A chemo-centric view of human health and disease

Miquel Duran-Frigola[a], David Rossell[b,c] and Patrick Aloy[a,d,*]

a. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB) Barcelona, c/ Baldiri Reixac 10-12, 08028 Barcelona, Spain

b. Biostatistics and Bioinformatics Unit, Institute for Research in Biomedicine (IRB) Barcelona, Spain

c. Department of Statistics, University of Warwick, Coventry, United Kingdom

d. Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain

*** Corresponding author:** Patrick Aloy. Institute for Research in Biomedicine (IRB) Barcelona, c/ Baldiri Reixac 10-12, 08028 Barcelona, Spain. Tel: +34 934039690; Email: patrick.aloy@irbbarcelona.org

Keywords: Fragment Mining, Disease Models, Disease Networks

Abstract

Efforts to compile the phenotypic effects of drugs and environmental chemicals offer the opportunity to adopt a chemo-centric view of human health. In this manuscript we consider thousands of chemicals and analyze their relationship with adverse and therapeutic responses. Our study includes molecules related to the etiology of 934 health threatening conditions and used to treat 835 diseases. We first identify chemical moieties that could be independently associated with each phenotypic effect, balancing interpretation and prediction efficiency to maximize the biological relevance of the reported molecular fragments. Using these fragments, we build accurate predictors for approximately 400 clinical phenotypes, finding many privileged and liable structures. Finally, we connect two diseases if they relate to a similar chemistry. The resulting networks of human conditions are able to predict disease comorbidities, as well as identifying potential drug side effects and opportunities for drug repositioning, and show a remarkable coincidence with clinical observations.

Humans, in their daily lives, are exposed to a great variety of chemicals, including drugs and environmental hazards. Therapeutic and adverse effects of these chemicals result from a complex interplay with the human body. It is now recognized that, in most cases, a reductionist viewpoint of such interplay is far from reality. Cumulative evidence shows that even the most thoughtfully specific drugs elicit promiscuous interaction profiles[1] and, accordingly, many adverse chemical events lack a compelling molecular explanation[2]. The emerging opinion is that systems biology strategies —that integrate several layers of detail and complexity— will be necessary to zoom out from a reductionist to a more holistic picture of pharmacology and toxicology[3].

As human biology continues to reveal itself more and more intricate, it is suggestive to realize that much information about the behavior of a chemical inside our bodies is encoded within a small molecule, with few bonds and atoms. Decoding correlations between the structure of a compound and its activity in biological systems has been a prolific research area, and the major goal of earliest pharmacologists[4]. Unfortunately, such a compound-centered view of phenotypes is blind to molecular mechanisms, lacking theoretical support and, therefore, requiring a considerable amount of bioactivity data. In particular, for humans, experiments to obtain this information cannot be conceived, and the bulk of chemical activity assays is placed several translational steps backward (i.e. at the level of single receptor binding), with the consequent reduction of the system complexity.

Recent advances in text-mining techniques and subsequent curation efforts are committed to compiling direct human response data from the knowledge accumulated through the years[5,6]. Here, we benefit from this enterprise to revisit the classical structure-activity relationship notion, this time for a vast and diverse list of human diseases. Concretely, we first delve into chemical structures to identify fragments that are associated with adverse or therapeutic responses. Then, we propose disease models based on these fragments, and assess their predictive efficiency. Finally, we use such models to relate diseases, providing a chemical map of human phenotypes.

Results and Discussion

Several resources exist that contain information on the interaction of small molecules with our health. Most notably, the Comparative Toxicogenomics Database (CTD)[5] is mainly

focused on environmental chemicals, and reports curated relationships with a comprehensive list of diseases. Moreover, it classifies disease annotations as 'Marker/Mechanism' (M) or 'Therapeutic' (T). M refers to a chemical that correlates with the disease (i.e. a marker) or may act in its etiology (i.e. a toxin), while T indicates that the chemical has a known or a potential therapeutic role in the condition (i.e. a drug). By analogy, hereafter we refer to adverse and therapeutic disease outcomes simply as M and T diseases, respectively.

Research worldwide is conducted at different levels of detail and, accordingly, CTD curators index publications with a hierarchical organization[7]. For instance, while some reports simply congregate 'Skin diseases', others are centered on 'Dermatomyositis', and even some are focused on a subtype of this condition called 'Amyopathic dermatomyositis'. Broad disease terms are obviously associated with more molecules (direct annotations plus those regarding child terms); however, they can involve diverse or more intricate mechanisms. As a consequence, extracting molecular rules for imprecise phenotypes may be as challenging as for very specific cases, where data are scarce. We have explored the disease hierarchy with the intuition that, in between general and specific disease concepts, there will be enough information to learn structure-activity relationships.

In total, 934 M and 835 T diseases could be analyzed after considering 8,881 molecules (Table 1). These diseases span the medical hierarchy endpoints, and thus are representative of the variety of known human conditions.

Chemical Fragment Mining

As a first step in the chemo-centric disease analysis, we sought to list chemical moieties that could be independently associated with the phenotype. Support for this idea is provided by examples of chemical scaffolds showing a strong correlation with bioactivity profiles[8,9]. Given, for example, an M set of molecules (i.e. biomarkers and toxins of a particular disease), we performed an exhaustive molecular fragmentation and, among the resulting fragments[a], we kept those that were over-represented with respect to compounds unrelated to the disease. We considered non-redundant sets of molecules to minimize annotation biases, and designed the statistical analysis so that the final selection of fragments was simplified in terms of substructural dependencies, without detrimenting posterior predictive models (see *Materials and Methods*). Exactly the same procedure was applied to therapeutic annotations, examining T molecules instead.

[a] In this work, the terms 'fragment', 'moiety', 'chemotype' and 'scaffold' are used interchangeably.

The median molecule broke into 5 fragments, ranging from a single piece up to 200. A total of 98,077 moieties were considered. After the significance analysis, both for M and T sets, we obtained around 200 over-represented fragments per disease, and for each fragment we found 4 associated diseases. Due to initial permissive statistical requirements, this constituted a *Low Confidence* (LC) set of ~$5 \cdot 10^4$ fragment—disease associations that was ideal for later achieving predictive power. When we controlled for the False Discovery Rate (FDR) below 1% and applied additional constraints (*Materials and Methods*), we obtained a subset of 7,411 *High Confidence* (HC) fragment—disease pairs (Table S1). These fragments are well represented in the known bioactive chemical space (Figure S1), and include both expected and novel moieties, emerging from diverse sets of molecules (Figure S2). Within HC pairs, a fragment was related to a median of 2 M or T diseases, and a disease was linked to 6 fragments (Figure 1A). At least one HC fragment could be found for 41% and 50% of M and T diseases, respectively (Table 1), providing a chemo-centric molecular description of phenotypes that is interpretable for the medicinal chemist, a property that has been recently vindicated in chemoinformatics[10].

Over-represented Fragments in the Chemical Space

Identified fragments exhibit a varied chemical repertoire (Figure 1B). HC moieties have a median size of 17 atoms, including 1 ring and 4 heteroatoms. Interestingly, 32% of the fragments follow the 'Rule of Three' (Ro3) (molecular weight (MW) < 300, number of hydrogen bond donors (HBD) ≤ 3, number of hydrogen acceptors (HBA) ≤ 3 and logP ≤ 3). Backwards studies found that fragments that accomplish these rules are good starting points to meet the Lipinski condition[11], or 'Rule of Five', that concerns bioavailability of oral drugs (i.e. MW < 500, HBD ≤ 5, HBA ≤ 5 and logP ≤ 5).

Activity-related fragments offer a simple way to compose customized chemical spaces. In Tables S1 and S2, they are given together with associated diseases, enabling the design of tailored chemical libraries. In general, while our collection covers a broad and representative spectrum of chemical features, it also reflects the diversity of CTD and most chemogenomics repositories[12] (Figure 1A, middle), which contain only a small number of well-represented scaffolds, and a large proportion of singular moieties. The balance between variety of fragments and coverage will depend on the needs. Similar to the case of kinase-focused libraries[13], we might want to achieve a thorough coverage of a narrow pharmacological space to address e.g. 'Anterograde Amnesia', where only 2 HC fragments represent 55% of the beneficial molecules. Sparse libraries would be preferable in cases like 'Chronic Obstructive Pulmonary Disease', where as many as 34 HC fragments can be

extracted from the corresponding 27 medicines, spanning 74% of the active space and requiring a higher diversity.

Accounting for this diversity is crucial in order to move away from chemical clichés[14]. The structural variety of known drugs[15] and, in general, of registered compounds is very low — the more frequently a scaffold has been used, the more likely it will be used again[16]. However, we have seen that our reported fragments not always emerge from well-studied moieties, yielding valuable novel chemotypes (Figure S2). Recently, it has been suggested that a large part of fragment space is indeed synthetically accessible, which also calls for a more exploratory chemistry[17]. If orphan regions of chemical space are to be populated, we propose that our findings could aid the charting of its biologically relevant, primordial regions.

Existence of Liable and Privileged Structures

When analyzing over-represented structures, the immediate question is whether fragments exist that are mostly associated with adverse events, while others are usually present in therapeutic molecules (Figure 2A). The former would correspond to problematic structures that should be avoided in, for instance, medicinal chemistry endeavors[18]. On the contrary, the latter are desired, privileged chemotypes of potential profit in the design of libraries for forward pharmacology practices like cell-based phenotypic screening[19].

As expected, it was slightly easier to detect privileged than liable structures (384 vs 367 liable HC fragments, respectively, over a total of 45,607 T and 72,804 M chemical—disease pairs considered (Table 1)). The medicinally relevant space is influenced by size constraints and ease of synthesis[20], and pharmaceutical research is often incremental. Liable fragments, which also occur in drugs and environmental chemicals, may have been abandoned or remained unperceived, and thus are less well represented (Wilcoxon's test (Wt) p-value $< 2.2 \cdot 10^{-16}$) (Figure S3). As a consequence, the LC liable fragment occurs in a higher proportion of M compound—disease pairs than the LC privileged fragment in T pairs (Wt p-value $8.0 \cdot 10^{-10}$), implying that it might be important across a range of phenotypes, although we can only capture the association with weak statistical signal. On the other hand, as expected, the trend is inverted for HC fragments (Wt p-value $1.4 \cdot 10^{-6}$), since only a thorough exploration of chemical space allows for extraction of strong structure-activity relationships.

Of particular interest are those privileged fragments that have not been successfully used in drug development yet. Out of the 367 fragments that could be considered as privileged (>80% T both in HC and LC sets), 40% were not present in any approved or experimental drug reported in the DrugBank[21] (note that CTD scope goes beyond drug molecules: 45% of

the compounds with T annotations were not found in DrugBank above a similarity cutoff of 0.8). In Figure 2B, for example, fragment **1** constitutes a fraction of the ergoline tetracycle. Mesulergine is a psychoactive compound of the ergoline class with a halted development due to adverse histological abnormalities in rats[22]. We speculate that **1**, that is present in 6 other molecules in CTD, could be kept and used to derive safer compounds outside the ergoline family. Fragment **2** is a propanolamine that we found useful to treat 'Cardiac Arrythmias' and could be further evolved into Alprenolol alternatives, a close analog in the market. Finally, **3** is the scaffold of Dexelvucitabine, a failed anti-HIV and anti-HBV agent that, while singular in structure, displays features similar to other desirable chemotypes, and is found over-represented in as many as 8 therapeutic indications —safer derivatives of Dexelvucitabine could be of potential interest.

Another group of interesting moieties, at least in retrospective, corresponds to those that are frequently included in drug molecules despite being mostly associated with adverse events. We recognize that, in general, drugs (usually prescribed for few indications) will indeed elicit many adverse reactions. However, in CTD the M:T annotation rate is quite balanced (less than 2:1), making >80% M a meaningful definition of a liable fragment. In the right panel of Figure 2B, structure **4** accounts for the prototypical hydrochlorothiazide, a class that includes methylclothiazide and cyclothiazide. Despite its popularity, we found a large number of adverse events associated to this class, ranging from 'Hypokalemia' to 'Arthritis'. As done elsewhere[23], **4** could undergo a scaffold-hopping exercise to find better analogs. Fragment **5**, present inside 13 medicines like Sufentanil, constituted a liable HC fragment for 5 conditions, including 'Sinus Arrythmia' and 'Muscle Hypertonia'. Similarly, **6** is part of several bronchodilator agents and resembles the ancestor Norepinephrine drug. We found 8 HC associations of **6** with inconvenient events such as 'Tachycardia' and 'Hypertension', suggesting that further generations of Norepinephrine successors are likely to remain unsafe.

Predictive Models

Although valuable, identifying the presence of a characteristic fragment in a molecule is usually not enough to accurately infer an association with a disease[24,25]. Very often, a combination or mutual exclusion of several moieties will determine the outcome. In general, predictive power and interpretability of structure-activity models are two different objectives that are difficult to achieve simultaneously. On this matter, a good tradeoff is offered by LC fragments, which are more frequent among disease-related molecules, and thus are promising variables for starting machine learning[24].

Given its reduced cost, fragment-based learning can be applied at virtually every step of the drug discovery pipeline, and offers a means to join chemoinformatics with expert opinion[26]. Its performance will largely depend on the specificity of the underlying biology, and the proper, delimited representation of the active chemical space. As a result, while detecting over-represented fragments gets easier for highly annotated, broad disease terms, predictive capability does not follow the same trend (Figure S4).

We built a fragment-based chemical classifier for each of the 934 M and 835 T diseases (Figure 3) using Random Forests (RFs). RFs allow detecting interactions between fragments, e.g. when the combination of two fragments has a therapeutic effect but each individual fragment does not. Table 1 provides a general view of the results. It shows, for instance, that point prediction performance metrics sensible to data imbalance (namely the positive predictive value and the F_1-score) take values close to zero. This is an expected observation given the pronounced imbalance of positive:unknown sampling (a median of 30:4,250). Also, note that sensitivity could be increased at the expense of the high specificity, and that the decision cutoff could slide at will so that e.g. G-mean is optimized (see *Materials and Methods*). The area under the ROC curve (AUC) measures the compromise between sensitivity and specificity at all possible cutoffs, and it is widely used to assess the performance of predictive models. Overall, 184 M and 216 T disease models exhibited a cross-validated AUC above 0.7. The successful models did not display a distinct chemistry (Figure S5), and covered 13% and 7% of the full medical hierarchy endpoints, respectively. Together, both results evidence our scarce knowledge of the relevant chemical space, and the difficulty to assess a priori if a region of it has been sufficiently exploited.

Therapeutic Effects are Better Predicted than Adverse Events

When analyzing accurate, plausible classifiers (AUC > 0.7), the first observation is that therapeutic outcomes are better modeled than adverse events, i.e. there is a larger proportion of T cases with AUC > 0.7 (Fisher's test (Ft) p-value 0.001, and Wt p-value $3 \cdot 10^{-8}$ for whole distributions) (Figure 3). Again, this arises from the fact that the therapeutic space is composed of incremental discoveries (Figure S3), and emphasizes the difficulty of the predictive toxicology task.

ROC curves on the right of Figure 3 correspond to satisfactory models of T diseases. 'Osteomyelitis', that refers to bone infections, is treated with antibiotics of well-used families (quinolones, cephalosporins, penicillins, etc.). Thus, it is easy to infer whether a molecule will be suitable for addressing such condition. A similar chemistry has been learned for 'Pseudomonas Infections', for instance. Analogous conclusions can be drawn for 'Paranoid

Schizophrenia', where e.g. benzodiazepines and phenothiazines are annotated, and for 'Supraventricular Tachycardia', a cardiovascular complication of which the aforementioned propanolamines are prominent examples.

Other chemicals, rather than treating, may trigger cardiovascular events. In fact, these are commonly alerted drug side effects. Pergolide, for instance, was withdrawn from the market due to heart issues —we predicted its association with 'Aortic Valve Insufficiencies' (this annotation was not available from CTD). A plausible model was also obtained for 'Mesenteric Valve Insufficiencies' (left ROC curves in Figure 3). In general, for heart events, even when the underlying biology remains intricate[27], there is a chemical signal that can aid prevention. In Figure 3, we also display the cross-validation of the 'Uterine Hemorrhage' model, and, regarding the same organ, that of 'Endometrial Neoplasms'.

Not All Types of Diseases are Equally Predictable

Following the last example above, we find support for the intuition that traveling the disease hierarchy from specific to broad terms can help to find informative chemical sets. Accordingly, while 'Neoplasms' are poorly understood as a whole (AUC 0.66), we obtained a number of accurate models for certain organs and types (Figure 4). In particular, we could solve many M cancer cases, while few successful T models existed. This illustrates that we know more of the chemistry of carcinogens and cancer markers than of the chemistry that is needed to cure it. A similar conclusion could be drawn for 'Male' and 'Female Urogenital Disorders'. On the contrary, we could provide several plausible classifiers for the treatment of 'Mental Disorders', meaning that the chemical space that addresses such conditions has been well exploited. Similarly, we have deep knowledge on treating 'Bacterial Infections and Mycoses' while, as expected, there is little chemistry that may facilitate them (the only example we found was 'Candidiasis', where most relevant structures corresponded to steroidal frameworks like glucocorticoids[28]). The rest of disease classes shared, in general, a balance between M and T plausible models. Remarkably, some disease classes were poorly modeled. We attempted, for instance, 41 M 'Eye Diseases', of which as few as 3 yielded a satisfactory classifier. Similarly, we only obtained a good predictor for 4 of the 28 T 'Endocrine System Disorders' (Figure S6).

Indeed, for a majority of diseases we lack an accurate model. We believe, however, that there is room for improving chemical classifiers based on literature mining. One important hindrance in training these classifiers is the absence of truly negative data (chemical—disease pairs that have been verified not to interact, as opposed to not having been observed so far). The so-called 'positive-unlabeled learning' tackles this issue and is now

being implemented in biomedicine[29]. However, in our hands, such methodologies[30,31] did not improve predictive power, most likely due to the sparseness and reduced size of the set of unknowns (Figure S7), an issue that, most likely, will be solved as disease—chemical annotations continue to increase[32]. Also, including physicochemical properties of compounds could be of enormous interest, particularly in the case of adverse events, where mechanisms of action may not be target-driven. Accordingly, the identification of toxicophores is usually thought of in metabolic and reactivity terms[18], since toxic effects can result from polar or nonpolar processes, uncoupling of oxidative phosphorylation, thiol-alkylation, etc. In this regard, reactivity prediction methods should be appropriate[33], particularly for nonspecific complications like tissue necrosis, carcinogenicity, or immune-mediated toxicities. Recently, a combination of structure and reactivity analysis was applied to select groups that shared structure and electronic state[34], and it was recommended that compounds undergo a structural clustering before the reactivity assessment, suggesting that our results could be readily complemented with reactivity profiles.

Disease Networks Based on Underlying Chemistry

In this study, we have analyzed each disease separately. However, results should be integrated to provide a general view. For this purpose, network representations are a prominent systems biology tool because they integrate relationships between different entities, facilitating contextualization and providing a general view[35,36]. In particular, disease networks help to assimilate the diversity of human conditions. In a seminal work, Goh et al. proposed that two diseases could be related if they share a genetic origin[37]. The resulting disease network was able to unveil biological modules and therefore offered a means to link the molecular and the organism levels.

Instead of connecting two diseases when the same genes participate in their etiology, we link them if they relate to a similar chemistry, i.e. when the molecules associated with the one are comparable to those associated with the other. The resulting chemo-centric map of human conditions is of singular interest for drug development, since it is focused on intervention, i.e. on disease relationships that are directly based on effector compounds.

The Disease Comorbidity Network

When we relate M disease models, the corresponding network is a comorbidity map, where two conditions are connected if the toxins and markers of the one are similar to those of the

other, implying that the two diseases could occur simultaneously. In practice, we screened all M molecule sets annotated to the 934 diseases against the 184 successful M models, and we related two diseases if the AUC of the cross-classification was higher than 0.7. This yielded a network of 12,610 edges (Tables 2 and S3). Interestingly, such a chemo-centric comorbidity map captured disease co-occurrences detected in the history of more than 30 million patients[38]: a medical semantics mapping found that a large number of our disease associations have indeed been observed in the clinics (9,788 matches, the corresponding contingency table yielded a Ft p-value of $4.5 \cdot 10^{-28}$), providing an excellent independent validation of our findings (see *Materials and Methods*). For instance, we predicted that molecules associated with 'Aortic Valve Insufficiency' are likely related to 'Neuroleptic Malignant Syndrome' (AUC 0.88). In turn, the 'Aortic Valve Insufficiency' model up-ranked 'Elimination Disorders' molecules (AUC 0.82) (Figure 5). In patients, not necessarily due to exposure to chemicals, these relationships have been observed with relative risks of 56.7 and 29.5, respectively[38]. Overall, together with e.g. studies of metabolic pathways[39], our results show that a chemical viewpoint is useful to account for the underlying molecular connection of human conditions.

The Drug Repositioning Network

Analogously, we may relate diseases based on T records and obtain a network that links two conditions when medicines for the first could also serve in the second. This so-called 'drug repositioning network' is appealing given the time and financial burdens of the drug discovery process. Currently, a number of computational approaches are taken in this direction[40], and even the simplest methods[41] are proposing remarkable opportunities. After screening the 835 T compound-disease pairs against the 216 good T models, we obtained a network of 14,590 edges (Tables 2 and S4). Some diseases like 'Hypertension' had a high in-degree (in this case, 235), meaning that they could be the repurposing opportunity of many indications, reflecting the clinical complexity of this physiological phenomenon associated with cardiovascular, endocrine and nervous system components. On the other hand, 'Urethral Diseases' displayed an out-degree of 137, i.e. its 11 medicines could have several other uses. When compared to a network drawn from approved indications of drugs[42], we observed a significant overlap (10,731 common edges, Ft p-value $3.4 \cdot 10^{-13}$), reinforcing the validity of our results. This network based on approved drugs represents the polypharmacy of medicines, and links two diseases if they are treated by a significant number of common drugs (see *Materials and Methods*). Even after a conservative semantic mapping, 3,859 of our repositioning opportunities were not found in such network, implying that they remain largely unexplored. Among these, we propose the use of 'Rhinitis'

therapeuticals like ketotifen for the treatment of 'Personality Disorders' (AUC 0.81), and the repurposing of antibronchitic drugs to treat 'Supraventricular Tachycardia' (AUC 0.81) (Figure 5).

The Drug Side Effect Network

Finally, linking T and M diseases yields a map that relates treatments to potential adverse events. As shown in Table 2, we screened the 835 T chemical—disease pairs to predict undesired side effects among the 184 M satisfactory models. The resulting network contained as many as 9,921 relationships (Tables 2 and S5). In this network, large peripheral nodes are particularly interesting: 'Seizures', for instance, has a well-defined therapeutic chemistry (AUC 0.71) related to as many as 255 molecules, and is not linked to any of the adverse events, suggesting that these treatments are rather safe. When we compared our predictions with side effects extracted from drug package labels[43], we also observed a significant coincidence (8,686 common associations, Ft p-value $6.9 \cdot 10^{-21}$), while still providing 1,235 novel predictions. One of them is the possible appearance of 'Serotonin Syndrome' after exposure to 'Hyperpituitarism' (e.g. carmoxirole) and 'Neointima' agents like nebivolol (AUC of 0.78 and 0.81, respectively) (Figure 5). Nebivolol, in fact, is metabolized by CYP450 2D6, resembling serotonin reuptake inhibitors —concomitant treatment with such inhibitors may lead to overdose[44]. Overall, these novel associations contribute to the completion of putative drug side effect profiles. In the last years, such profiles have shown useful to elucidate molecular events from phenotypic observations[45], in turn proving that a lot can still be learned from the always imperfect drug molecules[46].

Concluding Remarks

The current perception is that systems biology will aid the learning of drug action by rationalizing the influence that small molecules exert on our health[47]. In most cases, drug action is mediated through receptors, being of critical importance their identification. In a previous work[48], we reported protein targets shared among drugs with a common effect. Our approach was agnostic in the sense that it considered a vast chemical—protein interactome, and was therefore suitable to initiate a systems view. Although we recognize the relevance of target and off-target identification, we found this knowledge insufficient to anticipate side effects, in good agreement with the translational gap in drug discovery[49]. To complement this lack of knowledge, we also mined characteristic chemical moieties inside the drugs with the aim to surrogate phenomena that molecular biology is not yet able to consider, as done by others[50]. We learned that chemical structures treasure a remarkable predictive power,

although they are difficult to inspect given the small number of known drugs and their sparse distribution across the chemical space. Now, our results highlight that collecting and grouping molecules with enough insistency aids the modeling of phenotypic implications with no need to acknowledge all the underlying biological events. Several studies have proven the value of this chemo-centric view of biology. Most notably, such a view allowed for the prediction of ligand binding to protein targets with unresolved structures[51]. Databases like ChEMBL[52] and BindingDB[53], among others, have been essential to decipher relationships between chemical features and affinity, and a ligand-centered description of the binding event is now feasible[4]. In these databases, hundreds of thousands of distinct compounds are recorded. The ambition to relate chemical structures directly to human-body responses is, undoubtedly, a more challenging task, given the complicated intrinsic biology and the lack of compound records. We have shown that, even when only a few thousand molecules are available, it is already possible to gain some insights. Moreover, we anticipate that the number of well-modeled phenotypes could increase considerably in the upcoming years. Concretely, we estimate that the amount of accurate classifiers could be doubled if we would double the annotation of certain diseases (Figure S8). Approximately, increasing by 25% the number of chemical—disease records could result in this doubling of satisfactory models. To guide disease annotators, in Table S2 we detail which diseases fall on a learning plateau, be it because they are sufficiently apprehended or largely under-annotated, and which cases will benefit more from curation efforts[54]. Likewise, improving disease annotation will enable the modeling of more specific phenotyps: terms in this study are slightly broader than those commonly used in drug discovery, and these are, in turn, notably unspecific relative to the existing medical vocabulary (Figure S9).

To grow the body of chemical records, improvements in text chemical entity identification[55] and new knowledge discovery concepts[56] will be fundamental. Opposite to e.g. genomics, large-scale experimentation in chemistry has been conducted primarily by pharmaceutical industry and, traditionally, proprietary data have not been available to the community. Therefore, scientific literature is still a major support to publish chemical data. We expect that, with the advent of text-mining technologies, resources like CTD will continue to expand in size and scope. Moreover, current chemical—disease records are being gathered together with disease-related genes, which manifests that knowledge is being assembled at a fast pace towards a holistic view of biology. Only now, network-based tools to handle such complexity are flourishing[57], and urgently demand more chemistry awareness[58]. In this context, our study brings chemical cognizance to the systems level, fulfilling a need of translational sciences, and widening the applicability of network-based strategies.

Materials and Methods

Chemical Structures

Compound structures were obtained by querying the Chemical Identifier Resolver [cactus.nci.nih.gov] with CTD names. Additionally, we fetched the fraction of chemicals contributed by CTD to PubChem [pubchem.ncbi.nlm.nih.gov]. Organometallic compounds were excluded and inorganic salts were removed from mixtures. Substances with a molecular weight above 800 were also discarded, and stereochemical information was not considered. Figure 6 schemes the processing that these molecules underwent.

Exhaustive Fragmentation

We exhaustively fragmented each chemical structure through recursive bond breaks down to a minimum size of 5 atoms. We followed JChem's [www.chemaxon.org] CCQ fragmentation approach, based on cutting carbon—carbon bonds (CC) if at least one of the carbons is bound to a heteroatom (Q). Thus, CCQ rules do not modify functional groups. Aliphatic rings and aromatic systems were not cleaved either. The 5% of molecules that broke into more than 200 fragments were dismissed.

Disease Annotation of Chemicals

We fetched chemical—disease associations from CTD (January 2013)[5]. This knowledgebase includes a controlled vocabulary[7] that is based on the 'Diseases' branch of the National Library of Medicine's Medical Subject Headers (MeSH). MeSH hierarchy grows from broader to more specific disease terms, and molecules are annotated throughout. General concepts include annotations from the more specific ones.

To assign M and T molecules to each disease, we fetched curated ('Direct evidence') annotations from CTD. Ambiguous annotations (M and T, simultaneously) were removed. Molecules labeled in CTD as 'inferred' (through gene—disease triangulation[59]) were also discarded since they were confounding the obtainment of disease classifiers (Figure S10). The set of 'unknown' molecules corresponded to all of those entries that shared no relationship (neither curated nor inferred) with none of the terms in the corresponding branch of the disease vocabulary. Only diseases annotated with at least 10 molecules entered further analysis. In total, we kept 934 M and 835 T chemical-disease relationships.

Non-redundant Sets of Molecules

In order to obtain non-redundant sets of chemicals for each disease, we clustered a full pair-wise chemical similarity matrix. Chemical similarity was measured using topological

fingerprints in the RDKit [www.rdkit.org]. The resulting matrix underwent an unsupervised clustering with the Butina algorithm[60]. Clusters were flattened at a Tanimoto cutoff of 0.8, i.e. at a distance of 0.2 to the central molecule. Whenever a disease was associated with several chemicals in a cluster, the molecule with the highest accumulated similarity to the rest was kept as representative for the group. Analogously, we obtained non-redundant sets of disease-unrelated chemicals (unknowns).

Fragment Mining

Selection of LC Fragments Suitable for Machine Learning

For each M or T compound—disease pair, we outlined a matrix *W* listing small molecules in the rows and fragments in the columns. To fill in *W*, we screened each molecule *i* against all of the fragments. A molecule—fragment comparison was performed as follows. First, we broke compound *i* into fragments. The resulting set of fragments was then compared to fragment *j*. The score of this comparison corresponded to the highest Tanimoto similarity of MACCS fingerprints, and was kept in cell W_{ij}. MACCS keys are a set of questions about a 2D structure, and are thus useful to capture chemical features beyond simple topological matching. Using MACCS fingerprinting, we increased the power to detect relevant features, while diminishing the sparseness of *W*.

Then, the width of *W* was shrunk using statistical filtering. In the resulting matrix W^{LC}, for each column *j*, rows displaying a MACCS similarity > 0.8 were counted, and the significance of the over-representation of fragment *j* among molecules related to the disease was assessed using a right-tailed Fisher's exact test. Please note that the contingency table classifies 'positives' and 'unknowns' (instead of 'negatives'): this reduces statistical power, but should not affect the true positive rate (Figure S11). Those fragments with a p-value < 0.1 were retained, as recommended in [25]. Note that the selection of LC fragments underwent a final step that ensured an acceptable tradeoff between classification performance and statistical signal (see *Data Balancing* below).

Selection of High-Confidence Fragments

From LC fragments, we selected a subset of HC representatives. In *W*, these had to elicit an odds ratio ≥ 10, a minimum support of 3 molecules and a Benjamini-Hochberg adjusted p-value < 0.01. To report a diverse and representative set, we grouped those fragments that occurred in the same molecules. From each group, the fragment associated with more diseases was kept.

Chemo-centric Disease Models

Data Balancing

In general, few chemicals are known per disease, while the majority of chemicals is not related to it. We balanced W^{LC} using a combination of under-sampling and SMOTE over-sampling[61,62]. For each case in the minority class (i.e. chemicals annotated with a disease of interest), 5 new examples were created, up to a maximum of 1,000 instances. The majority class (i.e. 'unknown' cases) was under-sampled to achieve a 1:1 proportion with the minority class.

Then, columns in the balanced dataset ($W^{LC'}$) were hierarchically clustered using Fastcluster[63], and branches were pruned using DynamicTreeCut[64] with a minimum cluster size of 1. Inside each cluster, fragments were compared all-against-all to detect parent—child relationships. For a lineage of fragments, the one with the best initial over-representation p-value was retained. Overall, this led to matrices $W^{LC''}$ that had an even sampling through the rows and a simplified set of LC over-represented fragments in the columns.

Chemical Classifiers

$W^{LC''}$ matrices above are suitable for machine learning because they have a balanced class distribution, and a representative and reasonably distinct set of variables. Given its general robustness in the learning of structure-activity relationships[65], we chose to build chemical classifiers with the random forest algorithm. For this, we used the randomForest R-package[66], growing 10,000 trees and taking default values for the rest of parameters. Since each tree returns a decision, class probabilities were estimated from voting.

Cross-validation

As schemed in Figure 6, we performed a stratified 10-fold cross-validation of predictive models. Test and training sets were split before the LC fragment mining step (i.e. before the variable selection, and therefore previous to the data balancing). Performance metrics in Table 1 were obtained from the reassembled vector of test predictions.

Disease Networks

Network Construction

In a chemo-centric disease network, disease *A* is linked to disease *B* if molecules annotated to *A* are predicted to relate with *B*. Since we obtained M and T models, we can propose, at least, three different networks (Figure 5A and Table 2): (1) a comorbidity network, that links *A* to *B* if chemicals that cause *A* are predicted to cause *B*; (2) a drug repositioning network,

where chemicals employed to treat *A* may also be useful to treat *B*; and (3) a drug side effect network, that relates *A* to *B* when chemicals used in the treatment of *A* could cause *B*.

To infer an edge from *A* to *B*, we tested *A* curated chemicals together with a set of chemicals unrelated to *A* and *B* using the *B* random forest classifier. The strength of the association was assessed with the AUC of the cross-classification ROC plot, where molecules predicted to associate with *B* are checked for their association with *A*. Note that we removed easy cases by discarding disease pairs in the same branch of the medical hierarchy. To mine the examples discussed in Figure 5, we only considered those pairs that shared no chemicals, highlighting the importance of the fragment mining procedure.

Network Analysis

Comparison of the Comorbidity Network with a Clinical Disease Co-occurrence Network

A clinical disease network was obtained from Hudine[38], a comorbidity network that reports the relative risk (RR) of experiencing a disease when another disease is diagnosed. In Hudine, clinical reports are stored using the International Classification of Diseases, 9th revision (ICD-9). The mapping between MeSH and ICD-9 (3-digits code) terms was achieved using BioPortal's [bioportal.bioontology.org] UMLS concepts, and by best-matching MeSH and ICD-9 UMLS concepts with the UMLS-similarity Perl-package[67] (vector relatedness > 0.8). We assigned a significance p-value to the coincidence between our chemo-centric network and Hudine comorbidities (RR ≥ 20 or $\phi \geq 0.06$, as in [38]) by using a right-tailed Fisher's exact test. The corresponding confusion matrix classified predicted and unpredicted pairs, and pairs that were mapped and not mapped to Hudine.

In order to demonstrate the need for robust disease models, we also built a comorbidity network (same [*A*] and [*B*] sets) that linked *A* to *B* simply if at least 50% of *A* LC fragments were LC fragments of *B*. In addition to a reduction in the number of edges of two orders of magnitude, we observed no significant coincidence with the clinical network.

Comparison of the Drug Repositioning Network with a Drug Repositioning Network Derived from Known Drugs

Disease—disease associations were inferred based on drug indications[42]. Similar to [68], for a pair of diseases *A* and *B*, we filled a 2x2 confusion matrix counting the number of drugs that are used to treat both, one or none of the diseases. From this matrix, we obtained the two-tailed p-value of a Fisher's test and the Matthews correlation coefficient (MCC). *A* and *B* were linked in the drug repositioning network if p-value ≤ 0.05 and MCC ≥ 0.15[68]. Like above, node mapping was achieved using UMLS term similarities, and the significance of

the overlap with our results was evaluated analogously. Here again, we checked that the modeling step was important to provide significant results.

Comparison of the Side Effect Network with Side Effects Reported in Drug Labels

We collected a side effect network from [68]. This network represents side effects that occur frequently among approved drugs prescribed for a particular disease. As done for the comorbidity and the drug repositioning networks, we analyzed its coincidence with our chemo-centric map, and confirmed the convenience of disease models for building the network.

Acknowledgements

We thank Evarist Planet and Samira Jaeger (IRB) for helpful discussions. This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (BIO2010-22073), the European Commission under FP7 Grant Agreement 306240 (SyStemAge) and the European Research Council through the SysPharmAD grant (Agreement n^{o}: 201014). MD-F is a recipient of the Spanish FPU fellowship.

Author contributions

MD-F, DR and PA designed the study and wrote the manuscript. MD-F performed the experiments and analyzed the results.

Competing Financial Interests statement

The authors declare no competing financial interests.

References

1 Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. & Vidal, M. Drug-target network. *Nature biotechnology* **25**, 1119-1126, doi:10.1038/nbt1338 (2007).

2 Bauer-Mehren, A. *et al.* Automatic filtering and substantiation of drug safety signals. *PLoS computational biology* **8**, e1002457, doi:10.1371/journal.pcbi.1002457 (2012).

3 Pujol, A., Mosca, R., Farres, J. & Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences* **31**, 115-123, doi:10.1016/j.tips.2009.11.006 (2010).

4 Keiser, M. J., Irwin, J. J. & Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry* **49**, 10267-10276, doi:10.1021/bi101540g (2010).

5 Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2013. *Nucleic acids research* **41**, D1104-1114, doi:10.1093/nar/gks994 (2013).

6 Wishart, D. S. Chapter 3: Small molecules and disease. *PLoS computational biology* **8**, e1002805, doi:10.1371/journal.pcbi.1002805 (2012).

7 Davis, A. P., Wiegers, T. C., Rosenstein, M. C. & Mattingly, C. J. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database : the journal of biological databases and curation* **2012**, bar065, doi:10.1093/database/bar065 (2012).

8 Varin, T., Schuffenhauer, A., Ertl, P. & Renner, S. Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *Journal of chemical information and modeling* **51**, 1528-1538, doi:10.1021/ci2000924 (2011).

9 Wetzel, S. *et al.* Interactive exploration of chemical space with Scaffold Hunter. *Nature chemical biology* **5**, 581-583, doi:10.1038/nchembio.187 (2009).

10 Shultz, M. D. Setting expectations in molecular optimizations: strengths and limitations of commonly used composite parameters. *Bioorganic & Medicinal Chemistry Letters* **231**, 5980–5991, doi:http://dx.doi.org/10.1016/j.bmcl.2013.08.029 (2013).

11 Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug discovery today* **8**, 876-877 (2003).

12 Langdon, S. R., Brown, N. & Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *Journal of chemical information and modeling* **51**, 2174-2185, doi:10.1021/ci2001428 (2011).

13 Akritopoulou-Zanze, I. & Hajduk, P. J. Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. *Drug discovery today* **14**, 291-297, doi:10.1016/j.drudis.2008.12.002 (2009).

14 Lameijer, E. W., Kok, J. N., Back, T. & Ijzerman, A. P. Mining a chemical database for fragment co-occurrence: discovery of "chemical cliches". *Journal of chemical information and modeling* **46**, 553-562, doi:10.1021/ci050370c (2006).

15 Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **39**, 2887-2893, doi:10.1021/jm9602928 (1996).

16 Lipkus, A. H. *et al.* Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *The Journal of organic chemistry* **73**, 4443-4451, doi:10.1021/jo8001276 (2008).

17 Pitt, W. R., Parry, D. M., Perry, B. G. & Groom, C. R. Heteroaromatic rings of the future. *Journal of medicinal chemistry* **52**, 2952-2963, doi:10.1021/jm801513z (2009).

18 Williams, D. P. Toxicophores: investigations in drug safety. *Toxicology* **226**, 1-11, doi:10.1016/j.tox.2006.05.101 (2006).

19 Welsch, M. E., Snyder, S. A. & Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Current opinion in chemical biology* **14**, 347-361, doi:10.1016/j.cbpa.2010.02.018 (2010).

20 Wester, M. J. *et al.* Scaffold topologies. 2. Analysis of chemical databases. *Journal of chemical information and modeling* **48**, 1311-1324, doi:10.1021/ci700342h (2008).

21 Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035-1041, doi:10.1093/nar/gkq1126 (2011).

22 Dupont, E., Mikkelsen, B. & Jakobsen, J. Mesulergine in early Parkinson's disease: a double blind controlled trial. *Journal of neurology, neurosurgery, and psychiatry* **49**, 390-395 (1986).

23 Mavridis, L., Hudson, B. D. & Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *Journal of chemical information and modeling* **47**, 1787-1796, doi:10.1021/ci7001507 (2007).

24 Takigawa, I. & Mamitsuka, H. Graph mining: procedure, application to drug discovery and recent advances. *Drug discovery today* **18**, 50-57, doi:10.1016/j.drudis.2012.07.016 (2013).

25 Wang, Y. *et al.* Estimation of carcinogenicity using molecular fragments tree. *Journal of chemical information and modeling* **52**, 1994-2003, doi:10.1021/ci300266p (2012).

26 Greene, N., Judson, P. N., Langowski, J. J. & Marchant, C. A. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR and QSAR in environmental research* **10**, 299-314, doi:10.1080/10629369908039182 (1999).

27 Berger, S. I., Ma'ayan, A. & Iyengar, R. Systems pharmacology of arrhythmias. *Science signaling* **3**, ra30, doi:10.1126/scisignal.2000723 (2010).

28 Nakajima, A. *et al.* Investigation of glucocorticoid-induced side effects in patients with autoimmune diseases. *Journal of the Pharmaceutical Society of Japan* **129**, 445-450 (2009).

29 Yang, P., Li, X. L., Mei, J. P., Kwoh, C. K. & Ng, S. K. Positive-unlabeled learning for disease gene identification. *Bioinformatics* **28**, 2640-2647, doi:10.1093/bioinformatics/bts504 (2012).

30 Zhang, B. & Zuo, W. Reliable negative extracting based on kNN for learning from positive and unlabeled examples. *Journal of Computers* **4**, 94-101 (2009).

31 Liu, B., Lee, W. S., Yu, P. & Li, X. Partially supervised classification of text documents *ICML-02* (2002).

32 Davis, A. P. *et al.* A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database : the journal of biological databases and curation* **2013**, bat080, doi:10.1093/database/bat080 (2013).

33 Liew, C. Y., Pan, C., Tan, A., Ang, K. X. & Yap, C. W. QSAR classification of metabolic activation of chemicals into covalently reactive species. *Molecular diversity* **16**, 389-400, doi:10.1007/s11030-012-9364-3 (2012).

34 Casalegno, M. & Sello, G. Determination of toxicant mode of action by augmented top priority fragment class. *Journal of chemical information and modeling* **53**, 1113-1126, doi:10.1021/ci400130n (2013).

35 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics* **12**, 56-68, doi:10.1038/nrg2918 (2011).

36 Jacunski, A. & Tatonetti, N. Connecting the Dots: Applications of Network Medicine in Pharmacology and Disease. *Clinical pharmacology and therapeutics*, doi:10.1038/clpt.2013.168 (2013).

37 Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).

38 Hidalgo, C. A., Blumm, N., Barabasi, A. L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS computational biology* **5**, e1000353, doi:10.1371/journal.pcbi.1000353 (2009).

39 Lee, D. S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9880-9885, doi:10.1073/pnas.0802208105 (2008).

40 Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery* **3**, 673-683, doi:10.1038/nrd1468 (2004).

41 Chiang, A. P. & Butte, A. J. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics* **86**, 507-510, doi:10.1038/clpt.2009.103 (2009).

42 Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* **7**, 496, doi:10.1038/msb.2011.26 (2011).

43 Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* **6**, 343, doi:10.1038/msb.2009.98 (2010).

44 Gielen, W., Cleophas, T. J. & Agrawal, R. Nebivolol: a review of its clinical and pharmacological characteristics. *International journal of clinical pharmacology and therapeutics* **44**, 344-357 (2006).

45 Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263-266, doi:10.1126/science.1158140 (2008).

46 Duran-Frigola, M. & Aloy, P. Recycling side-effects into clinical markers for drug repositioning. *Genome medicine* **4**, 3, doi:10.1186/gm302 (2012).

47 Russell, R. B. & Aloy, P. Targeting and tinkering with interaction networks. *Nature chemical biology* **4**, 666-673, doi:10.1038/nchembio.119 (2008).

48 Duran-Frigola, M. & Aloy, P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chemistry & biology* **20**, 594-603, doi:10.1016/j.chembiol.2013.03.017 (2013).

49 Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nature reviews. Drug discovery* **10**, 428-438, doi:10.1038/nrd3405 (2011).

50 Audouze, K. *et al.* Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS computational biology* **6**, e1000788, doi:10.1371/journal.pcbi.1000788 (2010).

51 Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nature biotechnology* **25**, 197-206, doi:10.1038/nbt1284 (2007).

52 Bellis, L. J. *et al.* Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society transactions* **39**, 1365-1370, doi:10.1042/BST0391365 (2011).

53 Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research* **35**, D198-201, doi:10.1093/nar/gkl999 (2007).

54 Davis, A. P. *et al.* Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database. *PloS one* **8**, e58201, doi:10.1371/journal.pone.0058201 (2013).

55 Grego, T. & Couto, F. M. Enhancement of chemical entity identification in text using semantic similarity validation. *PloS one* **8**, e62984, doi:10.1371/journal.pone.0062984 (2013).

56 Ding, Y. *et al.* Entitymetrics: measuring the impact of entities. *PloS one* **8**, e71416, doi:10.1371/journal.pone.0071416 (2013).

57 Hu, Z. *et al.* VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic acids research* **41**, W225-231, doi:10.1093/nar/gkt401 (2013).

58 Oprea, T. I., May, E. E., Leitao, A. & Tropsha, A. Computational systems chemical biology. *Methods Mol Biol* **672**, 459-488, doi:10.1007/978-1-60761-839-3_18 (2011).

59 King, B. L., Davis, A. P., Rosenstein, M. C., Wiegers, T. C. & Mattingly, C. J. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PloS one* **7**, e46524, doi:10.1371/journal.pone.0046524 (2012).

60 Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences* **39**, 747-750, doi:10.1021/ci9803381 (1999).

61 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321-357 (2002).

62 Torgo, L. *Data Mining with R, learning with case studies*. (Chapman and Hall/CRC, 2010).

63 Müllner, D. Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software* **53**, 18 (2013).

64 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720, doi:10.1093/bioinformatics/btm563 (2008).

65 Wildenhain, J., Fitzgerald, N. & Tyers, M. MolClass: a web portal to interrogate diverse small molecule screen datasets with different computational models. *Bioinformatics* **28**, 2200-2201, doi:10.1093/bioinformatics/bts349 (2012).

66 Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).

67 McInnes, B. T., Pedersen, T. & Pakhomov, S. V. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2009**, 431-435 (2009).

68 Yang, L. & Agarwal, P. Systematic drug repositioning based on clinical side-effects. *PloS one* **6**, e28025, doi:10.1371/journal.pone.0028025 (2011).

Figure and Table Legends

Figure 1. Over-represented fragments. (A) Fragments per disease (left) and diseases per fragment (right), considering only the HC set. In the middle, a Voronoi diagram where each fragment is a shape with area and color proportional to the number of molecules that contain it (best match similarity > 0.8). (B) Chemical diversity. We display the cumulative distribution of the total number of atoms (left), the number of heteroatoms (middle), and the number of rings (right). Distributions are decorated with illustrative fragment structures. M and T chemical-disease relationships are shown in orange and green colors, respectively.

Figure 2. Privileged and liable structures. (A) Balance between privileged and liable structures, both for the HC and LC sets. % of M indicates the proportion of M associations for each fragment over its disease associations. (B) On the right, three scaffolds that, while being mostly liable, are included in drug molecules. On the left, fragments that are privileged and remain unsuccessful or unexplored as therapeutics. Next to each structure, top and bottom pie charts represent the number of diseases for which the fragment is LC- and HC-associated, respectively. Area of pie charts is proportional to the number of diseases. To select these examples, experimental and approved drug structures were extracted from Drugbank (July 2013) [21], and treated like CTD compounds.

Figure 3. Predictive models. In the middle, AUC distribution of M and T models. Area under a density region is proportional to the number of diseases. On the left and right panels, example ROC plots for M and T chemical-disease relationships, respectively.

Figure 4. Disease categories of successful models. M and T plausible disease models classified into high-level disease categories. Each circle represents an M or T disease model belonging to the corresponding category. Area of circles is proportional to the number of associated molecules in our dataset.

Figure 5. Disease networks. Disease comorbidity, drug repositioning and drug side effect networks. Examples discussed in the text are depicted with directed links on top of each network. To select these examples, we looked for strong correlations (see *Materials and Methods*) occurring between diseases in different categories. None of the cases share annotated chemicals, highlighting the value of our fragment-based models. Networks are displayed with a gravity layout, being node size proportional to the number of related chemicals. Network statistics can be found in Table 2.

Figure 6. Scheme of the method. Analysis protocol exemplified for an M disease of interest. (1) Annotated molecules are collected and split in training and test sets. (2) M training molecules are fragmented using CCQ rules. (3) *W* is built from the resulting fragments (columns) and the training set (rows) (stratified 10-fold cross-validation). *W* undergoes a significance filtering, a data balancing step, a column clustering and a pruning, resulting in $W^{LC'}$. (4) Columns of $W^{LC'}$ constitute the LC set of fragments; (5) further filtering considering substructural relationships and co-occurrence in molecules yields the HC set. (6) Using $W^{LC'}$, a random forest classifier is learned, and (7) tested against the test set. If the model performs with AUC > 0.7, it is considered of good quality. (8) Steps 1-7 are conducted for all M and T chemical-disease relationships. (9) Using plausible models, chemo-centric disease networks are constructed.

Table 1. Disease and fragment statistics

Analysis of M and T chemical-disease annotations, 'Total' column refers to the union of both categories. When applicable, median values are shown for count data, while mean values are shown for performance metrics. Point performance metrics are taken with default 0.5 cutoff in the random forest classifier. The cutoff could be slid along the classifier's outcome to get different point performances along the ROC space.

Table 2. Network statistics

General statistics of the chemo-centric disease networks.

Figures and Tables

Figure 1.

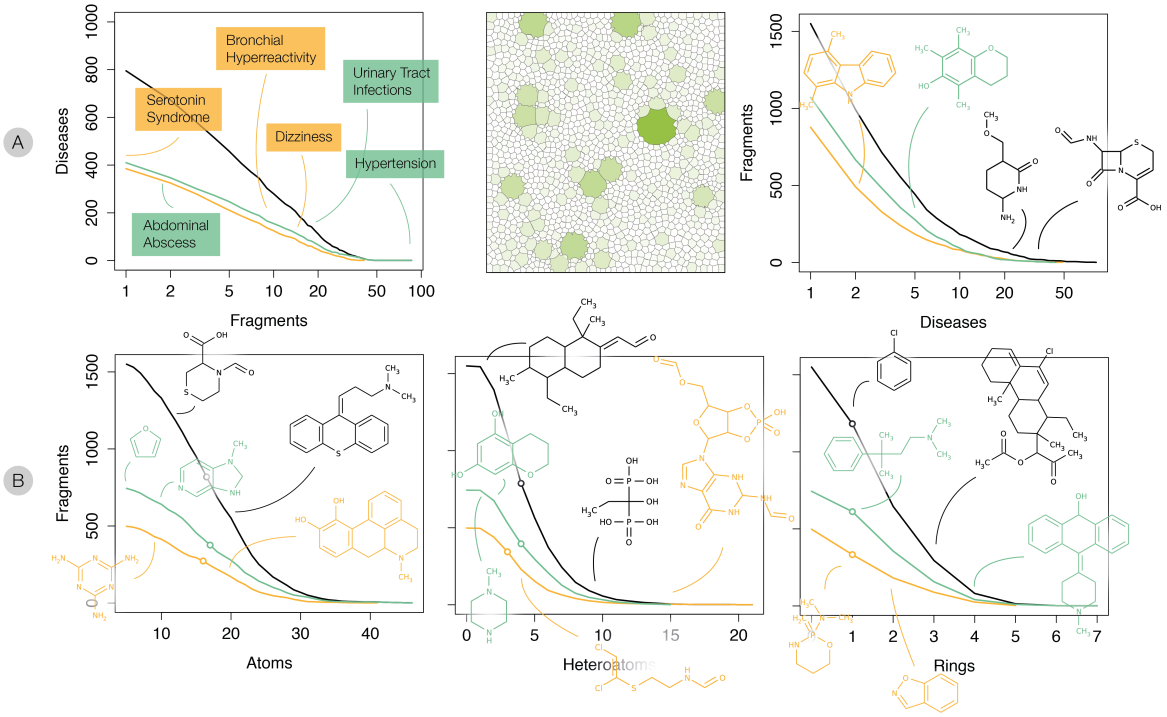

Figure 2.

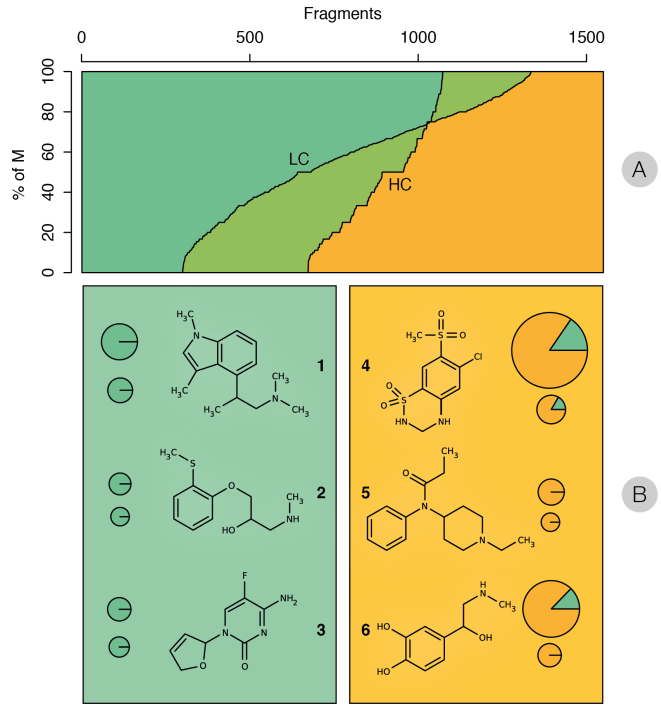

Figure 3.

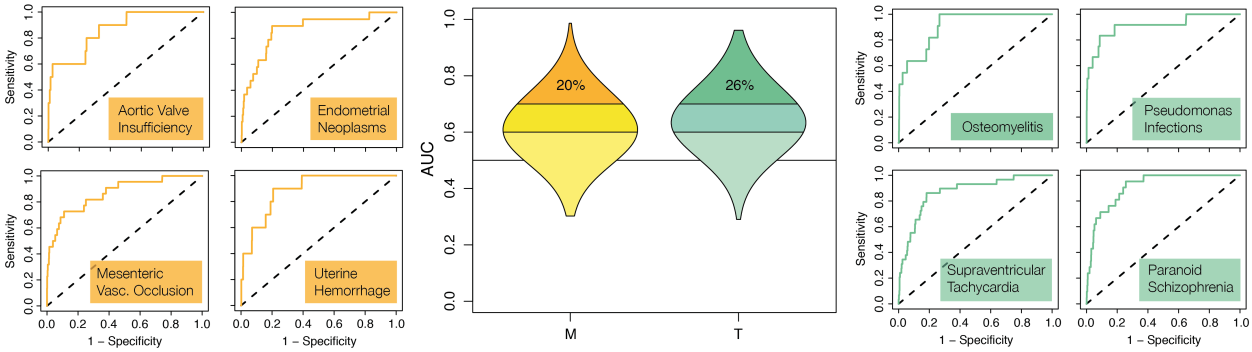

Figure 4.

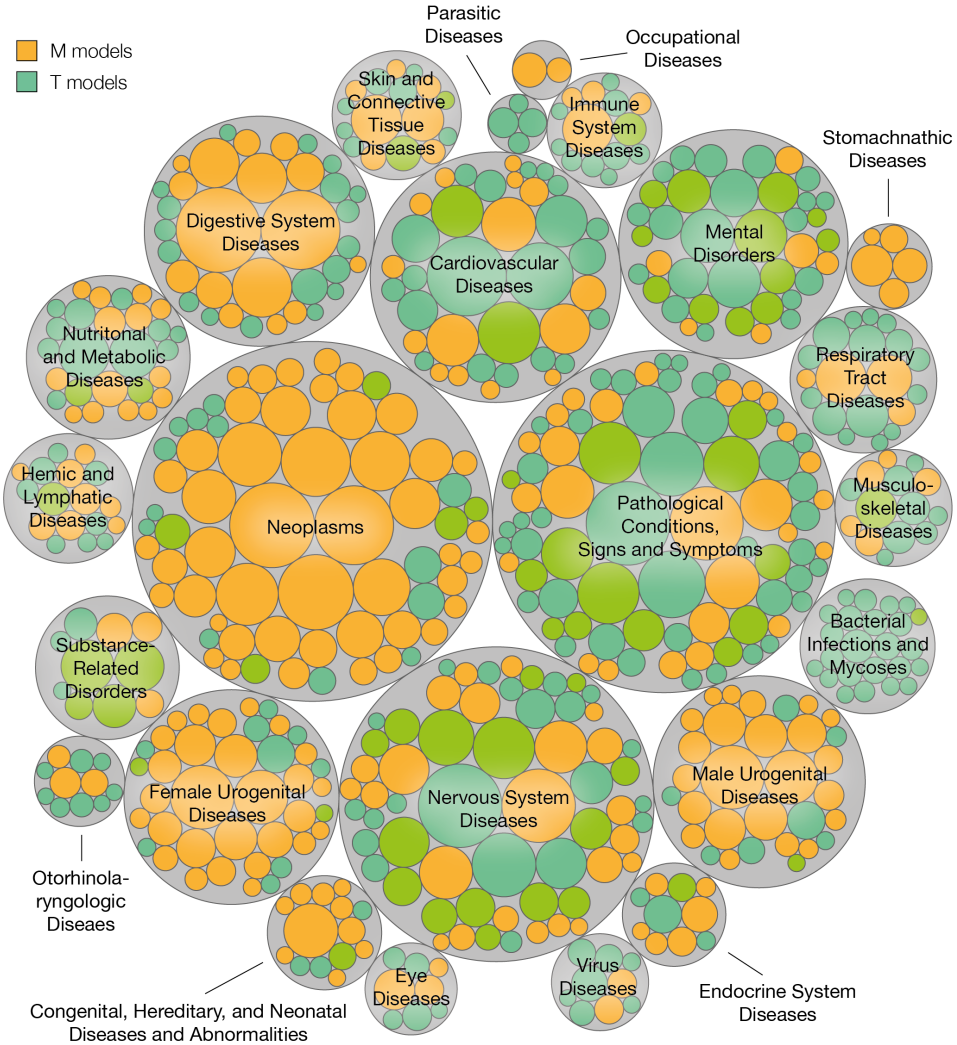

Figure 5.

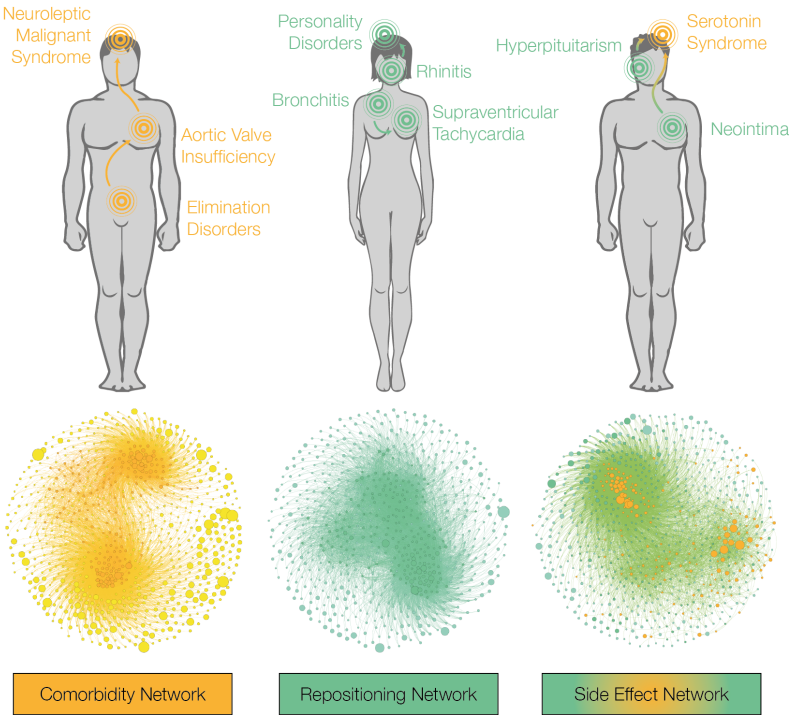

Figure 6.

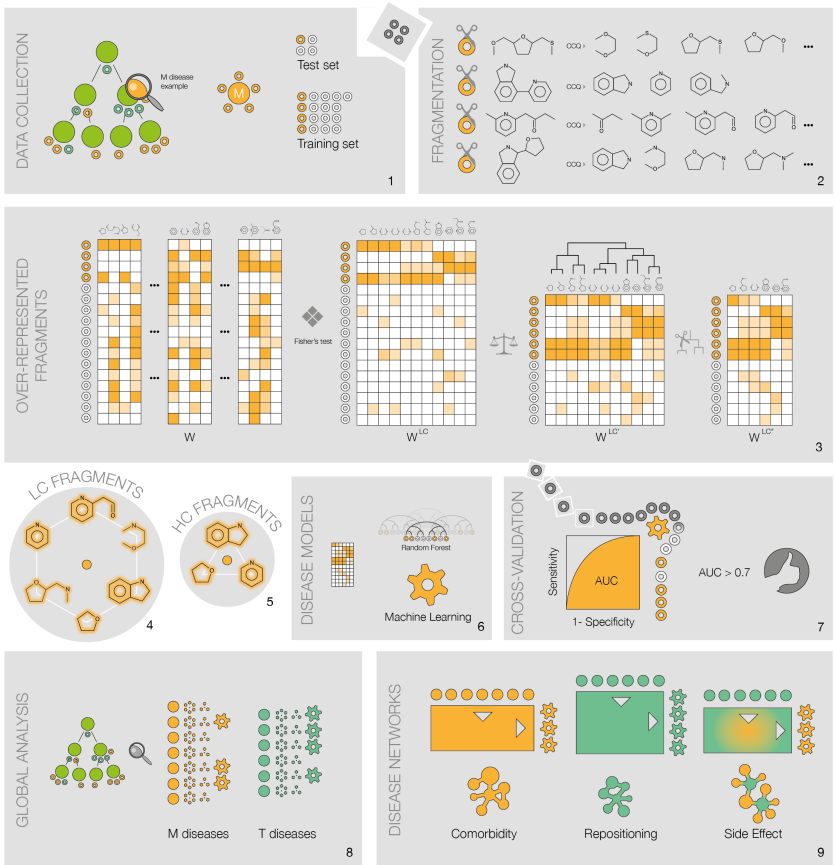

Table 1.

	M	T	Total
Diseases	934	835	1,176
Molecules per disease	36	25	30
LC fragments	23,135	28,325	37,809
HC fragments	910	1,107	1,550
LC fragments per disease	204.5	196.5	200.5
HC fragments per disease	5	6	6
Liable (M) and privileged (T) fragments	348	367	715
Diseases with ≥ 1 HC fragment	385	409	794
AUC	0.613	0.641	0.627
Specificity	0.878	0.882	0.880
Sensitivity	0.265	0.292	0.278
Balanced accuracy	0.571	0.588	0.579
Positive predictive value	0.032	0.023	0.029
G-mean	0.463	0.488	0.475
F_1-score	0.053	0.044	0.049
Diseases with AUC > 0.7	184	216	400

Table 2.

	Target Diseases	Source Diseases	Nodes	Directed Edges	In-degree	Out-degree	Undirected edges	Degree
Comorbidity	184 M	934 M	934	12,610	44.5	7	10,917	8
Repositioning	216 T	835 T	835	14,590	63	7	11,997	8
Side effect	184 M	835 T	1,019	9,921	31.5	2	9,921	8