Application Note

www.molinf.com

molecular informatics

doi.org/10.1002/minf.202100261

Metabolovigilance: Associating Drug Metabolites with Adverse Drug Reactions

Henry Tan[a] and Scott M. Reed*[a]

Abstract: The Metabolovigilance database (https://pharmacogenomics.clas.ucdenver.edu/pharmacogenomics/side-effect/) is a single repository of information on over 15,920 pharmaceuticals and the compounds expected to result from metabolism of these drugs. Metabolovigilance functions as both a web server, providing data directly to users and as a web application, applying user inputs to create logic statements that curate the data presented or downloaded. Using this tool, it is easy to collect information on drugs, their side effects, and the metabolites associated with specific side effects. Information on these compounds can be sorted based on physical properties of the drugs and their metabolites. All of this information can be viewed, sorted, and downloaded for use in other applications. This open-access tool will facilitate molecular studies on the causes of adverse drug reactions and is well suited to integrate with genomic data furthering the goals of personalized medicine.

Keywords: Metabolomics · pharmacogenomics · adverse drug reactions

The field of pharmacovigilance seeks to minimize the unintended harm drugs cause by collecting and analyzing data on adverse events that occur after drugs are prescribed. The approaching age of personalized medicine suggests a new and challenging job for the scientists engaged in pharmacovigilance as individuals can respond differently to drug therapy. Vast information is now available on the genetic heterogeneity of populations[1] and genetic variability has an influence on how drug efficacy differs between individuals. Many diseases are now recognized as being polygenic or omnigenic[2] in origin and many drugs that were thought to have a known mechanism of action are being discovered to have other pathways by which they function.[3] Furthermore, the field of metabolomics has provided a wealth of information about the chemical changes that occur as a drug is processed by the human body and gut microorganisms. There remains a need to make all this information readily available to scientists engaged in pharmacovigilance, personalized medicine, and drug design. Currently, obtaining the structures of drug metabolites and connecting their structure to specific side effects would require the installation of multiple software packages, the downloading and processing of large sets of data, advanced knowledge of chemistry and of data processing, and lots of time. Here we describe a tool that simplifies and speeds the study of the molecular causes of adverse drug reactions.

The data available through the Metabolovigilance website is stored using an open-source relational database management system and the database will be updated regularly. The initial list of drugs used to generate this database of metabolites was obtained from SIDER version 4.1.[4] SIDER contains information on 1430 drugs, 3000 different side effects, and 139,756 drug side effect pairings. The data in SIDER was obtained from pharmaceutical labels and is organized into side effect categories using the MedDRA dictionary (version 16.1). The web interface for SIDER 4.1 allows retrieving and displaying drug information from a list of side effects or a list of drugs but does not allow for selecting groups of drugs or side effects, or downloading curated information. Importantly, it does not provide information on possible metabolites of these drugs. The SIDER 4.1 dataset contains information on 5880 side effect categories pooled into 3000 entries using synonyms from the Unified Medical Language System (UMLS) Metathesaurus (version 2014AA). This data is obtained using natural language processing techniques applied to drug labels available from the FDA and the pharmaceutical companies and has been updated since its original publication in 2010.

The web application interface to the Metabolovigilance database was built using FAIR principles[5] with open-source software tools and the result is a freely distributable and modifiable tool that could be readily rebuilt around additional or alternate data sources. The application was written in Python with Django providing a Python web framework for making queries to the database. The JavaScript framework, VUE, was used for building the user interfaces. The software is readily modified to include additional data, sorting functions, and display methods. The process of

[a] *H. Tan, S. M. Reed*
Department of Chemistry, University of Colorado Denver
1151 Arapahoe St., Denver, CO 80204, USA
Phone: (303) 315-7634
E-mail: scott.reed@ucdenver.edu

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202100261

assembling the database and the schema underlying the site is described in the supporting information.

The metabolites that are stored in the Metabolovigilance database were generated using the open-access biotransformer tool v.1.1.5 available through the Metabolomic Innovation Centre.[6] Biotransformer combines a mixture of machine learning and knowledge-based approaches to make predictions of drug metabolites. Biotransformer combines predictions of human cytochrome P450 (CYP450)-catalyzed phase I metabolism, human gut microbial metabolism, phase II metabolism, and promiscuous enzymatic metabolism and more reliably predicts metabolites of compounds than commercially available software. These predictions are made based on a database of experimentally confirmed metabolic reactions, a reaction knowledgebase containing generic biotransformation rules, and an algorithm that implements both generic and transformer-specific tools for metabolite prediction and selection. Biotransformer performs well, for example it predicted 90% of experimentally verified CYP450 metabolites from a group of 60 compounds.[6] The precision is measured at 46% indicating that some compounds are generated for which there is no experimental validation. As such, the metabolites obtained should guide searches but not be taken as experimentally validated known structures. This is a reasonable accuracy for initiating a study on the possible molecular explanations for adverse drug reactions but independent verification of the metabolites should be considered.

We generated 14,955 metabolites by *in silico* metabolism prediction from the drugs available in SIDER 4.1. Of the 1430 drugs in the SIDER database, 965 drugs successfully generated metabolites. Others were too small or not prone to producing decomposition products in response to enzymatic functions. When possible, the *superbio* setting was used that includes all human metabolism function as well gut microbiota transformations. In cases where this was not possible, the *allhuman* function was used to identify metabolites. For each metabolite generated the following data was entered into the database; a smile code, as well as the the InChi code, which is a unique chemical identifier,[7] a hashed version of the InChi code, the InChiKey, the metabolite ID number, the reaction type used to produce the metabolite and its corresponding reaction ID number, the enzyme(s) used to produce the metabolite, the biosystem producing the metabolite (human or gut microbiota), and the biotransformer operation used to predict the metabolite (*allhuman* or *superbio*). This was appended with information about the precursor that generated the metabolite including its smile code, InChi code, and InChiKey. Because some metabolites can originate from multiple sources, duplicate entries were combined with an ampersand (*e.g.* Human&Gutmicro indicating both pathways could generate that metabolite). In combined entries, non-duplicative data was retained, for example, if multiple enzymes resulted in the same metabolite, both enzymes are listed in the enzyme field each separated by an ampersand.

Users of Metabolovigilance can select any number of side effects by entering a search term in the box on the top right or scrolling alphabetically through all side effects or a subset of side effects obtained after entering a search term (see supporting information for screen shots of each page). Each side effect that is selected by checking its box is combined into a cumulative search. By clicking submit, a query is generated that returns a list of all drugs that are reported to have any of the selected side effects. This query produces a second page with information on each drug. For each drug, multiple items are provided in a sortable list: the pubchem Chemical ID (CID) number, the stereoisomer ID number and the common drug name, both the UMLS name and concept ID number for the side effect, and the Anatomical Therapeutic Chemical (ATC) code. As multiple sources may provide different frequency information, a lower and upper bound for the frequency of occurrence is provided, if available, and notes in the % occurrence category can include qualitative descriptors (*e.g.* rare) or information on sources (*e.g.* post-marketing). Each of these 9 categories can be used to sort the list. A button labeled "csv" at the top left allows for downloading the list. A search box is provided at the top right and text entered there is matched to all the displayed fields and only drugs matching the text in a field are retained. Any filtering applied using the text box is applied to the displayed data instantaneously and is also reflected in the downloaded csv. Clicking on the name of the drug results in a new page with a list of all the side effects for that drug as found on the FDA website. A text search box is provided on that page to filter through the full list of side effects. Furthermore, a button labeled "csv" is provided that allows downloading of a file that contains CID number, Stereo-Isomer ID, Drug Name, UMLS Concept ID, Frequency (%), Frequency Lower Bound, Frequency Upper Bound, Side Effect, and ATC Code.

By selecting the "View Drugs Ranked" button, the user is sent to a "Drugs Ranked" page that retains and summarizes the results of the search. On this page, each drug name and CID number is listed and can be used to sort the list. This information is followed by the total count of the number of times the drug appears with the selected side effects in the previous page, the number of metabolites, and a link to "View Metabolites" for that individual drug. From the drugs ranked page, clicking on an individual drug ID number, results in download of a 3D structure of the drug from the pubchem API,[8] if available. Clicking on the drug name provides a link to the FDA list of side effects as described above. The search box provided at the top right is matched to all the fields and any filtering is applied to the displayed data instantaneously and is reflected in the downloaded data. A button at the top of the page allows for downloading a csv file that contains all the on-screen information

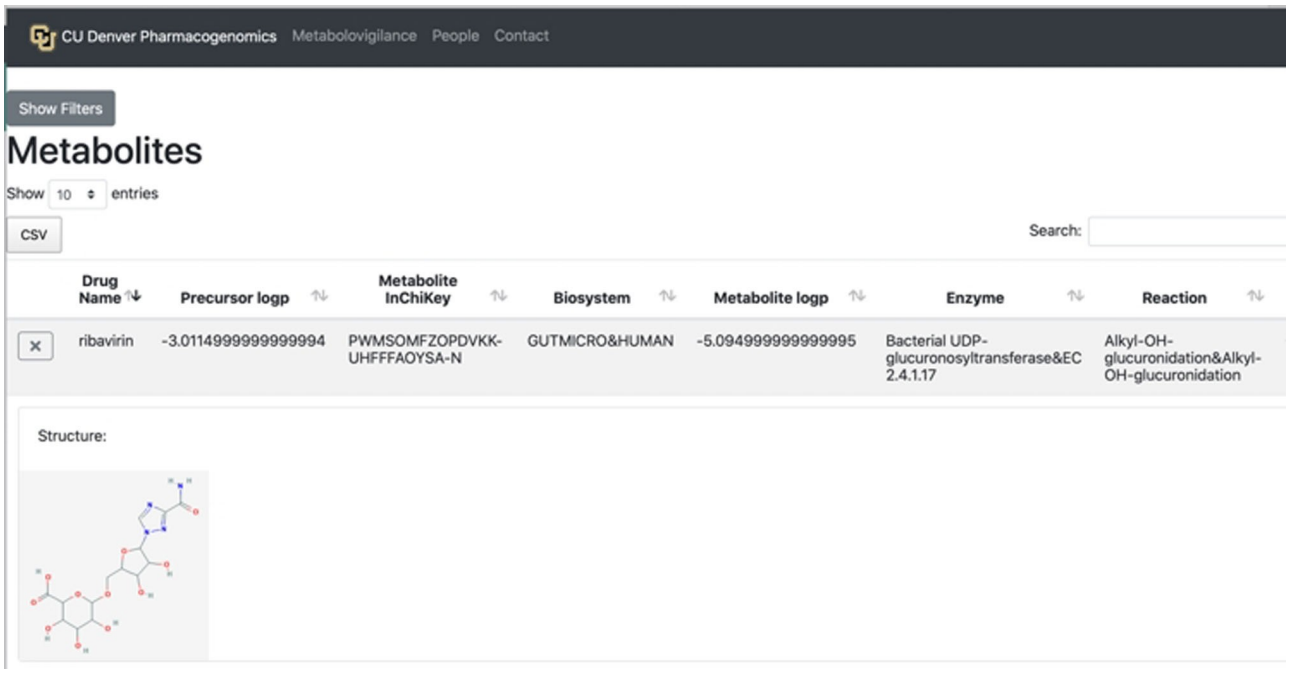

Figure 1. Screenshot of the Metabolovigilance site (https://pharmacogenomics.clas.ucdenver.edu/pharmacogenomics/side-effect/). The "Metabolites" view page showing a metabolite structure and properties of precursor and metabolite.

in a single download. The "View All Metabolites" button will generate all metabolites for all drugs on the list.

Once either the "View Metabolites" button next to a single drug's name is selected or the "View All Metabolites" button is clicked a view is returned containing information about both the metabolites and the precursors from which the metabolites originated (Figure 1). This includes the biosystem (human, gut microbiome, or both) expected to produce the metabolite, the enzymes involved in production of that metabolite, and the specific chemical reaction that converted the precursor into this metabolite. For the precursor, a solubility (logp) calculated using rdkit[9] is displayed. For the metabolites a logp value is also displayed as well as additional information about how that metabolite is expected to be produced *in vivo*.

A structure button is provided below each metabolite that produces an image of the metabolite structure using the NIH PUG REST system.[8] Because the metabolites are generated recursively, the precursor for a given metabolite may exist not only as a drug identified on the previous selection but can also be a later metabolite of that drug. Finally, for identification purposes, since these metabolites are not certain to have a designated CID number, a metabolite InChiKey and smile code are provided.

A "Show Filter" button at the top of the page for either the single or multiple metabolite view page reveals numerical range filters that allow curating the dataset prior to download. The filters provided are an upper and lower bound for the logp value for the metabolites as well as a lower and upper bound for the logp value for the precursor. A user may select for specific metabolites based on solubility, an expected target organ affinity, or a rate of secretion. In this way a user can generate a specific set of compounds that are likely to fit the criteria needed to explain a given side effect. A button at the top of the page allows for downloading a csv file that contains all the tabular data in a single download for all the metabolites on the page.

Metabolovigilance was built using FAIR principles[5] which are designed to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The code used to build the website is freely available on GitHub at https://github.com/scottmreed/CU-Denver-Pharmacogenomics-Website. No limits are placed on the reuse of material from this site or the reformulation of the application to include additional data. Compared to existing methods, this site makes the data substantially more accessible. While some drug information can be found on the SIDER website[4] the Metabolovigilance interface provides that information in a format that is much easier to access and sort and also provides metabolite information that is not available on the SIDER site. While the biotransformer software is freely available for download, the process of installing and running the software is cumbersome for a casual user and processing large numbers of compounds can take days. Biotransformer is also available as a web application for individual transformations of single compounds but each submitted compound is processed as submitted. In contrast, with Metabolovigilance, all the information on the metabolites has been prepared ahead of time, removing a

substantial bottleneck, and reducing the time and effort needed to obtain the information.

Of the 14,955 metabolites in Metabolovigilance, 7131 (48%) were generated by the *superbio* process of biotransformer, and 7824 (52%) originated from the human process. Of all the metabolites created, 12,698 were exclusively produced by a *allhuman* function, 1177 were exclusively produced by gut microbiota, and 1080 were formed by both human and gut microbiota functions. The logp values for the drug precursors (mean of 2.090) are slightly higher than the metabolite logp values (mean of 1.396) (see distribution in supporting information). Metabolovigilance is a tool that facilitates and speeds research on the molecular causes of drug side effects. The molecular information obtained from this tool is a starting point for experimental and computational studies and does not provide direct evidence for identification of specific molecules.

Table 1 illustrates the type of information that can be obtained from Metabolovigilance. Examining the 10 drugs with the highest number of side effects, we assembled a list of all the metabolites for each of these drugs (details in supporting info). Averaging the logp values and comparing them to the drug logp value reveals how different drugs are prone to different types of enzymatic reactions which have different effects on the properties of the metabolites. We expect users may also analyze patterns in side effects and molecular transformations that may have led to those side effects when designing new drugs for a target.

Table 1. List of top 10 drugs as ranked by number of side effects and properties of those drug's calculated metabolites.

Drug Name	Drug logp	Number Metabolites	Mean logp of Metabolites
tramadol	2.635	15	1.816
pimecrolimus	5.719	29	4.814
paroxetine	3.327	11	2.504
risperidone	3.590	17	2.817
tacrolimus	4.639	30	3.799
quetiapine	2.856	19	2.191
imatinib	4.590	21	3.554
docetaxel	3.260	26	2.279
topiramate	−0.395	0	N/A
citalopram	3.813	87	1.641

It has long been understood that genetics can influence the effect of and the metabolism of pharmaceuticals.[10] Recent advances in large scale studies of whole genomes[1] have made it possible to identify large numbers of genomic variants across nearly every gene opening new avenues to explore pharmacogenomic interactions. Recently over 98% of the human proteome was modeled by a neural network driven homology modeling program called AlphaFold.[11] Furthermore, the majority of known mutations in the human exome have been mapped to three-dimensional structures of proteins.[12] The combination of these advances makes feasible the mapping of low frequency adverse drug reactions to uncommon genetic variations. Efforts have been made previously to connect structural features of drugs to side effects[13,14] although the scope was modest and metabolites were not considered. The Metabolovigilance site will facilitate efforts to systematically connect genomic variation to adverse drug reactions.

Future additions to this project could include simulations of spectral data[7,15] for metabolites to allow for matching unknown spectra to possible chemical pathway origins. Additional plans include adding search tools for direct access to metabolites. Furthermore, work is in progress for connecting data from Metabolovigilance to genomic information with the goal of finding genetic and molecular explanations for rare event side effects.

Acknowledgements

HT acknowledges funding from the CU Denver Education Through Undergraduate Research and Creative Activities program. SMR thanks Woonghee Lee for a critical reading of the manuscript.

Conflict of Interest

None declared.

Data Availability Statement

This software is freely available for use and the underlying code is located in a publicly available repository at: https://github.com/scottmreed/CU-Denver-Pharmacogenomics-Website.

References

[1] M. Lek et al., *Nature* **2016**, *536*, 285–291.
[2] X. Liu, Y. I. Li, J. K. Pritchard, *Cell* **2019**, *177*, 1022–1034.
[3] A. Lin et al., *Sci. Transl. Med.* **2019**, *11*, 15451.
[4] M. Kuhn, I. Letunic, L. J. Jensen, P. Bork, *Nucleic Acids Res.* **2016**, *44*, D1075–D1079.
[5] M. D. Wilkinson et al., *Sci. Data* **2016**, *3*, 160018.
[6] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D. S. Wishart, *J. Cheminf.* **2019**, *11*, 2.
[7] a) H. Dashti, H. W. Westler, J. L. Markley, H. R. Eghbalnia, *Sci. Data* **2017**, *4*, 170073; b) H. Dashti, J. R. Wedell, W. H. Westler, J. L. Markley, H. R. Eghbalnia, *Sci. Data* **2019**, *6*, 190023.
[8] S. Kim, J. Chen, T. Cheng et al., *Nucleic Acids Res.* **2021**, *49*(D1), D1388–D1395.
[9] G. Landrum (2021) RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org. Accessed July 25, 2021.

[10] O. Osanlou, M. Pirmohamed, A. K. Daly, *Adv. Pharmacol.* **2018**, *83*, 155–190.
[11] K. Tunyasuvunakool, J. Adler, Z. Wu et al., *Nature* **2021**, *596*, 590–596.
[12] T. Khanna, G. Hanna, M. J. E. Sternberg, A. David, *Hum. Genet.* **2021**, *140*, 805–812.
[13] J. Scheiber, J. L. Jenkins, S. C. Sukuru, A. Bender, D. Mikhailov, M. Milik, K. Azzaoui, S. Whitebread, J. Hamon, L. Urban, M. Glick, J. W. Davies, *J. Med. Chem.* **2009**, *52*, 3103–3107.
[14] M. Kozyra, M. Ingelman-Sundberg, V. M. Lauschke, *Genet. Med.* **2017**, *19*, 20–29.
[15] M. Pupier, J. M. Nuzillard, J. Wist, N. E. Schlorer, S. Kuhn, M. Erdelyi, C. Steinbeck, A. J. Williams, C. Butts, T. D. W. Claridge, B. Mikhova, W. Robien, H. Dashti, H. R. Eghbalnia, C. Fares, C. Adam, P. Kessler, F. Moriaud, M. Elyashberg, D. Argyropoulos, M. Perez, P. Giraudeau, R. R. Gil, P. Trevorrow, D. Jeannerat, *Magn. Reson. Chem.* **2018**, *56*, 703–715.

Received: October 2, 2021
Accepted: January 3, 2022
Published online on January 22, 2022

