JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

In Silico Enzymatic Synthesis of a 400 000 Compound Biochemical Database for Nontargeted Metabolomics

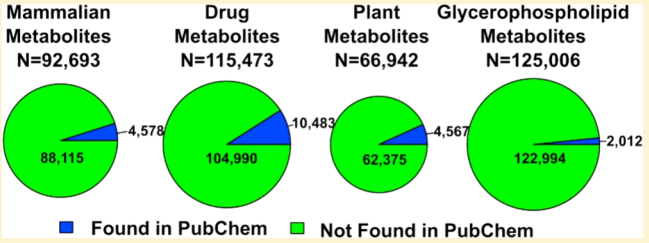
Lochana C. Menikarachchi,[†] Dennis W. Hill,[†] Mai A. Hamdalla,[‡] Ion I. Mandoiu,[‡] and David F. Grant*[,†]

[†]Department of Pharmaceutical Sciences, University of Connecticut, 69 North Eagleville Road, Storrs, Connecticut 06269, United States

[‡]Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Road, Unit 4155, Storrs, Connecticut 06269, United States

Supporting Information

ABSTRACT: Current methods of structure identification in mass-spectrometry-based nontargeted metabolomics rely on matching experimentally determined features of an unknown compound to those of candidate compounds contained in biochemical databases. A major limitation of this approach is the relatively small number of compounds currently included in these databases. If the correct structure is not present in a database, it cannot be identified, and if it cannot be identified, it cannot be included in a database. Thus, there is an urgent need to augment metabolomics databases with rationally designed biochemical structures using alternative means. Here we present the In Vivo/In Silico Metabolites Database (IIMDB), a database of in silico enzymatically synthesized metabolites, to partially address this problem. The database, which is available at http://metabolomics.pharm.uconn.edu/iimdb/, includes ~23 000 known compounds (mammalian metabolites, drugs, secondary plant metabolites, and glycerophospholipids) collected from existing biochemical databases plus more than 400 000 computationally generated human phase-I and phase-II metabolites of these known compounds. IIMDB features a user-friendly web interface and a programmer-friendly RESTful web service. Ninety-five percent of the computationally generated metabolites in IIMDB were not found in any existing database. However, 21 640 were identical to compounds already listed in PubChem, HMDB, KEGG, or HumanCyc. Furthermore, the vast majority of these in silico metabolites were scored as biological using BioSM, a software program that identifies biochemical structures in chemical structure space. These results suggest that in silico biochemical synthesis represents a viable approach for significantly augmenting biochemical databases for nontargeted metabolomics applications.



INTRODUCTION

Most nontargeted metabolomic studies use biochemical databases for structure determination.[1−5] These studies typically report the identification of fewer than 10% of the total number of compounds detected.[2−4] This consistently low percentage suggests that current biochemical databases do not contain most of the endogenous compounds routinely detected in biological samples. Although it is not known how many compounds exist in the human metabolome, one study has suggested that there are over 200 000 lipids alone.[6] This number, however, is likely an underestimate considering that there are over 8×10^6 microbial genes in the human microbiome.[7] Thus, with fewer than 70 000 compounds in current biochemical databases, there is an obvious paradox that severely limits the utility of nontargeted metabolomics research: if a structure is not included in a database it cannot be identified, and if it cannot be identified, it cannot be included in a database. Discovering, purifying, and identifying new biochemical compounds using classical analytical methods is a time-consuming, expensive, and laborious process, especially using human samples. An alternative approach is to supplement current databases with anticipated compounds (compounds likely to be found in humans but not yet identified). These anticipated compounds may include compounds consumed by humans, compounds to which humans are frequently exposed, and compounds that may be produced by biochemical pathways (human and microbial) in the human body. Many metabolomics databases have recognized the importance of including expected metabolites. Examples include computationally generated di- and tripeptide structures in the Metlin database,[8,9] computationally generated lipid structures in Lipid Maps,[10,11] and expected compounds (e.g., foods, food additives, environmental pollutants, etc.) in HMDB.[12,13] Inclusion of anticipated metabolites is especially important for mass-spectrometry-based metabolomics, where structure identification relies on all putative compounds being present in the database.

One approach for producing anticipated metabolites (unknown unknowns) would be to use in silico enzymatic synthesis. For this approach to be successful, the selected in silico enzymes would ideally have broad substrate specificity, as

Received: June 26, 2013
Published: August 30, 2013

dx.doi.org/10.1021/ci400368v | J. Chem. Inf. Model. 2013, 53, 2483−2492

these would potentially catalyze the metabolism of a variety of substrates (known and unknown) to produce novel products. Indeed, it has been suggested that broad enzyme specificity and/or side reactions might explain our incomplete knowledge of the human metabolome.[14] It is well-known that phase-I and phase-II enzymes are nonselective and are found in nearly all cells and tissues, having evolved from promiscuous ancestral enzymes.[15−17] Since these enzymes typically metabolize multiple drugs to give a variety of metabolites, we reasoned they might also metabolize multiple endogenous compounds to produce a variety of previously unknown products. Indeed, other investigators have identified multiple phase-I and phase-II metabolites of endogenous biochemicals in mammalian serum.[18]

Previous studies using in silico enzymatically synthesized databases include the University of Minnesota Biocatalysis/Biodegradation Database[19] (UM-BBD), the enzyme-catalyzed metabolic pathway predictor PathPred,[20] and the evidence-based metabolome library MyCompoundID.[21] The UM-BBD system uses a collection of microbial biodegradation pathways to predict one or more reaction steps. The PathPred server uses KEGG biochemical structure transformation patterns called RDM patterns.[22] This system focuses on predicting pathways for microbial biodegradation (based on 947 RDM patterns) and plant secondary metabolite biosynthesis (based on 1397 RDM patterns). The MyCompoundID database uses 76 literature-derived common metabolic transformations in the form of accurate mass transformations to identify unknown metabolites. This database includes 8021 known human endogenous compounds and their predicted metabolic products using one (375 809 metabolites) or two (10 583 901 metabolites) reaction steps.

Here we present the In Vivo/In Silico Metabolites Database (IIMDB), an easily searchable database comprising a non-redundant set of known biochemical "parents" collected from existing databases and their in silico phase-I and phase-II metabolites. The known parent compounds were obtained from HMDB,[12,13] KEGG,[23] HumanCyc,[24] PlantCyc,[25] Phenol Explorer,[26] Lipid Maps,[10,11] Drug Bank,[27] and the 1989 USAN and the USP Dictionary of Drug Names.[28] In silico metabolites were generated using phase-I and phase-II human biotransformation rules as implemented in the program Meteor 14.[29,30] Interestingly, more than 21 000 of these in silico-generated metabolites are already included in current databases, suggesting that this general approach is reasonable. The database features a user-friendly web interface and a programmer-friendly RESTful web service.

METHODS

Parent Data Sets. *Mammalian Compounds.* A data set of mammalian compounds was compiled by combining selected chemical structures in KEGG, HMDB, and HumanCyc. Compounds containing any element other than C, H, N, O, P, and S were eliminated. The data set was further limited to the 50−1000 Da molecular weight range. KEGG data were downloaded on April 23, 2011. Compounds belonging to at least one of the 63 known KEGG mammalian pathways were selected.[31] Data from HMDB (version 2.5) were downloaded on July 15, 2012. Data from HumanCyc (version 16.0) were downloaded on May 24, 2012. Duplicate compounds were eliminated by comparing the unique SMILES representations of the chemical structures. The final mammalian data set contained 1579 KEGG compounds, 5267 HMDB compounds, and 262 HumanCyc compounds.

Plant Metabolites. A data set of plant metabolites was compiled by combining plant metabolites from the KEGG database with polyphenols found in the Phenol Explorer database (downloaded on Oct 22, 2012). The plant data set was curated similarly to the mammalian data set. Any compound already contained in the mammalian data set was eliminated. The final data set contained 2765 KEGG compounds and 190 polyphenols.

Drugs. A data set of drugs was compiled by combining approved, illicit, and withdrawn drugs in Drug Bank 3.0 (downloaded on Jan 18, 2012) and drugs listed in the 1989 USAN and the USP Dictionary of Drug Names. Polymers, mixtures, single-element drugs (e.g., Fe), and compounds already contained in the mammalian and plant data sets were eliminated. The final data set contained 1412 compounds from Drug Bank and 4646 compounds from the 1989 USAN and the USP Dictionary of Drug Names.

Glycerophospholipids. Glycerophospholipid compounds were downloaded from the Lipid Maps database on April 23, 2012. The data set was curated similarly to the mammalian and plant data sets. Compounds already contained in the other data sets were eliminated. The final data set contained 6914 glycerophospholipids.

Structure Generation. In silico metabolites of parent compounds were generated with Meteor 14 (knowledge base version 14.0.0_09_02_2012) from Lhasa Ltd. Meteor is a knowledge-based expert system for predicting likely metabolites of a query chemical structure.[29,30,32] The Meteor system consists of a knowledge base of phase-I and phase-II biotransformation rules and a reasoning engine to determine the most likely metabolites. The list of phase-I and phase-II biotransformation types included in Meteor is given in Table S1 in the Supporting Information. The Meteor reasoning engine uses two types of rules, absolute and relative, to determine the more likely metabolites out of many possibilities.[32−34] Absolute reasoning rules include five levels of uncertainty, listed from most likely to least likely as "probable", "plausible," "equivocal," "doubted," and "improbable." Relative reasoning rules are used to determine the more likely reaction out of two competing biotransformations. The Meteor processing constraints listed in Table 1 were chosen to strike a balance between likelihood of occurrence and combinatorial explosion[29,35,36] of the results.

Table 1. Meteor Processing Constraints Used in Structure Generation

processing constraint	value
absolute reasoning level	plausible
relative reasoning level	top levels (2)
maximum number of steps in a pathway	4
species	human
phase option	do not grow from phase-II products
maximum total number of metabolites	100

Database Implementation and Access. OrientDB (version 1.3.0)[37] from Orient Technologies was used for the construction of the database. OrientDB is an open-source Java-based database management system (DBMS) with the features of both document and graph DBMSs. All of the chemical structures and associated data fields are stored in a single cluster (similar to a table in a relational database) named "Unique-

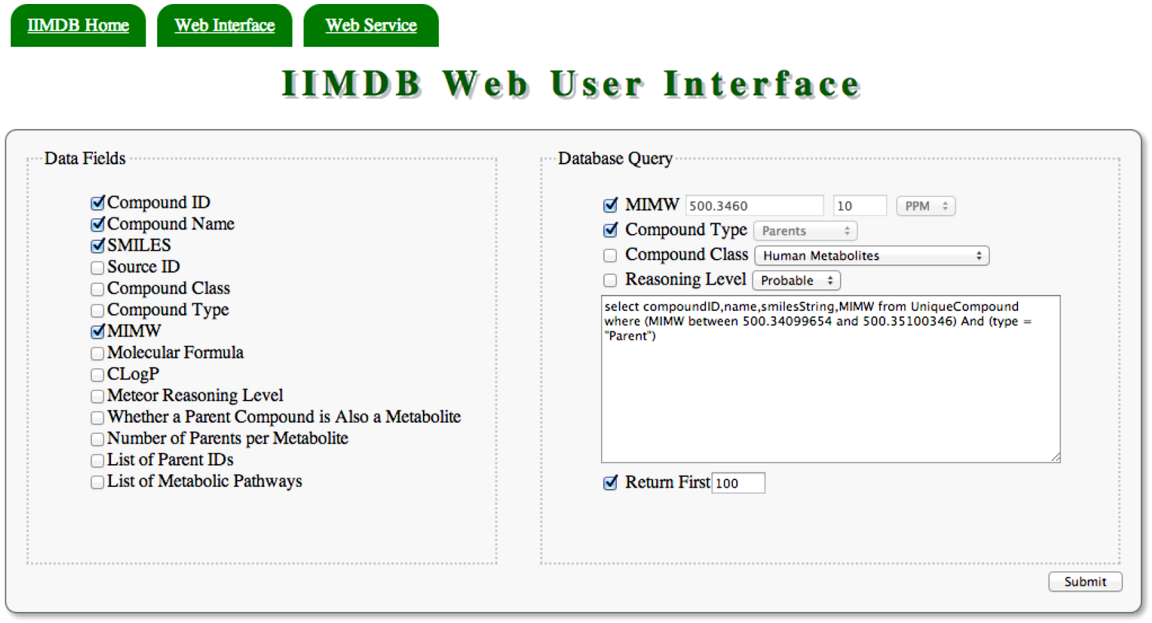


Figure 1. IIMDB web user interface.

Compound". The database is hosted on a Linux-based server running openSUSE 12.1. The database server is equipped with a 3.4 GHz Intel core i7 processor and 12 GB of RAM. IIMDB is available at http://metabolomics.pharm.uconn.edu/iimdb/. Access to IIMDB is provided via a password-protected web interface (Figure 1) and a web service. The Meteor-generated structures in the database are not freely available because of licensing restrictions (clause 6.1) in the Meteor licensing agreement. The end user is required to have a valid licensed copy of Meteor (purchasable from http://www.lhasalimited.org) to access IIMDB. However, all of the parent compounds used in this work are freely available in an Excel spreadsheet (xlsx format) provided in the Supporting Information. The web interface was built using HTML5, JavaScript, and JQuery. This web interface will operate on most HTML5-compatible web browsers such as Mozilla Firefox (recommended), Google Chrome, and Internet Explorer 9 or later. JavaScript must be enabled in the user's web browser.

Access to all data fields and most commonly used queries is provided through the web user interface. The actual database querying is done using OrientDB's own SQL-like query language. The end user's interaction with the checkboxes and drop-down menus on the interface is converted into an SQL expression and shown in the large text area to the right. This web interface can also be used as a tool to learn the underlying query language. The predefined queries generated with the web interface can be modified or extended by manually editing the text area. The query results are shown on a paginated data table. The data table includes options for sorting, full-text searching, and exporting data to CSV and PDF file formats. A step-by-step guide to viewing and converting structures is included in the Supporting Information. The programmatic access to IIMDB is provided via a RESTful web service. IIMDB allows read-only access to database records via OrientDB's built-in web service. The query URI has the following general format: http://metabolomics.pharm.uconn.edu/iimdb/query/iimdb/sql/"**SQL COMMAND**". For example, the command-line input for listing compound IDs, monoisotopic molecular weights, and SMILES strings of compounds that have monoisotopic molecular weights between 500.2450 and 500.4580 is:

```
http://metabolomics.pharm.uconn.edu/iimdb/query/iimdb/sql/select compoundID,MIMW,smilesString from UniqueCompound where MIMW between 500.2450 and 500.4580
```

ALogP Calculations. Three random samples of parent compounds (each containing 100 compounds) per data set were drawn from the mammalian, plant, drug, and glycerophospholipid data sets using the Knuth shuffle algorithm.[38] Duplicate structures in data sets were removed after combining random samples. The combined random samples comprised 298 mammalian parent compounds, 294 parent drugs, 288 parent plant compounds, and 295 parent glycerophospholipid compounds. AlogP values of parent compounds and their associated in silico metabolites in random samples were calculated using the web-based ALOGPS 2.1 algorithm.[39]

PubChem Search. All of the in silico-generated chemical structures were searched on the National Center for Biotechnology Information (NCBI) PubChem[40] database (the largest freely accessible compound database). The PubChem database searches were done between Oct 1, 2012, and Oct 31, 2012. The structure search was done with an in-house program that was built around PubChem's power user gateway (PUG). Any PubChem compound that had the same connectivity as the query compound or a tautomer of the query compound was considered a match.

BioSM Predictions. Biological Structure Matcher (BioSM) is a computational tool that uses graph matching and known mammalian metabolite structures to identify the biological likeness of a given structure.[31] Previous studies have shown that BioSM identifies endogenous metabolites with high accuracy. The BioSM algorithm was used to identify biological molecules in parent and in silico-generated structures. Since BioSM was trained to predict the biological likeness of chemical structures with molecular weights of 50−700, only this range was considered.

dx.doi.org/10.1021/ci400368v | J. Chem. Inf. Model. 2013, 53, 2483−2492

Table 2. Numbers of in Silico Metabolites for Different Classes of Parent Compounds

compound class	no. of parent compounds	no. of in silico metabolites	fold increase in database size	probable	plausible
drugs	6058	115 473	19	29%	71%
plant compounds	2955	66 942	23	36%	64%
mammalian compounds	7108	92 693	13	44%	56%
glycerophospholipids	6914	125 006	18	51%	49%

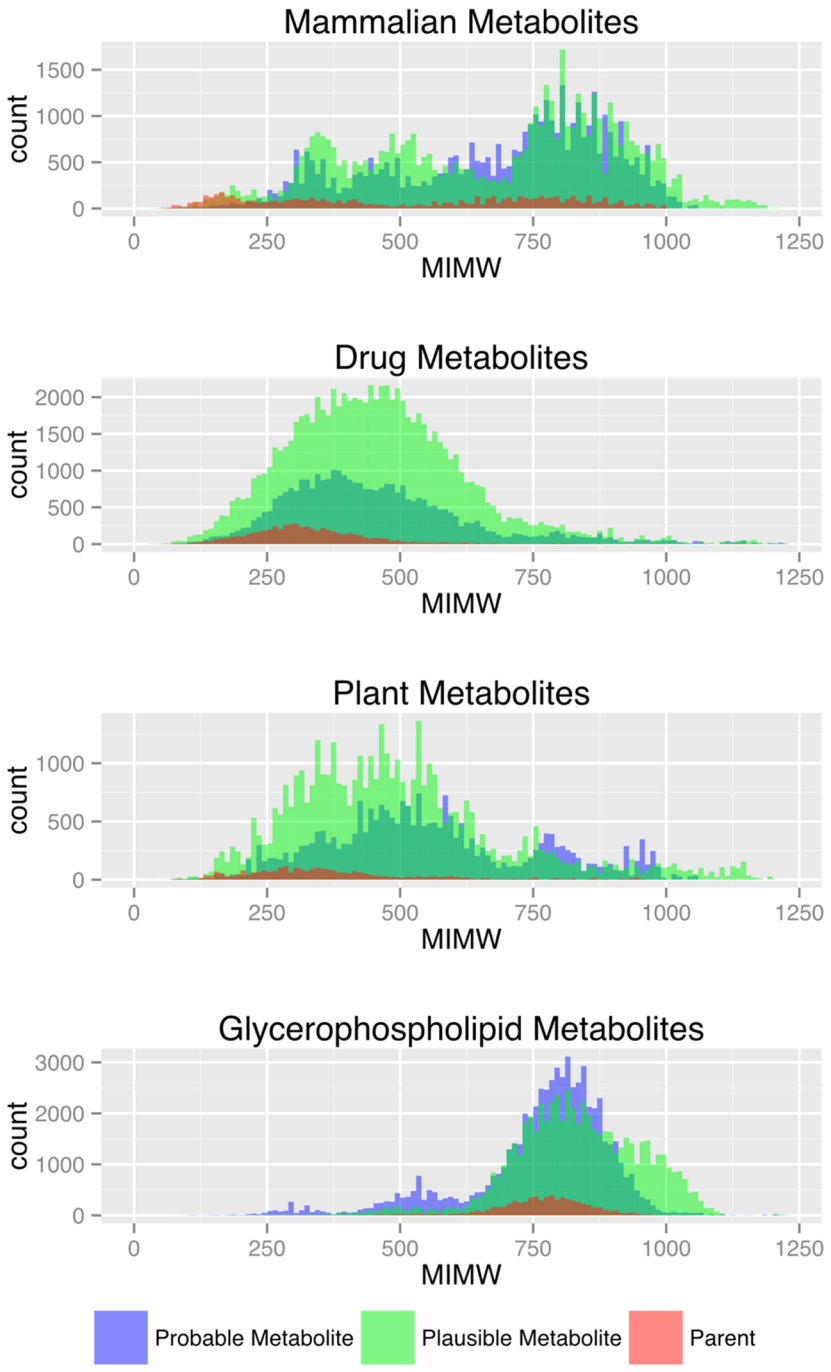


Figure 2. MIMWs for mammalian, drug, plant, and glycerophospholipid compounds. The histograms were generated with a bin size of 10 Da. Color blending has been used to illustrate the MIMW bins shared by different types of compounds. For example, dark green represents MIMW bins common to probable and plausible metabolites.

RESULTS AND DISCUSSION

Table 2 lists the number of Meteor-generated metabolites for each of the four different classes of parent compounds. The fourth column in Table 2 lists the fold increase in the number of database compounds for each class of metabolite (i.e., number of in silico compounds/number of parent compounds). Plant compounds produced the largest number of unique metabolites per parent compound, whereas mammalian

dx.doi.org/10.1021/ci400368v | J. Chem. Inf. Model. 2013, 53, 2483−2492

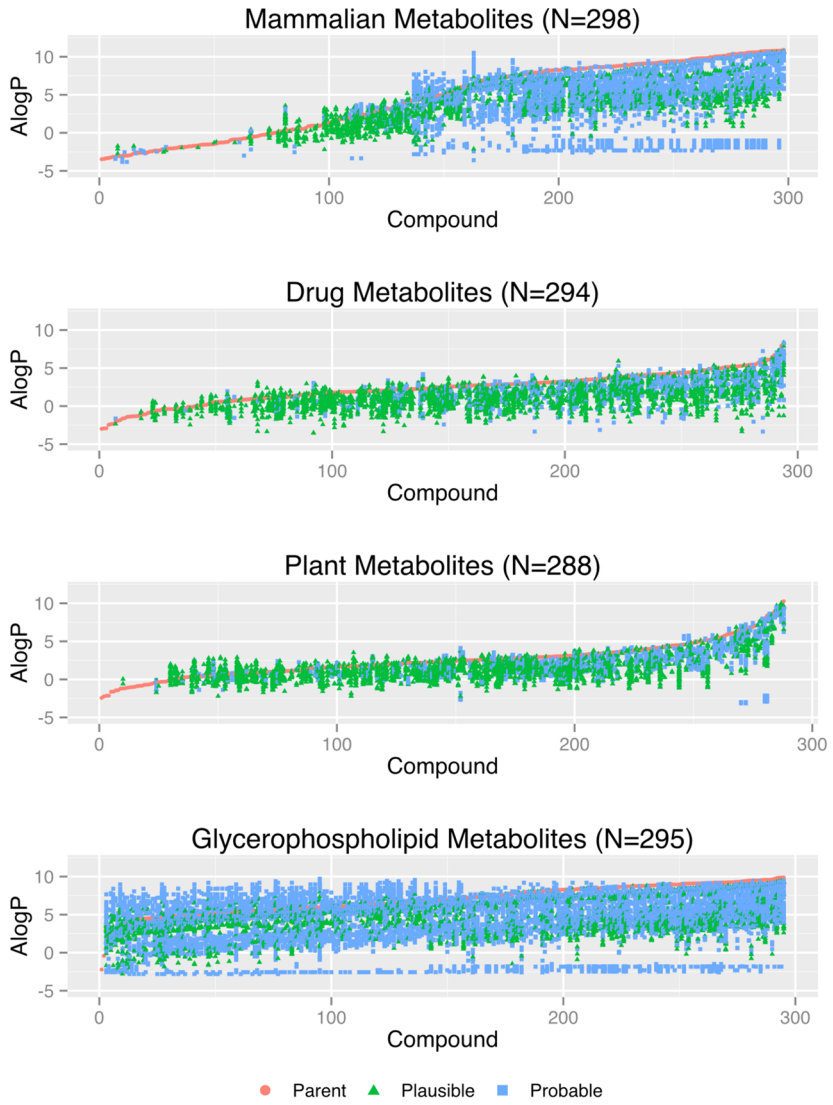


Figure 3. AlogP values for random samples. Each orange data point represents a parent compound. In silico metabolites of each parent compound are shown either below (more polar) or above (less polar) the parent that produced it.

compounds produced the fewest. On average, 18 metabolite structures were generated for each parent compound using the Meteor processing constraints given in Table 1.

Figure 2 shows the monoisotopic molecular weight (MIMW) distributions for mammalian, drug, plant, and glycerophospholipid parents and metabolites. Each individual plot in Figure 2 depicts the MIMWs of parents and probable and plausible metabolites as overlapping histograms. The MIMWs of mammalian parents were spread over a range of approximately 54−999 Da with a mean MIMW of 537 Da. The MIMWs of in silico mammalian metabolites span a range of approximately 31−1201 Da with a mean MIMW of 675 Da. Thus, phase-I and phase-II metabolic transformations resulted in an expansion of the MIMW range by −23 Da at the lower end and +202 Da at the upper end. The average MIMW of mammalian compounds was increased by approximately 138 Da upon metabolism. Similarly, the MIMW range of drugs was increased by approximately +32 Da with an average MIMW gain of 111 Da. The MIMW range of plant compounds was increased by −37 and +202 Da with an average gain of 142 Da. The glycerophospholipid metabolites showed the largest increase at the lower end of the MIMW range with an increase of approximately −225 Da. The upper end of the glycerophospholipids MIMW range showed a negative shift of 132 Da (i.e., in silico metabolism of higher-molecular-weight parents resulted in smaller metabolites), but on average, the MIMWs of glycerophospholipids were increased by 34 Da.

Figure 3 shows AlogP values for parents and metabolites in four random samples of approximately 300 compounds collected from the mammalian, drug, plant, and glycerophospholipid data sets. In most cases, the computationally generated metabolites were more polar (lower ALogP) than the parent compounds. These results are consistent with the established dogma that phase-I and especially phase-II biotransformation reactions produce metabolites with increased polarity and thus are more easily eliminated. However, in some cases less polar (higher ALogP) compounds were also observed. Kirchmair et al.[41] reported a similar observation in a recent study. They found an increase in computational logP values of 4−9% for phase-I and 8−13% for phase-II metabolic transformation products. These authors also suggested that metabolic reactions leading to more lipophilic molecules might be related to metabolism in skin, where an increase in logP allows metabolites to stay attached to lipids and be excreted through

dx.doi.org/10.1021/ci400368v | J. Chem. Inf. Model. 2013, 53, 2483−2492

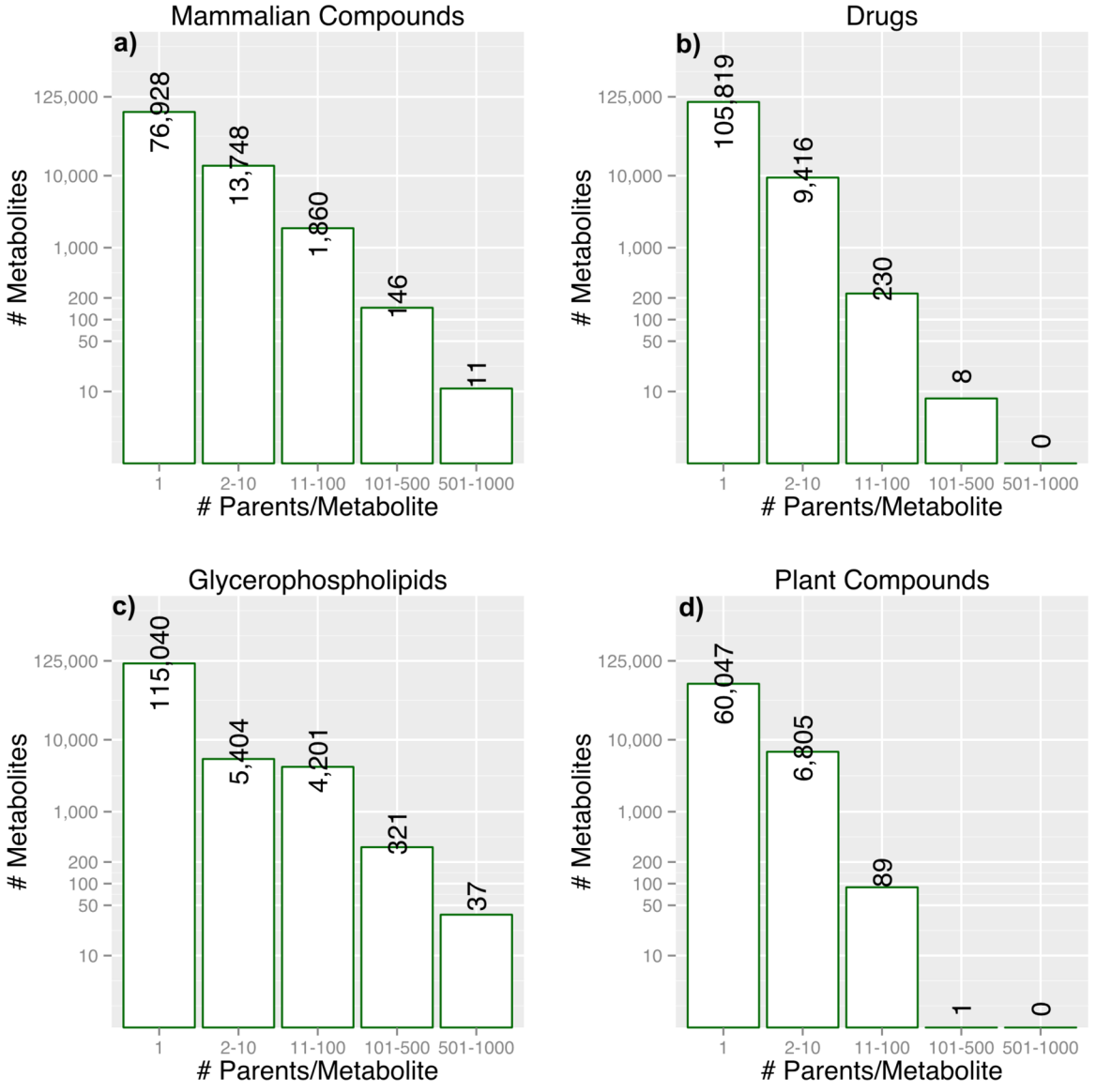


Figure 4. In silico metabolites produced by multiple parents.

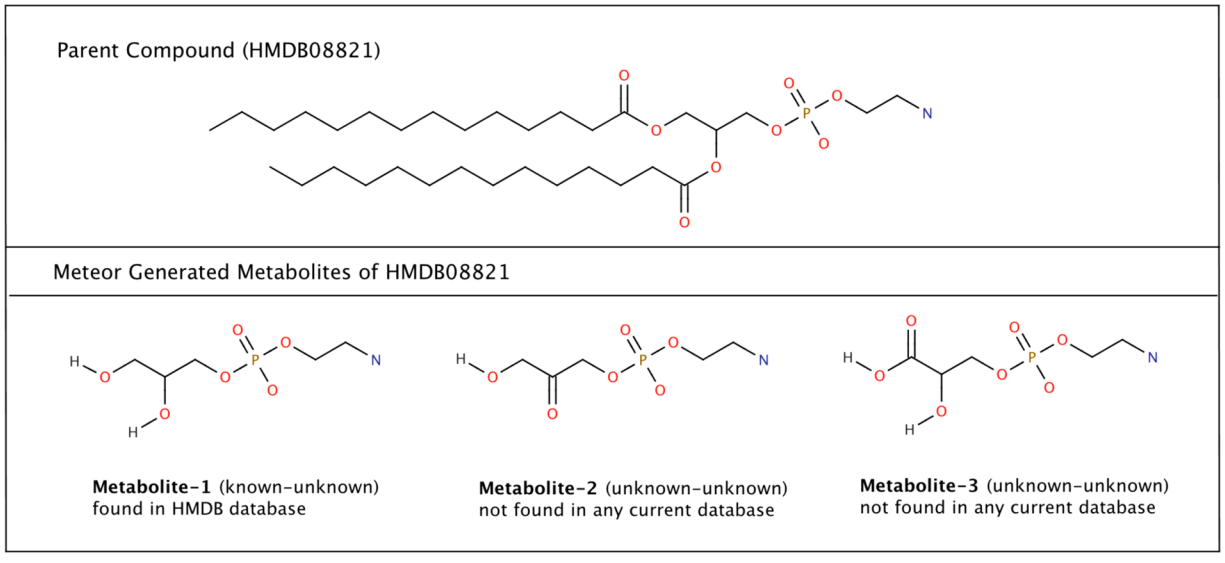


Figure 5. Three Meteor-generated metabolites of HMDB08821 (Metabolite-1 matched HMDB compound HMDB59660).

dx.doi.org/10.1021/ci400368v | J. Chem. Inf. Model. 2013, 53, 2483−2492

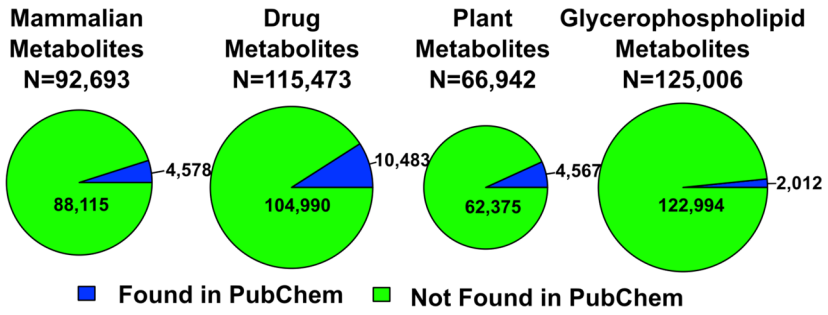


Figure 6. Meteor-generated metabolites found in PubChem. Any PubChem compound that had the same connectivity as the query compound or was a tautomer of the query compound was considered a match.

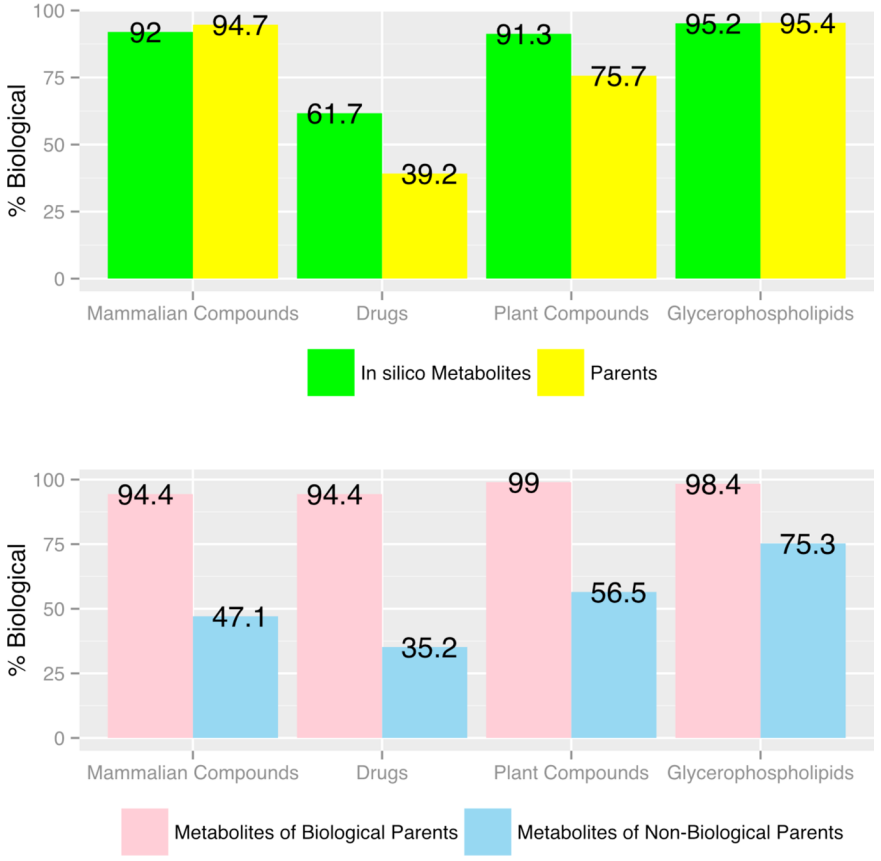


Figure 7. Biological structure matching with BioSM. Biological parents are parent structures predicted to be biological by BioSM.

desquamation of skin cells. A closer inspection of the chemical structures in Figure 3 reveals that the large majority of in silico metabolites with increased logP are either lipids or lipid-related molecules. In IIMDB, the calculated AlogP values can be used to restrict search results. This is especially useful when searching for metabolites in a certain ALogP range (e.g., more polar metabolites that might be found in urine).

The data in Figure 3 suggest that multiple mammalian and glycerophospholipid parent compounds were metabolized to form the same final product, since multiple in silico metabolites appear to have the same or very similar AlogP values. Indeed, a closer examination of the four data sets revealed that these metabolites were in fact identical but were produced from different parents. Figure 4 shows histograms presenting the numbers of in silico metabolites produced from various numbers of parents. For example, 11 in silico metabolites (last bin in Figure 4a) were produced from 501 to 1000 different mammalian parent structures. Interestingly, all 11 of these in silico metabolites were found in PubChem (accessed between Oct 1, 2012, and Oct 31, 2012); 72% of them were also found in HMDB 2.5, KEGG (downloaded on April 23, 2011), or HumanCyc 16.0. Similarly, 100% of the drug metabolites in the 101−500 bin, 73% of the glycerophospholipid metabolites in the 101−500 bin, and 61% of the plant metabolites in the 11−100 bin were found in PubChem. Acetic acid was produced from 196 different parent plant compounds (Figure 4d). These results suggest that metabolites in IIMDB that are generated from multiple parents were more likely to be present in the current database and/or previously found in vivo.

Figure 5 (top panel) gives an example of a parent mammalian metabolite [2-aminoethoxy-(2*R*)-2,3-bis-(tetradecanoyloxy)propoxyphosphinic acid: HMDB08821] metabolized by Meteor. Meteor predicted 50 metabolites of HMDB08821; 35 were not found in any database, eight were

dx.doi.org/10.1021/ci400368v | *J. Chem. Inf. Model.* 2013, 53, 2483−2492

found in HMDB, and 15 were found in PubChem. Figure 5 (bottom panel) shows three of the 50 metabolites. Metabolite-1 [2-aminoethoxy-(2*S*)-2,3-dihydroxypropoxyphosphinic acid: HMDB59660] was found in HMDB and has been identified in vivo,[42] while the other two compounds were not found in any database even though they are similar to Metabolite-1. Metabolite-1 was produced from HMDB08821 and 753 other parents (most are probably also glycerophospholipids). Metabolite-2 and Metabolite-3 were produced from HMDB08821 and 80 other parent metabolites. This result is consistent with what is shown in Figure 4: metabolites produced by multiple parents are more likely to be found in vivo and included in current databases. Selected Meteor-generated metabolites for three more examples (HMDB06335, HMDB12490, and HMDB00413) can be found in the Supporting Information.

As shown in Figure 6, most of the Meteor-generated metabolites are not found in PubChem or any other existing database. Out of 92 693 Meteor-generated mammalian metabolites, 4578 (~5%) are found in PubChem. Approximately 10% of the drug metabolites, 7% of the plant metabolites, and 2% of the glycerophospholipid metabolites are found in PubChem. Of the 4578 Meteor-generated mammalian metabolites found in PubChem, 1682 (1.81%) are also found in HMDB. Approximately 2% of the Meteor-generated mammalian metabolites (1756) matched a mammalian parent found in HMDB, KEGG, or HumanCyc. Thus, we found that approximately 25% of the 7108 mammalian parent compounds were produced by phase-I and phase-II in silico enzymatic metabolism of other parents. These results confirm that this method produces authentic biochemical metabolites.

The biological structure matching algorithm BioSM was also used to assess the potential biological relevance of augmenting current databases with computationally generated compounds. Both parent and in silico metabolites were classified as either biological or nonbiological (Figure 7). The results indicate that the biological likeness of the mammalian set of compounds compares closely with that of their in silico metabolites (94.7% vs 92%). If in silico metabolites of only those parents that are predicted to be biological are considered, the biological likenesses are nearly identical (94.7% vs 94.4%). Interestingly, 47.1% of the in silico metabolites generated from nonbiological mammalian parents (i.e., mammalian parents predicted to be nonbiological by BioSM) were predicted to be biological. The same trend was observed for all classes of compounds. A greater portion of in silico metabolites of both drugs and plant compounds were predicted to be biological than their parents. In the case of drugs, the number of compounds predicted to be biological increased by 22.5% upon metabolism. As shown in Figure 7 (lower panel), the biological drug metabolites (35.2%) generated from nonbiological drug parents (60.8%) account for the observed increase. In most cases, in silico metabolism increased the probability that a compound was scored as biological by BioSM. This seems especially likely when added functional groups are relatively large compared to the parent compound (e.g., glucuronidation).

IIMDB's imbedded RESTful web service allows easy integration with third-party applications. Any existing database-dependent metabolomics program can use IIMDB as an additional compound source. Obviously, in silico metabolites such as these cannot be annotated with experimental data such as MS/MS spectra or experimental retention times. However, quantitative structure−property relationship (QSPR)-based predictive models can be used to efficiently filter out irrelevant metabolites and retain candidates that match experimental values. In this method, in silico-generated candidate compounds whose predicted features lie outside the range of values allowed by the QSPR models are removed from consideration. Three such QSPR models (retention index, $ECOM_{50}$, and drift time) that can be used to filter out IIMDB metabolites are discussed in our previous work[43−48] and are implemented within the MolFind[46] software. The remaining candidate compounds can then be computationally fragmented and matched with experimental mass spectra to identify unknowns.

Meteor and other in silico metabolism programs are known to overpredict the number of possible metabolites when less restrictive constraints are used.[35,49] However, having a certain number of false positives (i.e., a larger biochemical database) is not a disadvantage if these can be filtered out efficiently. Our previous studies[46,48] showed that predictive models such as those in MolFind can filter out ~87% of candidate compounds in a PubChem bin (monoisotopic molecular weight ±10 ppm). The remaining candidates are ranked by comparing their predicted properties (retention index, $ECOM_{50}$, drift index, and simulated CID spectrum) with experimental data.

With the inclusion of computationally generated metabolites of additional structures in Lipid Maps and HMDB 3.0, the IIMDB could grow to approximately 2 million structures. Our previous work using BioSM showed that approximately 3 million compounds in PubChem are biological.[31] Since the majority of the compounds in IIMDB (~95%) are not found in PubChem, IIMDB will significantly augment these 3 million biological candidate structures to provide a useful resource for nontargeted metabolomics research.

SUMMARY AND OUTLOOK

In summary, IIMDB is a web-accessible, user- and programmer-friendly metabolite database for mass-spectrometry-based structure identification. IIMDB is also the largest small-molecule database of its kind, comprising 23 035 known and 400 414 computationally generated metabolites. The large majority of the in silico compounds are not found in existing databases such as PubChem. Furthermore, most of these compounds are predicted to be biological by BioSM. This article describes the status of the first version of IIMDB. We plan to significantly expand IIMDB by computationally metabolizing additional compounds found in HMDB 3.0 and other classes of lipids in the Lipid Maps structure database.

ASSOCIATED CONTENT

Supporting Information

A list of phase-I and phase-II biotransformation types in Meteor, three additional examples of Meteor-generated metabolites, a step-by-step guide to viewing and converting structures, and a list of all parent structures used in this work (XLSX). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: (860)486-4265. Fax: (860)486-5792. E-mail: david.grant@uconn.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Augustin Scalbert and Vanessa Neveu of the International Agency for Research on Cancer (IARC) for providing the polyphenol dataset used in this study. This research was funded by NIH Grant 1R01GM087714, Agriculture and Food Research Initiative Competitive Grant 2011-67016-30331 from the USDA National Institute of Food and Agriculture, and Award IIS-0916948 from NSF.

■ REFERENCES

(1) Loftus, N.; Barnes, A.; Ashton, S.; Michopoulos, F.; Theodoridis, G.; Wilson, I.; Ji, C.; Kaplowitz, N. Metabonomic investigation of liver profiles of nonpolar metabolites obtained from alcohol-dosed rats and mice using high mass accuracy MS^n analysis. *J. Proteome Res.* **2011**, *10*, 705−713.

(2) Hu, Y.; Yu, Z.; Yang, Z. J.; Zhu, G.; Fong, W. Comprehensive chemical analysis of Venenum Bufonis by using liquid chromatography/electrospray ionization tandem mass spectrometry. *J. Pharm. Biomed. Anal.* **2011**, *56*, 210−220.

(3) Baran, R.; Bowen, B. P.; Bouskill, N. J.; Brodie, E. L.; Yannone, S. M.; Northen, T. R. Metabolite Identification in *Synechococcus* sp. PCC 7002 Using Untargeted Stable Isotope Assisted Metabolite Profiling. *Anal. Chem.* **2010**, *82*, 9034−9042.

(4) Xu, F.; Zou, L.; Lin, Q.; Ong, C. N. Use of liquid chromatography/tandem mass spectrometry and online databases for identification of phosphocholines and lysophosphatidylcholines in human red blood cells. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 3243−3254.

(5) Yoo, B. C.; Kong, S.-Y.; Jang, S.-G.; Kim, K.-H.; Ahn, S.-A.; Park, W.-S.; Park, S.; Yun, T.; Eom, H.-S. Identification of hypoxanthine as a urine marker for non-Hodgkin lymphoma by low-mass-ion profiling. *BMC Cancer* **2010**, *10*, 1−9.

(6) Bou Khalil, M.; Hou, W.; Zhou, H.; Elisma, F.; Swayne, L. A.; Blanchard, A. P.; Yao, Z.; Bennett, S. A. L.; Figeys, D. Lipidomics era: Accomplishments and challenges. *Mass Spectrom. Rev.* **2010**, *29*, 877−929.

(7) Wallace, B. D.; Redinbo, M. R. The human microbiome is a source of therapeutic drug targets. *Curr. Opin. Chem. Biol.* **2013**, *17*, 379−384.

(8) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* **2012**, *30*, 826−828.

(9) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN: A metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(10) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527−D532.

(11) Sud, M.; Fahy, E.; Cotter, D.; Dennis, E. A.; Subramaniam, S. LIPID MAPS−Nature Lipidomics Gateway: An Online Resource for Students and Educators Interested in Lipids. *J. Chem. Educ.* **2012**, *89*, 291−292.

(12) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35*, D521−D526.

(13) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0: The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801−D807.

(14) Fiehn, O.; Barupal, D. K.; Kind, T. Extending biochemical databases by metabolomic surveys. *J. Biol. Chem.* **2011**, *286*, 23637−23643.

(15) Ekroos, M.; Sjögren, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682−13687.

(16) Nam, H.; Lewis, N. E.; Lerman, J. A.; Lee, D.-H.; Chang, R. L.; Kim, D.; Palsson, B. O. Network context and selection in the evolution to enzyme specificity. *Science* **2012**, *337*, 1101−1104.

(17) Carbonell, P.; Lecointre, G.; Faulon, J.-L. Origins of specificity and promiscuity in metabolic networks. *J. Biol. Chem.* **2011**, *286*, 43994−44004.

(18) Wikoff, W. R.; Anfora, A. T.; Liu, J.; Schultz, P. G.; Lesley, S. A.; Peters, E. C.; Siuzdak, G. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3698−3703.

(19) Gao, J.; Ellis, L. B. M.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: Improving Public Access. *Nucleic Acids Res.* **2010**, *38*, D488−D491.

(20) Moriya, Y.; Shigemizu, D.; Hattori, M.; Tokimatsu, T.; Kotera, M.; Goto, S.; Kanehisa, M. PathPred: An enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.* **2010**, *38*, W138−W143.

(21) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; Lin, G. MyCompoundID: Using an evidence-based metabolome library for metabolite identification. *Anal. Chem.* **2013**, *85*, 3401−3408.

(22) Faust, K.; Croes, D.; van Helden, J. Metabolic pathfinding using RPAIR annotation. *J. Mol. Biol.* **2009**, *388*, 390−414.

(23) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27−30.

(24) Romero, P.; Wagg, J.; Green, M. L.; Kaiser, D.; Krummenacker, M.; Karp, P. D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **2005**, *6*, R2.

(25) Chae, L.; Lee, I.; Shin, J.; Rhee, S. Y. Towards understanding how molecular networks evolve in plants. *Curr. Opin. Plant Biol.* **2012**, *15*, 177−184.

(26) Pérez-Jiménez, J.; Neveu, V.; Vos, F.; Scalbert, A. Systematic analysis of the content of 502 polyphenols in 452 foods and beverages: An application of the Phenol-Explorer database. *J. Agric. Food Chem.* **2010**, *58*, 4959−4969.

(27) 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.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.

(28) Heller, W. M.; Fleeger, C. A. *USAN and the USP Dictionary of Drug Names*; United States Pharmacopeial Convention: Rockville, MD, 1989; pp 1−761.

(29) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivery Rev.* **2002**, *54*, 407−415.

(30) Marchant, C. A.; Briggs, K. A.; Long, A. In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mech. Methods* **2008**, *18*, 177−187.

(31) Hamdalla, M. A.; Mandoiu, I. I.; Hill, D. W.; Rajasekaran, S.; Grant, D. F. BioSM: Metabolomics Tool for Identifying Endogenous Mammalian Biochemical Structures in Chemical Structure Space. *J. Chem. Inf. Model.* **2013**, *53*, 601−612.

(32) Button, W. G.; Judson, P. N.; Long, A.; Vessey, J. D. Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1371−1377.

(33) Judson, P. N.; Cooke, P. A.; Doerrer, N. G.; Greene, N.; Hanzlik, R. P.; Hardy, C.; Hartmann, A.; Hinchliffe, D.; Holder, J.; Müller, L.; Steger-Hartmann, T.; Rothfuss, A.; Smith, M.; Thomas, K.; Vessey, J. D.; Zeiger, E. Towards the creation of an international toxicology information centre. *Toxicology* **2005**, *213*, 117−128.

(34) Judson, P. Using Computer Reasoning about Qualitative and Quantitative Information to Predict Metabolism and Toxicity. In *Pharmacokinetic Profiling in Drug Research*; Wiley-VCH: Weinheim, Germany, 2006; pp 417−429.
(35) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617−648.
(36) Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S. Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics* **2011**, *27*, 1537−1545.
(37) *OrientDB*, version 1.3; Orient Technologies: London, 2012.
(38) Knuth, D. E. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd ed.; Addison-Wesley Longman: Boston, 1997; pp 1−170.
(39) Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(40) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Elsevier: Amsterdam, 2008; Vol. *8*, Chapter 12, pp 217−241.
(41) Kirchmair, J.; Howlett, A.; Peironcely, J. E.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van Buren, L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How do metabolites differ from their parent molecules and how are they excreted? *J. Chem. Inf. Model.* **2013**, *53*, 354−367.
(42) Yamamoto, K.; Yoon, K. D.; Ueda, K.; Hashimoto, M.; Sparrow, J. R. A novel bisretinoid of retina is an adduct on glycerophosphoethanolamine. *Invest. Ophthalmol. Visual Sci.* **2011**, *52*, 9084−9090.
(43) Hall, L. M.; Hall, L. H.; Kertesz, T. M.; Hill, D. W.; Sharp, T. R.; Oblak, E. Z.; Dong, Y. W.; Wishart, D. S.; Chen, M.-H.; Grant, D. F. Development of Ecom(50) and Retention Index Models for Nontargeted Metabolomics: Identification of 1,3-Dicyclohexylurea in Human Serum by HPLC/Mass Spectrometry. *J. Chem. Inf. Model.* **2012**, *52*, 1222−1237.
(44) Hill, D. W.; Baveghems, C. L.; Albaugh, D. R.; Kormos, T. M.; Lai, S.; Ng, H. K.; Grant, D. F. Correlation of Ecom50 values between mass spectrometers: Effect of collision cell radiofrequency voltage on calculated survival yield. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 2303−2310.
(45) Kertesz, T. M.; Hall, L. H.; Hill, D. W.; Grant, D. F. CE50: Quantifying collision induced dissociation energy for small molecule characterization and identification. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1759−1767.
(46) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; Wilder, J.; Grant, D. F. MolFind: A Software Package Enabling HPLC/MS-Based Identification of Unknown Chemical Structures. *Anal. Chem.* **2012**, *84*, 9388−9394.
(47) Kertesz, T. M.; Hill, D. W.; Albaugh, D. R.; Hall, L. H.; Hall, L. M.; Grant, D. F. Database searching for structural identification of metabolites in complex biofluids for mass spectrometry-based metabonomics. *Bioanalysis* **2009**, *1*, 1627−1643.
(48) Menikarachchi, L. C.; Hamdalla, M. A.; Hill, D. W.; Grant, D. F. Chemical Structure Identification in Metabolomics: Computational Modeling of Experimental Features. *Comput. Struct. Biotechnol. J.* **2013**, *5*, No. e201302005.
(49) Piechota, P.; Cronin, M. T. D.; Hewitt, M.; Madden, J. C. Pragmatic Approaches to Using Computational Methods To Predict Xenobiotic Metabolism. *J. Chem. Inf. Model.* **2013**, *53*, 1282−1293.