Towards an Integrated Compound to Compound Relatedness Measure

Maria Kissa*, Michael Schroeder and George Tsatsaronis

Biotechnology Center (BIOTEC) of the Technische Universität Dresden, Tatzberg 47-51, 01307 Dresden Germany.

ABSTRACT

The wealth of the publicly available data repositories related to chemical compounds and substances allows current research methodologies to integrate pieces of information across different resources. Typical compound-to-compound relatedness measures are based on structural commonalities between the compounds, sequential information for their targets or/and toxicological liabilities. In this paper, we take a step further towards compound-to-compound relatedness and integrate a significantly larger number of compound characteristics, including chemical properties, indications, related sequences, pathways, genes, toxicity, and other pharmacological information. The main novelty of the suggested methodology is the systematic data integration and the combination of several similarity measures, including a *string kernel*, *Jaccard* and *Tanimoto*. We evaluate the performance of the proposed relatedness measure through a manually curated benchmark dataset, also introduced in this work. Our results suggest that the proposed method generates meaningful associations among the tested compounds; examples of these associations are presented and discussed analytically.

Keywords: compound relatedness, integration of compound resources, substance characteristics, evaluation

1 INTRODUCTION

Over the past decade there has been a growing interest in research methodologies that aim at producing computationally associations between compounds. The range of applications varies and, among others, includes drug-target prediction (Perlman *et al.*, 2011), biological activity elucidation (Klebe *et al.*, 1994), computational drug design (Zhou *et al.*, 2010), identification of drug resistance factors (Gamo *et al.*, 2010), prediction of drugs' adverse events (Bender *et al.*, 2007) and indications (Gottlieb *et al.*, 2011), as well as drug repositioning (Li and Lu, 2012).

With regards to the methodological approaches followed, in the majority of the cases, the following aspects of the chemical compounds are considered: (i) adverse events, (ii) chemical features, (iii) sequence data, and, (iv) combinations of the aforementioned. More specifically, regarding the use of adverse event profiles, it has been shown to lead to the identification of shared targets between drugs (Campillos *et al.*, 2008). For the same purpose, drug actions and phenotypic information are also considered (Mizutani *et al.*, 2012). In addition, similarities between compounds have been inferred based on chemical features (Tabei *et al.*, 2012) or sequential information of drug targets (Yamanishi *et al.*, 2008). In parallel, the role of text mining in identifying related compounds for drug repositioning purposes, has also been found extremely important (Roberts and Hayes, 2008).

In this paper we adopt a wider perspective of the compound characteristics that may be utilized in a compound-to-compound relatedness measure[1]. More precisely, we integrate the information from numerous popular databases pertaining to compound data, and employ an extended set of similarity measures in order to compute compound relatedness in a wide range of considered dimensions. In total, 30 distinct compound characteristics are taken into consideration for the quantification of the relatedness between two given compounds. Notably, apart from the high data diversity, a significant bottleneck arising towards the systematic evaluation of the suggested relatedness measure was the lack, to the best of our knowledge, of a high quality benchmark dataset, which could be used to assess the reliability of the associations generated via the proposed methodology. Hence, an additional contribution of this work is the creation of such a manually curated benchmark dataset for the quality assessement of the suggested approach. In the following, we give an overview of the used resources (Section 2.1), a description of the created benchmark dataset (Section 2.2) and the considered compound characteristics (Section 3.1), as well as the employed similarity measures (Section 3.2). Finally, we discuss the potentiality of our method based on a case study of 10 known chemical compounds and their produced associations (Section 4).

2 MATERIALS

2.1 Data sources

The publicly available repositories comprising compound related data vary. Herein we attempt to exploit as much of this information as possible, by integrating data from five popular repositories, namely *DrugBank* (Knox *et al.*, 2011), *SIDER* (Kuhn *et al.*, 2010), *CTD* (Davis *et al.*, 2012; Wiegers *et al.*, 2009), *PharmaGKB* (Whirl-Carrillo *et al.*, 2012), and *STITCH* (Kuhn *et al.*, 2011).

DrugBank is a popular drug repository comprising approximately 6,700 drug entries regarding approved, experimental and nutraceutical drugs and their targets. The set of drugs included in this database serves as the drug set to which our methology is applied. Each drug record (*drugcard*) consists of structural, chemical and protein information regarding a drug and its targets. The respective fields constitute some of the features used by our method. Additionally, for each compound in *DrugBank* we retrieved the

*to whom correspondence should be addressed: maria.kissa@biotec.tu-dresden.de

[1] In the remaining of the paper, the terms *relatedness* and *similarity* might be used interchangeably to describe the notion of the similarity for a given pair of compounds.

Compound *I*	Compound *II*	Condition	Relation Type
Gabapentin	**Pregabalin**	neuropathic pain	derivatives
Buprenorphine	**Methadone**	opioid replacement therapy	repositioning
Lenalidomide	**Thalidomide**	mutiple myeloma	derivatives
Metformin	**Sitagliptin**	diabetes milletus	synergic
Ropinirole	**Methadone**	restless legs syndrome	repositioning
Thalidomide	**Vancomycin**	prurigo nodularis	repositioning
Thalidomide	**Vancomycin**	chronic bullous disease of childhood	repositioning

Table 1. The considered compound pairs and their relations, which are used in our case study.

corresponding adverse events information available in the *SIDER* database. *SIDER* currently counts roughly $1,000$ drugs and $4,200$ side effects.

Taking into account that a drug is not only characterised by its adverse effects but also from the diseases that it has been found to be associated with, we chose to include information from the *Comparative Toxicogenomics Database* (*CTD*). *CTD* provides detailed reports about the diseases that a drug is either found to be correlated to, have a therapeutic application, or a potential etiological role in its development. In addition, we include the respective pathway information provided by the *PharmaGKB* database, for the drugs that it was already made available. Finally, we employed *STITCH* to collect chemical-to-chemical interactions and chemical-to-protein relations. The overall size of the raw data that we collected adds up to 16.3 GB. Each of the data sources was stored in an independent database, and as a final step, an integrated database was produced in-house using as keys the *DrugBank* and *PubChem* (Bolton *et al.*, 2008) ids.

2.2 Benchmark Dataset

For the evaluation of the suggested measure, a gold standard set comprising pairs of drugs with similar properties is a prerequisite. Unfortunately, such a gold standard dataset does not exist. To overcome this issue, we manually built and curated such a dataset from the literature pertaining to drug repositioning. The dataset consists of roughly 150 drug pairs belonging to four distinct categories: (1) derivative compound, (2) known repositioned compounds, (3) compounds of synergic action, and, (4) compounds of opposite effects. For example, there are pairs wherein one of the two drugs is repositioned to a new disease and the other drug is the recommended medication for the treatment of this disease. Additionally, drugs that function synergically in a therapeutic application are being considered. Moreover, pairs of derivative drugs as well as pairs of drugs bringing the opposite effect (i.e., one drug treats a condition which is reported to be the adverse event of the second drug) are also taken into account.

For the purposes of our evaluation we applied our relatedness measure to a case study subset of the aforementioned data set. This set comprises 10 known compounds, namely: *Methadone*, *Buprenorphine*, *Vancomycin*, *Thalidomide*, *Lenalidomide*, *Ropinirole*, *Pregabalin*, *Gabapentin*, *Sitagliptin*, and *Metformin*. Their relation type that exists in the produced benchmark is shown in Table 1. In detail, *Pregabalin* was designed as a potent successor to *Gabapentin* for the treatment of *neuropathic pain*, among other conditions. *Buprenorphine*, an analgesic compound, was found to have a new application in the *opioid replacement* therapy, so far controlled with compounds such as *Methadone*. When *Thalidomide*'s therapeutic potential against *multiple myeloma* was discovered, *Lenalidomide* was proposed as an effective derivative without the teratogenic liabilities of its analog. Towards the therapy of *diabetes mellitus*, *Sitagliptin* and *Metformin* are co-administered. *Ropinirole* was found able to substantially replace *Methadone* in the therapy of *restless legs syndrome*, while *Thalidomide*'s new applications included treatment of conditions such as *prurigo nodularis* and *chronic bullous disease of childhood*; conditions for which *Vancomycin* has been also administered.

Similarity	Characteristic	Similarity	Characteristic
Jaccard	Interacting Drug	**Numeric**	logP
	Salts		logS
	Adverse Event		pka
	Indications		Polar Surface Area
	Pathway		Refractivity
	Interacting Gene		Polarizability
	Diseases		Rotatable Bond Count
	Related Protein		logp/hydrophobicity
String Kernel	Drug Description		Mass
	Mechanisms of Action		caco2 Permeability
	Pharmacology	**BLAST**	targets
	Protein Binding		carriers
	Toxicity		transporter
	Metabolism		enzymes
Tanimoto	3D structure of the compound		DNA sequences

Table 2. The considered compound characteristics grouped according to the employed similarity measure.

3 METHODS

3.1 Compound Characteristics

Following the integration of the resources described in Section 2.1, we present here the compound characteristics which are taken into account for the computation of the compound relatedness values. These are shown in Table 2, and they are grouped according to the similarity measure that is used for each of the characteristics. The similarity measures used are explained analytically in the next section.

As the table shows, we have collected for each compound a total of 30 characteristics, which may be split into five categories: (1) fields for which we have a set of string values for each compound, where the values are stemming from a finite and predefined vocabulary, e.g., the interacting drugs for each compound, or the pathways which the compound affects; for a pair of compounds, their relatedness in these fields may be measured using the *Jaccard similarity coefficient*, (2) fields for which we have numeric values, e.g., logP or the polarizability of the compound; in these fields the relatedness may be measured after normalizing the numeric values for each field in the range $[0, 1]$, and then, given a pair of such values, the distance may be computed, through which the similarity may be inferred by subtracting from the unit, (3) fields for which we have free textual descriptions; in these fields the relatedness may be measured using a string kernel that considers $n-$grams, (4) fields for which we have the *FASTA* sequences of related target, carrier, transporter or enzyme proteins, as well as the respective *DNA* sequences; the relatedness in these fields may be computed using *BLAST*(Altschul *et al.*, 1990), but since for each compound we have a set of such sequences, we employ a measure called *THESUS*(Varlamis *et al.*, 2004) to generalize the similarity to the sets of a given pair of compounds, (5) the *3D* structural representation of the compounds, in which case we compute the *Tanimoto* similarity for a given pair of compounds in *3D*.

3.2 Similarity Measures

In the following, we give the details of the five different employed similarity measures, and we discuss how we combine them to measure the overall relatedness for a given pair of compounds considering all of the 30 compound characteristics. For the description of the measures we will use the notation shown in Table 3.

Symbol	Description
f_i	A field (characteristic) of a compound among the 30 of Table 2
F_1	The set of the *Jaccard* similarity compound fields
F_2	The set of the string kernel similarity compound fields
F_3	The set of the numeric similarity compound fields
F_4	The set of the *BLAST* similarity compound fields
C_1, C_2	Two input compounds
F_{C_1}, F_{C_2}	All the fields of C_1 and C_2 respectively for which values are not null
$f_{(i,C_1)}, f_{(i,C_2)}$	The set of values of field f_i for C_1 and C_2 respectively
$\|f_{(i,C_1)}\|, \|f_{(i,C_2)}\|$	The number of elements (values) that field f_i contains for C_1 and C_2 respectively
C	The set of all compounds

Table 3. Summary of notation used for the description of the similarity measures.

3.2.1 Jaccard Similarity Coefficient*:* Given the fields $f_i \in F_1$, we can compute the similarity in each such field between C_1 and C_2, as shown in the following equation. This similarity is always in the range $[0, 1]$.

$$J(f_{(i,C_1)}, f_{(i,C_2)}) = \frac{|f_{(i,C_1)} \cap f_{(i,C_2)}|}{|f_{(i,C_1)} \cup f_{(i,C_2)}|} \quad (1)$$

3.2.2 String Kernel*:* Given the fields $f_i \in F_2$, we can compute the similarity in each such field between C_1 and C_2, using the string kernel suggested by Lodhi et al. (2002). The used kernel receives as input two texts and computes their inner product in the feature space generated by all character subsequences of maximum length n ($n-$grams). We use $n = 5$, and the implementation provided by *Mallet*[2]. The resulting similarity between two input texts is always in the range $[0, 1]$.

3.2.3 Numeric Distance as Similarity*:* Given the fields $f_i \in F_3$, we can compute the similarity in each such field between C_1 and C_2, by normalizing first the values $f_{(i,C_1)}$ and $f_{(i,C_2)}$ in the range $[0, 1]$, as shown in Equation 2 (for C_1, but the same can be applied for C_2), and then, computing the similarity as shown in Equation 3. The resulting similarity between two numeric values is always in the range $[0, 1]$, because their distance is also in the same range.

$$f'_{(i,C_1)} = \frac{f_{(i,C_1)} - MIN_{C_j \in C}(f_{(i,C_j)})}{MAX_{C_j \in C}(f_{(i,C_j)}) - MIN_{C_j \in C}(f_{(i,C_j)})} \quad (2)$$

$$NumSim(f'_{(i,C_1)}, f'_{(i,C_2)}) = 1 - |f'_{(i,C_1)} - f'_{(i,C_2)}| \quad (3)$$

where $|f'_{(i,C_1)} - f'_{(i,C_2)}|$ symbolizes the absolute value of the Euclidean distance between $f'_{(i,C_1)}$ and $f'_{(i,C_2)}$.

3.2.4 Tanimoto*:* Given the one and only *3D* structural field for C_1 and C_2, we compute the similarity between them using the *Tanimoto* similarity measure applied on the *3D* fingerprints of the compounds. For the purposes of our implementation, we are using the *Marvin suite* API[3].The resulting *Tanimoto* similarity between the two compounds is always in the range $[0, 1]$, as the measure is equivalent to the *Jaccard* similarity index.

3.2.5 THESUS on BLAST*:* Given the fields $f_i \in F_4$, we can compute the similarity in each such field between C_1 and C_2, by computing first the *NCBI's BLAST* sequence alignment[4] for every pair of sequences belonging to $f_{(i,C_1)}$ and $f_{(i,C_2)}$. Then, we compute the overall similarity between the two sets $f_{(i,C_1)}$ and $f_{(i,C_2)}$ by applying the *THESUS* set similarity measure (Varlamis *et al.*, 2004). In summary, *THESUS* finds for each sequence in $f_{(i,C_1)}$, the maximum *BLAST* similarity with any sequence in $f_{(i,C_2)}$. Once it computes all the $|f_{(i,C_1)}|$ maximum similarities, it does the same reversely, for each sequence in $f_{(i,C_2)}$. Finally, it averages the sum of all the computed maximum similarities (in total there are $|f_{(i,C_1)}| + |f_{(i,C_2)}|$ such). Since the *BLAST* similarities are in the range $[0, 1]$, in this case *THESUS* also returns an overall similarity in the same range.

3.2.6 Overall Relatedness*:* Finally, we compute the overall relatedness between C_1 and C_2 by averaging the individual similarity scores on all common fields between F_{C_1} and F_{C_2}.

4 RESULTS

The results of our experimental evaluation using the 10 drugs dataset described in Table 1 follows. First, we computed all the pairwise relatedness scores between all 10 drugs, i.e., a total of 45 pair relatedness scores. Next, for each drug we produced the ranking of the remaining 9 based on the relatedness scores, and measured for each one the *R-precision*, where R is the number of related drugs that we should identify for the examined drug based on our dataset. Overall, for 7 out of the 10 drugs, *R-precision* was found 100%, for 2 out of the 10 it was found 50% and for only one was found 0%. This shows that the suggested methodology produced very meaningful rankings in the majority of the cases.

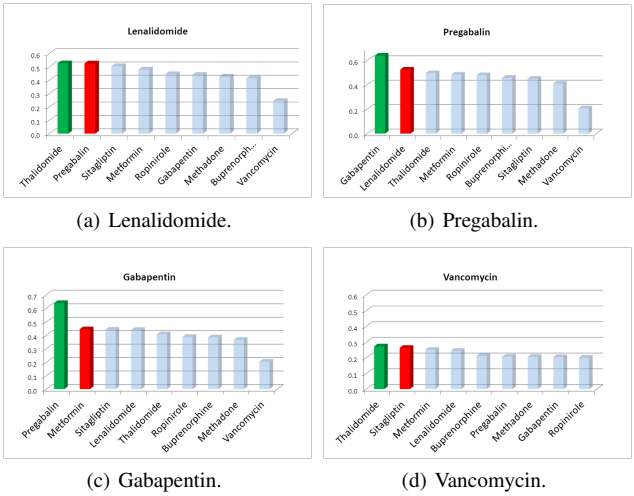
In addition, we explored whether there is any relation between any of the drugs and its top-ranked related drugs which we might have missed in our dataset. The results of this manual curation were very encouraging. In the majority of the cases, though the top-ranked was in our dataset, we identified a missing relation between the second ranked drug as well. We analyze this process for 4 out of the 10 drugs in the following, with a visualization shown in Figure 1. In the case of *Lenalidomide*, the top-ranked (*Thalidomide*) is its analog, which we knew from our dataset. *Lenalidomide* was designed as an analog to avoid the teratogenic effects of *Thalidomide*. The second drug is the anti-epileptic compound *Pregabalin*. This drug is reported to treat myeloma neuropathy, a symptom of myeloma to which *Lenalidomide* is used as a primary treatment, a relation which we did not know it existed. Regarding *Pregabalin*, its known predecessor *Gabapentin* received the highest similarity score. *Lenalidomide* follows, as described previously, which shows consistency between the results. Moving to *Gabapentin*, *Pregabalin* ranks first, and *Methformin* follows. Interestingly, it is reported that co-administration of these compounds may result in reversible hearing loss, therefore it is highly probable that they share some contradictory functionalities. Lastly, *Vancomycin* has been found similar to *Thalidomide* and *Sitagliptin*. With regards to *Thalidomide*, the reason is obvious, since that was the compound repositioned for the condition *prurigo nodularis*, to which *Vancomycin* constitutes also a therapeutic agent. As far as *Sitagliptin* is concerned, this compound along with *Vancomycin* is used in the treatment of renal failure, a relation we also did not know it existed.

From the aforementioned discussion we can conclude that the suggested measure seems to produce meaningful associations between the compounds. We also observed that the same pipeline might also be used towards a (semi-)automated approach for the production of a benchmark set in order to evaluate drug or compound similarity measures. However, there are still many parameters to be considered in order to apply the suggested measure in a larger scale and produce rankings for thousands of related compounds.

[2] `http://mallet.cs.umass.edu/`

[3] `http://www.chemaxon.com/products/marvin/`

[4] We use a cutoff of 1 on the e-value, and we store the maximum identity found after the alignment.



(a) Lenalidomide. (b) Pregabalin.

(c) Gabapentin. (d) Vancomycin.

Fig. 1. Rankings of the top most related drugs to *Lenalidomide* (Fig. 1(a)), *Pregabalin* (Fig. 1(b)), *Gabapentin* (Fig. 1(c)), and *Vancomycin* (Fig. 1(d)). Highlighted with green, are the relations covered by our dataset. With red are the relations we did not know they actually existed.

Our measure currently assumes the same weight for every of the 30 considered compound characteristics. Naturally, this needs to change specifically to the application for which the relatedness measure is applied, e.g., by learning the respective weights of the characteristics for this application.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a novel relatedness measure for chemical compounds. The novelty of the measure lies in the utilization of 30 different compound characteristics stemming from the integration of several popular databases, namely *DrugBank*, *SIDER*, *CTD*, *PharmaGKB*, and *STITCH*. The measure uses five different similarity measures to compute the overall compound-to-compound relatedness, depending on the nature of the values of each of the 30 characteristics considered. For the purposes of our evaluation, we created a benchmark dataset of compound pairs for which there exists a strong reported relation, i.e., derivative compounds, known repositioning examples, synergistic drugs, or compounds that act similarly but with opposite effects. Our evaluation in a set of 10 compounds showed that the produced ranked associations of the suggested measure are meaningful and consistent. As a future work, we are planning to explore several alternatives to the used similarity measures, such as semantic smoothing kernels for the text characteristics, and to expand our evaluation in a significantly larger number of examined compounds, and conduct an analysis per relation type of the benchmark compound pairs.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

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

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*, **4**.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug Target Identification Using Side-Effect Similarity. *Science*, **321**(5886), 263–266.

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., and Mattingly, C. J. (2012). The comparative toxicogenomics database: update 2013. *Nucleic Acids Research*, **41**(D1), D1104–D1114.

Gamo, F.-J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.-L., Vanderwall, D. E., Green, D. V. S., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R., and Garcia-Bustos, J. F. (2010). Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**(7296), 305–310.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, **7**(1).

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *Journal of Medicinal Chemistry*, **37**, 4130–4146.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, **39**(Database issue), 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. *Molecular Systems Biology*, **6**(1).

Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., and Bork, P. (2011). STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Research*, **40**(D1), D876–D880.

Li, J. and Lu, Z. (2012). A new method for computational drug repositioning using drug pairwise similarity. In *BIBM*, pages 1–4.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. J. C. H. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, **2**, 419–444.

Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating drug-protein interaction network with drug side effects. *Bioinformatics*, **28**(18), i522–i528.

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, **18**(2), 133–145.

Roberts, P. M. and Hayes, W. S. (2008). Information needs and the role of text mining in drug development. In *Pacific Symposium on Biocomputing*, pages 592–603.

Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., and Yamanishi, Y. (2012). Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. *Bioinformatics*, **28**(18), i487–i494.

Varlamis, I., Vazirgiannis, M., Halkidi, M., and Nguyen, B. (2004). Thesus, a closer view on web content management enhanced with link semantics. *IEEE Trans. Knowl. Data Eng.*, **16**(6), 685–700.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., Altman, R. B., and Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*, **92**(4), 414–417.

Wiegers, T. C., Davis, A. P., Cohen, K. B., Hirschman, L., and Mattingly, C. J. (2009). Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC Bioinformatics*, **10**, 326.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**(13), i232–i240.

Zhou, T., Huang, D., and Caflisch, A. (2010). Quantum mechanical methods for drug design. *Current Topics in Medicinal Chemistry*, **10**(1), 33–45.