# Predicting serious rare adverse reactions of novel chemicals

Aleksandar Poleksic[*,†] and Lei Xie[*,‡,§]

[†]Department of Computer Science, University of Northern Iowa, Cedar Falls, Iowa 50614, USA

[‡]Department of Computer Science, Hunter College, The City University of New York, New York, New York 10065, USA

[§]Ph.D. Program in Computer Science, Biochemistry and Biology, The Graduate Center, The City University of New York, New York, New York 10065, USA

*Correspondence to poleksic@cs.uni.edu or lei.xie@hunter.cuny.edu.

**Abstract:** *Adverse drug reactions (ADRs) are one of the main causes of death and a major financial burden on the world's economy. Due to the limitations of the animal model, computational prediction of severe ADRs is invaluable. However, current state-of-the-art computational methods for prediction of rare and serious ADRs do not yield significantly better results than random guessing. We present a novel computational method, based on the theory of "compressed sensing", that can accurately and reliably predict serious side-effects of candidate and market drugs. Not only is our method able to infer new chemical-ADR associations using existing noisy, biased, and incomplete databases, but our data also demonstrates that the accuracy of our approach in predicting a serious adverse reaction (ADR) for a candidate drug increases with increasing knowledge of other ADRs associated with the drug. Namely, as the candidate drug moves up the different stages of clinical trials, the prediction accuracy of our method will increase accordingly. Thus, the compressed sensing based computational method reported here represents a major advance in predicting severe rare ADRs, and may facilitate reducing the time and cost of drug discovery and development.*

Adverse drug reactions (ADRs) are one of the main burdens in modern drug discovery [1]. Rare and serious ADRs are responsible for failed drug discovery pipelines and for drug market withdrawals. Cumulative costs of the management of ADRs have been estimated at more than 30 billion per year in the US alone [2]. Clinical impact, including the emergency department visits and prolonged hospital stay, account for a large portion of health care cost. Up to one-third of emergency visits by older adults are due to ADRs [3], while more than one third of ADRs in the pediatric population are potentially life threatening [4]. According to a nationwide Swedish study, ADRs rank 7$^{th}$ among all causes of death [5]. The figures from US studies are even more alarming as they

place ADRs as the 4th most common cause of death, ahead of diabetes, pulmonary disease, AIDS, pneumonia, general accidents, and automobile accidents [6].

A serious and rare ADR often surfaces years or even decades after the drug has been approved. An inability to predict these events leads to complications in diseases and treatments, which can have long term consequences and fatal outcomes. Drug pipeline failures and post-marketing drug withdrawals result in loss of effective compounds (those for which the benefit-to-harm balance is unfavorable) and loss of revenue by the drug manufacturer. A methodology capable of predicting ADRs, long before the drug reaches the market or even before the drug is withdrawn from the market, would significantly enhance the drug discovery process and improve human health.

The existing computational approaches for ADR prediction are unreliable, in particular when it comes to predicting rare and serious ADRs. Further progress in the field is hindered by the lack of clean, complete and readily accessible databases on drug-ADR associations as well as the inability of current methodologies to deal with sparse and noisy data.

We show that the so-called “compressed sensing” technique, from the digital signal processing field, can be easily adapted and used to predict drug-ADR associations with unmatched accuracy. Originally proposed to solve problems arising in coding and data acquisition, compressed sensing (CS) has proved to be an efficient way of recovering any type of signal from a few and erroneous samples [7,8,9]. In the framework of ADR prediction, the “signal” can be thought of as the set of all drug-ADR associations (those already known and those yet to be found). The “sample” represents known (reported) associations, identified and stored in the existing drug-ADR databases, such as SIDER [10]. The key observation is that the sample, defined this way, is both sparse and noisy, due to the well-known difficulty of identifying ADRs during clinical trials and post-marketing studies. Therefore, just like the problems in imaging and face recognition, or problems in optical systems research or wireless networking, the drug-ADR association prediction problem is highly amenable to “compressed sensing” solution.

## RESULTS

The compress sensing (CS) algorithm is able to reliably infer new chemical-ADR associations using existing noisy, biased, and incomplete data. Not only is CS highly tolerant to database errors (mislabeled drug-ADR associations), but it also handles sparse data (yet unknown/unrecorded associations) very well.

To date, only several computational methods have been developed for predicting adverse reactions of drugs. We compared our compressed sensing (CS) approach to two recent and (according to published findings) top performing algorithms. The first algorithm is based on supervised "multi-label" learning" (ML) [11], while the second algorithm is based on sparse "canonical correlation analysis" (CCA) [12].

In order to gain insight into the overall progress in the field, we also included in our benchmark a naïve, reference method. This straightforward and simple to implement method, which we call REF, sets the probability that a given drug will give rise to a particular ADR to the overall promiscuity of that ADR. In other words, the probability of a side-effect *A* being associated to any drug is the same for all drugs and is set to the overall frequency of occurrence of *A*.

To compare compressed sensing (CS) to the other three methods, we ran a classical blind benchmarking experiment, consisting of 5 rounds of 10-fold cross validation on the set of all drug-ADR associations from the SIDER database. SIDER (version 4.1) contains drug-ADR association data for 1430 FDA approved drugs and 5868 ADRs. Each method was run using its default parameters. We employed two well-known performance measures, namely the area under the ROC curve (AUC) and the area under the PR curve (AUPR). The ROC (Receiver Operating Characteristic) curve represents the relationship between the false-positive and the true-positive rate while the PR (Precision-Recall) curve represents the relationship between the sensitivity (true positive rate or recall) and the positive predictive value (precision).

### *Existing methods are not much better than random guess*

To assess methods' accuracies on rare ADRs, we ran multiple, independent and statistically robust, cross-validation experiments, one for each selected ADR promiscuity cutoff (12, 25, 50, 100, 200, 400, 800, ∞), where the "promiscuity" of an ADR is defined as the number of FDA approved drugs that are known to cause the ADR (Fig. 1). For each promiscuity cutoff *x*, a cross-validation experiment was performed on the set of drug-ADR pairs in which the ADR's promiscuity does not exceed *x* (*x*-axis of Fig. 1).

While our analysis confirms the published accuracy of current methods, it also provides very useful insights into the performance of the State-of-the-Art algorithms. We summarize them as follows:

(i) The existing algorithms are unable to predict serious side-effects. While the published accuracies of current methods are more or less satisfactory (balanced and unbalanced AUC and balanced AUPR ~0.9; unbalanced AUPR ~0.35), they should be interpreted properly, since they only represent the average accuracies computed for all ADRs combined (the right side of Fig. 1). The cumulative accuracies are driven strongly by easy predictions of frequent and innocuous ADRs, those of little interest in drug discovery. For rare and serious ADRs, the accuracy of current algorithms quickly approaches the accuracy a purely random classifier (AUC ~0.5).

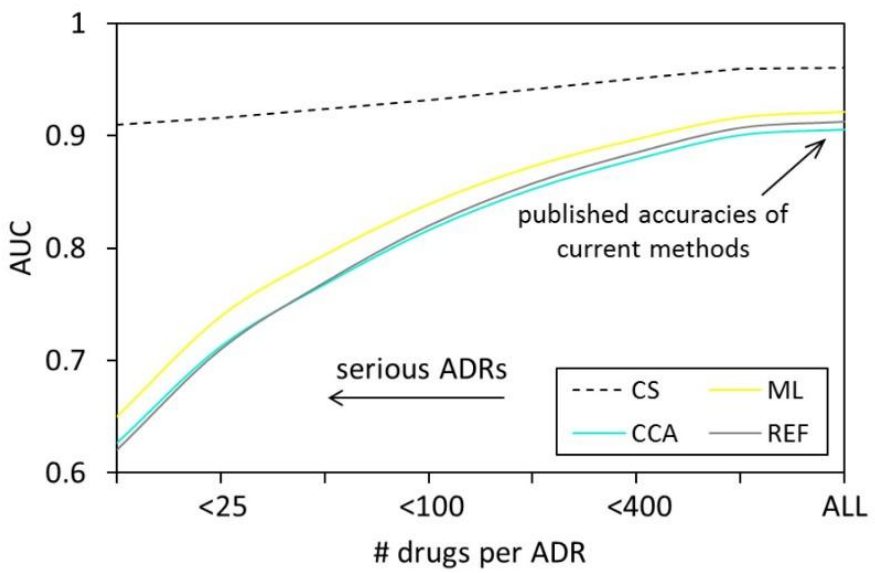


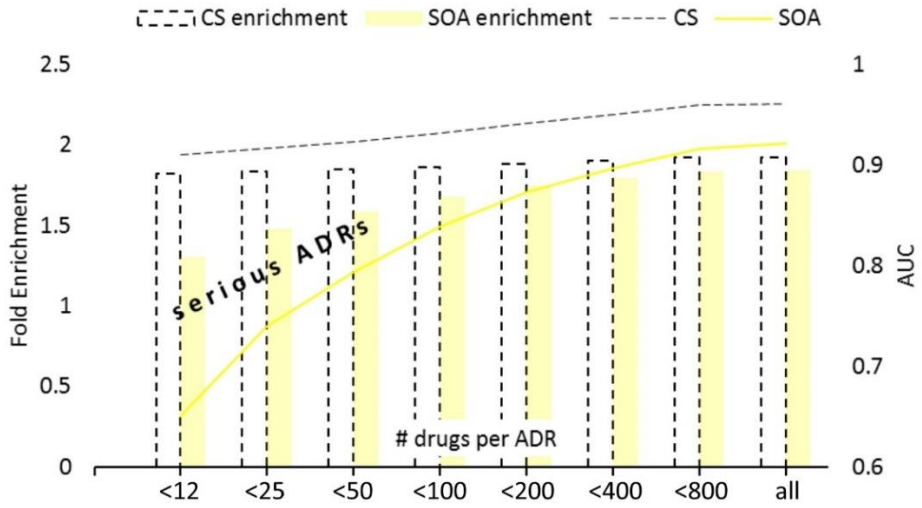

**Fig. 1. State-of-the-art in ADR prediction and the value added by compressed sensing.** CS: Compressed Sensing; ML: Multi-Labeling; CCA: Canonical Correlation Analysis; REF: naïve (reference) method. The values on the x-axis represent ADR promiscuities. The $y$-axis represents the performance metrics, defined as the area under the ROC curve (AUC). The results were obtained using a statistically rigorous cross-validation experiment (STDEV too small to show) on the SIDER database.

(ii) To date, the progress in the field of computational ADR prediction has been dismal at best. To assert this claim, it is enough to glance over the line that traces the performance of the REF method in Fig 1. Going beyond this simple approach and implementing more sophisticated techniques, such as multi-label learning and canonical correlation analysis, yields a disappointing, low diminishing return.

### *CS significantly improves the ADR prediction*

Fig. 1 shows that "compressed sensing" (CS) easily overcomes current obstacles in predicting drug-ADR associations. Our method is so efficient in extracting relevant data from noisy, biased and incomplete databases that its performance in predicting severe ADR (left part of Fig. 1) matches or even exceeds the cumulative performances of current methods on all side-effects combined (right part of Fig. 1).

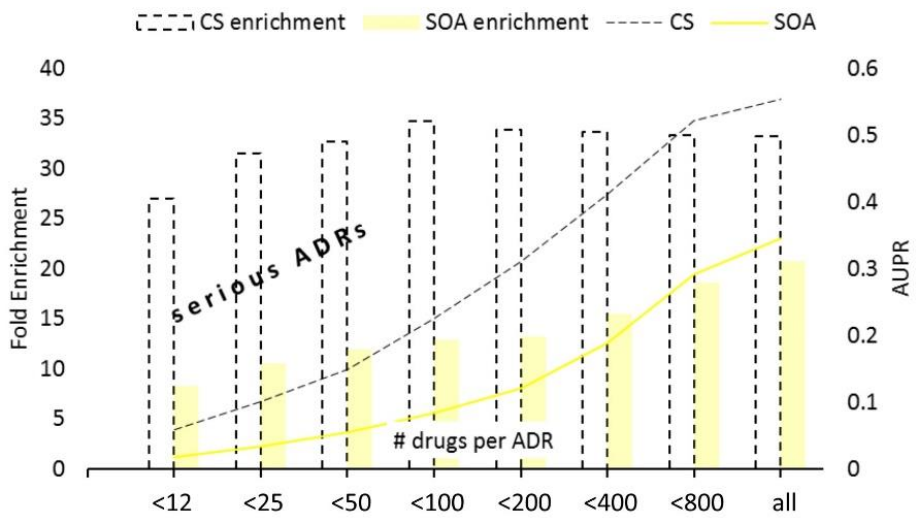
A more detailed and more illustrative performance analysis is presented in figures 2 and 3. For clarity of presentation, figures 2 and 3 show a head-to-head comparison of compressed sensing (CS) against the single, best performing method (we call it "State-of-the-Art", abbreviated SOA) with respect to the particular measure used (AUC or AUPR). Aside from showing the raw scores, figures 2 and 3 illustrate the "fold enrichment" offered by CS and SOA. The "fold enrichment" represents the improvement in a method's performance over the random predictor (one that generates prediction scores uniformly at random). In other words, the "fold enrichment" shows how many times is the method's AUC (or AUPR) better than the AUC (respectively, AUPR) obtained by the purely random classifier.




| | <12 | <25 | <50 | <100 | <200 | <400 | <800 | all |
|---|---|---|---|---|---|---|---|---|
| **CS** | 0.91 | 0.92 | 0.92 | 0.93 | 0.94 | 0.95 | 0.96 | 0.96 |
| **SOA** | 0.65 | 0.74 | 0.80 | 0.84 | 0.87 | 0.90 | 0.92 | 0.92 |

**Fig 2. Value added by compressed sensing (CS) in the AUC benchmark.** Yellow color traces the performance of the State-of-the-Art (SOA) method, defined as an existing method that exhibits the best performance in the AUC test (according to Fig. 1, SOA = ML). Dashed lines trace the performance of compressed sensing (CS). The data table beneath the graph gives the actual AUC values for CS and SOA.

Fold enrichment is particularly useful when interpreting the AUPR scores because, in contrast to the intuitive AUC scores, the AUPR scores depend on the property of the test set. More specifically, the AUPR scores obtained from the purely random classifier are equal to the fraction of condition positives in the test set. As seen in figures 2 and 3, CS enriches prediction of ADRs at an almost uniform rate, irrespective of the ADR promiscuity and the type of test performed (AUC or AUPR). For extremely rare ADRs, associated with less than 12 FDA drugs (such as carcinogenicity or death), the performance of CS, as measured by AUPR, is about 27 times better than the performance of the random classifier, while the performance of SOA is only about 8 times better. For the more frequent serious ADRs, such as neurotoxicity or cardiotoxicity (which, according to SIDER, are associated to < 50 FDA approved drugs) the fold enrichments of CS and SOA are 34 and 12, respectively.




| | <12 | <25 | <50 | <100 | <200 | <400 | <800 | all |
|---|---|---|---|---|---|---|---|---|
| **CS** | 0.06 | 0.10 | 0.15 | 0.22 | 0.31 | 0.41 | 0.52 | 0.55 |
| **SOA** | 0.02 | 0.03 | 0.05 | 0.08 | 0.12 | 0.19 | 0.29 | 0.35 |

**Fig. 3. Value added by compressed sensing (CS) in the AUPR benchmark.** In contrast to AUC, the actual AUPR values (shown in the data table beneath the graph) are more difficult to interpret. To put them in the context, we note that the AUPR of a random classifier represents the fraction of conditional positives in the data set ($\sum cond.pos / (\sum cond.pos + \sum cond.neg)$) We also note that the performance of both CS and SOA in this benchmark is probably much better than shown, due to unreliability and sparseness of data in the SIDER database. An illustrative example is provided in the Figure's 4 caption.

*Compressed sensing learns on the fly*

Aside from the large value added in predicting rare ADRs in the cross-validation benchmark, compressed sensing (CS) has another unique and practical benefit. As seen in figures 4, 5, and 6, its performance in

predicting ADRs for a particular chemical improves with the increasing knowledge of other ADRs associated with the chemical. In practice, this means that the ability of compressed sensing to predict a serious ADR for a candidate chemical increases as the drug progresses through various stages of clinical trials. Other methods are unable to take advantage of this information. This comes as no surprise to us, since a closer look into the ML algorithm reveals that, when predicting whether a drug *D* is likely to cause an ADR *A*, ML utilizes the information on other drugs that cause the side-effect *A* but not the information on other ADRs associated to *D*.

We tested the performance of compressed sensing (CS) in predicting selected serious and rare side-effects, including hepatotoxicity, cardiotoxicity, carcinogenicity, neurotoxicity, as well as immune thrombocytopenia, leukopenia, anaemia and death. Those eight ADRs are the main cause for drug market withdrawals [13]. Starting with hepatotoxicity, we selected all drugs known to cause hepatotoxicity ("cases") and the same number of randomly selected drugs that are known not to cause hepatotoxicity ("controls"). Fig. 4 illustrates the differences in normalized raw scores obtained by the compressed sensing (CS) and the other two methods on "case" and "control" drugs.

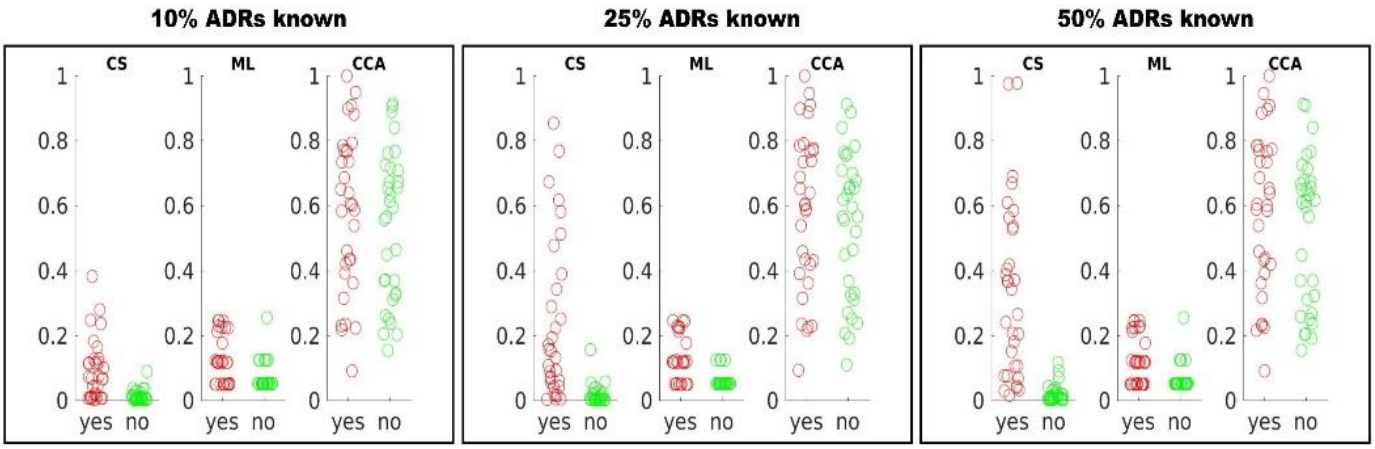



**Fig. 4. Predicting hepatotoxicity of drugs.** Red circles ("cases") represent drugs known to cause hepatotoxicity. Green circles ("controls") represent a randomly chosen set of drugs known not to cause hepatotoxicity. The vertical axis gives normalized prediction scores. Only compressed sensing (CS) scores can differentiate between the two sets of drugs. We note that the methods' predictions might be better than suggested in the figures. For instance, the "control" outlier (green circle), corresponding to the drug *minoxidil*, stands out by its high score in the middle figure for the CS method. Despite being classified as a non-hepatotoxicity drug in SIDER, *minoxidil* is, according to FDA case reports, known to cause hepatotoxicity in some patients, particularly those over the age of 60.

The performance statistics (AUC and AUPR) obtained from the algorithms' raw scores (after averaging the raw scores over a dozen of randomly chosen sets of "control" drugs) is presented in Fig. 5. Summary

performance data for cardiotoxicity, hepatotoxicity, and neurotoxicity is presented in Fig. 6. The results for the remaining five ADRs show an almost identical trend.

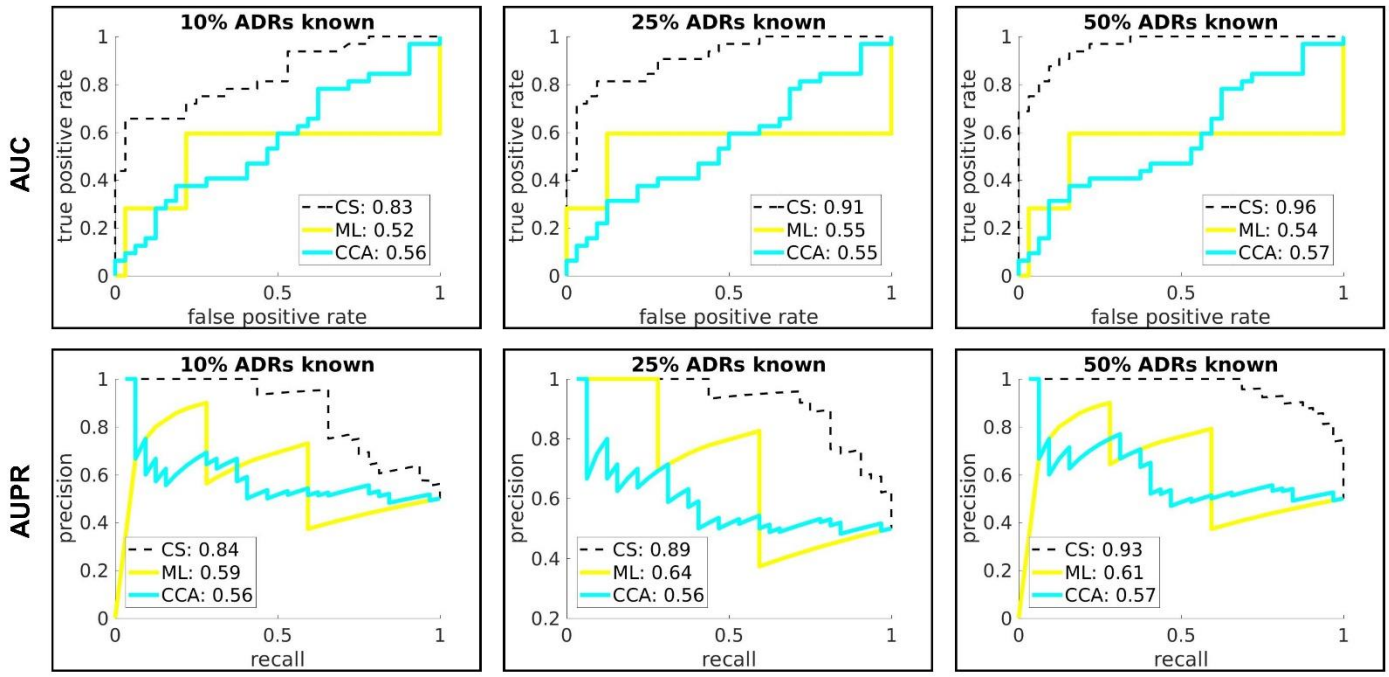



**Fig. 5. Accuracy of hepatotoxicity predictions.** The ROC (top) and PR (bottom) curves are generated based upon the raw scores obtained on "case" and "control" drugs. We performed a number of different tests, each time letting the methods under study (CS, ML, CCA) access different amount of information (10%, 25%, 50%) on other, non-hepatotoxicity ADRs associated to "case" and "control" drugs, thus mimicking methods' accuracy and reliability during clinical trials. We use "balanced" AUPR for better visualization. Unbalanced AUPR scores are easily obtained by multiplying the balanced ones by the fraction of condition positives in the test set.

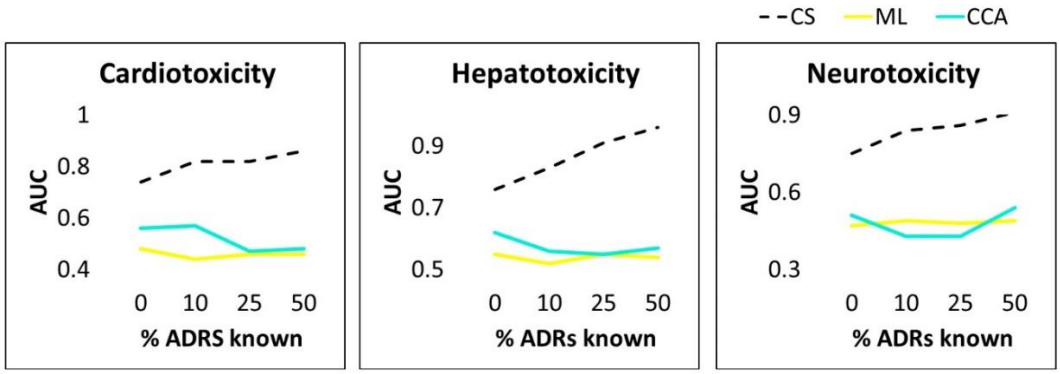



**Fig. 6. Predicting ADRs most responsible for drug market withdrawals.** The x-axis represents the percentage (0%, 10%, 25%, 50%) of already known ADRs for the drug. The y-axis represents the AUC values. The mean AUC values shown in the figures are obtained over multiple runs on "control" drugs to achieve robust statistics (STDEV too small to show).

As seen in Figure 6, compressed sensing has significant advantage over the other methods, even in cases of novel chemicals, with no known ADRs. The performance advantage of CS increases further with the increasing number of ADRs discovered for the drugs. A more comprehensive benchmark, illustrating the power of CS in detecting ADRs for drugs with no known rare ADRs is illustrated in Fig. 7.

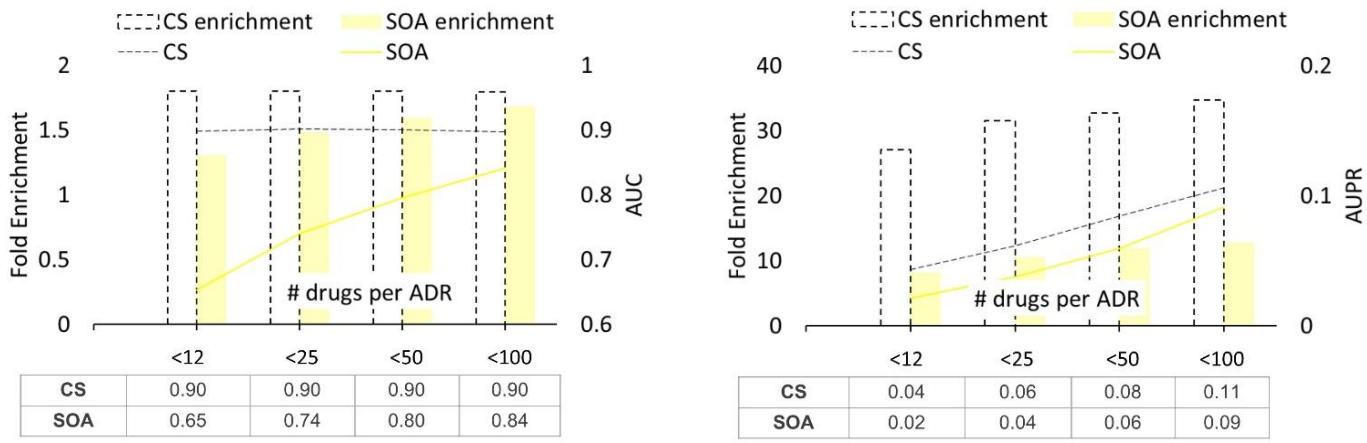



**Fig 7. Value added by compressed sensing (CS) in the AUC (left) and AUPR (right) test for rare ADRs.** The actual AUC and AUPR scores represent the mean values obtained from 5 rounds of the 10-fold cross-validation test on the set of "new" drugs; those with all rare ADRs masked out (hidden). STDEV values are too small to show.

## METHODS

Starting from a known (in practice, noisy and incomplete) binary matrix of drug-ADR associations $R = (r_{i,j})$, the pairwise ADR similarity matrix $M = (m_{i,j})$ and the pairwise drug similarity matrix $N = (n_{i,j})$, our algorithm outputs the "latent" ADR and drug preferences $F = (f_{i,j})$ and $G = (g_{i,j})$ by minimizing the loss function

$$\sum_{i,j} w_{i,j} \left\{ ln\left(1 + e^{f_i g_j^T}\right) - (r_{i,j} + q_{i,j}) f_i g_j^T \right\} + \lambda_r (\|F\|_2^2 + \|G\|_2^2) + \lambda_M tr(F'(D_M - M)F) + \lambda_N tr(G'(D_N - N)G) \quad (1)$$

where $tr$ and $D_M$ are "matrix trace" and "degree", respectively. The lambdas ($\lambda$) are optimizable parameters. The output matrix of drug-ADR associations is computed as $P = \exp(FG') / (1 + \exp(FG'))$ (Fig. 8).

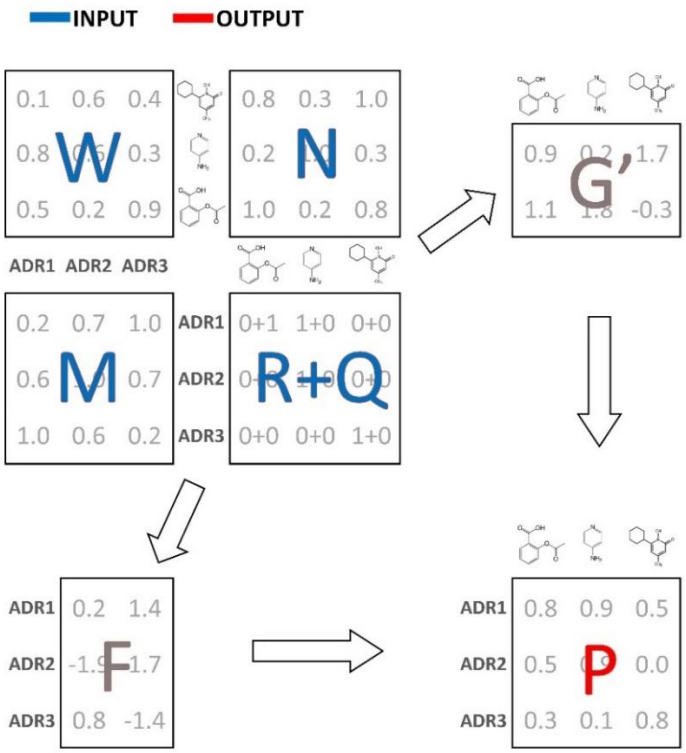



**Fig. 8. Algorithm flow.** R: known drug-ADR associations (sample); M: pairwise ADR similarity matrix; N: pairwise drug similarity matrix; W: drug-ADR frequencies; Q: impute values; F: latent ADR preferences; G: latent drug preferences; P: output drug-ADR probabilities (recovered signal). Algorithm's input and output are shown in blue and red colors, respectively.

The first two terms in (1) drive the "signal recovery" process (matrix completion), whereas the last two terms mandate that similar drugs have similar side-effects and vice versa. Although our method is capable of factoring in the drug-ADR frequency values $w_{i,j}$ and the drug-ADR impute values $q_{i,j}$, this information is currently not been taken advantage of and $w_{i,j}$ and $q_{i,j}$ are set to 1's and 0's, respectively.

The matrices of latent ADR and drug preferences ($F$ and $G$, respectively) are found during the standard minimization procedure. For the sake of brevity, we skip technical details (see [14]), but emphasize that the key idea behind our approach is to demand that $F$ and $G$ are small in one dimension. That way, the output matrix $P$ of predicted interaction probabilities (recovered signal) must be of small rank and, in turn, free of noise. An efficient optimization of the objective function (1) is achieved using a stochastic gradient descent method [15].

While the pairwise drug similarity scores ($N$) are computed using the classical Jaccard index [16], the notion of pairwise ADR similarity scores (along with the notion of frequencies and impute values) is unique to our method

and significantly improves the prediction accuracy. Our pairwise ADR similarity scores are defined as the average of semantic and relatedness measures (*path* and *lesk*, respectively) and are computed by running the *umls-similarity* software [17] on MedDRA vocabulary [18].

While we have originally developed and published the analytical framework (1) for the drug-target interaction problem [14], we subsequently noticed that the compressed sensing is much more amenable to predicting adverse drug reactions (ADRs). The reason is two-fold:

i. In contrast to drug-target interaction problem, where the baseline data is already clean but incomplete, the drug-ADR association data is both, incomplete and noisy. Compressed sensing is particularly suited to deal with erroneous (noisy) data.

ii. In contrast to minor benefits seen when using the protein pairwise similarity matrix $N$ in the context of drug-target interaction prediction problem, utilizing ADR-ADR similarity matrix $N$ (which is unique to our method) seems to significantly improve the accuracy of ADR predictions.

*Future high-quality databases will improve the performance*

The difficulty in identifying ADRs during clinical trials and the complexity of parsing side-effect data from drug package inserts and post-marketing data files result in incomplete and noisy databases of drug-ADR associations. Cleaner and more comprehensive databases represent a straightforward way of improving the performance of prediction methods. For instance, we were able to increase the accuracy of our method in predicting drug-induced liver injury by replacing SIDER hepatotoxicity associations by those stored in LTKB-BD [19]. While SIDER is an automatically generated and comprehensive source, containing almost 140K drug-ADR associations, LTKB-BD represents an expert classification of only 287 drugs with respect to drug-induced liver injury. Even a quick comparison of the two sets of drugs (comparing only drugs that have easily identifiable Pubchem CIDs) reveals significant classification differences between two sources. For instance, 38 out of 137 “most hepatotoxicity” drugs as well as 56 out of 85 “less hepatotoxicity” drugs in LTKB-BD were labeled as “no known hepatotoxicity” in SIDER.

We also believe that much more accurate predictions of drug-ADR associations can be made by utilizing not only cleaner data on drug-ADR associations but also gender-, age- and demographics-specific drug-ADR associations, drug-dose specific associations, and data on side-effects arising from combination drugs [20].

*Incorporating observed drug-ADR frequencies*

Finally, we note that the objective function (1) allows one to take advantage of the frequencies of known drug-ADR associations and the drug-ADR impute values. Each $w_{i,j}$ in the loss function (1) represents the frequency at which the drug $i$ causes the side-effect $j$, while each $q_{i,j}$ can be used to explicitly specify the likelihood of a drug-ADR association. To explain how the impute values can be useful in predicting drug-ADR associations, consider, for instance, an ambiguous case of a newly discovered drug-ADR association that has not yet been recorded in the database ($r_{i,j} = 0$). This new knowledge can be easily incorporated into our method by setting $q_{i,j} = 1$, while adjusting the corresponding weight $w_{i,j}$ to account for any uncertainty in the imputed value.

Our method does not use the weight and impute value functionality, due to unavailability of comprehensive drug-ADR frequency information in the SIDER database. This might change in the future, as more comprehensive databases, containing frequency information, become available.

*A note on computational efficiency.*

Our method's running time is comparable to that of CCA but worse than the running time of ML. A straightforward parallel implementation can make compressed sensing practical, even in large scale studies.

## DISCUSSION

Adverse drug reactions (ADRs) play a major role in drug discovery and human health. Despite significant efforts made over the last decade, the progress in developing computational tools capable of predicting serious side-effects of novel chemicals and market drugs has been dismal at best. No current computational method is able to predict whether a novel and promising compound will eventually cause hepatotoxicity, carcinogenicity, cardiotoxicity, neurotoxicity, immune reaction thrombocytopenia, leukopenia, anaemia or any other harmful and potentially fatal ADRs. Moreover, advances in the area of drug-ADR association prediction are hindered by a

lack of clean and comprehensive databases that store drug-ADR associations and by the difficulty of current methods to deal with noisy and sparse information.

Using the “compressed sensing” framework from the digital signal processing field we developed a computational method that can reliably infer new chemical-ADR associations using existing noisy, biased, and incomplete databases. Originally proposed to address problems in electrical engineering, nowadays “compressed sensing” is a reliable way of recovering any type of signal from a relatively small number of (often erroneous) samples. In the setting of drug-ADR association prediction, the signal can be thought of as the set of all true (known and yet to be discovered) drug-ADR associations. The sample represents known drug-ADR associations, those stored in existing drug-ADR association databases. Because the current databases of drug-ADR associations are sparse and noisy, the drug-ADR association prediction problem is highly amenable to the compressed sensing solution.

Not only is our method able to detect rare ADRs associated with novel chemicals, but also our data demonstrates that the accuracy of compressed sensing in predicting a serious ADR for a candidate drug increases with increasing knowledge of other ADRs associated with the drug. In practice, this means that, as the candidate drug moves up the different stages of clinical trials, the prediction accuracy of our method will increase accordingly.

Compressed sensing represents an important first step in the development of a fully automated and accurate computational method for predicting serious Adverse Drug Reactions (ADRs). Ultimately, accurate and reliable prediction of ADRs will accelerate drug discovery and reduce the risks of drug treatment.

## REFERENCES


[1] Bouvy, J. C., De Bruin, M. L., & Koopmanschap, M. A. (2015). Epidemiology of adverse drug reactions in Europe: a review of recent observational studies. *Drug safety*, *38*(5), 437-453.

[2] Sultana, J., Cutroneo, P., & Trifirò, G. (2013). Clinical and economic burden of adverse drug reactions. Journal of Pharmacology and Pharmacotherapeutics, 4(5), 73.

[3] Budnitz, D. S., Shehab, N., Kegler, S. R., & Richards, C. L. (2007). Medication use leading to emergency department visits for adverse drug events in older adults. Annals of internal medicine, 147(11), 755-765.

[4] Impicciatore, P., Choonara, I., Clarkson, A., Provasi, D., Pandolfini, C., & Bonati, M. (2001). Incidence of adverse drug reactions in paediatric in/out-patients: a systematic review and meta-analysis of prospective studies. British journal of clinical pharmacology, 52(1), 77-83.

[5] Wester, K., Jönsson, A. K., Spigset, O., Druid, H., & Hägg, S. (2008). Incidence of fatal adverse drug reactions: a population based study. British journal of clinical pharmacology, 65(4), 573-579.

[6] Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama, 279(15), 1200-1205.

[7] Donoho, D. L. (2006). Compressed sensing. IEEE Transactions on information theory, 52(4), 1289-1306.

[8] Candès, E. J. (2006, August). Compressive sampling. In Proceedings of the international congress of mathematicians (Vol. 3, pp. 1433-1452).

[9] Candes, E. J., Romberg, J. K., & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. Communications on pure and applied mathematics, 59(8), 1207-1223.

[10] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. Molecular systems biology, 6(1), 343.

[11] Zhang, W., Liu, F., Luo, L., & Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. BMC bioinformatics, 16(1), 365.

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

[13] Onakpoya, I. J., Heneghan, C. J., & Aronson, J. K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC medicine, 14(1), 10.

[14] Lim, H., Gray, P., Xie, L., & Poleksic, A. (2016). Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Scientific Reports*, *6*.

[15] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121-2159.

[16] Rogers DJ, Tanimoto TT. (1960) A Computer Program for Classifying Plants. *Science* 132 (3434): 1115–1118.

[17] McInnes, B. T., Pedersen, T., & Pakhomov, S. V. (2009). UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In AMIA Annual Symposium Proceedings (Vol. 2009, p. 431). American Medical Informatics Association.

[18] Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). Drug safety, 20(2), 109-117.

[19] Chen, M., Vijay, V., Shi, Q., Liu, Z., Fang, H., & Tong, W. (2011). FDA-approved drug labeling for the study of drug-induced liver injury. Drug discovery today, 16(15), 697-703.

[20] Tatonetti, N. P., Patrick, P. Y., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science translational medicine*, *4*(125), 125ra31-125ra31.