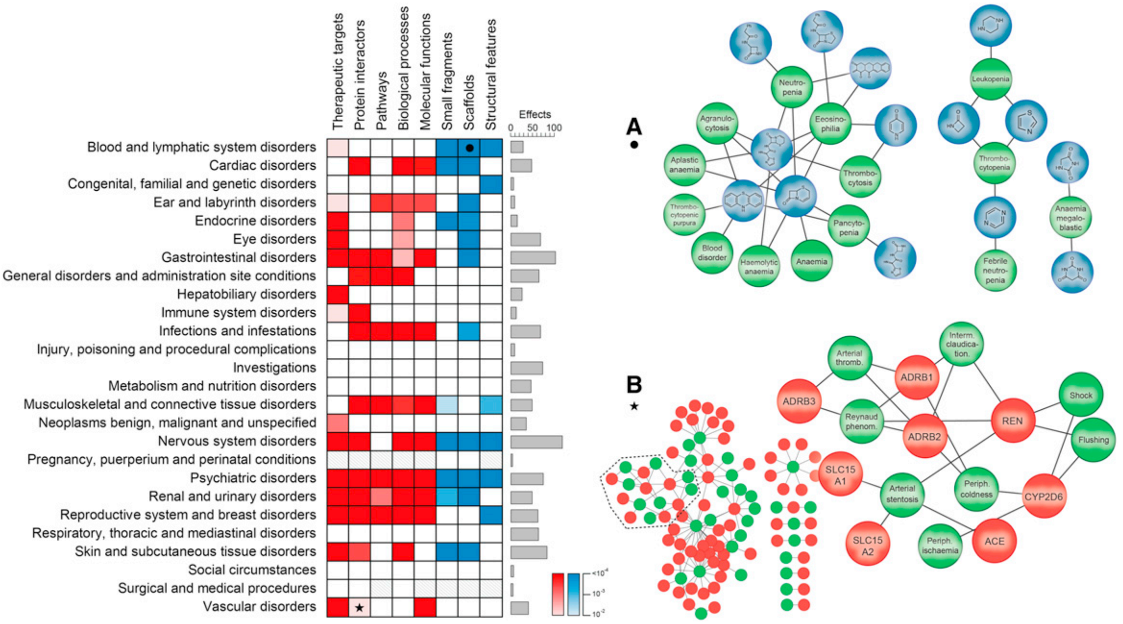

Figure 5. Systemic Clustering of Biological and Chemical Details

(A and B) Filled cells correspond to SOCs annotated to SEs with overrepresented features. The clustering of SEs within each SOC is given a significance *P-value*. As illustrative examples, we highlight chemical scaffolds related to blood and lymphatic disorders (A) and proteins associated to vascular disorders (B).

overrepresentation of drugs that interact with the 5-hydroxytriptamine receptor 5-HT2A and the dopamine receptor DRD2. Also, we found that a significant proportion of the molecules that cause this SE contain the privileged piperazine moiety (Figure 4B). There are 44 molecules that target 5-HT2A and/or DRD2, and 33 of those contain the piperazine ring, but only 6 of them are known to cause the syndrome. Hence, models built with these features separately failed to correctly identify such drugs. However, by screening for 5-HT2A and/or DRD2 modulators that contain a piperazine ring, we could classify drugs with 80% precision and 67% recall. For other SEs, we observed redundant signals: 25 out of the 31 drugs that cause skin striae contain the steroidal scaffold depicted in Figure 4B. In parallel, we also found skin striae related to glucocorticoid signaling, which is mediated by steroidal hormones. Here, both the chemical and the biological views refer to the modulation of hormonal receptors in the skin, well known to be involved in striae formation (Blanchet et al., 2012).

Overall, 57% of the comprehensive classifiers were obtained by mixing biological and chemical details. Accordingly, we propose that chemical information, usually ciphered and limited, could be used to account for unspecific or intricate phenomena unperceived by classifiers based on the molecular biology of drugs.

A Systemic Organization

We finally assessed to what extent the identified feature-SE relationships are common within SEs occurring in the same system or organ class (SOC), since this could be a good indicator to guide toxicological tests without the need for full body systems. With this aim, we used all of the SE-feature associations (Table S2) and calculated the overlap of enriched features among the SEs annotated to each SOC, as defined in the MedDRA terminology (http://medrramsso.com). As shown in Figure 5, psychiatric disorders, for instance, map to characteristic regions both on the chemical and the biological spaces. On the other hand, for metabolism and nutritional disorders our findings are sparse, suggesting that a diverse spectrum of drug-protein interactions and chemical moieties may elicit them. Interestingly, issues in the blood and lymphatic systems are poorly organized by the biological view but are related to a well-defined chemistry (Figure 5A), in good agreement with the isolated chemical cluster found in Scheiber et al. (2009). Conversely, there was a collection of targets that mostly correlated with vascular disorders. In Figure 5B, we zoom into a cluster of vascular events related to adrenergic receptors (ADRB1/2/3), the renin-angiotensin system (REN and ACE) and drug metabolism (CYP2D6 and SLC15A1/2). This protein-centered view is compatible with the recent claim that molecular biology can model satisfactorily cardiovascular complications (Berger and Iyengar, 2011).

SIGNIFICANCE

We have performed an exhaustive search for chemical and biological molecular traits in a diverse collection of human phenotypic responses to drug treatments. Taking an agnostic approach, we have been able to recover knowledge accumulated over the years and, most importantly, suggest molecular mechanisms mediating drug SEs that are poorly understood. However, although we discovered biological traits related to most SEs, we were only able to suggest a comprehensive model for a relatively small proportion of them. Notably, our results also highlight that, for half of the SEs with underlying biological causes, the

adverse and the therapeutic mechanisms of action correlate. On the other hand, a chemocentric approach seemed more assertive, even though chemical features were harder to flag. We found that roughly 6% of the SEs are exclusively associated to the chemistry of the compounds, perhaps implying that the drug action is unspecific. As expected, our analyses emphasize the need to blend biology and chemistry to provide molecular explanations of complex SEs.

Overall, we provide sets of detailed biological and chemical features that are strongly overrepresented among the compounds causing a given event. We have presented some illustrative examples, but, to get the most out of our analysis, the overrepresented features detailed in Table S2 should be considered along the drug discovery process. Additionally, our findings confirm that reductionist approaches are often insufficient to anticipate drug SEs (Berger and Iyengar, 2011), where genetic variability is known to play an important role (Becquemont, 2009; Gurwitz and Motulsky, 2007). We anticipate that, for a significant number of events, the emerging field of systems biology will help to understand how complex signals propagate through cellular networks, eventually unmasking silent phenotypes.

EXPERIMENTAL PROCEDURES

The experimental pipeline is schemed in Figure 2. First, data is collected. Then enrichment analyses are performed exhaustively for all of the SEs and for each of the molecular features. The overrepresentation profile for each of the SEs constitutes the bulk of our results. Finally, we perform a prediction step to assess the comprehensiveness of our findings.

Data Sets

Drug Side Effects

We collected information on drugs and their SEs from SidER (v. 2) (Kuhn et al., 2010), a resource of compiled drug package inserts that uses the MedDRA dictionary, which is accepted as a standard, clinically validated terminology for SE reporting. SEs were fetched at the ‘‘preferred term’’ (PT) level and labeled with the corresponding ‘‘system organ classes’’ (SOCs) from MedDRA (v. 14.1). We discarded SEs caused by less than 5 drugs to ensure statistical robustness, obtaining a space of clinical events composed of a set D of 992 drugs related to a set S of 1,626 MedDRA PTs. Drug SEs were considered regardless of their frequency.

Biological Space

Therapeutic Targets

The DrugBank database (Knox et al., 2011) combines detailed drug data with comprehensive drug target information, providing a rich picture of the current knowledge on the pharmacology of drugs. Accordingly, the human targets with known pharmacological action in DrugBank constitute a plausible annotation of therapeutic targets. Drugs were mapped to the DrugBank database (v. 3) using their PubChem (http://pubchem.ncbi.nlm.nih.gov) compound identifiers. The mapping was further covered using name and structure matching. We achieved the latter by encoding InChIKeys after stripping salts, removing hydrogen atoms, and merging stereochemistry with OpenBabel (v. 2.3.1) (O’Boyle et al., 2011). Overall, we fetched 88 human proteins targeted by at least 5 drugs in D; this same criterion was used for the rest of the feature types below.

Protein Interactors

The STITCH database (Kuhn et al., 2012) is an aggregate repository that captures as much as possible of the publicly available knowledge on protein-chemical interactions. We matched the compounds in D to the flat compounds in STITCH (v. 3) and retrieved their human protein interactors. We required a confidence score higher than 0.7 supported by either databases or experiments. In total, we collected 702 gene products from STITCH.

Pathways

To link drugs and human pathways, we mapped STITCH proteins to KEGG (Kanehisa and Goto, 2000) using the KEGG.db Bioconductor package (v. 2.9) (Gentleman et al., 2004). 203 pathways entered the analysis.

Biological Processes and Molecular Functions

The full gene ontology (GO) was downloaded (data v. 1.1.2491) and separated into three directed acyclic graphs (DAGs), namely, ‘‘biological process’’, ‘‘molecular function’’, and ‘‘cellular component’’, considering only ‘‘is a’’ relationships. We annotated STITCH proteins to the leaves of the ‘‘biological process’’ and ‘‘molecular function’’ subontologies using the Ensembl cross-references of the biomaRt Bioconductor package (Gentleman et al., 2004). The subgraphs spanning all of the protein annotations from the leaves to the roots had a size of 6,353 and 2,123 nodes, respectively.

Chemical Space

Small Fragments

PubChem provides 2D structural descriptions of molecules through CACTVS fingerprints. CACTVS fingerprints denote the presence or absence of 881 structural features. Although not strictly limited to small fragments, CACTVS representations embrace a wide spectrum of structural concepts, such as element counts, types of ring systems, or atom pairings. In total, we considered 552 CACTVS fragments.

Scaffolds

A scaffold network was generated as described in Varin et al. (2011). A scaffold network of chemicals consists of molecular scaffolds and smaller parent scaffolds generated by the pruning of rings (Bemis and Murcko, 1996), effectively leading to a map of substructure relationships. The resulting network from PubChem structures of D formed a DAG of 197 nodes, with minimal scaffolds as roots and the molecules without terminal side chains as leaves.

Structural Terms

ChEBI (de Matos et al., 2010) is a dictionary of molecular entities that includes groups and classes of structures (e.g., imatinib is annotated in ChEBI with ‘‘diazine’’ and ‘‘piperazine’’ terms). The ChEBI ontology (data v. 90) was downloaded and converted to a DAG by removing cyclic relationships (‘‘is tautomer of’’, ‘‘is enantiomer of’’, etc.) and merging the others into a single ‘‘is a’’-like relationship (Ferreira and Couto, 2010). The strategy to map D onto ChEBI was analogous to that onto DrugBank. We spanned the ‘‘chemical entity’’ subgraph as it was done for the GO terms, obtaining a network of 467 nodes.

Analyses

Feature Enrichments

For each SE, we aimed at profiles of overrepresented molecular features of the eight types above (‘‘therapeutic targets’’, ‘‘small fragments’’, etc.). We performed multiple univariate right-tailed Fisher’s exact tests in each feature type. Univariate Fisher’s test were calculated under the null hypothesis that drugs that cause the SE are not more likely to contain the molecular feature of interest. Only features belonging to at least one of the SE-related drugs were tested, and we applied a Bonferroni multiple testing correction to keep the family-wise error rate below 0.05.

In the case of features with an inherent tree-like structure (i.e., GO terms, scaffolds, and ChEBI structural terms), the ‘‘Elim’’ method was implemented to investigate from the most specific features to the most general ones (Alexa et al., 2006). Briefly, the algorithm starts processing the nodes from the leaves and iteratively travels to parental nodes. Because nodes from the same level share no edge, they can be investigated independently, and the Fisher’s tests above may be performed. When a node is processed, the drugs responsible for a significant enrichment signal in a previous step are removed from the test, thus prioritizing specific overrepresentations and decorrelating the DAG structure to provide distinct findings.

Performance of Findings as Classifiers

The discriminative potential of overrepresented features was examined by using decision tree classifiers based on Gini’s impurity as implemented in the Scikit-learn Python package (Pedregosa et al., 2011). We evaluated the accuracy of the classification by means of the F_1-score after a leave-one-out (LOO) cross-validation, which respects sample proportions (see the Supplemental Experimental Procedures for further information). In a LOO cross-validation, the SE model is repeatedly refit leaving out one drug and then used to derive the prediction for such left-out compound. After a prediction is derived for all of the drugs, the precision can be calculated as the number of true positives over the total number of positive predictions, whereas the recall scales the true positives over the total number of drugs known to cause the SE and is thus a measure of sensitivity. The F_1 score balances the precision p and the recall r, i.e., focuses on the true positive rate:

$$F_1 = 2 \times \frac{p \times r}{p + r} \qquad \text{(Equation 1)}$$

In order to investigate whether a given SE might be better described from a composition of perspectives, we also proposed classification schemes for all possible combinations of feature types. Combined models were built from overrepresented features of each of the selected types. Eventually, the model with a higher F_1 was selected for each SE, and the information gain upon classification was measured using Shannon entropies. Here, the probability of anticipating a SE for a given drug when using the model is compared to the probability of anticipating the event solely based on SE occurrence.

Similarity within System Organ Classes

For each MedDRA SOC, we collected the associated SEs and computed their pair-wise similarity based on the previously determined enrichment signals, i.e., two SEs are similar if they have similar overrepresentation profiles. For independent features, we used the Jaccard coefficient of similarity. For features organized in a DAG, we used the set-set Lin's semantic similarity (Pesquita et al., 2009), where the proportion of drugs annotated to each of the overrepresented features was used to derive the information content required by Lin's measure. Feature-feature similarities were based on the information content of the most informative common ancestor and ultimately used to compute the similarity between sets of features.

The similarity of SEs within a SOC was then regarded as the average of the pair-wise SE similarities, and we estimated the significance of this measurement by bootstrapping 10^4 samples of SE sets.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, Tables S1–S4, and Figures S1–S4 and can be found with this article online at http://dx.doi.org/10.1016/j.chembiol.2013.03.017.

ACKNOWLEDGEMENTS

We would like to thank David Rossell, Samira Jaeger, Andreas Zanzoni, and Modesto Orozco (IRB Barcelona) for helpful discussions and critically reading our manuscript. This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (BIO2010-22073) and the European Commission under FP7 grant agreement 223101 (AntiPathoGN). M.D.-F. is a recipient of a Spanish FPU fellowship. M.D.-F. and P.A. designed the study, analyzed the results, and wrote the manuscript. M.D.-F. performed the experiments.

Received: November 14, 2012
Revised: March 8, 2013
Accepted: March 25, 2013
Published: April 18, 2013

REFERENCES

Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

Arnold, W.N., and Loftus, L.S. (1991). Xanthopsia and van Gogh's yellow palette. Eye (Lond.) *5*, 503–510.

Bauer-Mehren, A., van Mullingen, E.M., Avillach, P., Carrascosa Mdel, C., Garcia-Serna, R., Pinero, J., Singh, B., Lopes, P., Oliveira, J.L., Diallo, G., et al. (2012). Automatic filtering and substantiation of drug safety signals. PLoS Comput. Biol. *8*, e1002457.

Becquemont, L. (2009). Pharmacogenomics of adverse drug reactions: practical applications and perspectives. Pharmacogenomics *10*, 961–969.

Bemis, G.W., and Murcko, M.A. (1996). The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. *39*, 2887–2893.

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

Berger, S.I., and Iyengar, R. (2011). Role of systems pharmacology in understanding drug adverse events. Wiley Interdiscip. Rev. Syst. Biol. Med. *3*, 129–135.

Blanchet, P.J., Parent, M.T., Rompre, P.H., and Levesque, D. (2012). Relevance of animal models to human tardive dyskinesia. Behav. Brain Funct. *8*, 12.

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

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

Clayton-Smith, J., Walters, S., Hobson, E., Burkitt-Wright, E., Smith, R., Toutain, A., Amiel, J., Lyonnet, S., Mansour, S., Fitzpatrick, D., et al. (2009). Xq28 duplication presenting with intestinal and bladder dysfunction and a distinctive facial appearance. Eur. J. Hum. Genet. *17*, 434–443.

Davis, A.P., Murphy, C.G., Johnson, R., Lay, J.M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Rosenstein, M.C., Wiegers, T.C., et al. (2013). The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. *41*, D1104–D1114.

de Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical entities of biological interest: an update. Nucleic Acids Res. *38*, D249–D254.

Ferreira, J.D., and Couto, F.M. (2010). Semantic similarity for automatic classification of chemical compounds. PLoS Comput. Biol. *6*, pii: e1000937.

Fura, A. (2006). Role of pharmacologically active metabolites in drug discovery and development. Drug Discov. Today *11*, 133–142.

Garcia-Serna, R., Ursu, O., Oprea, T.I., and Mestres, J. (2010). iPHACE: integrative navigation in pharmacological space. Bioinformatics *26*, 985–986.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

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

Gottlieb, A., Stein, G.Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. *7*, 496.

Gurwitz, D., and Motulsky, A.G. (2007). 'Drug reactions, enzymes, and biochemical genetics': 50 years later. Pharmacogenomics *8*, 1479–1484.

Hillenmeyer, M.E., Ericson, E., Davis, R.W., Nislow, C., Koller, D., and Giaever, G. (2010). Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. Genome Biol. *11*, R30.

Holzer, P. (2009). Opioid receptors in the gastrointestinal tract. Regul. Pept. *155*, 11–17.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

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

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., et al. (2009). Predicting new molecular targets for known drugs. Nature *462*, 175–181.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. *39*, D1035–D1041.

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

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J., and Bork, P. (2012). STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res. *40*, D876–D880.

Lee, S., Lee, K.H., Song, M., and Lee, D. (2011). Building the process-drug-side effect network to discover the relationship between biological processes and side effects. BMC Bioinformatics *12*(*Suppl 2*), S2.

Lin, R., Karpa, K., Kabbani, N., Goldman-Rakic, P., and Levenson, R. (2001). Dopamine D2 and D3 receptors are linked to the actin cytoskeleton via interaction with filamin A. Proc. Natl. Acad. Sci. USA *98*, 5258–5263.

Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Cote, S., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. Nature *486*, 361–367.

McQuibban, G.A., Gong, J.H., Wong, J.P., Wallace, J.L., Clark-Lewis, I., and Overall, C.M. (2002). Matrix metalloproteinase processing of monocyte chemoattractant proteins generates CC chemokine receptor antagonists with anti-inflammatory properties in vivo. Blood *100*, 1160–1167.

Mestres, J., Martin-Couce, L., Gregori-Puigjane, E., Cases, M., and Boyer, S. (2006). Ligand-based approach to in silico pharmacology: nuclear receptor profiling. J. Chem. Inf. Model. *46*, 2725–2736.

Mestres, J., Gregori-Puigjane, E., Valverde, S., and Sole, R.V. (2008). Data completeness—the Achilles heel of drug-target networks. Nat. Biotechnol. *26*, 983–984.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: an open chemical toolbox. J. Cheminform. *3*, 33.

Oprea, T.I., and Gottfries, J. (2001). Chemography: the art of navigating in chemical space. J. Comb. Chem. *3*, 157–166.

Pauwels, E., Stoven, V., and Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics *12*, 169.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Pesquita, C., Faria, D., Falcao, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. PLoS Comput. Biol. *5*, e1000443.

Scheiber, J., Jenkins, J.L., Sukuru, S.C., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., et al. (2009). Mapping adverse drug reactions in chemical space. J. Med. Chem. *52*, 3103–3107.

Senanayake, N., and Roman, G.C. (1992). Disorders of neuromuscular transmission due to natural environmental toxins. J. Neurol. Sci. *107*, 1–13.

Shuvy, M., Abedat, S., Beeri, R., Valitsky, M., Daher, S., Kott-Gutkowski, M., Gal-Moscovici, A., Sosna, J., Rajamannan, N.M., and Lotan, C. (2011). Raloxifene attenuates Gas6 and apoptosis in experimental aortic valve disease in renal failure. Am. J. Physiol. Heart Circ. Physiol. *300*, H1829–H1840.

Soler-López, M., Zanzoni, A., Lluis, R., Stelzl, U., and Aloy, P. (2011). Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. Genome Res. *21*, 364–376.

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

Trifilieff, A., Da Silva, A., and Gies, J.P. (1993). Kinins and respiratory tract diseases. Eur. Respir. J. *6*, 576–587.

Tune, B.M., and Fravert, D. (1980). Mechanisms of cephalosporin nephrotoxicity: a comparison of cephaloridine and cephaloglycin. Kidney Int. *18*, 591–600.

Vallon, R., Muller, R., Moosmayer, D., Gerlach, E., and Angel, P. (1997). The catalytic domain of activated collagenase I (MMP-1) is absolutely required for interaction with its specific inhibitor, tissue inhibitor of metalloproteinases-1 (TIMP-1). Eur. J. Biochem. *244*, 81–88.

Varin, T., Schuffenhauer, A., Ertl, P., and Renner, S. (2011). Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. J. Chem. Inf. Model. *51*, 1528–1538.

Vasquez-Pinto, L.M., Nantel, F., Sirois, P., and Jancar, S. (2010). Bradykinin B(1) receptor antagonist R954 inhibits eosinophil activation/proliferation/migration and increases TGF-beta and VEGF in a murine model of asthma. Neuropeptides *44*, 107–113.

Vedani, A., Dobler, M., and Smiesko, M. (2012). VirtualToxLab - a platform for estimating the toxic potential of drugs, chemicals and natural products. Toxicol. Appl. Pharmacol. *261*, 142–153.

Wallach, I., Jaitly, N., and Lilien, R. (2010). A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. PLoS ONE *5*, e12063.

Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T.I., Mutzel, P., and Waldmann, H. (2009). Interactive exploration of chemical space with Scaffold Hunter. Nat. Chem. Biol. *5*, 581–583.

Williams, D.P., and Park, B.K. (2003). Idiosyncratic toxicity: the role of toxicophores and bioactivation. Drug Discov. Today *8*, 1044–1050.

Wu, T.Y., Jen, M.H., Bottle, A., Molokhia, M., Aylin, P., Bell, D., and Majeed, A. (2010). Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. J. R. Soc. Med. *103*, 239–250.