Improving Metabolomics Coverage and Standardization

By

Parker Ladd Bremer

A DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA DAVIS

Approved:

____

Oliver Fiehn, Chair

____

Delmar Larsen

____

Justin Siegel

Committee in Charge

2023

Dedication

To my son Arthur, his brothers, and his sisters.

Abstract

Progress in metabolomics has brought the field from investigations of pre-selected compound lists and limited sample size toward comprehensive compound exploration of large sample size. This shift in focus demanded corresponding advances in informatics areas that we explore in this dissertation, such as in-silico compound identification tools, metabolomics meta-analysis, and metabolomics repository design.

In Chapter 1, we focus on compound identification. Compound identification is traditionally treated as an information retrieval problem, where unknown compounds are identified by comparing their observed signals to the signals of chemical standards. Unfortunately, the metabolome contains significantly more compounds than standards, so there is a desire to computationally expand the space of indexed signals. Here, we benchmark a tool, CFM-ID, that predicts the signal of a compound based on its structure. We show that there is much progress needed in this area by determining that CFM-ID's predictions could be readily replicated via heuristic rules that focus on structure. Extrapolating these ideas emphasizes the need for increased machine learning model training set sizes and standardization due to the complexity of the physics and statistical mechanics that mass spectrometry signals reflect.

In Chapter 2, we focus on meta-analysis of metabolomics studies. We believe that the synchronization of many independent datasets will allow for biological insights of high confidence and/or high generality. To this end, we developed a tool named BinDiscover, which allows for rapid hypothesis generation by enabling user-directed exploration of over 150,000 samples processed at the West Coast Metabolomics Center. We believe that this tool improves existing repository meta-analysis for several reasons. First, it is programmatic in nature, which allows for meta-analysis on a timescale of minutes rather than months. Second, the meta-analysis that it

enables is focused on sample metadata rather than study hypotheses, which dramatically expands the number of investigations that can be conducted. Third, it is dramatically easier to use than existing options. Finally, it showcases our novel procedure, ontologically-grouped-differential analysis, which allows for the convenient comparison of categories of samples (e.g., mammals digestive system organs vs. bacterial cells) in order to produce tractable amounts high-confidence results.

In Chapter 3, we focus on repository design. We strongly believe that enabling the programmatic meta-analysis developed on in Chapter 2 onto a larger-scale, community-contributed repositories of metabolomics data will enable massive clinical progress. To this end, we developed a tool that standardizes sample metadata. At current, user-submitted sample metadata matrices preclude programmatic meta-analysis because they suffer from the looseness and complexity of natural language. Our multistep standardization tool employs machine learning models embedded into an intuitive frontend to ensure that only high-quality sample descriptions are lodged into repositories.

Finally, in the appendix, we share several projects spanning the topics of the main chapters. In the first part, we share ClusterBase, which is a computational platform that uses network analysis to organize and annotate spectral data from metabolomics studies. In the second part, we share an automatic compound-ID workflow that harnessed the online CFM-ID tool. Finally, in the third part, we describe a machine learning approach to predicting spectral intensities that can augment quantum mechanically predicted spectra.

Acknowledgements

I would like to thank my principal investigator, Professor Oliver Fiehn. I respect his commitment to scientific excellence. This commitment drove me to improve the analyses and tools that I developed, which in turn challenged me to develop living software. Correspondingly, this forced me to improve my ability to communicate concisely and clearly. I also respect his unrivalled work ethic. His extraordinary effort funded my time at UC Davis, which allowed me to focus on my skills and research. Moreover, his lab's excellent reputation has opened doors for me in other scientific organizations. I have achieved my professional goals because of his work.

I would like to thank my wife, Dr. Jamie Gleason. I love her for the emotional and scientific support that she has provided since I met her. I love that she balances my life by making sure that I take the time to enjoy the beauty around me. She is one of the strongest women whom I have ever met and I am truly lucky to have her stand by me.

I would like to thank my father, Rob, and my mother, Lynn, for the opportunities that their love and work provided. I am only beginning to understand the years of love and exhaustion that go into raising one's child.

I would like to thank my brother, Grant, for his support and companionship. I was able to complete this Ph.D. program because I could tag-in my best friend to handle half of the efforts of supporting our family.

I would like to thank Dr. Brian DeFelice for the opportunities that his hard-work and genuine demeanor provided. He helped me get an internship and then full-time employment doing scientific software development, which is a dream-come-true.

I would like to thank Gert Wohlgemuth for the feedback and ideas that his extraordinarily broad knowledge of the software and DevOps fields provided. Additionally, I would like to thank

him for the specific tools and software that I employed as well as the energy that he put into deployment and updates for me.

I would like to thank the entire Fiehn lab for the infrastructure on top of which my work is derived. The research group has provided amazing feedback over the last several years, the core lab analyzed the thousands of samples which I used in my projects, and the IT team developed numerous software and infrastructure on which my work is based. I would like to thank certain individuals, as well, in no particular order. I would like to thank Dr. Arpana Vaniya for her excellent feedback throughout my time in the Fiehnlab as well as the warm social environment that her presence creates. I would like to thank Dr. Tobias Kind for his thoughtful informatics perspective and suggestions. I would like to thank Dr. Shunyang Wang for his partnership on several projects and good-faith recommendations for employment.

Table of Contents

Chapter 1: How Well Can We Predict Mass Spectra from Structures?

Reproduced from "How Well Can We Predict Mass Spectra from Structures? Benchmarking Competitive Fragmentation Modeling for Metabolite Identification on Untrained Tandem Mass Spectra" by Parker Ladd Bremer, Arpana Vaniya, Tobias Kind, Shunyang Wang, and Oliver Fiehn in the Journal of Chemical Information and Modeling.

1.1 Abstract

Competitive Fragmentation Modelling for Metabolite Identification (CFM-ID) is a machine learning tool to predict in silico tandem mass spectra (MS/MS) for known or suspected metabolites for which chemical reference standards are not available. As a machine learning tool, it relies on both an underlying statistical model and an explicit training set that encompasses experimental mass spectra for specific compounds. Such mass spectra depend on specific parameters such as collision energies, instrument types, and adducts which are accumulated in libraries. Yet, ultimately prediction tools that are meant to cover wide expanses of entities must be validated on cases that were not included in the initial training and testing sets. Hence, we here benchmarked the performance of CFM-ID 4.0 to correctly predict MS/MS spectra for spectra that were not included in the CFM-ID training set and for different mass spectrometry conditions. We used 609,456 experimental tandem spectra from the NIST20 mass spectral library that were newly added to the previous NIST17 library version.

We found that CFM-ID's highest energy prediction output would maximize the capacity for library generation. Matching the experimental collision energy with CFM-ID's prediction energy produced the best results, even for HCD-Orbitrap instruments. For benzenoids, better MS/MS predictions were achieved than for heterocyclic compounds. However, when exploring CFM-ID's performance on 8,305 compounds at 40 eV HCD-Orbitrap collision energy, >90% of the 20/80 split test compounds showed <700 MS/MS similarity score. Instead of a stand-alone tool,

CFM-ID 4.0 might be useful to boost candidate structures in the greater context of identification workflows.

1.2 Introduction

The expanse of metabolites observed in humans, plants, and other forms of life is enormous. The Human Metabolome Database (HMDB) alone currently contains well over 100,000 documented metabolites and the total plant metabolome is believed to span over one-million compounds.1,2 In liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)-based metabolomics, a compound in a sample is commonly annotated by comparing their experimental mass spectra to reference mass spectra that are contained in a mass spectral libraries.3 Classically, these libraries are developed by acquiring mass spectra from authentic analytical standards. In practice however, reference mass spectra are available for only a small fraction of the metabolome.4,5 The coverage of compounds in PubChem that have associated mass spectra is estimated to be less than 1%.4 Therefore, millions of compounds do not have associated experimental mass spectra and, moreover, most of them are not commercially available. Hence, mass spectra for these compounds must be predicted by in silico tools to facilitate compound identification in untargeted metabolomics.6 Predicted reference MS/MS spectra are in untargeted metabolomics because it is estimated that more than 80% of unknown MS/MS spectra remain unidentified.

Numerous computational tools have been developed for compound identification or structure elucidation.7 The three basic approaches are: (1) rule-based fragmentation tools,8 for which fragmentation trends are identified by either classic organic chemistry based rules such as hydrogen-rearrangement rules9 or literature based reaction rules.10 (2) Quantum chemistry tools,11,12 in which first principle theory is applied to simulate fragmentation of a compound of interest. Quantum chemistry tools such as quantum-chemical electron ionization mass spectra (QCEIMS) are generally applied to electron ionization spectra, but there have been recent works to predict ESI-MS spectra.13 (3) Machine learning tools,14,15 for which statistical models are parameterized to generate spectra based on compound and spectrum relationships. These tools produce millions of in silico reference mass spectra relatively quickly and easily in hope to alleviate the pressing demand for reference MS/MS spectra. The success of enhancing experimental libraries with in silico libraries has been demonstrated, however, it is also clear that as stand-alone tools, they are not sufficient.16 Other machine learning tools attempt to predict chemical structures or chemical fingerprints from spectra. Examples are CSI:FingerID, the structure classifier Canopus or ChemDistiller.17–19

CFM-ID 4.0, the tool tested in this publication, is a machine learning software based on a stochastic homogeneous Markov process, with additional hard-coded fragmentation rules for certain classes of compounds such as complex lipids.8 Therefore, it is important to highlight that in this paper we examine the underlying statistical model in conjunction with its default training set. However, CFM-ID comes with the capacity to reparametrize according to whatever example set the user might provide. CFM-ID was trained on a set of 12,165 Q-TOF fragmentation spectra for the [M+H]+ adducts and 6,120 MS/MS spectra for the [M-H]- adducts, covering collision energies of 10 eV, 20 eV, and 40 eV.4 Accordingly, CFM-ID predicts spectra for these collision energies for any given input compound.

The chemical space of the metabolome is more expansive than any training set. The higher accessibility of high accuracy mass spectrometers today enables the use of training sets that are representative of both orbital ion trap and Q-TOF mass spectrometers equally. We therefore tested CFM-ID’s prediction capabilities for compounds, fragmentation methods, and collision energies that it has not yet encountered. To accomplish this, we predicted spectra for the highly curated and reliable NIST20 MS/MS library, which contains compounds that are not included in CFM-ID's training set that were measured on both Q-TOF and orbital ion trap instruments.

1.3 Methods

The workflow for our methods is shown in Figure 1. We used the highly curated NIST20 library from the U.S. National Institute of Standards and Technology (NIST) as input of spectra and molecules into the benchmarking test.20 Compounds found in NIST17 21 or the CFM-ID training set were removed from NIST20 library set. The remaining chemical structures were used to predict MS/MS spectra using the CFM-ID 4.04 and the Mass Spectrum Rule-Based Fragmenter (MSRB) 1.1.3 software programs that were provided in Docker image format from the David Wishart laboratory (University of Alberta, Canada).22 The software performance was evaluated by matching predictions against experimental NIST20 library MS/MS spectra using the unweighted dot product with a mass tolerance of 10 ppm and excluding all ions within 2 Da of precursors. All spectra were normalized to relative abundance before calculating mass spectral similarities. Compound structures were classified according to the Wishart laboratory ClassyFire tool using the batch version implemented at http://cfb.fiehnlab.ucdavis.edu.23 To test our similarity-prediction model, the Vaniya/Fiehn Natural Product Library set of Q-Exactive HF orbital ion trap accurate mass MS/MS spectra (VFNPL) was freely downloaded from the Massbank of North America (https://massbank.us). For all chemical structure datasets, CACTVS molecular fingerprints were obtained using the PubChem web tool.24 All analyses were conducted using custom python scripts.

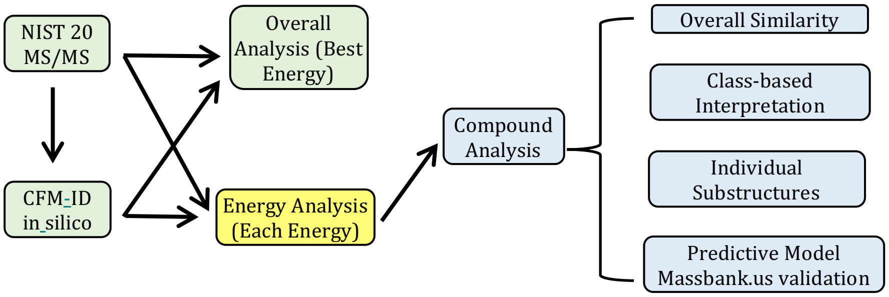

NIST 20 MS/MS
CFM-ID in silico
Overall Analysis (Best Energy)
Energy Analysis (Each Energy)
Compound Analysis
Overall Similarity
Class-based Interpretation
Individual Substructures
Predictive Model Massbank.us validation

Figure 1: Overall method workflow

1.4 Results and Discussion

1.4.1 Selecting Experimental MS/MS Spectra

The NIST20 MS/MS library is composed of 27,613 compounds that generated 1,026,712 MS/MS mass spectra. This library is commercially available to the public and is released in three-year intervals after extensive curation. Only spectra for the most-often observed [M+H]+ and [M-H]- adducts were used to yield a consistent and large benchmarking data set. Compared to the 2017 release (NIST17 library), there was a significant increase with 15,961 compounds and 609,456 spectra newly added. A few NIST20 molecules were already used in CFM-ID training libraries and were consequently removed, leaving 15,328 and 15,494 compounds in [M+H]+ and [M-H]- the benchmarking set, respectively. While CFM-ID was solely trained on Q-TOF mass spectra, we included Q-TOF as well as orbital ion trap spectra. Orbital ion trap spectra included both higher energy collisional dissociation (HCD) and collision induced dissociation (CID) fragmentations.

1.4.2 Creating the CFM-ID Library

For these filtered NIST20 compounds, a CFM-ID 4.0.4 spectral library was created that was patched with CFM-ID predictions for molecules for which a rule-based upgrade model was available, MSRB 1.1.3. The MSRB-Fragmenter patch is an add-on tool that predicts spectra based on rules. The CFM-ID webtool shows users rule-based predictions when available, instead of machine-learning based predictions. Therefore, to replicate user experience, we utilized the MSRB predictions when possible.4 In total, the MSRB-Fragmenter yielded 834 spectra for 278 compounds for [M+H]+ adducts and 822 spectra for 274 compounds for [M-H]- adducts.

Table 1: MS/MS spectra from the NIST20 library used tobenchmark CFM-ID software.

Adduct and type of fragmentation	Number of tested spectra
$[M+H]^+$, Orbitrap HCD	157,407
$[M-H]^-$, Orbitrap HCD	71,026
$[M+H]^+$, Orbitrap CID	12,295
$[M-H]^-$, Orbitrap CID	6,333
$[M+H]^+$, Q-TOF MS/MS	1,111
$[M-H]^-$, Q-TOF MS/MS	35

1.4.3 Overall CFM-ID Performance

We aimed at benchmarking the performance of CFM-ID on spectra that were not included in either training, testing, or validating CFM-ID software.25 CFM-ID version 4.0 was created in early 2020. For that reason, we utilized the NIST20 MS/MS library that was released in June 2020 and removed all compounds that were present in NIST17 or the CFM-ID 4.0 training set. For each remaining compound, we generated CFM-ID predictions for three collision-induced dissociation energies, 10, 20, 40 eV. After removing CFM-ID training compounds, NIST17 compounds, and uncommon adducts, 248,207 spectra remained. For each spectrum, we obtained the dot product similarity score with all three energy predictions for CFM-ID. We did not include any peak within 2 Da of the precursor ion because the precursor ion signifies the intact molecule and must be considered as orthogonal to MS/MS fragment spectra, and because the intensity of precursor ions vary a lot between instrument types and collision energies. For each experimental spectrum, we saved only the score with the greatest similarity among its three comparisons.

We hypothesized that that the quality of CFM-ID predictions of these spectra might depend on (a) instrument type and type of collision induced-fragmentation, (b) adduct type (a complexity which we limited by constraining to only protonated and deprotonated molecules), (c) collision energy and finally, the actual compound structure (defined by InChI Codes which were hashed as InChIKeys). We first partitioned 248,207 NIST20 mass spectra into six groups defined by instrument type and adduct type as given in Table 1.

When subjecting these molecules to in silico fragmentation by CFM-ID 4.04 and benchmarking these spectra against the NIST20 experimental mass spectra, we were surprised to see a clear dichotomy of matches in a histogram plot (Figure 2), with very disparate frequencies of a number of compounds that excellently matched to experimental mass spectra (at dot-score similarity >950) and many more compounds that did not show satisfying MS/MS similarities (<50 dot-score similarity). Between these two boundaries we found a nearly flat distribution of few other compounds. For Q-TOF spectra, the low total number of compounds may have hampered finding any good MS/MS matches at all.

1.4.4 Impact of Collision Energy on CFM-ID Performance

Next, we analyzed the impact of collision energies. We first focused on the 157,407 protonated MS/MS spectra fragmented in HCD-mode using orbital ion traps and compared these to the 1,111 mass spectra in positive ESI mode obtained by a Q-TOF mass spectrometer. In contrast to the overall analysis in Figure 2 that focused on the best MS/MS match across all experimental and in silico collision energies, here we kept all individual MS/MS dot-score similarities separate that matched each experimental spectrum against the simulated CFM-ID spectra for each of the three CFM-ID predictions. We binned all experimental collision energies into 1 eV bins, ranging from 1 to 45 eV for Q-TOF spectra and 1-70 eV for orbital ion trap mass spectra (Figure 3). For orbital ion traps, energy data differed within the NIST20 library, and we therefore selected only one specific instrument type (the Thermo Finnigan Elite Orbitrap data) to be able to utilize uniform energy descriptors. For the full range of energies calculated for this

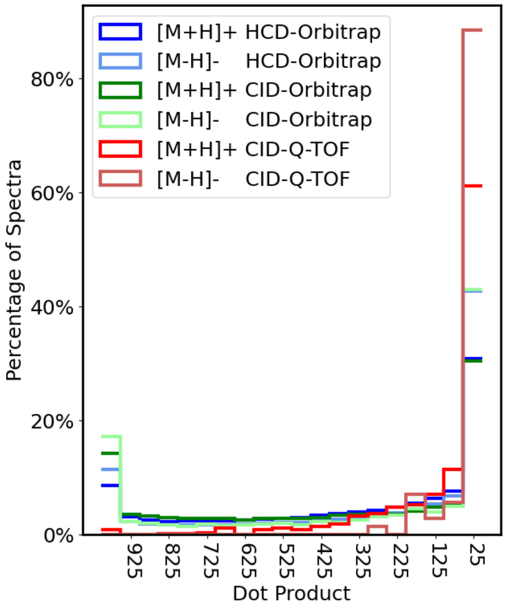


Figure 2: Overall CFM-ID performance measured by dot products between experimental NIST20 MS/MS spectra and CFM-ID predictions for the same compound and adduct. The dot product was taken between experimental spectra and the three CFM-ID predictions, regardless of fragmentation method or settings. The best-scoring dot-product among the three comparisons was recorded and the total list was partitioned into six groups according to fragmentation conditions and adduct.

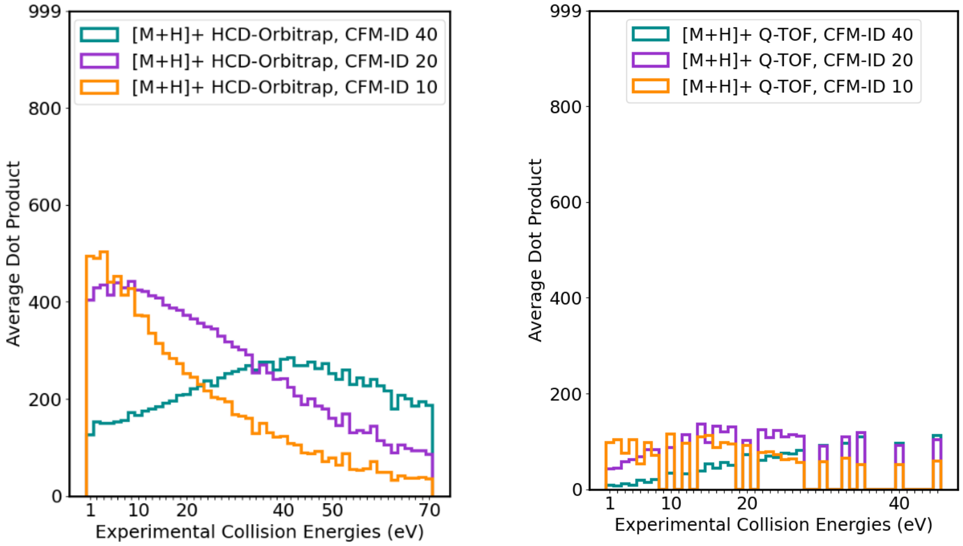


Figure 3: Histograms of dot-score similarities for $[M+H]^+$ molecules between experimental versus predicted MS/MS spectra, by experimental collision energies**.**
Left (a): 1,111 experimental Q-TOF spectra from the NIST20 library.
Right (b): 86,747 Thermo Finnigan Elite Orbital Ion Trap spectra from the NIST20 library.

instrument type, we generated 200 bins, but found a dramatic dip in the number of spectra beyond the first 50 bins (up to 70 eV) to which we therefore limited the analyses. We conclude that CFM-ID performs poorly for the Q-TOF mass spectra from the NIST20 library that were not publicly available during CFM-ID 4.0 software development. We did not find any relationship of dot-score similarities of predicted versus experimental spectra, neither with respect to the experimental energies nor when analyzed for the different simulated energies at 10-40 eV.

For the Elite Orbitrap mass spectra, we yielded a more nuanced result. While averaged MS/MS dot-score similarities remained well below the mark of 600 scores, a threshold that is often used to annotate compounds in experimental MS/MS investigations, we still saw an increase in higher-ranking dot-score similarities depending on the collision energies. For simulated low collision energies at 10-20 eV in CFM-ID (orange and purple graphs in Figure 3b), much better dot-scores were achieved for experimental spectra at <10 eV or <20 eV than at >40 eV collision energies. Vice versa, CFM-ID spectra simulated for 40 eV collision energy showed best dot-score similarities around 40 eV experimental collision energies. Based on these observations, we conclude that CFM-ID is best used for Orbitrap spectra that match in silico with experimental collision energies. However, very often experimental MS/MS spectra at 10-20eV showed very simplistic mass spectra with very little fragmentation, which we interpret as the main reason why average dot-score similarities reached higher maxima than experimental versus predicted MS/MS spectra at 40 eV. In practice, low energy MS/MS spectra only yield uninformative neutral losses such as water or ammonium losses. Hence, for the purpose of annotating unknown compounds with in silico libraries, experimental and in silico spectra at 40 eV should be more useful.

Orbital ion traps collision energies are often given in relative normalized collision energies (%NCE). To refer %NCE values to energies given in eV, we used information from metadata given in the NIST20 library for collision energies for the Thermo Finnigan Elite Orbitrap instrument contained both eV and %NCE information. Applied Orbitrap energies are represented as proportions of an optimal energy that scales (linearly) with the precursor mass. This proportion is typically written as “%NCE”.

$$\text{(Applied eV)} = \text{(Optimal eV)} * \text{(\%NCE)}$$

and

$$\text{(Optimal eV)} \propto \text{(Precursor mass)}$$

therefore

$$\text{(Applied eV)} \propto \text{(Precursor mass)} * \text{(\%NCE)}$$

The applied eV was used as x-axis in Figure 3b. Hence, histograms give very similar results if eV values are known, of if they are displayed as Precursor mass * %NCE (Supplement S1).

For other instrument types, such as the Thermo Fisher Lumos instrument, a different constant C in the proportionality would be needed. For this reason, we did not include all Orbitrap NIST20 spectra, but only spectra from this specific instrument type. Overall, it is clear that one cannot simply use %NCE values that are typically reported for orbital ion traps instruments and report definitive eV values across all instrument types.

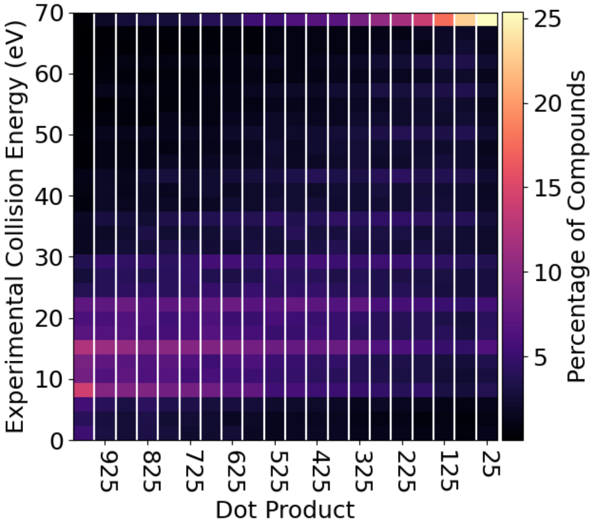


Figure 4: Histogram of $[M+H]^+$/HCD-Orbitrap experimental collision energy against CFM-ID predictions. Each column was normalized to the sum of spectra in that bin of dot-product scores

We wondered why most spectra predictions gave either excellent at >900 similarity or dismal results at <100 similarity. We used the best-scoring CFM-ID energy for each molecule and analyzed the percentage of all 86,747 molecules for [M+H]+ adducts for the Thermo Finnigan Elite orbital ion trap mass spectrometer that yielded acceptable dot-score similarities between CFM-ID predictions and HCD-experimental MS/MS spectra (Figure 4). In this analysis, it becomes clear that very good predictions were found for a comparatively large population of very low experimental collision energies, while very poor MS/MS predictions consisted of a comparatively large population of very high experimental collision energies. The "best-predictions" (>950) were bolstered by experimental collision energies close to 1 eV. Hence, the vast majority of the "best predicted spectra" resulted from a systematic bias of matching very simple MS/MS fragmentation spectra with simple predictions.

1.4.5 Impact of Molecule Structure on CFM-ID Performance

Next, we investigated the impact of structure on CFM-ID predictability. To remove observed systematic bias from mismatched energies, we limited the analyses of MS/MS spectra to the 8,035 molecules that were assigned with explicit eV units in the NIST20 library between 35-45 eV for the Thermo Finnigan Orbitrap. Figure 5 show that for >90% of these compounds, MS/MS similarity dot-scores of <700 were yielded, even when choosing the optimal 40 eV setting in CFM-ID predictions for HCD Orbitrap spectra. Yet, for about 10% of these molecules, decent MS/MS spectra could be simulated with dot-scores >600, and in some cases even >800 dot-score similarities.

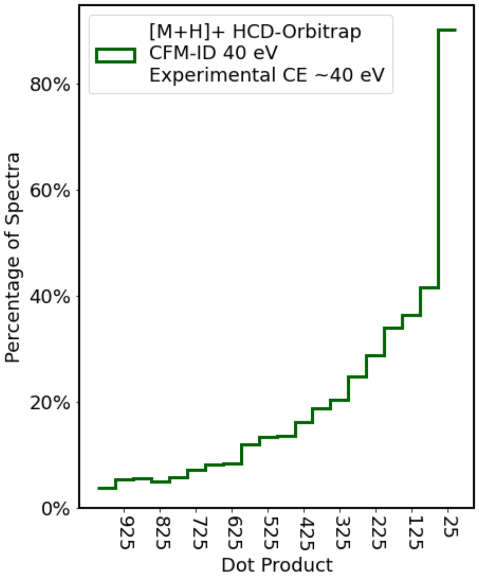


Figure 5: Histogram of 8,035 $[M+H]^+$/HCD-Orbitrap compounds with experimental collision energies 35-45 eV and simulated CFM-ID energy 40.

We therefore used this subset of data to explore the impact of chemical structure on CFM-ID predictability of MS/MS spectra. We first hypothesized that compounds with a greater similarity to the CFM-ID training set might yield better dot-score MS/MS similarities. To this end, we acquired CACTVS fingerprints using the PubChem REST API for 4,040 molecules of the training set (that was disclosed by the authors of the CFM-ID software), and applied these to 8,298 chemical fingerprints for the 35-45 eV HCD spectra molecules for [M+H]+ adducts in the Orbitrap NIST20 database.26 With all chemical fingerprints combined, we created a 2 dimensional reduction embedding of fingerprints using Uniform Manifold Approximation and Projection (UMAP), Figure 6.27 We also examined dimensionality reduction using PCA and t-SNE. Pairwise comparison of PCA's dimensions as well as t-SNE projections yielded the same clustering of well-performing compounds (Supplements S2, S3). Chemical fingerprints of molecules with lo+w dot-score MS/MS similarities were expected to be found far away from the training data. We found that compounds with very poor MS/MS dot scores (dark blue) showed UMAP structural overlaps to the same degree as compounds with good dot scores. Hence, chemical similarity to the training data itself did not predict the ability to correctly simulate MS/MS spectra in CFM-ID. Instead, we found clusters of good predictions (yellow dots), suggesting a success of CFM-ID for very specific chemical classes, but not for others. To this end, we classified all 8,298 molecules by the ClassyFire algorithm into chemical SuperClasses and analyzed the proportion of dot-score similarities for the top-6 SuperClasses (Figure 7). It became clear that well-predicted compounds in CFM-ID at >900 dot score similarities were very likely to be benzenoids, while the poorly predicted compounds at <600 dot scores were likely to be organoheterocyclics. The overall proportion of chemical compounds were heavily biased towards these two SuperClasses, precluding definitive comments about other chemical structures.

Intrigued by the notion that specific compound types were well-predicted and specific compound types were poorly-predicted, we sought to achieve a higher-resolution view on chemical substructures. Here, we used a random forest approach to identify fingerprint bits with the capability to distinguish between well-predicted and poorly-predicted compounds and then later in an attempt to predict CFM-ID's capability to predict spectra, using a binary classification scheme with dot-score similarity of 700 as watershed mark between good and poorly predictable substructures.

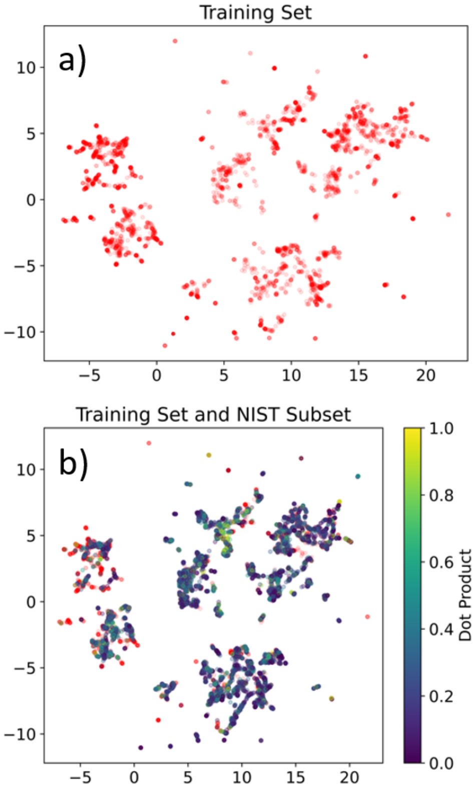


Figure 6: 2D UMAP embedding of CFM-ID positive training fingerprints and $[M+H]^+$/HCD-Orbitrap fingerprints. Upper panel (a) Training data set. Lower panel (b) 8,298 molecules with 35-45eV $[M+H]^+$ MS/MS spectra superimposed onto the training data (red dots). Yellow/blue color scheme indicates the normalized dot product values 0-1000 between 0 and 1

This simplistic binary scheme was performed to allow the RF model to learn specific chemical features that had a high impact on overall good CFM-ID scoring, instead of using regression models that might focus on differentiating among the more-sampled, lower MS/MS similarity dot scores. We chose the model that maximized precision because precisions is most important for building libraries of predicted MS/MS spectra. To identify features, we selected the top-50 chemical fingerprint bits that showed the greatest capacity to distinguish between good- and worse MS/MS predictions. We examined the distributions for compounds for each chemical fingerprint bit in heatmaps and give results for the top-substructure fingerprints in Table 2, Table 3 and Supplement S2. Using the chemical fingerprint bit 185 ("two rings of membership 6") and bit 143 ("at least 1 ring of size 5") explicitly reproduced the result of the superclass analysis. Hence, both the fingerprint analysis and the ClassyFire SuperClass analysis showed that CFM-ID maintained the trained ability to predict MS/MS spectra for simple aromatic molecules that consisted of carbon-only rings. However, this training did not extend to other cyclic structures such as small ring systems with heteroatoms for which CFM-ID predictions failed. Using a

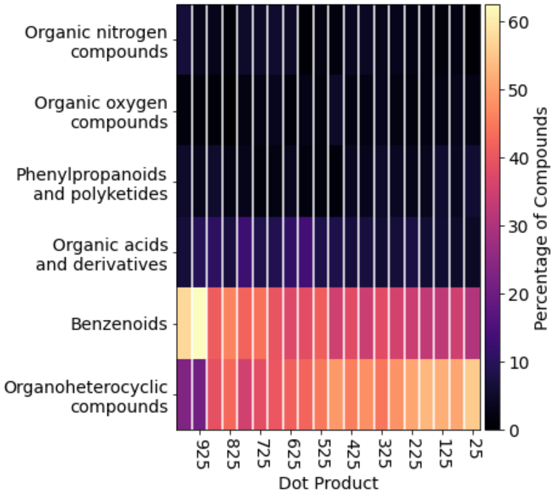


Figure 7: ClassyFire-defined chemical superclasses vs. MS/MS dot product similarity for Orbitrap HCD spectra $[M+H]^+$ between 35-45 eV. Each binned column of dot product is sum-normalized.

train/test split as 20%/80%, chosen randomly from the NIST20 dataset, we found that more than 90% of the structures yielded <700 dot-score similarities to the corresponding experimental spectra (see confusion matrix Supplement S3). Yet, for 20 of the 23 benzenoids included in this withheld testing set gave >700 dot score similarities of confidence that the model can be used to select subsets of proposed compounds for which one can generate an in silico library.

Table 2: Substructures associated with >700 dot score similarities by CFM-ID

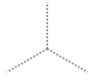
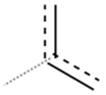
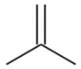
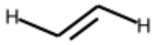
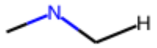
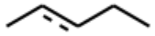
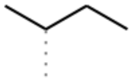
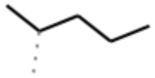
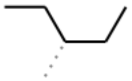
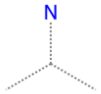
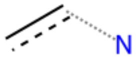
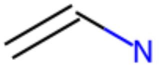
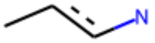
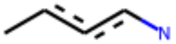
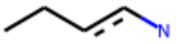
Bit Number	SMILES/SMARTS	Visualization
185	At least 2 rings of size 6	N/A
333	C(~C)(~C)(~C)	
345	C(~C)(~H)(~N)	N/A
356	C(~C)(:C)(:C)	
365	C(~H)(~N)	N/A
430	C(-C)(-C)(=C)	
516	[#1]-C=C-[#1]	
540	C-N-C-[#1]	
688	C-C:C-C-C	
708	C-C(C)-C-C	
709	C-C(C)-C-C-C	
710	C-C-C(C)-C-C	

Table 3: Substructures associated with <700 dot score similarities by CFM-ID

Bit Number	SMILES/SMARTS	Visualization
19	>= 2 O	N/A
143	At least 1 ring of size 5	N/A
340	C(~C)(~C)(~N)	
374	C(~H)(~H)(~H)	N/A
376	C(~N)(:C)	
449	C(-N)(=C)	
545	N-C:C-C	
600	N-C:C:C-C	
665	N-C:C-C-C	

To confirm how generalized this model is, we sought an orthogonal test set for which we used the Vaniya-Fiehn Natural Product Library within the public MassBank.us repository. Because our collision energy analysis for CFM-ID strongly suggested that matching the %NCE for Orbital Ion Trap instrument was extremely important, we removed all compounds for which we could not obtain or calculate an equivalent %NCE to match the CFM-ID "40 eV collision energy". This constraint left 226 compounds to be tested using the CFM-ID 40 eV prediction. When removing all ions within 2 Da of the precursor ion, only 6 of the 226 tested natural product compounds yielded a >700 dot score (Supplement S3), confirming that CFM-ID has very limited prediction ability for correct MS/MS spectra beyond simple benzenoid structures.

1.5 Conclusions

It is important that machine learning-based prediction models are tested and benchmarked by independent analyses on datasets that were not available during model building. Here, we tested mass spectra from NIST20 and MassBank.us (MassBank of North America) to probe the accuracy for which CFM-ID 4.0 was able to predict spectra from structure, a holy grail in tools for use in untargeted metabolomics or exposome research. As a standalone too, CFM-ID's performance provides only few spectra with high MS/MS similarity scores when validated against experimental spectra. However, even with low dot-score similarities, tools like CFM-ID might be worthwhile to be used in the context of compound identification workflows to boost some structures over alternative chemicals, as has been shown in the CASMI 2016 contest.16 For example, CFM-ID could be used to predict fragmentation at 40 eV at which richer fragmentations occur that are useful for compound identifications. For HCD spectra in orbital ion trap mass spectrometers we observed some structural clusters of good MS/MS predictability. While it is not possible to match CFM-ID to a specific %NCE, CFM-ID collision energies in eV are proportional to the product of %NCE and precursor mass of the compound. Based on these results, it seems reasonable that for improvement of MS/MS in-silico prediction from structures, Q-TOF and HCD experimental spectra may be combined to expand the space of training sets.

During our benchmarking tests we found that the accuracy of CFM-ID 4.0 predictions depended on specific chemical substructures, but not on the similarity of tested structures to the structural space in the training set. Hence, we can conclude that at current, machine learning for direct MS/MS predictions in CFM-ID did not work for most compound classes, except for the ClassyFire SuperClass of benzenoids. Nevertheless, if CFM-ID 4.0 is cautiously used in conjunction with compound-identification workflows, it may improve overall compound ID scores. 16,28 We hope that in the coming years the standardization of metabolomics repositories will enable massive datasets to drive the progress of machine learning methods to predict mass spectra from chemical structures.

1.6 Data and Software Availability

The code used in this manuscript is available at https://github.com/plbremer/cfmid_2. The CFM-ID docker images are available at https://hub.docker.com/repository/docker/wishartlab/cfmid. The NIST20 and NIST17 datasets are available for purchase at https://www.nist.gov/programsprojects/nist20-updates-nist-tandem-and-electron-ionizationspectral-libraries. The VFNPL is freely available at https://massbank.us/.

1.7 References

(1) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: The Human Metabolome Database for 2018. Nucleic Acids Res. 2018, 46 (D1), D608–D617. https://doi.org/10.1093/nar/gkx1089.

(2) Rai, A.; Saito, K.; Yamazaki, M. Integrated Omics Analysis of Specialized Metabolism in Medicinal Plants. Plant J. Cell Mol. Biol. 2017, 90 (4), 764–787. https://doi.org/10.1111/tpj.13485.

(3) Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics | Analytical Chemistry. https://pubs.acs.org/doi/abs/10.1021/acs.analchem.5b04491 (accessed 2021-03-04).

(4) Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D. S. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. Metabolites 2019, 9 (4). https://doi.org/10.3390/metabo9040072.

(5) Go, Y.-M.; Walker, D. I.; Liang, Y.; Uppal, K.; Soltow, Q. A.; Tran, V.; Strobel, F.; Quyyumi, A. A.; Ziegler, T. R.; Pennell, K. D.; Miller, G. W.; Jones, D. P. Reference Standardization for Mass Spectrometry and High-Resolution Metabolomics Applications to Exposome Research. Toxicol. Sci. 2015, 148 (2), 531–543. https://doi.org/10.1093/toxsci/kfv198.

(6) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. Untargeted Metabolomics Strategies – Challenges and Emerging Directions. J. Am. Soc. Mass Spectrom. 2016, 27 (12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y.

(7) Krettler, C. A.; Thallinger, G. G. A Map of Mass Spectrometry-Based in Silico Fragmentation Prediction and Compound Identification in Metabolomics. Brief. Bioinform. 2021, 22 (5). https://doi.org/10.1093/bib/bbab073.

(8) Allen, F.; Greiner, R.; Wishart, D. Competitive Fragmentation Modeling of ESI-MS/MS Spectra for Putative Metabolite Identification. Metabolomics 2015, 11 (1), 98–110. https://doi.org/10.1007/s11306-014-0676-4.

(9) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Anal. Chem. 2016, 88 (16), 7946–7958. https://doi.org/10.1021/acs.analchem.6b00770.

(10) Powering Confident Insights - Explore Your Small-Molecule Data to Its Core. 12.

(11) Ásgeirsson, V.; A. Bauer, C.; Grimme, S. Quantum Chemical Calculation of Electron Ionization Mass Spectra for General Organic and Inorganic Molecules. Chem. Sci. 2017, 8 (7), 4879–4895. https://doi.org/10.1039/C7SC00601B.

(12) Wang, S.; Kind, T.; Tantillo, D. J.; Fiehn, O. Predicting in Silico Electron Ionization Mass Spectra Using Quantum Chemistry. J. Cheminformatics 2020, 12 (1), 63. https://doi.org/10.1186/s13321-020-00470-3.

(13) Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S. Quantum Chemistry Calculations for Metabolomics. Chem. Rev. 2021, 121 (10). https://doi.org/10.1021/acs.chemrev.0c00901.

(14) Wei, J. N.; Belanger, D.; Adams, R. P.; Sculley, D. Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. ACS Cent. Sci. 2019, 5 (4), 700–708. https://doi.org/10.1021/acscentsci.9b00085.

(15) Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites 2020, 10 (6). https://doi.org/10.3390/metabo10060243.

(16) Blaženović, I.; Kind, T.; Torbašinović, H.; Obrenović, S.; Mehta, S. S.; Tsugawa, H.; Wermuth, T.; Schauer, N.; Jahn, M.; Biedendieck, R.; Jahn, D.; Fiehn, O. Comprehensive Comparison of in Silico MS/MS Fragmentation Tools of the CASMI Contest: Database Boosting Is Needed to Achieve 93% Accuracy. J. Cheminformatics 2017, 9 (1), 32. https://doi.org/10.1186/s13321-017-0219-x.

(17) Searching molecular structure databases with tandem mass spectra using CSI:FingerID | PNAS. https://www.pnas.org/content/112/41/12580 (accessed 2021-03-04).

(18) Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S. Systematic Classification of

Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. Nat. Biotechnol. 2021, 39 (4), 462–471. https://doi.org/10.1038/s41587-020-0740-8.

(19) Laponogov, I.; Sadawi, N.; Galea, D.; Mirnezami, R.; Veselkov, K. A. ChemDistiller: An Engine for Metabolite Annotation in Mass Spectrometry. Bioinformatics 2018, 34 (12), 2096–2102. https://doi.org/10.1093/bioinformatics/bty080.

(20) NIST 20 MS/MS Library (2020). https://www.sisweb.com/software/nist-msms.htm#2 (accessed 2021-03-04).

(21) Stein, S. E. NIST 17 MS/MS LIbrary (2017), 2017. https://doi.org/10.18434/T4H594.

(22) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. Anal. Chem. 2021, 93 (34), 11692–11700. https://doi.org/10.1021/acs.analchem.1c01465.

(23) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. J. Cheminformatics 2016, 8 (1), 61. https://doi.org/10.1186/s13321-016-0174-y.

(24) PubChem/CACTVS Fingerprints. https://pubchemdocs.ncbi.nlm.nih.gov/data-specification (accessed 2021-09-08).

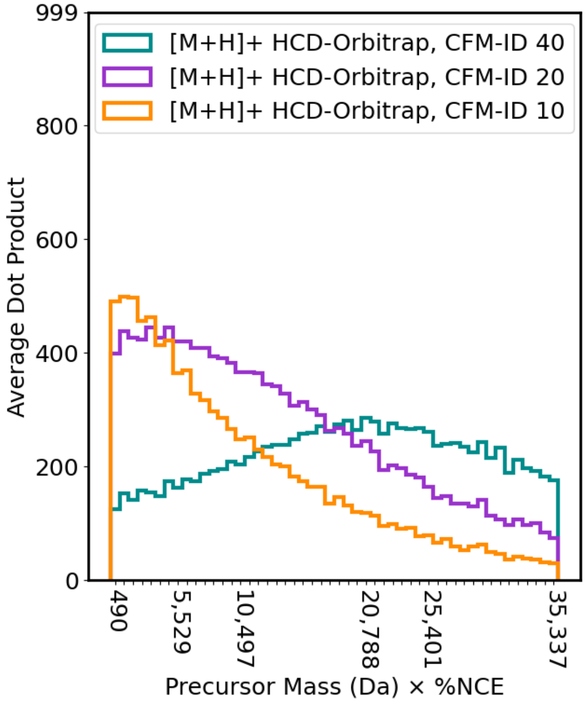
(25) Chao, A.; Al-Ghoul, H.; McEachran, A. D.; Balabin, I.; Transue, T.; Cathey, T.; Grossman, J. N.; Singh, R. R.; Ulrich, E. M.; Williams, A. J.; Sobus, J. R. In Silico MS/MS Spectra for Identifying Unknowns: A Critical Examination Using CFM-ID Algorithms and ENTACT Mixture Samples. Anal. Bioanal. Chem. 2020, 412 (6), 1303–1315. https://doi.org/10.1007/s00216-019-02351-7.

(26) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. J. Chem. Inf. Comput. Sci. 1994, 34 (1), 109–116. https://doi.org/10.1021/ci00017a013.

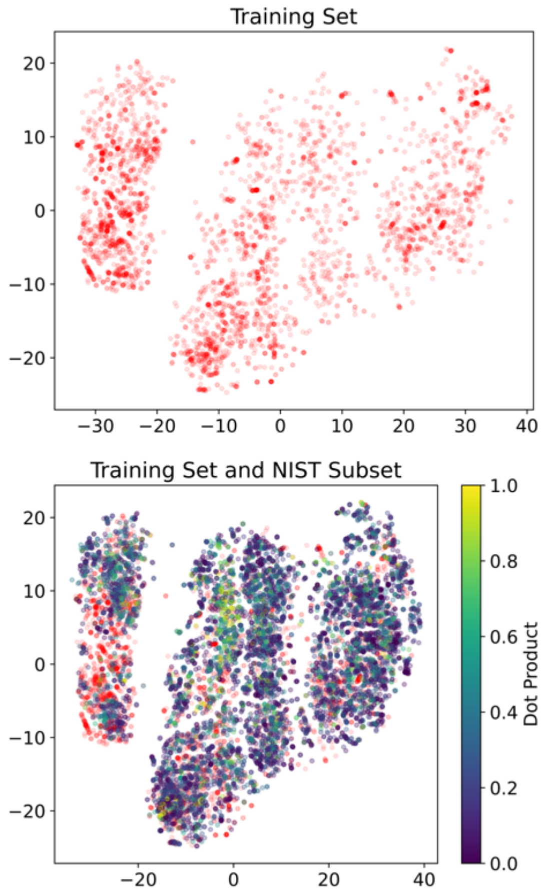
(27) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat 2020.

(28) Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Dührkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Böcker, S.; Rousu, J.; Shen, H.; Tsugawa, H.; Sajed, T.; Fiehn, O.; Ghesquière, B.; Neumann, S. Critical Assessment of Small Molecule Identification 2016: Automated Methods. J. Cheminformatics 2017, 9 (1), 22. https://doi.org/10.1186/s13321-017-0207-1.

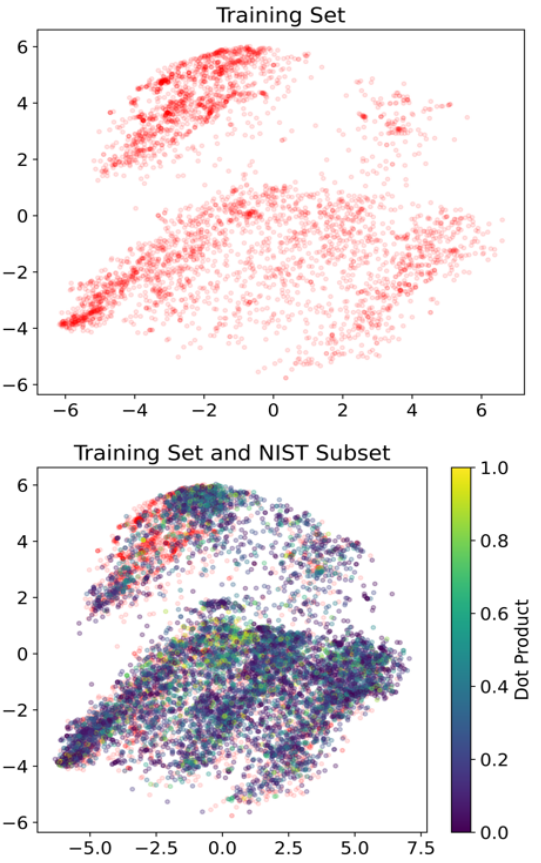
1.8 Supplemental



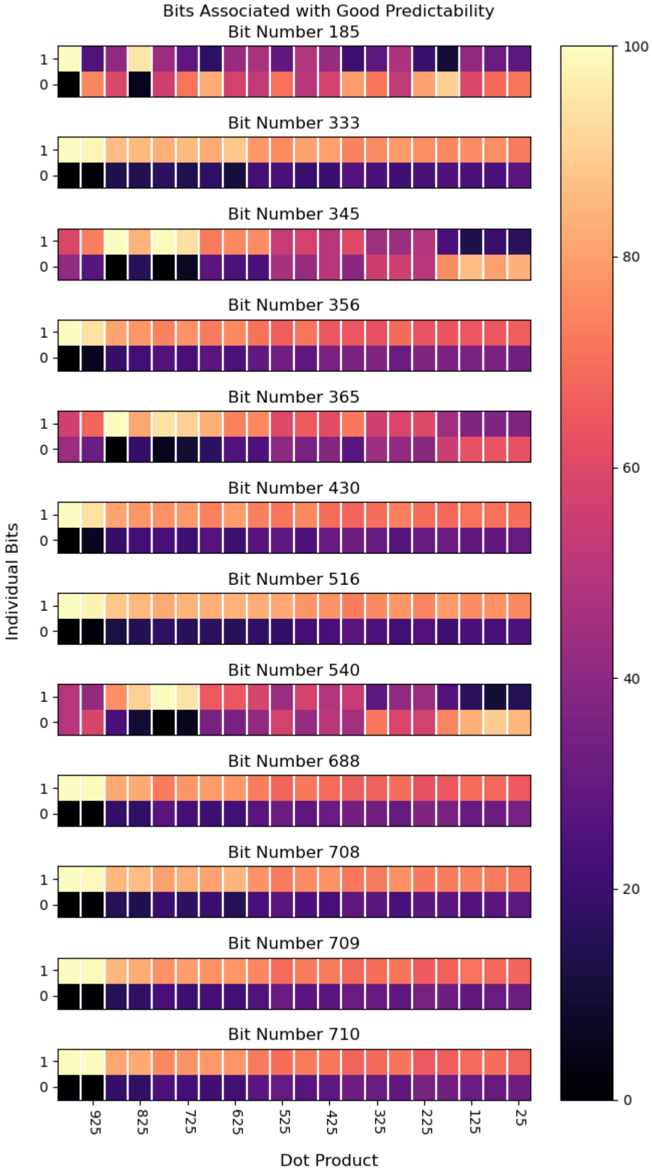
Supplement S1. Histograms of dot-score similarities for $[M+H]^+$ molecules between experimental versus predicted MS/MS spectra, using Precursor Mass × %NCE as a surrogate for explicit collision energies. Note the slight difference in curvature compared to Figure 3b.



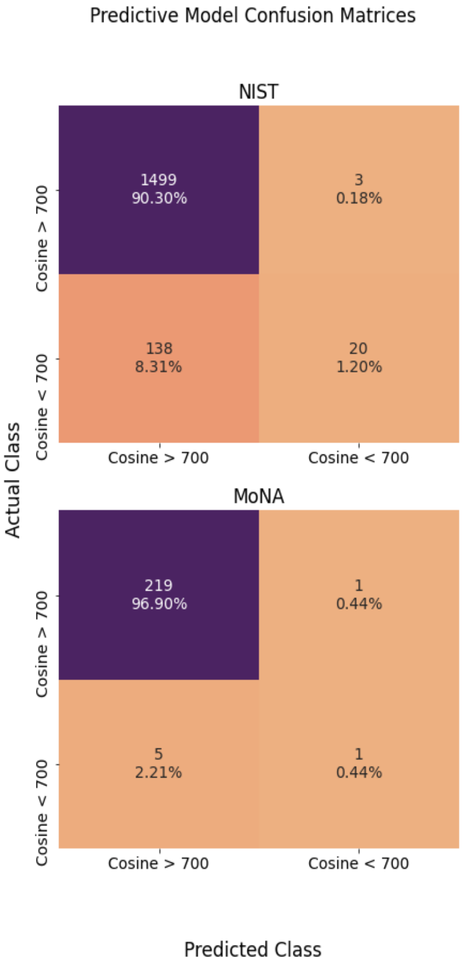
Supplement 2: 2D t-SNE embedding of CFM-ID positive training fingerprints and $[M+H]^+$/HCD-Orbitrap fingerprints. Upper panel (a) Training data set. Lower panel (b) 8,298 molecules with 35-45eV $[M+H]^+$ MS/MS spectra superimposed onto the training data (red dots). Yellow/blue color scheme indicates the normalized dot product values 0-1000 between 0 and 1.



Supplement 3: First and second dimension of PCA transformed CFM-ID positive training fingerprints and $[M+H]^+$/HCD-Orbitrap fingerprints. Upper panel (a) Training data set. Lower panel (b) 8,298 molecules with 35-45eV $[M+H]^+$ MS/MS spectra superimposed onto the training data (red dots). Yellow/blue color scheme indicates the normalized dot product values 0-1000 between 0 and 1.



Supplement S4: Histograms of bit-distributions across dot product score for each bit that was deemed to be associated with better predictability.



Supplement S5: Confusion matrices of applying CFM-ID 4.0 for compound class prediction on 20% withheld NIST20 spectra (top) and 226 natural products from the VFNPL dataset in MassBank.us (bottom).

Chapter 2: BinDiscover’s Sample-Oriented and Programmatic Meta-analysis Applied to 156,000 GC-TOF MS Metabolome Samples

Reproduced from “BinDiscover’s Sample-Oriented and Programmatic Meta-analysis Applied to 156,000 GC-TOF MS Metabolome Samples” by Parker Ladd Bremer, Gert Wohlgemuth, and Oliver Fiehn, in the Journal of Cheminformatics.

2.1 Abstract

Metabolomics by gas chromatography / mass spectrometry (GC/MS) provides a standardized and reliable platform for understanding small molecule biology. Since 2005, the West Coast Metabolomics Center at the University of California at Davis has collated GC/MS metabolomics data from over 156,000 samples and 2,000 studies into the standardized BinBase database. We believe that the observations from these samples will provide meaningful insight to biologists and that our data treatment and webtool will provide insight to others who seek to standardize disparate metabolomics studies.

We here developed an easy-to-use query interface, BinDiscover, to enable intuitive, rapid hypothesis generation for biologists based on these metabolomic samples. BinDiscover creates observation summaries and graphics across a broad range of species, organs, diseases, and compounds. Throughout the components of BinDiscover, we emphasize the use of ontologies to aggregate large groups of samples based on the proximity of their metadata within these ontologies. This adjacency allows for the simultaneous exploration of entire categories such as “rodents”, “digestive tract”, or “amino acids”. The ontologies are particularly relevant for BinDiscover’s ontologically grouped differential analysis, which, like other components of BinDiscover, creates clear graphs and summary statistics across compounds and biological metadata. We exemplify BinDiscover’s extensive applicability in three showcases across biological domains.

2.2 Introduction

Metabolomics databases can serve a variety of purposes. Some databases compile spectral libraries into repositories that users can download and incorporate into their identification workflows. Examples include MassBank of North America (https://massbank.us) [1] or Global Natural Products Social Molecular Networking (GNPS) [2]. Other examples include study-centric databases that store the metadata and observations of user-submitted studies, including the Metabolomics Workbench, [3] MetaboLights, and ReDU [3–5] databases. Others, such as the Human Metabolite Database (HMDB) [6], can be loosely described as information compilers, as they synthesize information from a range of sources. Finally, but not exhaustively, are compilation databases, that aggregate multiple, smaller, databases. A recent example of this is the COCONUT (COlleCtion of Open Natural ProdUcTs) database for natural products [7]. One of the most tantalizing research directions in metabolomics is harmonizing the archipelago of datasets in order to create a critical mass of synergistic data that can be used to achieve broad understanding of biology [8]. Within the MetabolomicsWorkbench database, users can query metabolite-centric comparisons with Venn diagrams and metabolite ratios data tables. While such queries are easily performed for individual compounds, navigating interfaces for bulk queries across different study designs is best performed via application programming interfaces (APIs) that require computational expertise that not all scientists have. HMDB compiles information from disparate sources. HMDB and related databases from the same laboratory are reliable because the information is manually curated in painstaking efforts. For both HMDB and MetabolomicsWorkbench queries, meta-analysis on bulk metabolite queries suffers because it is a retrospective attempt to harmonize compound-centric information sets across multiple biological study designs. Too much biological metadata is lost in translating either text sources (as in HMDB) or cryptic and unstructured sample/treatment naming schemes (as used during

MetabolomicsWorkbench uploads). At least for compound names, MetabolomicsWorkbench employs a database-internal naming scheme, RefMet. However, neither confidence levels for compound annotations nor concentration values are known for MetabolomicsWorkbench, due to the complexity and variety of instrument conditions.

Within individual laboratories, data may be more harmonized due to use of a specific type of instrumentation under defined protocols. Here, tools like meta-XCMS [9] or Amanida [10] allow for the generation of results that come from multiple studies. However, such tools expect a specific input data format, and such data files are not homogeneous even within a laboratory when different individuals process metabolomics raw data. Hence, even on a laboratory level, gathering data in a systematic way to render compiled results accessible to meta-analyses tools is not straightforward. Hence, classic meta-analysis is performed on a higher abstract level such as pathways or reducing to sets of synonymous names [11], instead of queries of bulk metabolite tables.

We recognize the challenge of aggregating results derived across labs and methods. We therefore posit that standardization of protocols is key to useful cross-study comparisons and queries, for both study metadata and data acquisition processes. Here, standard operating procedures are more mature in GC-MS metabolomics compared to LC-MS/MS. At UC Davis, we operate a unified, automated workflow to process metabolomics data since 2005, called BinBase. We here took a snapshot of all data processed until winter 2021 to enable large scale, multi-study meta analyses to investigate the data. We term this tool BinDiscover. It is a webtool to enable users to perform meta-analysis within minutes to extract data trends and propose hypotheses. Rather than simply comparing two types of metadata (e.g. two different species with the same organ), we assigned all metadata into ontologies to empower broad comparisons such as phylo organizations or *ontologically grouped differential analysis* (OGDA). OGDA queries transform broad questions into sets of smaller categories and then combine statistical result outputs into graphs.

2.3 Methods

The BinDiscover database draws spectral and compound information from the GC-Binbase database [12, 13]. GC-Binbase uses a bucket sort approach, where new peaks from the chromatographic runs of samples are either matched to previously annotated groupings or identified as new compounds. This bucket sort is algorithmic, with a retention index tolerance of 2,000 Fiehn RI units (approximately 2 seconds) and a weighted dot product similarity >600. All compound annotations have been manually conducted and curated over the past 20 years. Additional details such as automatic recognition of 'isomeric interferences', 'peak purity', 'peak apex ions', 'unique ion', 'signal/noise' and further parameters indicating data quality are used within GC-BinBase as output by the vendor's ChromaTOF software that was used for MS-deconvolution [12, 13]. Spectra presented in BinDiscover are consensus spectra that constantly improve spectra quality for all individual mass spectra that are assigned to a Bin (a mass spectrum with a specific unique ion and a specific retention index).

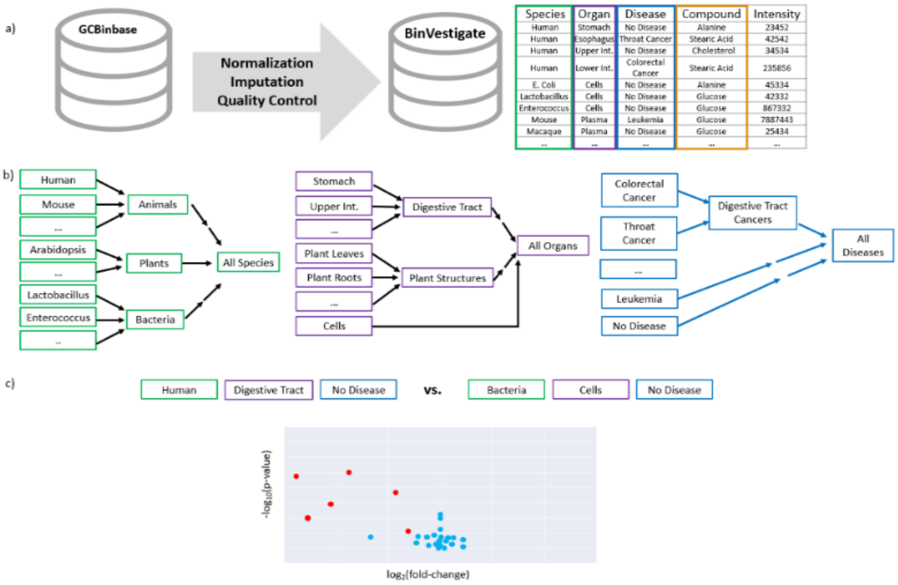
The BinDiscover database draws spectral and compound information from the GCBinbase database.[12, 13] GCBinbase uses a bucket sort approach, where new peaks from the chromatographic runs of samples are either matched to previously annotated groupings or identified as new compounds. This bucket sort is algorithmic, with a retention index tolerance of 2,000 FiehnRI units (approximately 2 seconds) and a dot product similarity of 600. New annotation group identifications have been manually conducted over the past 20 years. Spectra presented in BinDiscover are consensus spectra which are the average among all those individuals belonging to a bin.

For generating the BinDiscover database, all analyses were conducted using custom python scripts that are available in Github (see Data Availability). We heavily employed statistics routines from SciPy and the network analysis framework from NetworkX. Development was performed locally before full-data transformation on a 64 core, 128-GB RAM Amazon Web Services (AWS) virtual machine. The BinDiscover output database is deposited on a Postgres database managed by AWS. The API employed the Flask library and the frontend relied heavily on Plotly/Dash. The API and frontend were containerized with docker and deployed on AWS Elastic Beanstalk.

2.4 Results

2.4.1 BinBase is an Automatic Data Processing Database for GC-TOF Mass Spectrometry

At the UC Davis West Coast Metabolomics Center, primary metabolites are studied for 18 years using identical workflows for data acquisition and data processing using gas chromatography-time of flight mass spectrometry (GC-TOF MS). At current, five GC-TOF MS instruments are in operation. Standard operating procedures have been published extensively and have been locked and remained unchanged since 2005. Data were aligned by a set of fatty acid methyl ester internal standards, forming a stable retention index. Co-eluting mass spectra were deconvoluted and automatically de-noised by the instruments' software. This software also provided a range of metadata on the quality of data reports, from peak purity to isomeric interferences, absolute and relative ion intensities, and unique ions that best described the presence of specific metabolites within the proximity of other compounds. All this metadata was utilized by a multi-level filtering algorithm to generate a comprehensive database for both known and unidentified metabolites, called BinBase. To query biological metadata for cross-study analyses, we downloaded all data from BinBase in December 2021. This data comprised 156,174 samples that were processed into 18,290 Bins, i.e. unique mass spectra at specific retention times that used specified quantification ions. Bins included 773 identified metabolites, 39 known chemical artifacts (like polysiloxanes that originate during the GC-TOF MS process) and 15,843 spectra that were not annotated as specific chemicals. The remaining bins were accounted for by, over the course of 17 years of use, algorithmic artifacts that led to multiple bins which were merged into single metabolite values during data exports. Some Bins are associated with the same biological metabolite due to incomplete chemical trimethylsilylation, as has been reported before. We generated a workflow to investigate the biological associations for each Bin, called BinDiscover. A simplified workflow is shown in Figure 1. GC-TOF MS Compound identifications were performed within the BinBase administrative graphical user interface (GUI) (BinView) using both mass spectral spectral similarity and retention index difference between library spectra and calculated retention times. For compound identification, the FiehnLib library [12] was used in conjunction with MassBank.us and NIST20 spectra [14]. Kovats retention index values (based on alkane elution order) were automatically normalized to Fiehn retention indices that are based on fatty acid methyl ester (FAME) elution order.



Species	Organ	Disease	Compound	Intensity
Human	Stomach	No Disease	Alanine	23452
Human	Esophagus	Throat Cancer	Stearic Acid	42542
Human	Upper Int.	No Disease	Cholesterol	34534
Human	Lower Int.	Colorectal Cancer	Stearic Acid	235856
E. Coli	Cells	No Disease	Alanine	45334
Lactobacillus	Cells	No Disease	Glucose	42332
Enterococcus	Cells	No Disease	Glucose	867332
Mouse	Plasma	Leukemia	Glucose	7887443
Macaque	Plasma	No Disease	Glucose	25434
...	..	...	..	...

Figure 1: Overall workflow for BinDiscover database queries
(a) BinBase records observations from 156,174 metabolomic samples run on a GC-TOF mass spectrometer from 2005-2021. Corresponding biological metadata were curated and the resulting annotation table formed the basis of the exploratory webtool BinDiscover. (b) BinDiscover associates metabolite intensities across species, organs, and diseases. Established ontologies are used to order biological metadata for queries. For metabolites, we used the ClassyFire ontology to enable compound class-level queries. (c) Biological metadata are associated with all samples and are represented and can be queried via different ontology levels, such as “digestive system” or “bacteria”. Species, organ and disease ontologies are highlighted by colors.

2.4.2 Wrangling and Transforming Metabolomic and Biological Metadata

Each Bin is associated with biological information with respect to all studies when it was positively detected. Biological metadata were curated as detailed below, mapping sample metadata to established ontologies. We used three ontologies: 1) the National Center for Biotechnology Information (NCBI) taxonomy for species [15, 16], 2) the Medical Subject Headings (MeSH) taxonomy for organs and diseases [17], and the ClassyFire ontology for compounds [18]. In total, we used and input of 1,696 metadata combinations, defined as specific organ/species/disease triad. Across all samples, a total of 55,261,308 observed metabolites were associated with Bins, along with the full spectra and intensities of the quantification ions for each specific Bin. Each sample in BinBase is associated with information on the corresponding biological study that was conducted. Studies included both published and unpublished experiments, as data were gathered for both in-house academic purposes over the past 17 years, as well as for extramural fee-for-service projects. Biological metadata was entered into the small version of SetupX, called miniX [13]. Clients entered minimal information such as species, organs, short abstracts and sample labels that contained text for specific aspect of study designs. Since there is no universal algorithm to capture all details of biological designs in coherent and machine readable forms, the biological metadata necessarily remained heterogeneous. We therefore had to curate biological metadata and transform and normalize ion intensities.

The first step was to remove technical variance that arose from using four GC-TOF mass spectrometers and varying instrument conditions over the last 17 years. Across all studies, the exact same concentrations of FAME internal standards were used, offering us the opportunity to use FAME retention index markers as a surrogate value for instrument performance for each specific sample. Hence, we normalized metabolite intensities in each sample by the sum of the FAME ion intensities. We validated that FAME intensities showed correlations greater than 0.8 across all samples, demonstrating that they also reflected differences in GC-MS injection conditions. Next, we automatically identified problematic samples and excluded those from BinDiscover. To do this, we removed samples with poor FAME patterns, as defined as extremely high or low FAME intensity values. In addition, we removed entire biological metadata triads if they showed more than 20% failed FAME samples (Supplemental Figure S1), or if there were fewer than 10 samples in total for a specific biological metadata triad class. This data wrangling ensured that outliers did not have outsized effects on average metabolite intensities for any specific biological class. In this way, we balanced maximizing metadata coverage and maintaining statistical reliability. The distribution of sample counts is shown in Figure 2.

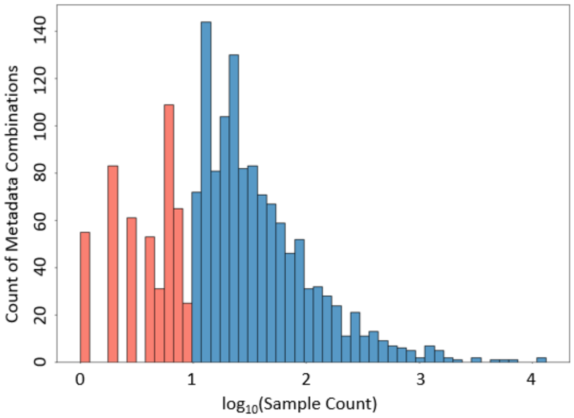


Figure 2: Sample count for all combinations of biological metadata triads. Triads with fewer than 10 samples (red) were removed to increase statistical reliability.

Next, we curated and combined metadata combinations to map metadata to established ontologies and to correct for misspellings. Metadata were manually entered into miniX over the last 17 years, leading to an array of metadata combinations for 'homo sapiens', 'homo sapien', 'Human', 'human', spellings with extra spaces or tabs, and different synonyms for either species or organs. All strings were transformed into formal ontology entries, accounting for the largest reduction of metadata combinations. Overall, 515 metadata combinations remained, concomitant with a 23.3% reduction of specimen to a total number of 119,783 samples. The next type of data wrangling accounted for correcting intensity values for unique bins. Here, we first combined bins that were best represented by a single unique metabolite. Such double bins arose over the course of 17 years because of multiple derivatization forms (with or without trimethylsilylation of amino groups) or because of incorrect retention time index calculations due to overloaded chromatograms. To obtain a single intensity for each compound for each sample, we preferentially drew intensities from the most-populated bin. If that bin was not detected, we scaled the intensity of the next-most-populated bin according to the average intensity ratio between the two bins. Overall, we retained 16,616 bins to be associated with the metadata combinations (773 metabolites with known chemical structure, and 15,843 unknowns).

Lastly, we had to impute missing values. Here, we considered four scenarios (Supplemental Figure S2). (1) A specific bin might be truly absent from a sample, and perhaps even from a full metadata combination. Indeed, most bins were absent from most biological specimens, for biological reasons. However, when calculating intensity ratios of bins between organs or species, ratio fold-changes become infinite when compounds are absent from one organ or species but present in the other. (2) On the other hand, a bin might be absent is a sample due to random errors, such as thresholds in peak detection algorithms. For example, as reported before, our BinBase algorithm uses conservative thresholds for spectral quality based on signal intensity. If a peak failed weighted dot-score similarity thresholds of 700, that bin would not be declared to be found in that sample, and missed in the BinBase database. Manual investigations or recursive backfilling might find such peak, but those approaches are not tractable. We call such peaks missing at random (MAR), while truly missing compounds (for biological reasons) can be thought of as missing not at random (MNAR). (3) Most peaks are not found 100% of all samples in a specific metadata combination, or 0% detected, i.e. always absent, but somewhere in-between. Imputing the minimum intensities for missing data has been shown to work well for MAR metabolomics data using vectors of samples or vectors of features [19]. However, if a bin is largely absent for a specific metadata combination (i.e. very rarely detected), a single outlier could grossly inflate the overall distribution. Therefore, we imputed the percentage of presence, multiplied by the minimum value of detected peaks (bins) for each metadata combination. In this way, if nearly all samples have annotations, then we simply impute the minimum. If nearly all samples lack annotations, then we impute a small number that is close to the noise level and will conserve the semi-quantitative fold change. This approach also provides a solution to the uncommon, but challenging case of ~50% present, where the data neither clearly represent MAR nor MNAR cases. (4) Lastly, if a bin is completely absent, there is no minimum value. In this case, we imputed a value such that the average for any 0% MDC will appear on the left edge of the average distribution for that compound across all metadata combinations (such that differential analysis would show an increase from the 0% case). Hence, for all bins and all metadata combination, a value is given, often as a small noise term. After normalizing, imputing, and curating distributions for all bins and all metadata combinations, we calculated derivatives of the bin intensities to empower comparisons and queries of metabolome-wide metadata combinations. Here, we calculated the averages, medians, and ratios of intensity values and stored the resulting dataset in a PostgreSQL BinDiscover database. We also computed the Welch t-test on pairs of log-transformed pairs of distributions. We chose log-transformed data here instead of directly using Welch t tests due to the known phenomenon of typically non-Gaussian distributions of metabolite values. The results of fold-change and significance calculations were stored, rather than the underlying distributions, in order to dramatically speed-up the return of query results in real-time for user queries.

2.4.3 Ontologically Grouped Differential Analysis

We here introduce ontologically grouped differential analysis (OGDA) to extract generalizations hidden within the complex data in Omics databases. In metabolomics as well as proteomics or genomics databases, studies performed by biologists or biomedical scientists comprise complex study designs that ultimately can be described in biological metadata that are associated with each sample. We summarized the biological metadata that was available to us using Medical Subject Header (MeSH) ontologies, ClassyFire chemical ontologies and NCBI species ontologies. Hence, all sample metadata were tabularized into ontological sets. OGDA then exploits ontologies to select sets based on their taxonomic proximities. In this way, samples from many studies can be compared on different ontological hierarchies on a database-wide level. Hence, intractable lists of results get transformed into condensed lists to base further analysis.

To exemplify the power of this approach, we randomly used three use cases involving queries on organ levels across species, queries across species, queries on a human disease level, and queries on metabolite levels. Figure 3 demonstrates how ontologically grouped differential analyses calculations are performed. Here, a nutritional researcher might be interested in

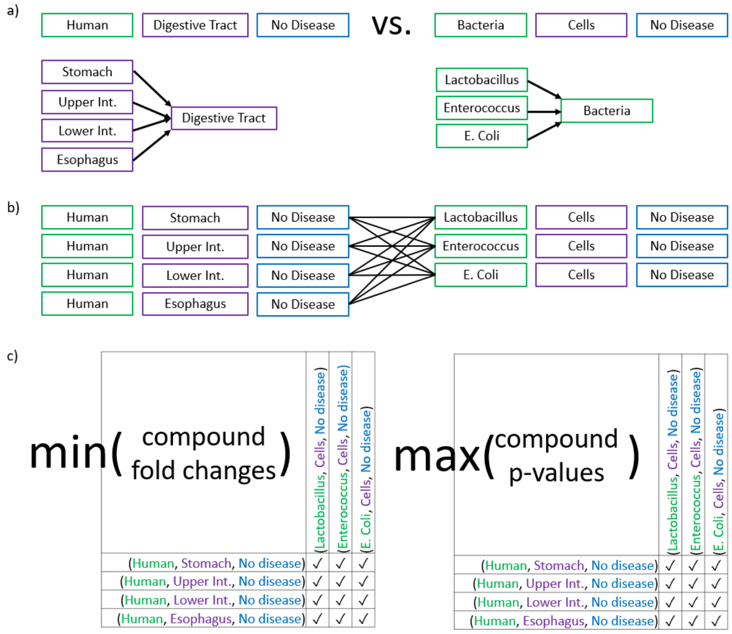

a)
Human
Digestive Tract
No Disease
vs.
Bacteria
Cells
No Disease
Stomach
Upper Int.
Lower Int.
Esophagus
Digestive Tract
Lactobacillus
Enterococcus
E. Coli
Bacteria
b)
Human
Stomach
No Disease
Human
Upper Int.
No Disease
Human
Lower Int.
No Disease
Human
Esophagus
No Disease
Lactobacillus
Cells
No Disease
Enterococcus
Cells
No Disease
E. Coli
Cells
No Disease
c)
min(compound fold changes)
(Lactobacillus, Cells, No disease)
(Enterococcus, Cells, No disease)
(E. Coli, Cells, No disease)
(Human, Stomach, No disease) ✓ ✓ ✓
(Human, Upper Int., No disease) ✓ ✓ ✓
(Human, Lower Int., No disease) ✓ ✓ ✓
(Human, Esophagus, No disease) ✓ ✓ ✓
max(compound p-values)
(Lactobacillus, Cells, No disease)
(Enterococcus, Cells, No disease)
(E. Coli, Cells, No disease)
(Human, Stomach, No disease) ✓ ✓ ✓
(Human, Upper Int., No disease) ✓ ✓ ✓
(Human, Lower Int., No disease) ✓ ✓ ✓
(Human, Esophagus, No disease) ✓ ✓ ✓

Figure 3: Schema for Ontologically Grouped Differential Analysis. Example query human digestive tract versus bacterial metabolomes.

a) All BinBase samples with metadata that ontologically map to (Human, Digestive System without Disease) were compared to samples that mapped to (Bacteria Cells without Disease).

b) Such ontology-based summary queries yield a set of biological metadata combinations that are then subjected to pairwise differential analysis.

c) For each compound, pairwise differential analysis yields a matrix of p-values and a matrix of fold changes that can be conservatively described by the maximum *p*-value and minimum fold-change, respectively. Therefore, only one point is visualized per compound in downstream volcano plots.

querying the metabolomic differences between microbial cells (bacteria) and metabolites that are found in the human digestive tract. Hence, the example query would use the BinDiscover ontology triads [(Human, Digestive Tract, no Disease) vs. (Bacteria, Cells, no Disease)] on a very generic term level (Digestive Tract and Bacteria) that by themselves would not be found in the study metadata. Yet, such words and abstractions are commonly used and understood in the literature.

To process this request, BinDiscover transforms the given request into an equivalent request that utilizes all relevant and available samples within BinBase. The ontology search yields all samples that associated with 'Digestive Tract' or 'Bacteria' and obtains a set of all nodes that are ontologically related to the requested hierarchical level ("belongs to"). Details are given in Supplemental Table S1. Importantly, stool (human feces) does not belong to the MeSH ontology of digestive system, but to the ontology "fluids and secretions". Hence, human stool samples were not included in this specific query. We then summarize all samples and transform the higher ontology level request into a list of related metadata combinations. The metabolomes of all BinBase samples that are summarized to the query groups defined in this manner are then subjected to pairwise statistical analyses. For each pair, BinDiscover creates classic results of a list of Welch-test statistical p-values and corresponding fold-changes between the two query sets. Therefore, if we have n combinations for one ontology sample set and m combinations for the comparator sample set, we yield n*m fold-changes and p values for each metabolite. The results can then be rethought of as an n*m fold change matrix and an n*m p-value matrix for every compound (Figure 3c).

Next, BinDiscover simplifies these compound matrices to exactly one aggregated p-value and associated fold-change for each compound. To extract overarching trends across the database, we conservatively estimate results for each compound across all n*m pairs. For example, if at least one bacterium showed significant higher levels of a metabolite than any human gut organ, but other bacteria would not be significant, this metabolite would not be summarized as an overall significant difference between bacterial metabolism and human gut samples. To maintain this level of conservative constraint, we therefore used the maximum p-value for each compound and the minimum fold change as boundaries. If statistical tests were overall significant, but n*m pairs showed both positive and negative fold-changes, BinDiscover represents the fold change as 0. For the example query shown in Figure 3, we ultimately did not find any chemically identified bacterial metabolite that was significantly different and at higher levels than detected in human gut metabolomes. However, the query retrieved 15 significant metabolites that were found in increased levels in human digestive system organs (Supplemental Table S2). These compounds can be summarized into vitamins, lipids, sterols, and amino acid derivatives. These metabolites are indeed not known to be directly produced by bacteria but relate to human food metabolism in a broad sense, confirming the validity of BinDiscover queries to match classic information that could be derived from scientific literature. When we conducted tests for the 773 structurally identified compounds, we obtained results in 26 seconds, at a rate of approximately 1 second per metadata combination query. When we repeated the analyses for 15,843 unknown compounds, BinDiscover retrieved results in 8 minutes and 40 seconds, at a rate of 6.9 seconds per query. Overall, we found 74 unknown compounds to be at significantly higher levels in bacteria, and 0 compounds in higher concentrations in human organs.

It is worth noting that ontologically grouped differential analysis between samples with distinct extraction methods offers important quantitative results that presence/absence analysis would overlook, however, these results must be interpreted with care. In general, we can expect that comparisons among groups of species or diseases, no matter how distant in their ontologies (human blood vs. fish blood), would be quantitative because the mass or volume of sample and other analytical method parameters remain fixed. However, comparisons that aggregate organs may involve the grouping of organs that have orthogonal extraction methods, e.g., plasma volume compared to tissue mass. We believe that such comparisons are important to allow because, for example, basic presence/absence analysis would represent the small amount of sucrose in human blood in the same way as the large amount of sucrose in plants. However, there is some bias that comes from each extraction methods, so those results having larger fold changes can be interpreted with higher confidence. Indeed, one of the goals of ontologically grouped differential analysis is to conservatively minimize the fold changes for each compounds among the set of requested organs in order to increase confidence in these quantitative findings.

2.4.4 Case Study 1 –Exploring Food Metabolomes

Metabolomics is a hypothesis generating tool. Databases must prove their usefulness by serving specific queries. We here provide four use cases to highlight how biologists or biomedical scientists might use the BinDiscover webtool. To enable rapid exploration of the metabolome data on differences between species, organs and diseases, users define ontologically grouped differential analysis on biological metadata, or explore data from a compound-centric pool. The webtool relies on commonly accepted statistics and clear graphics to obtain rapid insights into major metabolic differences in biological comparisons (Figure 4).

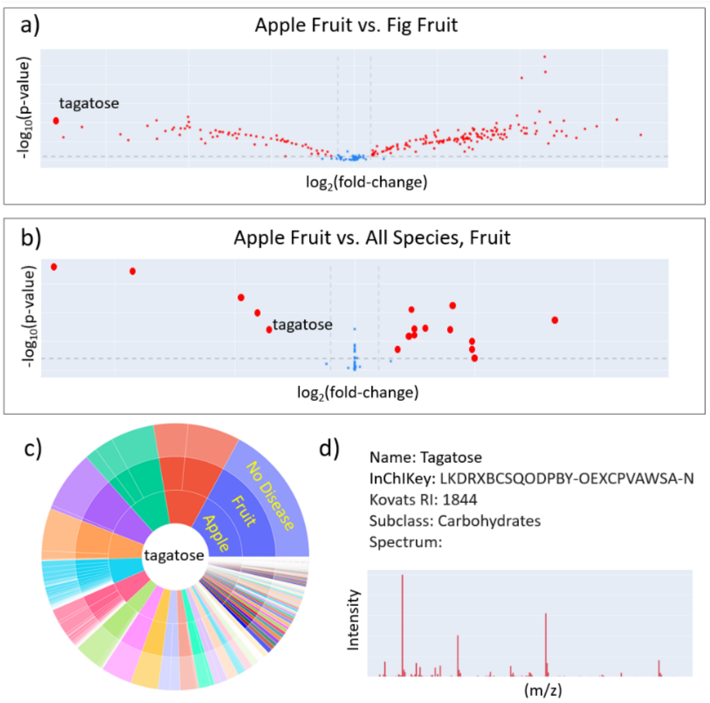


Figure 4: Queries in BinDiscover give novel biological insights.

a) Comparing the metabolome of a specific organ across two different species, here: apple vs. fig fruits, yields many differences.
b) Comparing that specific organ (apple fruit) against the same organ of all species constrains overall differences to a few metabolites.
c) One differential apple metabolite, tagatose, was then queried and found to be the most abundant in apple fruits compared to all other species/organ combinations across the metabolome database.
d) Chemical information for tagatose is then given as mass spectrum, quantification mass, international chemical identifier, retention index and chemical class ontology.

We first envisioned a nutritional researcher exploring this tool. Food metabolomes and dietary biomarkers are increasingly recognized as important contributor to disease [20, 21]. As a starting point, a researcher might wonder why “an apple a day keeps the doctor away”? The user might choose to compare an apple to any other fruit, in this case a fig (Figure 4a). Such a comparison is valid and produces a large amount of information comparing these two fruits. When hovering over the online graph (Figure 4a), each dot represents an individual compound. Tagatose is highlighted here as the metabolite that showed the largest difference in apple over fig fruits. At this point, the user might want to increase the query and compare apple fruits to all fruits in the BinDiscover database (currently 26 fruits). In this way, researchers find out which metabolites are uniquely increased, or decreased, in apples compared to all other fruits. Interestingly, this query still showed tagatose to be found in higher levels in apples than in other fruits (Figure 4b), with notably fewer total metabolic differences compared to the differential analyses of the apple/fig pair. The online data tables that correspond to the visual charts show all differential metabolites and guide users to compound-specific follow up queries. Here, the envisioned nutritionist user would find a sunburst diagram and chemical metadata (Figures 4c, 4d). The sunburst diagram shows that indeed, tagatose showed the highest intensity in apple fruits across all species/organ/disease metadata combinations. Such finding may be interesting because tagatose, despite containing 92% of the sweetness of sucrose, provides only approximately 1/3 of the calories compared to sucrose [22]. Moreover, tagatose does not increase insulin in patients with Type-2 diabetes [23]. Researchers might use this finding as starting point for additional research, e.g. apple genomic tools to increase tagatose contents in other fruits or even in apple cultivars.

2.4.5 Case Study 2 – Cancer Metabolism

Next, we envisioned a cancer biologist interested using BinDiscover. Here, we highlight how repeatedly utilizing the BinDiscover differential analysis tool empowers isolating both identified and unknown compounds that distinguish cancer metabolic phenotypes from corresponding non-malignant analogs, and how different cancer cells and tissues would reveal specific alterations that are not prevalent in other cancers. Specifically, for proof of principle, we obtained three metadata combinations for lung, liver and pancreas cancers, each compared against their non-malignant counterparts. In each utilization, we obtained a set of compounds. By taking the intersection of the resultant sets, a cancer biologist may find compounds that are differentially regulated in all cancer types (Figure 5a), and compounds that would be specific for each cancer type. We found 11 identified compounds that intersected with all cancers, such as increases in glutamine, dehydrated glutamine, n-acetylglutamate, and methylmalonic acid (Supplemental Data S1). These compounds can be associated with tricarboxylic acid (TCA) cycle activity, specifically for anaplerotic reactions supplanting carbon into the TCA cycle. For example, excess glutamine is known to be heavily used in cancerous cells in particular via glutamine dehydrogenase to generate glutamic acid, which is then converted to alpha-ketoglutarate [24]. Similarly, the branched-chain amino acid degradation product methylmalonic acid is converted to the TCA metabolite succinyl-CoA in an anaplerotic reaction, as cancer cells are deprived of mitochondrial acetyl-CoA due to lowered activity of pyruvate dehydrogenase. Another typical cancer biomarker found by this combined BinDiscover differential analysis was increased pyrophosphate, which is associated with increased kinase activity and cell growth [25]. Additionally, we explored apparent compounds that might distinguish the three cancer types investigated here (Supplemental Data S1).

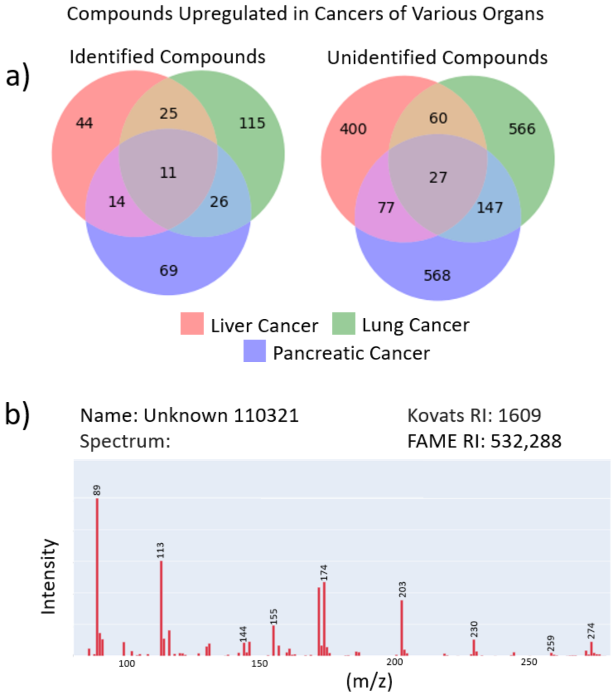


Figure 5: Sequential queries extract unknown metabolites associated with cancer metabolism.
(a) Integrating results from three BinDiscover queries comparing liver, lung and pancreas cancer studies with and without cancer yields three sets of compounds. Results are separated here between identified and unknown compounds.
(b) BinDiscover gives spectra and chemical metadata to enable chemists to utilize unknown compounds in their own studies, either for targeting these compounds in their own studies or for compound identification. Here, unknown 110321 is displayed.

For example, in pancreatic cancers we observed increased amounts of all four forms of tocopherol, also known as Vitamin E. Vitamin E has been proposed to be associated with decreased pancreatic cancer risk, in opposite to our findings [26]. We also noticed several dipeptides to be increased specifically in pancreatic cancer studies, such as cystine, homocystine, and dialanine (Supplemental Data S1)., indicating enhanced import of peptides as supplement nutrients or increased proteolysis. For lung cancer studies, we noted specific increased levels in alpha-keto acids such as 2-ketoisocaproic acid and 2-ketoisovaleric acid along with corresponding alpha-hydroxy acids like 2-hydroxyvaleric acid and 2-hydroxyglutaric acid (Supplemental Data S1). These compounds are usually associated with increased use of amino acid degradation. Lung cancer studies were also marked by elevated acetylations, including N-acetyl-glycine, -mannosamine, -serine, -aspartate and –putrescine (Supplemental Data S1). The latter two compounds have previously been proposed as biomarkers of lung cancer progression [27, 28]. For liver cancer, the most apparent specific trend that was absent in prostate- or lung cancer studies was the abundance of glycolytic intermediates galactose-6-phosphate, fructose-6-phosphate, fructose-1,6-bisphosphate, 3-phosphoglycerate, 2-phosphoglycerate, and phosphoenolpyruvate, along with the pentose phosphate cycle metabolite ribulose-5-phosphate, and generic sugar phosphates inositol-4-monophosphate and N-acetylglucosamine-6-phosphate (Supplemental Data S1). An increased glycolytic flux is not only well-known for liver cells [29] but also a generic hallmark of cancer and, according to studies available in BinBase, much elevated in liver cancers compared to lung- or pancreatic cancers. Apart from classic known metabolites, chemists and metabolomic researchers might assist cancer researchers in finding novel clues towards metabolic dysregulation in cancer. Here, we found more than 1,500 unidentified compounds that were specific for the three cancer types, and 27 unknown compounds that were commonly differentially regulated in all cases (Supplemental Data S2). The chemical metadata for a randomly chosen example from the 27 common dysregulated compounds, unknown 110321, is shown in Figure 5b. As BinBase gives both spectra, quantification ions and retention indices, other metabolomics researchers can readily use that information to target these unidentified cancer biomarkers in their studies. Secondly, spectra of novel biomarkers serve as starting point for compound identification. Compound 110321 shows a range of even-numbered fragment ions such as m/z 144, 172, 174, which are typical of primary amines, plus high m/z ion clusters around m/z 274 and m/z 230 which also point to the presence of nitrogen moieties. The spectrum lacks m/z 117, a typical fragment for carboxylic acids and sugars. The retention indices reveal a compound that has a boiling point similar to other amino acids, and hence, compound 110321 can be classified as a primary amine with additional functional groups such as a secondary amine. With chemical ionization/accurate mass spectrometry, the full structure would then become identifiable [30].

2.4.6 Diversity of Bacterial Metabolism

A microbiologist might use BinDiscover to study bacterial metabolism across species, for example, as background for synthetic biology supplanting traditional synthetic routes [31]. Likewise, the gut microbiome is gaining focus as the source of many endogenous metabolites as well as the origin of phenotypes in pharmaceutical testing [32]. The diversity of potential of bacterial metabolic function is of interest, and we therefore generated a clustered heatmap as phylo-metabolomic tool in BinDiscover (Figure 6). These phylo-metabolomic heatmaps utilize the chemotaxonomic presence of all detected metabolites (in columns) against the specified combination of taxa (in rows) using hierarchical clustering.

Such heatmaps can be used to delineate specific outlier species, as shown for highlighted section #1 in Figure 6a for methylomonas denitrificans which uses methane metabolism as its carbon source. A detailed BinDiscover comparison of this species against all other bacteria (Figure 6b) revealed much elevated production of squalene [33] and inosine-5-phosphate concomitant with reduced ribose biosynthesis. Section #2 in Figure 6a highlighted a cluster of compounds that were unique to synechococcus elongatus, a blue-green photosynthetic algae, that produces the pigment trans-phytol in addition to various alkanes that were absent in all other bacteria in BinDiscover. Section #3 contained ubiquitously present metabolites such as fatty acids, amino acids, and nucleic acids , which therefore did not contribute to bacterial classifications. Finally, Section #4 marked a section of metabolites that linked the human mouth

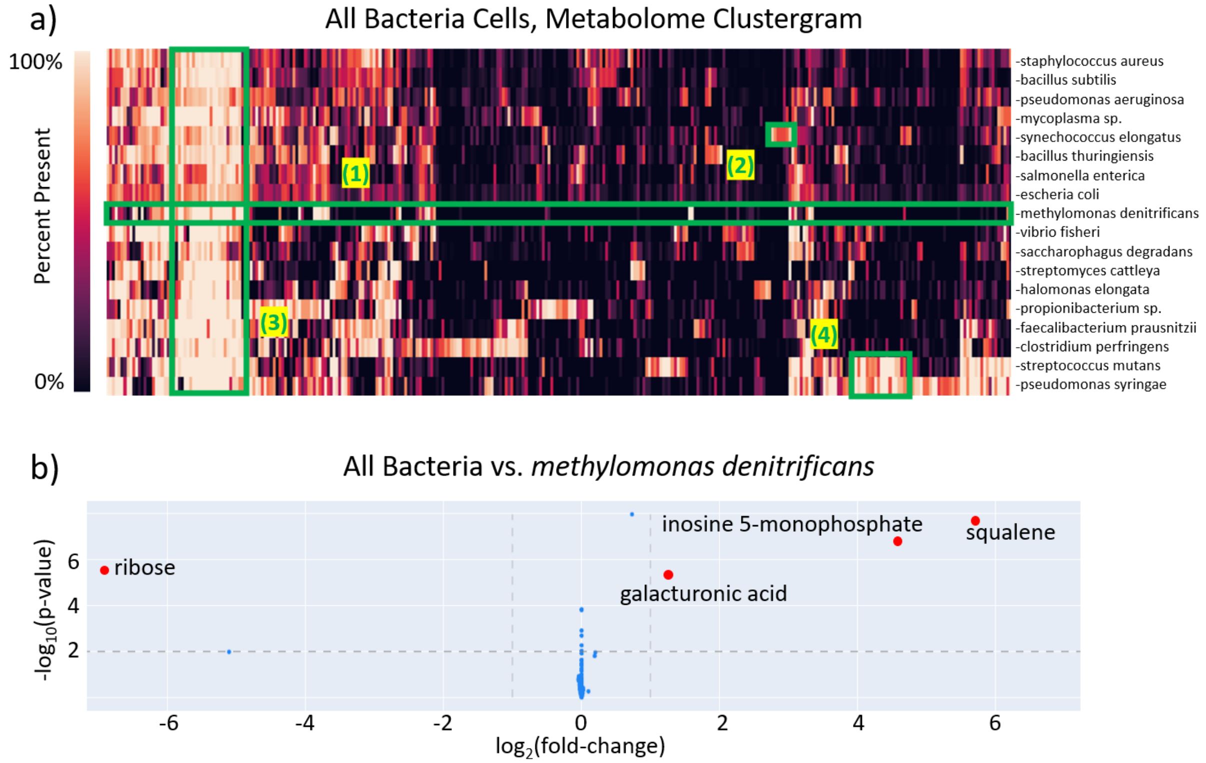


Figure 6: Comparison of the gas chromatography metabolomes of bacteria in BinDiscover.
(a) A heatmap of all metabolites in BinDiscover against all available bacteria species. Matrix entry color is determined by percent presence of that metabolite in that species. Four regions of interest (1)-(4) are highlighted in green and discussed in the text. (b) A differential comparison of metabolomic abundances in bacteria species against the methane-metabolizing species *methylomonas denitrificans*.

bacterium streptococcus mutans and the plant pathogen pseudomonas syringae. Observed metabolites included tryptamine and indole-3-acetate, which have been included in publications studying the host-pathogen relationship [34, 35]. In general, the diversity present in these bacterial metabolomes reflects the niches that are to be expected [36]. We focus on biological concepts in the case studies presented here because we BinDiscover itself is intrinsically informatics-oriented. An additional case study that is more oriented toward cheminformatics where we showcase the relevance of unknown compounds is shown in Supplemental Figure 4.

2.5 Discussion

BinDiscover effectively enables rapid meta-analysis of metabolomics information with the objective of ease of use for biological scientists, focusing on both capability and breadth of metabolome coverage. However, post-hoc retrieval and harmonization of biological sample metadata were challenging. To our knowledge, there are scant examples of usable interfaces that correctly map biological study designs, covering not only species and organs, but also treatments, time courses or disease phenotype dimensions of study designs. Hence, two of the most important issues concerning to sample metadata were the inconsistency of metadata terminology used when capturing biology study information in our miniX study design DB and the omission of fine-grained biological study design details. Inconsistent metadata terminology describes the informality by which samples were labeled by biologists who were sending studies to the UC Davis West Coast Metabolomics Center over the past 18 years. While for domain experts, a word such as “C57BL/6” might sufficiently describe a specific mouse wildtype, even for this classic example there are different laboratory strains such as B6J (or B6/J) for mice from the Jackson laboratory, and similar strain variants from other laboratories. The same is true across other biological domains, from cell types to fine grained descriptions of tissues and organs. Closely related to this omission of details was the difficulty to capture the essence of biological studies, such as the use of specific gene knockouts or drug treatments. Information was sometimes delivered by biologists in text formats and through sample lists, but usually domain-specific acronyms were used that were intractable to compile retrospectively throughout the diversity of 2,000 different studies in our GC/MS database.

An alternative approach to programmatically capture study design details might use named-entity recognition combined with NoSQL/GraphDB records. An entity recognition system might start with a vocabulary of known ontologies, but would need to be capable to expand an internally consistent vocabulary to capture arbitrary descriptions. While a graph approach allows for robust and dynamic descriptions of samples and their relationships, the named-entity recognition avoids problematic curation. Yet, a graph-based interface would present significant complexities for users, especially biologists who are asked to submit their study information. Initial efforts led to frustrations and overwhelmed potential users. An alternative approach to capture study metadata is to pre-define motifs of study designs and coerce study design details into those motifs. While many fine-grained study details (and, hence, sample metadata resolution) get lost in coarse motif-based GUI forms, such tools may dramatically simplify the procedure for biological clients. While not comprehensive, reducing the burden on researchers can dramatically increase the likelihood that individuals will contribute these details when using metabolomics (or other –omics) services.

Importantly, ontologically grouped differential analysis offers important quantitative results that simple presence/absence analysis ignores. For example, sucrose is present and detected at low amounts by untargeted GC-MS metabolomics in human blood. However, it would be wrong to conclude that sucrose is a major constituent in human samples, compared to plant samples. Here, semi-quantitative assessments are possible in GC-MS based metabolomics for two reasons: (a) Electron ionization at 70 eV is standardized in GC-MS for 60 years, and it does not suffer from suppression by co-eluting compounds, unlike electrospray processes used by LC-MS/MS. (b) Extraction, derivatization, injection, detection and data processing methods at UC Davis have been standardized to assure that chromatograms were never overloaded (i.e. avoiding peak saturations), but also never blank (ensuring that the most abundant peaks in specific samples were reaching detector saturation). Hence, semi-quantification was assured by both data acquisition and data processing procedures, including using the exact same concentration of (fatty acid methyl ester) internal standards over the past 18 years. Nevertheless, of course despite these precautions, quantitative must be interpreted with care. For example, comparisons across organs may include biofluids versus tissues, i.e. different units of biomass. In addition, different solvent extraction efficiencies across different tissues or biofluids may introduce bias. Hence, quantitative comparisons that yield large fold-change differences can be interpreted with higher confidence than small differences. Indeed, one of the goals of ontologically grouped differential analysis is to conservatively minimize the fold changes for each compounds among the set of requested organs in order to increase confidence in these quantitative findings. However, for biological metadata combinations with few samples and few studies, quantitative comparisons are less robust than for differential analyses for which there were thousands of sample data available in BinBase.

Additionally, BinDiscover is built on top of a snapshot of BinBase data as they were in December 2021. As BinBase continues to expand, novel compounds get added. For example, in November 2022, we reliably detected the presence of carboxymethylcysteine (Figure S3) for the first time, in a study analyzing bovine muscle tissues, treated with inhibitors against oxidative phosphorylation complexes. Hence, compounds in BinBase that were formally recorded at later dates might have been present infrequently or at low abundance, and were therefore not sufficiently validated for induction into BinBase. To overcome such metadata incongruencies, BinDiscover focuses on high-level analyses of species and organ queries using ontological differential analysis, sunburst diagrams, and phylo-metabolomic trees. Users obtain the number of samples for each query and metadata combination with the notion that the estimation of median metabolite levels gets more robust the more samples are included in comparisons. Even more specific metadata comparisons may provide insights into metabolic differences if users focus on compounds with sufficiently large fold changes.

BinDiscover aims at hypothesis-generating and data exploration. We are motivated to discover unexpected findings, and, contextually, are relatively unconcerned about false-positives (type I error). Similarly, we do not use Fisher's method to aggregate p-values when combining metadata combinations, because we here compare completely different hypotheses in each pairwise comparison, using ontologically grouped differential analysis.

Future versions of BinDiscover may become incrementally updated by data from new studies including from public contribution. A related tool, ADAP-KDB, perpetually retrieves and updates a user-explorable consensus library of spectra from the MetabolomicsWorkbench.[37] ADAP-KDB does not use static snapshots and focuses on a community-contributed source of data, but it is clearly spectrum-centric and assisted by the de-facto standards in GC-MS. We hope that there will be community-wide efforts to further standardize standard operating procedures for metadata definitions, sample extraction, data acquisition, and data processing to confidently include broader contributions from the community into GC-Binbase.

It is critical that meta-analysis systems for metabolomics focus on *samples, not on studies*. In this way, metadata of samples can be repurposed for new biological comparisons, conducted from a library of analyzed samples. At current, meta-analysis often relies on combining studies that had approximately the same intention, which dramatically reduces the ways in which data can be re-used. As a part of this grand unification of metabolomics data, we hope that standardization in metabolomics will improve. The inclusion of *internal standard kits* as matrix spikes into samples before extraction could serve as a check of instrument state as well as allow for semi-quantitative, on-the-fly calibrations that would dramatically improve the level of confidence in sample-to-sample integration.

2.6 Conclusion

BinDiscover is a webtool based on a 156,000 sample GC-TOF database that has accumulated data since 2005. We curated this dataset by removing samples that failed quality control checks, imputing missing values, and mapping the metadata as well as identified metabolites to established ontologies. We showed that our webtool enables rapid hypothesis generation and trend extraction in order to transform machine-sized databases into human-sized, actionable simplifications. Our tool provides components that enable the examination of large swaths of data simultaneously as well as the ability to focus on individual compounds. We enable the comparison of multiple types of species and organs using chemotaxonomy trees and ontologically grouped differential analysis, but also the visualization of single compounds with sunburst diagrams or chemical metadata. One novel approach to data analysis, ontologically grouped differential analysis, uses external ontologies, such as the NCBI species taxonomy or MeSH hierarchy, to create groups of samples that match generic terms. The logic of ontologically grouped differential analysis can be applied to arbitrary metadata or features, as long as a corresponding ontology exists, so we believe that it has applicability for other -omics as well. Hence, queries can be grouped along the ontology axes, for example, to compare "rodent blood" against "human blood" or similar broad groupings. Metabolomics is now mature enough to empower re-using data deposited in large scale databases derived from standardized methods, with the explicit aim to perform meta-analyses across disparate studies. We strongly emphasize the importance of metabolome standardization initiatives that are critically needed for cross-study and cross-species data comparisons. Indeed, this type of sample-centric data collection could form training sets for large scale phenotype-predicting machine learning models. We found that one of the most challenging aspects in the creation of this metanalysis tool was curating and harmonizing the swaths of metadata submitted by biologist clients. We envision working toward simplified, yet powerful metadata capture systems.

2.7 Data and Software Availability

All code is available at https://github.com/metabolomics-us/bindiscover. The complete distributions are available at https://zenodo.org/record/7982901. The derived data are available via https://bindiscover.metabolomics.us.

2.8 References

1. MassBank of North America. https://massbank.us/. Accessed 24 Oct 2022

2. Wang M, Carver JJ, Phelan VV, et al (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597

3. Metabolomics Workbench : Home. https://www.metabolomicsworkbench.org/. Accessed 24 Oct 2022

4. Haug K, Cochrane K, Nainala VC, et al (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res 48:D440–D444. https://doi.org/10.1093/nar/gkz1019

5. ReDU: a framework to find and reanalyze public mass spectrometry data | Nature Methods. https://www.nature.com/articles/s41592-020-0916-7. Accessed 5 Jun 2023

6. Wishart DS, Guo A, Oler E, et al (2022) HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res 50:D622–D631. https://doi.org/10.1093/nar/gkab1062

7. Sorokina M, Merseburger P, Rajan K, et al (2021) COCONUT online: Collection of Open Natural Products database. J Cheminformatics 13:2. https://doi.org/10.1186/s13321-020-00478-9

8. Mak TD, Goudarzi M, Laiakis EC, Stein SE (2020) Disparate Metabolomics Data Reassembler: A Novel Algorithm for Agglomerating Incongruent LC-MS Metabolomics Datasets. Anal Chem 92:5231–5239. https://doi.org/10.1021/acs.analchem.9b05763

9. Tautenhahn R, Patti GJ, Kalisiak E, et al (2011) metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data. Anal Chem 83:696–700. https://doi.org/10.1021/ac102980g

10. Llambrich M, Correig E, Gumà J, et al (2022) Amanida: an R package for meta-analysis of metabolomics non-integral data. Bioinformatics 38:583–585. https://doi.org/10.1093/bioinformatics/btab591

11. Goveia J, Pircher A, Conradi L-C, et al (2016) Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. EMBO Mol Med 8:1134–1142. https://doi.org/10.15252/emmm.201606798

12. Kind T, Wohlgemuth G, Lee DY, et al (2009) FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas

chromatography/mass spectrometry. Anal Chem 81:10038–10048. https://doi.org/10.1021/ac9019522

13. Scholz M, Fiehn O (2007) SetupX--a public study design database for metabolomic projects. Pac Symp Biocomput Pac Symp Biocomput 169–180

14. NIST 20 MS/MS Library (2020). https://www.sisweb.com/software/nist-msms.htm#2. Accessed 4 Mar 2021

15. Sayers EW, Cavanaugh M, Clark K, et al (2019) GenBank. Nucleic Acids Res 47:D94–D99. https://doi.org/10.1093/nar/gky989

16. Schoch CL, Ciufo S, Domrachev M, et al (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J Biol Databases Curation 2020:baaa062. https://doi.org/10.1093/database/baaa062

17. Rogers FB (1963) Medical subject headings. Bull Med Libr Assoc 51:114–116

18. Djoumbou Feunang Y, Eisner R, Knox C, et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminformatics 8:61. https://doi.org/10.1186/s13321-016-0174-y

19. Kokla M, Virtanen J, Kolehmainen M, et al (2019) Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics 20:492. https://doi.org/10.1186/s12859-019-3110-0

20. Scalbert A, Brennan L, Manach C, et al (2014) The food metabolome: a window over dietary exposure. Am J Clin Nutr 99:1286–1308. https://doi.org/10.3945/ajcn.113.076133

21. FooDB. https://foodb.ca/about. Accessed 28 Nov 2022

22. BeMiller JN (2019) 19 - Carbohydrate and Noncarbohydrate Sweeteners. In: BeMiller JN (ed) Carbohydrate Chemistry for Food Scientists (Third Edition). AACC International Press, pp 371–399

23. Donner TW, Wilber JF, Ostrowski D (1999) D-tagatose, a novel hexose: acute effects on carbohydrate tolerance in subjects with and without type 2 diabetes. Diabetes Obes Metab 1:285–291. https://doi.org/10.1046/j.1463-1326.1999.00039.x

24. Martínez-Reyes I, Chandel NS (2021) Cancer metabolism: looking forward. Nat Rev Cancer 21:669–680. https://doi.org/10.1038/s41568-021-00378-6

25. Rao F, Xu J, Fu C, et al (2015) Inositol pyrophosphates promote tumor growth and metastasis by antagonizing liver kinase B1. Proc Natl Acad Sci 112:1773–1778. https://doi.org/10.1073/pnas.1424642112

26. Peng L, Liu X, Lu Q, et al (2015) Vitamin E Intake and Pancreatic Cancer Risk: A Meta-Analysis of Observational Studies. Med Sci Monit Int Med J Exp Clin Res 21:1249–1255. https://doi.org/10.12659/MSM.893792

27. Lou T-F, Sethuraman D, Dospoy P, et al (2016) Cancer-Specific Production of N-Acetylaspartate via NAT8L Overexpression in Non-Small Cell Lung Cancer and Its Potential as a Circulating Biomarker. Cancer Prev Res Phila Pa 9:43–52. https://doi.org/10.1158/1940-6207.CAPR-14-0287

28. Liu R, Li P, Bi CW, et al (2017) Plasma N-acetylputrescine, cadaverine and 1,3-diaminopropane: potential biomarkers of lung cancer used to evaluate the efficacy of anticancer drugs. Oncotarget 8:88575–88585. https://doi.org/10.18632/oncotarget.19304

29. Rajas F, Gautier-Stein A, Mithieux G (2019) Glucose-6 Phosphate, a Central Hub for Liver Carbohydrate Metabolism. Metabolites 9:282. https://doi.org/10.3390/metabo9120282

30. Lai Z, Tsugawa H, Wohlgemuth G, et al (2018) Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods 15:53–56. https://doi.org/10.1038/nmeth.4512

31. Gandhi SG (2019) Chapter 8 - Synthetic Biology for Production of Commercially Important Natural Product Small Molecules. In: Singh SP, Pandey A, Du G, Kumar S (eds) Current Developments in Biotechnology and Bioengineering. Elsevier, pp 189–205

32. Tsunoda SM, Gonzales C, Jarmusch AK, et al (2021) Contribution of the Gut Microbiome to Drug Disposition, Pharmacokinetic and Pharmacodynamic Variability. Clin Pharmacokinet 60:971–984. https://doi.org/10.1007/s40262-021-01032-y

33. de la Torre A, Metivier A, Chu F, et al (2015) Genome-scale metabolic reconstructions and theoretical investigation of methane conversion in Methylomicrobium buryatense strain 5G(B1). Microb Cell Factories 14:188. https://doi.org/10.1186/s12934-015-0377-3

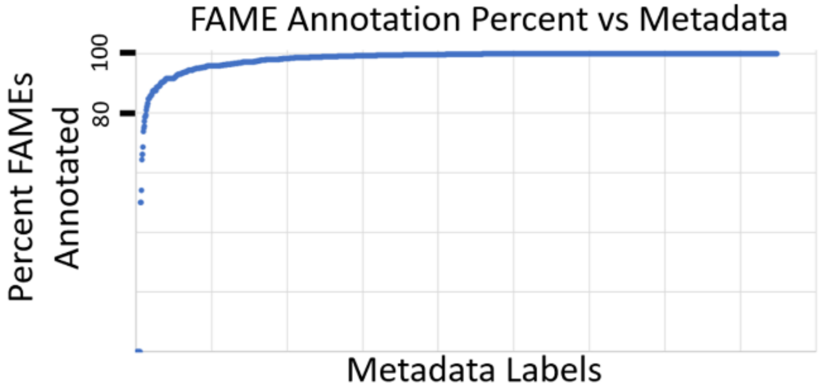
34. Edlund A, Garg N, Mohimani H, et al (2017) Metabolic Fingerprints from the Human Oral Microbiome Reveal a Vast Knowledge Gap of Secreted Small Peptidic Molecules. mSystems 2:e00058-17. https://doi.org/10.1128/mSystems.00058-17

35. McClerklin SA, Lee SG, Harper CP, et al (2018) Indole-3-acetaldehyde dehydrogenase-dependent auxin synthesis contributes to virulence of Pseudomonas syringae strain DC3000. PLOS Pathog 14:e1006811. https://doi.org/10.1371/journal.ppat.1006811

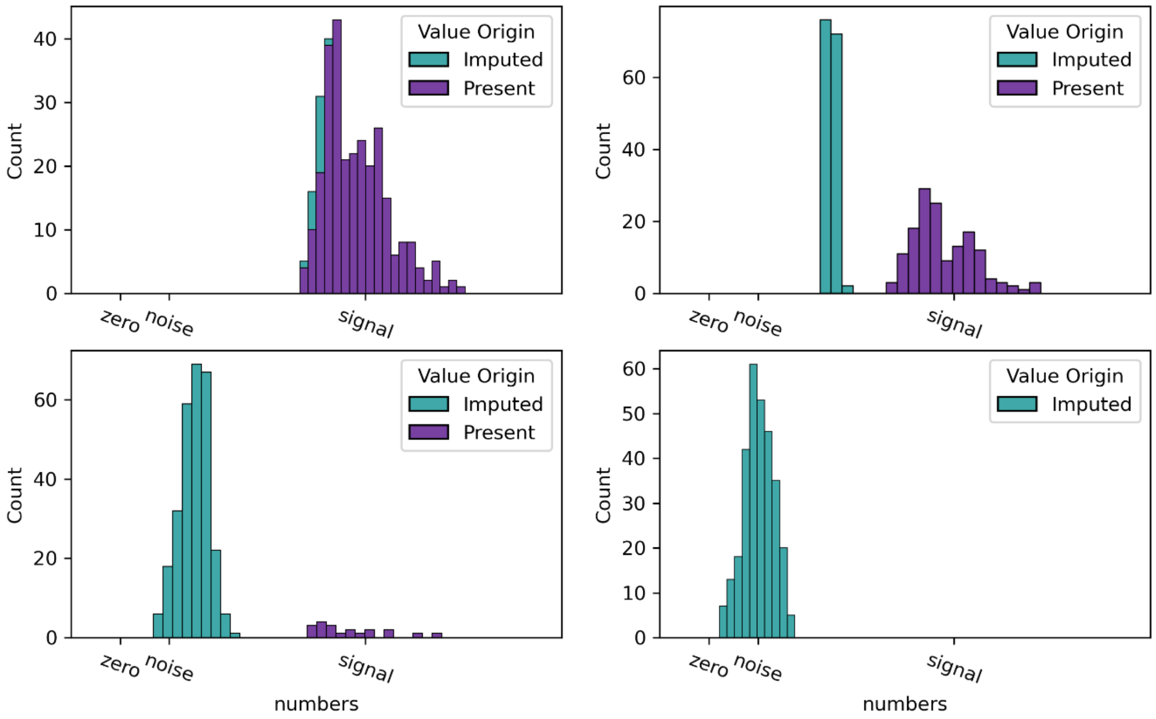
36. Gargallo-Garriga A, Sardans J, Granda V, et al (2020) Different “metabolomic niches” of the highly diverse tree species of the French Guiana rainforests. Sci Rep 10:6937. https://doi.org/10.1038/s41598-020-63891-y

37. Smirnov A, Liao Y, Fahy E, et al (2021) ADAP-KDB: A Spectral Knowledgebase for Tracking and Prioritizing Unknown GC–MS Spectra in the NIH's Metabolomics Data Repository. Anal Chem 93:12213–12220. https://doi.org/10.1021/acs.analchem.1c00355

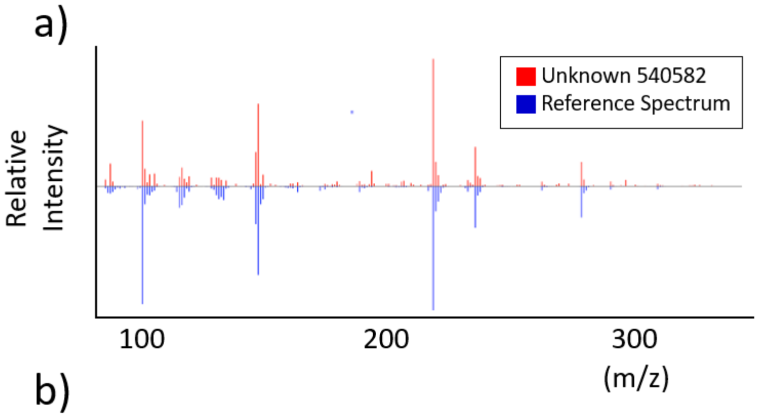
2.9 Supplemental



Supplemental Figure S1: Frequency of FAME detections across all BinBase study samples (1,696 metadata triplet combinations). Samples were removed if they belonged to metadata combinations with a FAME annotation frequency less than 80%.



Supplemental Figure S2: Concept graphs of imputed data (to replace missing data) versus experimentally detected data. These concept graphs use simulated data, not data obtained from BinBase, to illustrate four different scenarios of how missing metabolite data in BinBase metadata combinations might be overcome. (a) **Top left.** Data missing at random might have been missed by experimental causes, such as data processing thresholds or instrument malfunctions. (b) **Bottom left.** Data missing not at random, but with many missing data and few detected data for metadata combinations. Examples could be for metabolites that were generally found at low levels and for which experimental limits-of-detections caused data missingness in many samples, but not in all samples. Another cause for data missingness in this scenario is metabolites that are synthesized or detected only in specific conditions, such as pharmaceutical drugs in human plasma that may be found in high levels in some subjects, but very low or absent in most others. (c) **Top right**. Data missing due to unexplained differences in study design parameters that impact absence or presence of metabolite not at random. Example for such rare cases could be different animal feeds used in rat plasma metabolome studies, or age-related metabolites that were present in one study but not in another. Typically, such not-at-random gross missingness might only be found in metadata combinations that have a small total sample count. (d) **Bottom right.** Metabolites that were completely absent in specific metadata combinations. Yet, to compute fold-changes for differential analyses, instrument noise levels are used to impute missing data around defined variance**.**



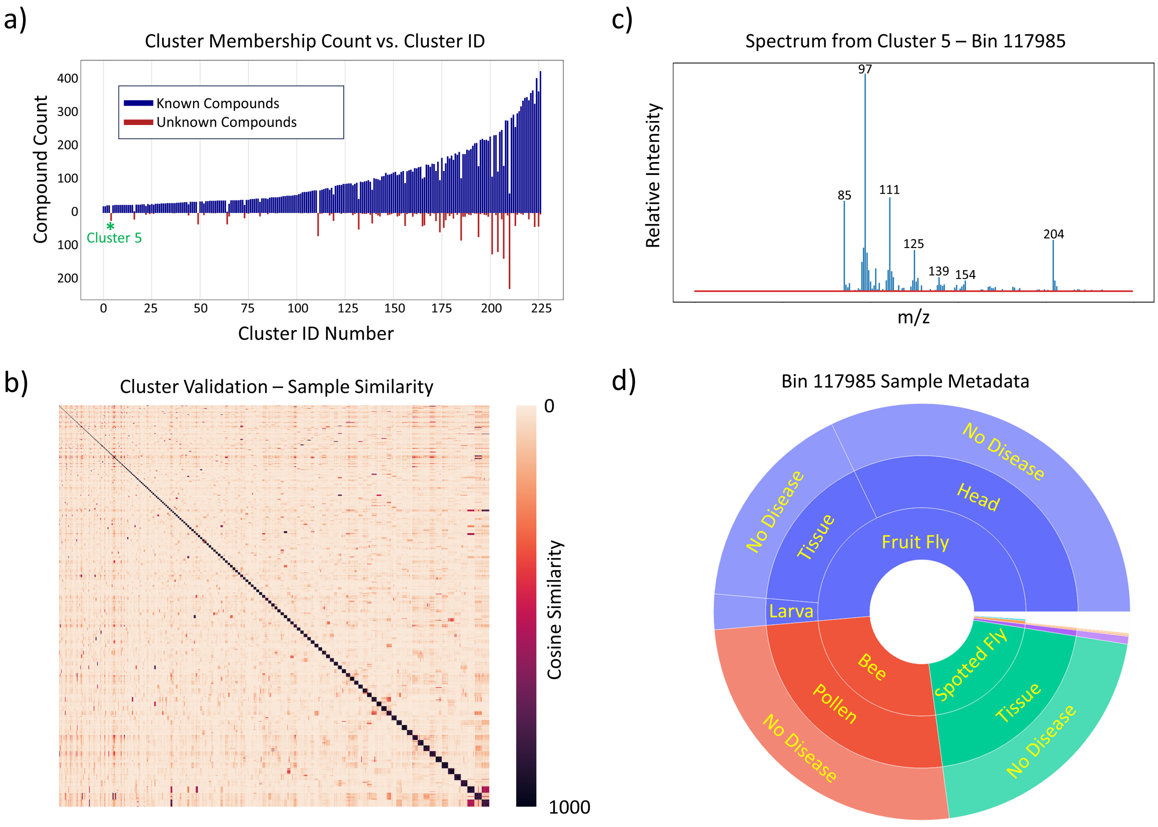
b)

Name: S-carboxymethylcysteine
InChIKey: GBFLZEXEOZUWRN-UHFFFAOYSA-N
Bin ID: 540582
Kovats RI: 1819
FAME RI: 626,424
Quant Mass: 218

Supplemental Figure S3: S-carboxymethylcysteine, a compound detected in November 2022 in bovine skeletal muscle and automatically added to BinBase.

a) Experimentally observed spectrum (red) compared to reference library spectrum (blue) and identified by both mass spectral and retention index similarity.

b) Chemical compound information for S-carboxymethylcysteine, including the international chemical identifier hash key (InChI), the BinBase identifier, the FAME-based retention index 626424, the Kovats calculated retention index, and the quantification mass.



Supplemental Figure S4: Discovering Unknown Unknowns in the GC/MS metabolome.

a) The spectra of unknowns in BinDiscover and NIST17 GC/MS were clustered using DBSCAN with a minimum cluster membership of 25 and a cosine similarity of 950 (set high to identify distinct archetypes). This yielded ~225 clusters containing both known and unknown compounds. It was observed that certain clusters were comprised (almost) entirely of unknown spectra, which means that there were no similar spectra in NIST17. If we assume that NIST17 is a comprehensive authority on GC/MS spectra, then the compounds associated with these spectra remain outside of those measured and could not be identified using a library-retrieval approach.
b) Clusters were validated by visualizing the similarity between spectra when spectra were sorted by cluster membership. Visually distinct squares arose along the diagonal, which indicates high cluster purity.
c) An example bin's spectrum from cluster 5 is visualized.
d) The associated metadata for the same bin is visualized. This compound appears to originate in plants and/or insects that interact with fruiting plants.

Supplemental Table S1: Example metadata combinations associated with ontological queries, used in Figure 1.

From or to	Metadata Triplet	Sample Count
from	homo sapiens - Pancreas - No Disease	217
from	homo sapiens - Duodenum - No Disease	183
from	homo sapiens - Liver - No Disease	471
to	lactobacillales - Cells - No Disease	16
to	pseudomonas syringae - Cells - No Disease	16
to	synechococcus elongatus - Cells - No Disease	414
to	vibrio fischeri - Cells - No Disease	13
to	saccharophagus degradans - Cells - No Disease	55
to	helicobacter pylori - Cells - No Disease	24
to	salmonella enterica - Cells - No Disease	66
to	chromobacterium - Cells - No Disease	24
to	ralstonia eutropha - Cells - No Disease	17
to	streptomyces cattleya - Cells - No Disease	11
to	bacillus subtilis - Cells - No Disease	88
to	staphylococcus aureus - Cells - No Disease	78
to	mycoplasma - Cells - No Disease	12
to	faecalibacterium prausnitzii - Cells - No Disease	66
to	synechococcus - Cells - No Disease	27
to	clostridium perfringens - Cells - No Disease	36
to	streptococcus mutans - Cells - No Disease	18
to	pseudomonas aeruginosa - Cells - No Disease	163
to	escherichia coli - Cells - No Disease	1313
to	propionibacterium - Cells - No Disease	24
to	clostridium - Cells - No Disease	36
to	methylomonas denitrificans - Cells - No Disease	12
to	halomonas elongata - Cells - No Disease	11
to	bacillus thuringiensis - Cells - No Disease	170

Supplemental Table S2: Significant compounds resulting from the query in Supplemental Table S1

Compound Name	InChIKey	$\log_2$(fold-change)	p-Value
zymosterol	CGSJXLIKVBJVRY-XTGBIJOFSA-N	-4.1	9.01E-20
ascorbic acid	CIWBSHSKHKDKBQ-JLAZNSOCSA-N	-7.4	1.38E-40
5-hydroxy-3-indoleacetic acid	DUUGKQCEGZLZNO-UHFFFAOYSA-N	-6.4	7.19E-55
(5E)-isovitamin D3	LMBGVVOJTGHJNP-FVUVGDFOSA-N	-3.2	8.02E-03
docosahexaenoic acid	MBMBGCFOFBJSGT-KUBAVDMBSA-N	-8.8	1.35E-94
cholesterone	NYOXRYYXRWJDKP-GYKMGIIDSA-N	-4.9	6.39E-16
hexadecylglycerol	OOWQBDFWEXAXPB-UHFFFAOYSA-N	-3.0	9.32E-30
tocopherol gamma-	QUEDXNHFTDJVIY-DQCZWYHMSA-N	-2.3	1.21E-37
epicholestanol	QYIXCDOBOSTCEI-FBVYSKEZSA-N	-7.1	4.38E-50
campesterol	SGNBVLSWZMBQTH-PODYLUTMSA-N	-4.7	1.32E-19
2-monoolein	UPWGQKDVAURUGE-KTKRTIGZSA-N	-2.7	9.75E-27
hypotaurine	VVIUBCNYACGLLV-UHFFFAOYSA-N	-4.2	1.54E-33
D-erythro-sphingosine	WWUZIQQURGPMPG-KRWOKUGFSA-N	-1.2	1.87E-03
N-methylglutamic acid	XLBVNMSMFQMKEY-BYPYZUCNSA-N	-4.1	3.60E-21
arachidonic acid	YZXBAPSDXZZRGB-DOFZRALJSA-N	-5.4	5.61E-24

Chapter 3: SMetaS: A Sample Metadata Standardizer for Metabolomics

Reproduced from "SMetaS: A Sample Metadata Standardizer for Metabolomics" by Parker Ladd Bremer and Oliver Fiehn, in *Metabolites.*

3.1 Abstract

Metabolomics has advanced to an extent where it is desired to standardize and compare data across individual studies. While past work in standardization has focused on data acquisition, data processing, and data storage aspects, metabolomics databases are useless without ontology-based description of biological samples and study designs. We here introduce a user-centric tool to automatically standardize sample metadata. Using such a tool in frontends for metabolomic databases will dramatically increase the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of data, specifically for data reuse and finding datasets that share comparable sets of metadata, e.g., study meta-analyses, cross-species analyses or large scale metabolomic atlases.

SMetaS (Sample Metadata Standardizer) combines a classic database with an API and frontend and is provided in a containerized environment. The tool has two user-centric components. In the first component, the user designs a sample-metadata matrix and fills the cells using natural language terminology. In the second component, the tool transforms the completed matrix by replacing free-text terms with terms from fixed vocabularies. This transformation is designed to maximize simplicity and is guided by, among other strategies, synonym matching and typographical fixing in an n-grams/nearest neighbors model approach. The tool enables downstream analysis of submitted studies and samples via string equality for FAIR retrospective use.

3.2 Introduction

Motivation for Sample Metadata Standardization

There is growing interest in standardizing metabolomics data [1–4]. Such standardization could lead to dramatic increases in biological and biomedical applicability of metabolomics. For example, decreasing the workload involved in the meta-analysis of metabolomics datasets would validate metabolomics' reproducibility overall, in addition to providing conclusions for those systems which are specifically studied [5,6]. Another way is the development of a vast metabolomics dataset, which would serve as the input to large machine learning models [7]. These multivariate models could aid or even supplant hypothesis-driven biology in the same way that statistical language models reproduce language despite the absence of a comprehensive language theory [8].

There are multiple areas where metabolomics standardization is being improved [9,10]. Areas that we do not address in this work include observational/chemical data acquisition and data processing. In this area, there has been much progress, and it is hoped that efforts toward (semi) quantitation and homogenization of methods will occur.

We here focus on the development of tools to standardize the metadata that describes samples. Our goal is to enable meta-analysis that occurs on a sample-level and is programmatic. This *programmatic meta-analysis* means the ability of computers and their users to aggregate samples very quickly and very easily. We envision users to be able to aggregate through samples by checking the equality of strings (e.g., "species"= "*mus musculus*") rather than aggregation using natural language tasks. Likewise, we employ ontological relationships (e.g. "X is a type of Y" relationships) to group sample metadata to query database on different levels of abstraction.

By emphasizing the sample-level for meta-analysis, we dramatically increase the number of ways that samples can be compared. In traditional meta-analysis, researchers are constrained to explore only the original intentions of authors', i.e., based on study design factors and hypotheses. If, instead, samples are labeled by every column header and corresponding values (such as body mass index, sex, age), then researchers could reuse those samples to explore any number of new hypotheses therein as a potential on-the-fly-factor (e.g., comparing metabolomes of specific organs across age groups, or organs across diseases).

Sample Metadata in -Omics

Ultimately all -omics analyses are based on samples. Challenges to capturing sample metadata from other fields may therefore inform solutions for metabolomics. There is a growing interest in the reuse of sample data for understanding reproducibility of findings, validation of hypotheses or as input into machine learning models. Many projects, databases, or consortia operate by formalizing and mandating metadata standards [11–13][14]. While the intent for project-wise metadata standardization is an appealing first step, the dispersion of authority creates challenges for system and data interoperability. It is very difficult to merge databases with different standards in a traceable and logical manner. Today, there are over 1,000 metadata standards lodged at https://fairsharing.org [15]. Formalizing and harmonizing these standards is an area of active research, and sophisticated informatics schemes have been proposed to reduce this bottleneck [15][16].

Perhaps an even greater challenge is the latency of biological and biomedical communities to adopt metadata standards and to adhere to reporting guidelines. There are at least three obstacles: (a) Definitions of 'minimum requirements' and 'best practice' surely change over time and between sub-communities. (b) Individual biologists or biomedical researchers do not have immediate benefits or incentives to adhere to metadata standards. This problem may be viewed as a variant of the 'tragedy of commons' [17]. (c) Many metadata upload tools are written with underlying database architectures in mind, not with user friendliness. In industry, user friendliness for web interfaces is a primary objective. In academia, user friendliness of front ends is claimed, but not tested or proven.

Hence, classic databases and sample submission interfaces expect users to submit samples and their metadata in good-faith. The ideas of reusing data have been commonplace since the early 2000s, however, with little progress so far [18]. Attempts to reuse genomics data for Covid-19 analysis revealed that, despite relatively simple requirements, over 77% of 12,000 Covid sequencing experiments lacked location metadata [17]. Similar findings have been reported for metabolomics sample metadata [19]. Finally, as we have recently learned from our own BinBase metabolomics database [20], retroactively assigning machine-ready metadata is either very tedious or demands much more research[20,21].

SMetaS development was therefore focused on the user-facing aspects of sample metadata. We avoided developing yet another standard, and rather created a tool that others can adapt and utilize within their own pipelines. Second, SMetaS intends to simplify the process of creating machine-ready metadata for the non-coding scientist. We take out the user awareness of standards and ontologies but instead employ these as backend for programmatic curation of user-based metadata. We focus on presenting a familiar tabular format, which might increase the fraction of scientists who are willing and able to describe their samples in some detail.

Sample Metadata in Metabolomics – Tool Critiques

Ultimately, successful programmatic meta-analysis, especially on a community-wide level, becomes an engineering problem. Design choices can improve or hinder a tool's capacity to facilitate programmatic meta-analysis. We explore and critique three community tools that focus on sample metadata.

Tool Critique - Metabolomics Workbench

On the Metabolomics Workbench platform, samples are submitted as part of a study [22]. A user chooses a single type of sample (human, plant, material, etc.), and then the user is exposed to a set of sample metadata categories with free text fields that depend on the chosen sample type (e.g., plant will yield "watering" options, human will not). Additionally, they are exposed to a copy/paste .tsv parser for accompanying (and possibly redundant with freetext options) tabular sample metadata as the study uploader conceives of the samples.

The MetabolomicsWorkbench offers some design choices that support programmatic meta-analysis. The displayed fields for steps are based on the selections of previous steps, which reduces visual complexity for the user.

MetabolomicsWorkbench also makes design choices that are not favorable for programmatic meta-analysis. The most unfavorable is the usage of freetext for sample metadata values (e.g., in the specific headers and in the additional matrix). Until natural language models become much more advanced, this design choice prohibits inter-study programmatic meta-analysis. Similarly, there are too many metadata categories offered to users. Full-reproducibility-level detail should be reserved for the associated publication. Moreover, the exposure to too many fields can overwhelm users and reduce interest or willingness in participation. Finally, the assignment of sample-specific values to the study as a whole, e.g., annotating a study as a "human study", precludes studies with multiple sample types. We believe this represents a strong argument for the assignment of metadata values to each sample individually, rather than making the strong assumption that some metadata value will apply to all samples in a particular study (it is very easy to conceive of studies that compare people, food, and bacteria, for example). We believe that a good phrase to describe the assignment of descriptions to samples is *sample-level granularity*.

Tool Critique - ReDU

In ReDU, sample metadata are submitted retrospectively to be associated with spectral information that has already been uploaded to the GNPS/MassIVE platform [23]. Users copy a Google Sheets template which offers a fixed set of sample metadata categories. For each sample, for each category, users select a value from a finite set of options provided in a dropdown or on another sheet. Completed metadata files can be checked with a graphical tool and then uploaded.

ReDU offers some design choices that support usage and programmatic meta-analysis. First, the constrained vocabularies enable programmatic meta-analysis via string equality rather than entity recognition. Second, the sample-level granularity naturally expands meta-analysis capabilities by enabling a looser selection of samples across studies, rather than forcing meta-analysis to focus on the small subset of comparisons that can be made on those studies with the same hypothesis. Third, Google sheets is a familiar tool to many users, which minimizes complexity, thereby encouraging use. Fourth, there is ongoing update capability because the metadata categories/headers can be expanded via Github requests.

ReDU makes some design choices that hinders programmatic meta-analysis. The most important involves the specific metadata categories that are available and the corresponding vocabularies. The included columns seem rare, such as Altitude or TermsOfPosition. Also, the overlap in category meaning, such as comorbidity and disease, necessitates merges in downstream curation before final analyses. This complexity hinders programmatic meta-analysis because future programmers will have to explore and compare vocabulary spaces. Finally, the constrained vocabularies terms are not expandable in an easy way, which precludes sample submission.

Tool Critique - MetaboLights

In MetaboLights, samples are submitted as part of a study [24]. Users are walked through a series of steps that include submitting study-wide attributes and experimental methodology. Ultimately, users are exposed to a step where samples are described via an uploadable or buildable sample metadata matrix.

There are some design aspects that favor programmatic meta-analysis. The step-by-step walkthrough simplifies the submission process which encourages use. Likewise, the emphasis on sample-level resolution is essential to FAIR programmatic meta-analysis. Additionally, the connection of metadata categories/headers to ontology terms generates a constrained vocabulary that would enable string equality comparisons and support downstream ontological analysis.

Unfortunately, Metabolights has several problems that hinder programmatic meta-analysis. All ontologies are accessible at any place in the sample matrix. Because the same idea can have different forms in different ontologies, tedious downstream merges are required to remove inconsistencies. Likewise, terms can be entered as freetext, which leads to the same problem. Finally, upload to the matrix interface requires the use of an ftp service, which greatly discourages use.

SMetaS

Based on the above discussions, our goal is to create a tool that enables sample-oriented and programmatic meta-analysis if used in a front end for study submissions to metabolomic databases. Such tools necessarily must remain a compromise between asking users to detail every aspect of a study (e.g., the exact composition of chow in studies of animal models), versus the time and efforts users are willing to spend for sample or data submissions. Such tools offer programmatic relationships to existing standards, but do not complexify the space of already existing standards. Indeed, we designed a tool that captures the essence, but not the total complexity, of a sample's nature, while giving users the option to add more details if they are inclined to do so. This mixture of mandatory and voluntary metadata is auto-curated and recorded into a database, which can then be linked to the observed metabolomics of a sample in downstream analyses.

3.3 Methods

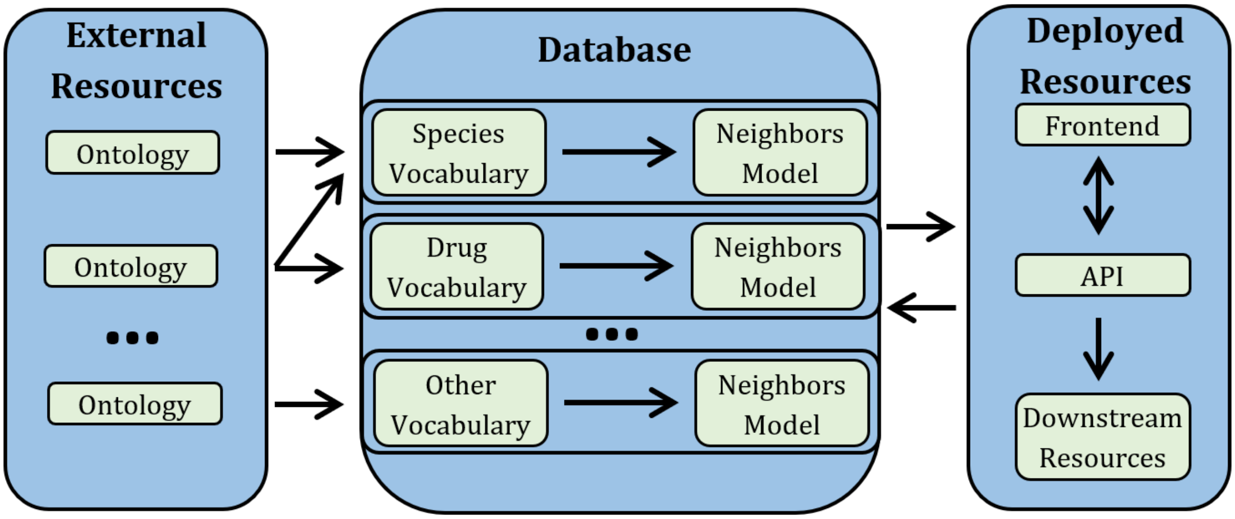


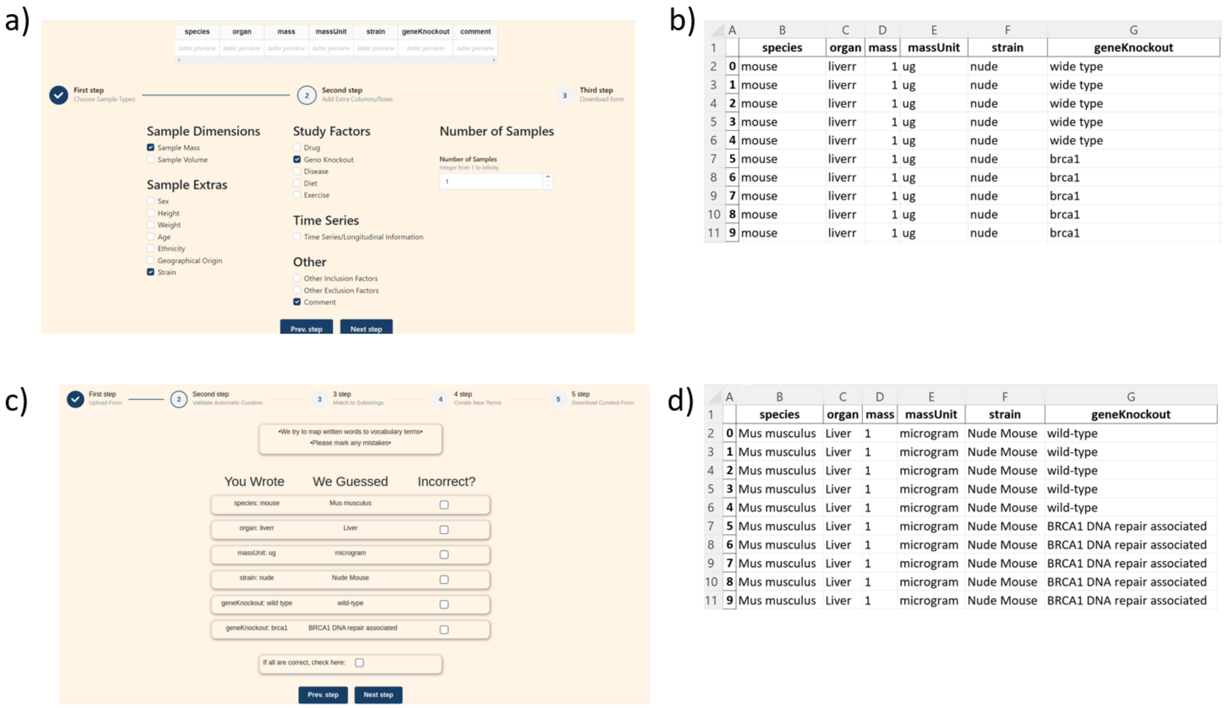
Figure 1: Workflow of the Metadata Standardizer
First, each metadata column type (species, organ, drug, etc.) has a starting vocabulary derived by combining/subsetting existing ontologies/vocabularies, making sure that the intersection of any two vocabularies is 0. Next, for each vocabulary, we generate additional resources that facilitate ease-of-use for sample submitters (e.g., nearest neighbor models that map synonyms/typos to the correct term). Finally, we make the vocabularies and associated resources as the backend to a user-friendly frontend. These vocabularies and models are expandable if new terms are desired by users. The vocabularies and models are also available as an API directly. A more detailed workflow is available as **Supplemental Figure 1**.

For generating the vocabularies and associated models, we made extensive use of custom python scripts that are available in Github (see Data Availability). We heavily employed snakemake, networkx, pandas, scikit-learn, and other libraries [25–27]. The API and frontend were also generated using custom python scripts that are available in Github (see Data Availability). We heavily employed Flask, Dash, and Docker. Development and creation were performed locally on a personal computer.

3.4 Results

3.4.1 Overview

The primary result of this work is a tool that facilitates the standardization of sample metadata for downstream programmatic analysis. Basic usage for SMetaS is illustrated in **Figure 2**. Here, the user first chooses metadata that are associated with their samples. There is no capability to specify "factors" because that is an artificial constraint that can be readily applied downstream. User selections generate a downloadable csv file for which each row is a sample and each column is a metadata attribute. We chose this format because all scientist users are familiar with such basic worksheets and know how to manipulate these documents. Cells can remain empty if that attribute does not apply. Users then reupload their csv files, and interact



species	organ	mass	massUnit	strain	geneKnockout	comment
table preview	table preview	table preview	table preview	table preview	table preview	table preview

Sample Dimensions
☑ Sample Mass
□ Sample Volume

Sample Extras
□ Sex
□ Height
□ Weight
□ Age
□ Ethnicity
□ Geographical Origin
☑ Strain

Study Factors
□ Drug
☑ Geno Knockout
□ Disease
□ Diet
□ Exercise

Time Series
□ Time Series/Longitudinal Information

Other
□ Other Inclusion Factors
□ Other Exclusion Factors
☑ Comment

Number of Samples

Number of Samples
Integer from 1 to infinity
1

	A	B	C	D	E	F	G
1		species	organ	mass	massUnit	strain	geneKnockout
2	0	mouse	liverr	1	ug	nude	wide type
3	1	mouse	liverr	1	ug	nude	wide type
4	2	mouse	liverr	1	ug	nude	wide type
5	3	mouse	liverr	1	ug	nude	wide type
6	4	mouse	liverr	1	ug	nude	wide type
7	5	mouse	liverr	1	ug	nude	brca1
8	6	mouse	liverr	1	ug	nude	brca1
9	7	mouse	liverr	1	ug	nude	brca1
10	8	mouse	liverr	1	ug	nude	brca1
11	9	mouse	liverr	1	ug	nude	brca1

•We try to map written words to vocabulary terms•
•Please mark any mistakes•

You Wrote	We Guessed	Incorrect?
species: mouse	Mus musculus	□
organ: liverr	Liver	□
massUnit: ug	microgram	□
strain: nude	Nude Mouse	□
geneKnockout: wild type	wild-type	□
geneKnockout: brca1	BRCA1 DNA repair associated	□

If all are correct, check here: □

	A	B	C	D	E	F	G
1		species	organ	mass	massUnit	strain	geneKnockout
2	0	Mus musculus	Liver	1	microgram	Nude Mouse	wild-type
3	1	Mus musculus	Liver	1	microgram	Nude Mouse	wild-type
4	2	Mus musculus	Liver	1	microgram	Nude Mouse	wild-type
5	3	Mus musculus	Liver	1	microgram	Nude Mouse	wild-type
6	4	Mus musculus	Liver	1	microgram	Nude Mouse	wild-type
7	5	Mus musculus	Liver	1	microgram	Nude Mouse	BRCA1 DNA repair associated
8	6	Mus musculus	Liver	1	microgram	Nude Mouse	BRCA1 DNA repair associated
9	7	Mus musculus	Liver	1	microgram	Nude Mouse	BRCA1 DNA repair associated
10	8	Mus musculus	Liver	1	microgram	Nude Mouse	BRCA1 DNA repair associated
11	9	Mus musculus	Liver	1	microgram	Nude Mouse	BRCA1 DNA repair associated

Figure 2: Walkthrough of user experience.
a) the first component of the tool is a walkthrough that allows users to design a sample metadata matrix. b) an example metadata matrix prior to standardization. c) the second component of the tool is a walkthrough that allows users to that curate that submission. d) the same submission with terms standardized in order to dramatically simplify meta-analysis.

with the SMetaS transformation process, which converts freetext, natural language entries into a formalized and standardized representation.

This tool's most noteworthy design choices that facilitate this are shown in **Table 1**. The tool is provided as a container that is directly runnable. Associated code is available for the vocabulary pipeline as well as downstream API/frontend. Documentation is available as well.

Table 1: SMetaS design principles.

Number	**Design Principle**
1	headers with orthogonal vocabularies
2	vocabularies with non-redundant terms
3	inclusion of a synonym set for each "main term" to facilitate the loose expression of a term
4	vocabularies/models that expand to incorporate new terms easily submitted by users
5	machine learning models that increase speed-of-use and make the program typo tolerant
6	a deference for simplicity when possible. We believe that user apathy/disinterest is as much a problem as any technical challenge

3.4.2 Extended Description of Components

Our tool is comprised of two main components. In the first component, users walk through a short series of steps where they generate a csv spreadsheet onto which they transcribe their sample metadata. Users can select core sample types (tissue, cells, etc.) as well as additional sample attributes. Users then fill out their created spreadsheet locally and resubmit it to the second part of the tool.

In the second component, the user submission is transformed into an equivalent sample metadata matrix that is ready for programmatic meta-analysis at a later date. To do this, there are several steps that are made at the user's submission. By the end of these passes, it is guaranteed that all terms will be transformed into an existing term or become new terms in the corresponding vocabulary.

In the first pass, a term-frequency-inverse-document-frequency (tf-idf) vectorizer and nearest neighbors model (nnm) automatically curates user-submitted strings [28][29]. This vectorizer works by transforming a given string into a numeric vector and then finding the vocabulary term with the most similar vector. The components of the numeric vector are decided during the database construction step in **Figure 1/Supplemental Figure 1**. The vector has components of all three character combinations present at least once. For example, *mus musculus* would generate (m,u,s), (u,s,' '), (s,' ',m), (' ',m,u), (u,s,c), etc. while *arabidopsis thaliana* would generate (a,r,a), (r,a,b), etc., and the union of each term's set generates the total set of vector components. The magnitudes of each cell in the term/component matrix are based on presence of that component in a term after weighting component magnitudes according to number of appearances within that term (term frequency) and rareness of that letter triplet across all terms (inverse document frequency). New words are coerced into this pre-determined space and cosine similarity determines the distance between vocabulary terms and user-provided terms.

The tf-idf vectorizer and nnm curate strings such as species (e.g. mice, mouse, M. musculus and similar terms), organ names, drugs names, units, or other metadata. Metadata that are intrinsically unique to a sample (e.g., magnitudes of height or drug amount) are not curated. The derivations of the initial controlled vocabularies from official ontologies are described in **Table 2**. The first pass is expected to deal with the bulk of user-submitted strings.

In the second pass, terms that were not able to be mapped in the first pass can be transformed using a substring search. This might happen if the term was accidentally misspelled or an unknown synonym or abbreviation was used. Both the first and second pass map sets of strings to “main terms”, e.g., “*mus musculus*”, “mouse”, “mice”, or “house mouse” would all be mapped to “*mus musculus*”.

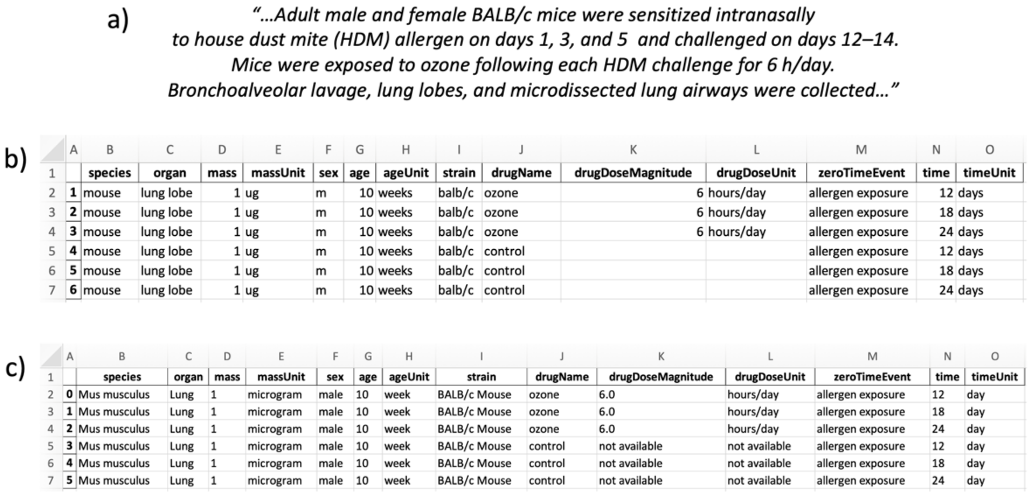
Finally, in the third pass, users can confirm to add new terms that were not present in the associated header’s vocabulary, for example, for organ, species or experimental intervention that were not included in the large, standardized community vocabularies that we employ (see below). In this way, users add new strings to our underlying ontologies to update and renew the system over time to increase the likelihood that next users will find matching selections (e.g., for new cell types, drugs, etc.). Users are given a freetext input box preloaded with the observed term. Once confirmed, these freetext terms are added to the corresponding vocabulary for future users, and corresponding models will be retrained on this expanded vocabulary.

Users receive a standardized csv file to be used to submit a study for metabolomic data acquisitions (e.g., at the UC Davis West Coast Metabolomics Center), or to submit metabolomic data to a common repository (e.g., for the MetabolomicsWorkbench) to enable programmatic meta-analysis. The system includes a programmatic access point for submitted studies and authors for convenient integration into existing pipelines.

3.4.3 Use case

We here provide an explicit example based on a study performed at the West Coast Metabolomics Center involving the effect of ozone on metabolism in the lung [30]. In **Figure 3**, we show an excert from the study abstract, the freetext representation created with our tool, and finally the curated representation created with our tool.

Briefly, male and female adult mice were exposed to house dust mite allergen, then exposed to ozone, and finally sacrificed and differences in lung metabolism were compared to untreated control mice by metabolomics assays. Most of the information given in the publication **(Figure 3a)** was captured by SMetaS, but not all details of the study design (**Figure 3b)**. We recorded basic descriptions of the lungs (organ, mass, massUnit), the mice from which they were derived (species, sex, age, ageUnit, strain), the treatment of ozone exposure represented as a drug (drugName, drugDoseMagnitude, and drugDoseUnit), and the time-series aspect after exposure to the allergen (zeroTimeEvent, time, and timeUnit). We lose information such as the intranasal delivery, ozone chamber details, and high-detail lung lobe locations. We recognize that there are other valid ways to represent this study. For example, it would have been possible to represent the allergen exposure as another drug. Indeed, creating unambiguous instructions for metadata representation is an area of active research [15]



"…Adult male and female BALB/c mice were sensitized intranasally to house dust mite (HDM) allergen on days 1, 3, and 5 and challenged on days 12–14. Mice were exposed to ozone following each HDM challenge for 6 h/day. Bronchoalveolar lavage, lung lobes, and microdissected lung airways were collected…"

	species	organ	mass	massUnit	sex	age	ageUnit	strain	drugName	drugDoseMagnitude	drugDoseUnit	zeroTimeEvent	time	timeUnit
1	mouse	lung lobe	1	ug	m	10	weeks	balb/c	ozone	6	hours/day	allergen exposure	12	days
2	mouse	lung lobe	1	ug	m	10	weeks	balb/c	ozone	6	hours/day	allergen exposure	18	days
3	mouse	lung lobe	1	ug	m	10	weeks	balb/c	ozone	6	hours/day	allergen exposure	24	days
4	mouse	lung lobe	1	ug	m	10	weeks	balb/c	control			allergen exposure	12	days
5	mouse	lung lobe	1	ug	m	10	weeks	balb/c	control			allergen exposure	18	days
6	mouse	lung lobe	1	ug	m	10	weeks	balb/c	control			allergen exposure	24	days

	species	organ	mass	massUnit	sex	age	ageUnit	strain	drugName	drugDoseMagnitude	drugDoseUnit	zeroTimeEvent	time	timeUnit
0	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	ozone	6.0	hours/day	allergen exposure	12	day
1	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	ozone	6.0	hours/day	allergen exposure	18	day
2	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	ozone	6.0	hours/day	allergen exposure	24	day
3	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	control	not available	not available	allergen exposure	12	day
4	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	control	not available	not available	allergen exposure	18	day
5	Mus musculus	Lung	1	microgram	male	10	week	BALB/c Mouse	control	not available	not available	allergen exposure	24	day

Figure 3: Example use case of SMetaS.
a) Excerpt of a published study abstract *[30]* b) SMetaS matrix representation of information from the abstract and methods section *[30]* c) SMetaS curation of freetext terms of the matrix representation.

Importantly, SMetaS transforms freetext strings to formalized nomenclature (**Figure 3c**) by mapping to pre-existing terms**.** All non-numeric strings were already contained in the initial ontology-derived vocabularies except for 'allergen exposure', 'ozone', and 'hours/day' and freetext strings were successfully. These three metadata strings were then added to their corresponding vocabularies (zeroTimeEvent, drugName, drugDoseUnit) for future users.

3.4.4 Construction of Vocabularies

Sample metadata standardizers should incorporate vocabularies. As expanded on in the discussion, there are several important properties of vocabularies. SMetaS relies on a constrained, non-overlapping, non-redundant, and expandable vocabulary for each metadata category. Constrained vocabularies allow for equality testing via string equality rather than entity recognition. For example, for databases supported by SMetaS, all samples associated with mice are mapped to "*mus musculus*". For databases that accept free-text without automatic curation, users must devise post-hoc models or elaborate search criteria to collect those mice samples. Such post-hoc data curations easily creates errors, dramatically increases the workload and decreases overall data quality.

We limited the number of downloaded vocabularies and ontologies to large, mature community repositories such as MeSH [31], NCBI [32,33], Cellosaurus [34], NCIT [35] and FDA [36]. These non-redundant vocabularies avoid string overlaps and therefore abolish the need for complicated downstream merges between headers and terms. Finally, expandable vocabularies minimize system maintenance and allow for perpetual updates when users submit, for example, a new drug, genetic variant, animal model, etc.

To generate vocabularies that maintain our first three principles, we accessed a set of ontologies and vocabularies relevant to metabolomics study samples, listed in **Table S1**. For each header/category, we extracted sets of non-overlapping "main" vocabulary terms, each of which had a set of 0 to n synonyms derived from the same sources. For example, "*mus musculus*" would be a main term, while "mouse", "mice", or "house mouse" would be synonyms. The selections for vocabulary origins were made based on internal discussion. The headers and vocabularies are summarized in **Table 2**.

Table 2: The metadata categories, term counts, and definition of initial vocabularies. The Initial Vocabulary Description column describes what subsections of formal ontologies comprise each vocabulary, initially.

Grouping	Metadata Category	Term Count	Initial Vocabulary Description
Core Sample Type	species	724,962	NCBI ontology less -rank ‘strain’ -parent node scientific name contained ‘environmental sample’ -parent node scientific name contained ‘unclassified’ -rank ‘no rank’ that contained ‘/’ -rank ‘species’ containing numerical characters -rank ‘species’ containing ‘vector’
	organ	11,494	MeSH ontology heading ‘A’ and lower
	cellLine	247,365	Cellosaurus ontology
	material	2,056	MeSH ontology: -heading ‘D20’ and lower -heading ‘G16’ and lower
Sample Description	massUnit	49	Unit Ontology: -heading UO0000002 and lower
	volumeUnit	79	Unit Ontology: -heading UO0000095 and lower
	sex	3	All sexes
	heightUnit	48	Unit Ontology: -heading UO0000001 and lower
	weightUnit	49	Unit Ontology: -heading UO0000002 and lower
	ageUnit	22	Unit Ontology: -heading UO0000003 and lower
	ethnicity	1,057	NCIT Ontology: -header C17049 and lower
	geographicalOrigin	799	MeSH ontology: -header Z01 and lower -header G16.500.275 and lower
	strain	2,282	NCIT Ontology: -header C14250 and lower except those terms which exist in the NCBI ontology or are descendants of Gene header in NCIT ontology

Study Factors	drugName	9,537	FDA drug vocabulary
	drugDoseUnit	753	Unit Ontology
	geneKnockout	141,605	NCBI human gene vocabulary
	disease	36,378	MeSH ontology: -header C and lower
	diet	1,164	MeSH heading G07.203 and lower
	exercise	569	MeSH heading I03 and lower, MeSH heading G11.427.410.698 and lower
Time Series	zeroTimeEvent	69,321	NCIT: ontology: -header C43431 and lower
	timeUnit	22	All Units
Other	inclusion	0	None in initial vocabulary
	exclusion	0	None in initial vocabulary
	comment	0	None in initial vocabulary

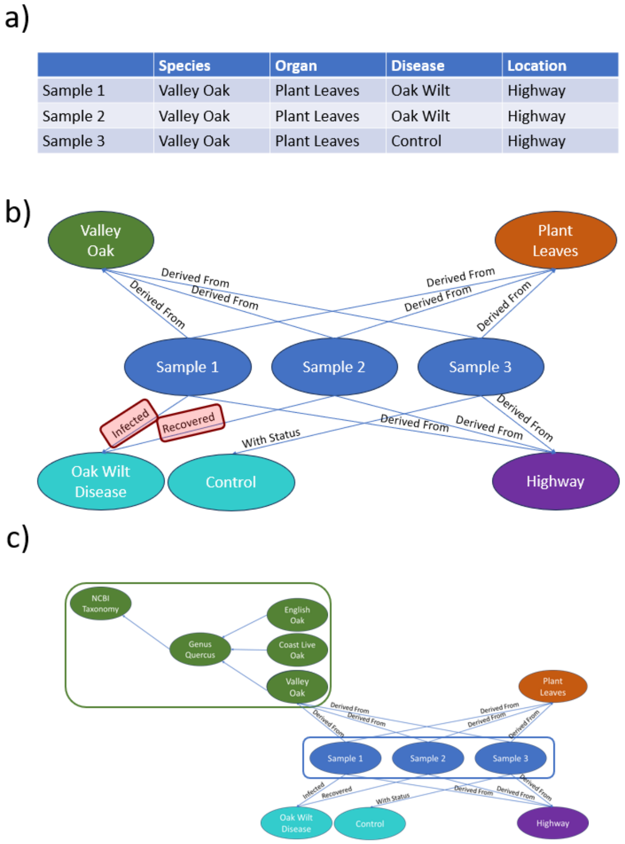
3.5 Discussion

There are several shortcomings of our design. Intrinsic problems are derived from the decision to use a tabular approach to retaining study metadata information. No matter what set of metadata categories that we offer to users, eventually there may be a study that cannot be fully described using that schema. Study design space is extraordinarily broad, and we have encountered surprising factors such as “proximity to parking lots” in studies submitted to our metabolomics service core at UC Davis. No tractable tabular approach with a fixed set of metadata could capture such design. Instead, we focus on capturing the essence of a sample rather than its full intricacies. In doing so, we assume that metadata categories that we do not include do not define different populations in the statistical sense.

We expect to improve the sample description space of SMetaS based on samples submitted to the West Coast Metabolomics Center because our service core analyzes over 30,000 samples/year. SMetaS offers the metadata categories ‘comment’, ‘inclusion criteria’, and ‘exclusion criteria’ as alternatives to categories that are explicitly provided. Over time, we will accumulate data on metadata objects that are common enough to warrant creating explicit categories. Such commentary metadata also give us insights which exiting categories or units *do* cause discrepancies in study descriptors that are insufficiently covered yet. For example, in metagenomics, nuanced aspects of environment or host descriptors are necessary for valid interpretations [17]. Hence, future versions of SMetaS will include additional categories/vocabularies derived from submitted data.

The tabular approach in SMetaS is also limiting the relationship between descriptors and samples. For example, we assume unique “is a” relationships between a sample and its listed species. While our approach can handle multiple species (for the user, simply a delimiter is needed), the precise meaning of a list of species is not specified. To add additional relationship types (such as co-cultures) would mean increasing the complexity of the written string, which is philosophically prohibited in our fixed vocabulary approach (of course, we cannot prevent users from submitting messy custom terms). Future improvements would maintain sample metadata in node/edge graphs, where the edges afford the opportunity to programmatically store more detailed relationships between samples and descriptors. We illustrate this in the transition from **Figure 4a** to **Figure 4b**.

Storing sample metadata as graphs would simplify querying data in hierarchical and ontology-based meta-analysis. As illustrated in **Figure 4c**, we see that ontologies map neatly with sample sets when using a graph approach and we can imagine this implementation affording very natural queries.



	Species	Organ	Disease	Location
Sample 1	Valley Oak	Plant Leaves	Oak Wilt	Highway
Sample 2	Valley Oak	Plant Leaves	Oak Wilt	Highway
Sample 3	Valley Oak	Plant Leaves	Control	Highway

Figure 3: Comparing tabular to graph storage for sample metadata.
a) a simple schema for several samples is shown b) the same schema is expressed as a graph, with additional complexity programmatically embedded into nodes c) the same schema as a graph where the ontology for the species node is included. This allows for the natural hierarchical analysis of metadata.

While tempting, the notion of a parser that automatically derives sample metadata from natural language is likely not applicable to existing publications or existing datasets for which the assignment of unique properties to specific samples is not clear.

In general, we recognize some of the shortcomings that are built into our system. We accept these as an inevitable consequence of our goal to render this tool intuitive for users.

Challenges and complexities have been documented when users deposit studies into repositories [19]. While our tool aims to conform to user expectations with a front-end that is intuitive enough to be completed without further manuals or communications, our tool has not yet been deployed as mandatory study submission frontend for a metabolomic repository. User feedback is currently sought from users of the UC Davis metabolomics service core with its approximately 300 clients/year. Upon completion, an updated frontend will feed into the new UC Davis LC-BinBase system, replacing the outdated miniX (SetupX) system that was in operation for 18 years [37].

3.6 Conclusion

We have created SMetaS, which standardizes submitted sample descriptions to enable downstream programmatic meta-analysis. This tool is readily deployable in a fully reproducible way for core laboratories or larger repositories.

We are interested in standardizing metadata to programmatically utilize metadata. We envision at least two ways of doing this in masse. The first is agglomerative analysis similar to those available in our BinDiscover tool [20]. There, we allowed for the exploration of combinations of metadata to be visualized and explored according to user specified nodes on ontological hierarchies. We can imagine expanding upon this concept and programmatically probe which data patterns persist when comparing studies and samples on their highest ontological parent nodes, i.e., what are the largest generalizations that can be made? Validations of such meta-analyses would be based on published ground truths (e.g., absence of cholesterol in the plant kingdom), giving credibility to new hypotheses to be revealed by large scale database queries.

The second way that we hope to utilize this tool is as the foundation for a comprehensive metabolomics atlas. We hope that this atlas could be one of many large, normalized datasets provided to a large machine learning model that circumvents hypotheses altogether to provide clinical or therapeutic predictions in opposition to or in conjunction with theory provided by domain experts.

3.7 Data and Software Availability

All code is available at https://github.com/metabolomics-us/metadatastandardizer. We make extended documentation available at https://metabolomics-us.github.io/metadatastandardizer/. The documentation reviews deployment instructions for this tool on Amazon Web Services as well as a detailed walkthrough for using the backend.

3.8 References

1. Guo, J.; Yu, H.; Xing, S.; Huan, T. Addressing Big Data Challenges in Mass Spectrometry-Based Metabolomics. *Chemical Communications* **2022**, *58*, 9979–9990, doi:10.1039/D2CC03598G.

2. Kirwan, J.A. Translating Metabolomics into Clinical Practice. *Nat Rev Bioeng* **2023**, *1*, 228–229, doi:10.1038/s44222-023-00023-x.

3. Forcisi, S.; Moritz, F.; Thompson, C.J.; Kanawati, B.; Uhl, J.; Afonso, C.; Bader, C.D.; Barsch, A.; Boughton, B.A.; Chu, R.K.; et al. Large-Scale Interlaboratory DI-FT-ICR MS Comparability Study Employing Various Systems. *J. Am. Soc. Mass Spectrom.* **2022**, *33*, 2203–2214, doi:10.1021/jasms.2c00082.

4. Dias, D.A.; Koal, T. Progress in Metabolomics Standardisation and Its Significance in Future Clinical Laboratory Medicine. *EJIFCC* **2016**, *27*, 331–343.

5. Martínez-Reyes, I.; Chandel, N.S. Cancer Metabolism: Looking Forward. *Nat Rev Cancer* **2021**, *21*, 669–680, doi:10.1038/s41568-021-00378-6.

6. Goveia, J.; Pircher, A.; Conradi, L.-C.; Kalucka, J.; Lagani, V.; Dewerchin, M.; Eelen, G.; DeBerardinis, R.J.; Wilson, I.D.; Carmeliet, P. Meta-Analysis of Clinical Metabolic Profiling Studies in Cancer: Challenges and Opportunities. *EMBO Molecular Medicine* **2016**, *8*, 1134–1142, doi:10.15252/emmm.201606798.

7. Eisenstein, M. Machine Learning Powers Biobank-Driven Drug Discovery. *Nature Biotechnology* **2022**, *40*, 1303–1305, doi:10.1038/s41587-022-01457-1.

8. Large Language Models Demonstrate the Potential of Statistical Learning in Language - Contreras Kallens - 2023 - Cognitive Science - Wiley Online Library Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13256 (accessed on 11 July 2023).

9. Spicer, R.A.; Salek, R.; Steinbeck, C. A Decade after the Metabolomics Standards Initiative It's Time for a Revision. *Sci Data* **2017**, *4*, 170138, doi:10.1038/sdata.2017.138.

10. Long, N.P.; Nghi, T.D.; Kang, Y.P.; Anh, N.H.; Kim, H.M.; Park, S.K.; Kwon, S.W. Toward a Standardized Strategy of Clinical Metabolomics for the Advancement of Precision Medicine. *Metabolites* **2020**, *10*, 51, doi:10.3390/metabo10020051.

11. Field, D.; Garrity, G.; Gray, T.; Morrison, N.; Selengut, J.; Sterk, P.; Tatusova, T.; Thomson, N.; Allen, M.J.; Angiuoli, S.V.; et al. The Minimum Information about a Genome Sequence (MIGS) Specification. *Nat Biotechnol* **2008**, *26*, 541–547, doi:10.1038/nbt1360.

12. Perez-Riverol, Y. Toward a Sample Metadata Standard in Public Proteomics Repositories. *J. Proteome Res.* **2020**, *19*, 3906–3909, doi:10.1021/acs.jproteome.0c00376.

13. Specimen and Sample Metadata Standards for ... | Wellcome Open Research Available online: https://wellcomeopenresearch.org/articles/7-187/v1?src=rss (accessed on 22 July 2023).

14. Sasse, J.; Darms, J.; Fluck, J. Semantic Metadata Annotation Services in the Biomedical Domain—A Literature Review. *Applied Sciences* **2022**, *12*, 796, doi:10.3390/app12020796.

15. Batista, D.; Gonzalez-Beltran, A.; Sansone, S.-A.; Rocca-Serra, P. Machine Actionable Metadata Models. *Sci Data* **2022**, *9*, 592, doi:10.1038/s41597-022-01707-6.

16. Moxon, S.; Solbrig, H.; Unni, D.; Jiao, D.; Bruskiewich, R.; Balhoff, J.; Vaidya, G.; Duncan, W.; Hegde, H.; Miller, M.; et al. The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. In Proceedings of the CEUR Workshop Proceedings; CEUR-WS, 2021; Vol. 3073, pp. 148–151.

17. Schriml, L.M.; Chuvochina, M.; Davies, N.; Eloe-Fadrosh, E.A.; Finn, R.D.; Hugenholtz, P.; Hunter, C.I.; Hurwitz, B.L.; Kyrpides, N.C.; Meyer, F.; et al. COVID-19 Pandemic Reveals the Peril of Ignoring Metadata Standards. *Sci Data* **2020**, *7*, 188, doi:10.1038/s41597-020-0524-5.

18. Nichols, B.N.; Ghosh, S.S.; Auer, T.; Grabowski, T.; Maumet, C.; Keator, D.; Martone, M.E.; Pohl, K.M.; Poline, J.-B. Linked Data in Neuroscience: Applications, Benefits, and Challenges 2016, 053934.

19. Ferreira, J.D.; Inácio, B.; Salek, R.M.; Couto, F.M. Assessing Public Metabolomics Metadata, Towards Improving Quality. *Journal of Integrative Bioinformatics* **2017**, *14*, doi:10.1515/jib-2017-0054.

20. Bremer, P.L.; Wohlgemuth, G.; Fiehn, O. The BinDiscover Database: A Biology-Focused Meta-Analysis Tool for 156,000 GC–TOF MS Metabolome Samples. *Journal of Cheminformatics* **2023**, *15*, 66, doi:10.1186/s13321-023-00734-8.

21. Hawkins, N.T.; Maldaver, M.; Yannakopoulos, A.; Guare, L.A.; Krishnan, A. Systematic Tissue Annotations of Genomics Samples by Modeling Unstructured Metadata. *Nat Commun* **2022**, *13*, 6736, doi:10.1038/s41467-022-34435-x.

22. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K.S.; et al. Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools. *Nucleic Acids Res* **2016**, *44*, D463–D470, doi:10.1093/nar/gkv1042.

23. ReDU: A Framework to Find and Reanalyze Public Mass Spectrometry Data | Nature Methods Available online: https://www.nature.com/articles/s41592-020-0916-7 (accessed on 5 June 2023).

24. Haug, K.; Cochrane, K.; Nainala, V.C.; Williams, M.; Chang, J.; Jayaseelan, K.V.; O'Donovan, C. MetaboLights: A Resource Evolving in Response to the Needs of Its Scientific Community. *Nucleic Acids Research* **2020**, *48*, D440–D444, doi:10.1093/nar/gkz1019.

25. Mölder, F.; Jablonski, K.P.; Letcher, B.; Hall, M.B.; Tomkins-Tinch, C.H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S.O.; Kanitz, A.; et al. Sustainable Data Analysis with Snakemake 2021.

26. Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using Networkx*; Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2008;

27. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

28. SPARCK JONES, K. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation* **1972**, *28*, 11–21, doi:10.1108/eb026526.

29. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27, doi:10.1109/TIT.1967.1053964.

30. Stevens, N.C.; Brown, V.J.; Domanico, M.C.; Edwards, P.C.; Van Winkle, L.S.; Fiehn, O. Alteration of Glycosphingolipid Metabolism by Ozone Is Associated with Exacerbation of Allergic Asthma Characteristics in Mice. *Toxicological Sciences* **2023**, *191*, 79–89, doi:10.1093/toxsci/kfac117.

31. Rogers, F.B. Medical Subject Headings. *Bull Med Libr Assoc* **1963**, *51*, 114–116.

32. GenBank | Nucleic Acids Research | Oxford Academic Available online: https://academic.oup.com/nar/article/47/D1/D94/5144964 (accessed on 11 July 2023).

33. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database (Oxford)* **2020**, *2020*, baaa062, doi:10.1093/database/baaa062.

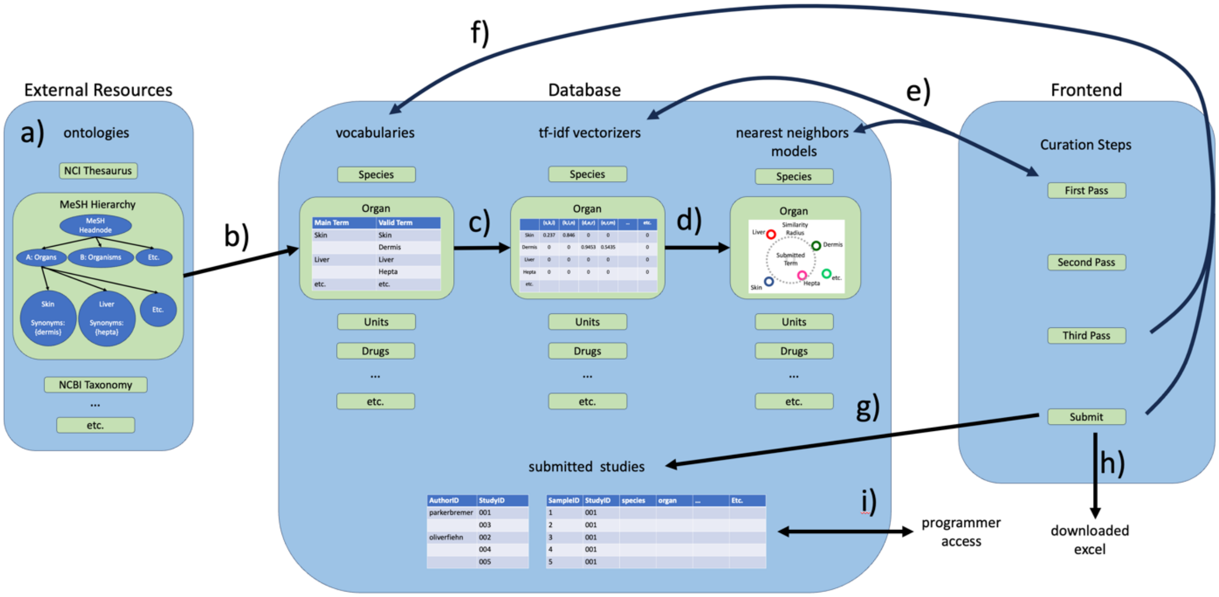
34. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech* **2018**, *29*, 25–38, doi:10.7171/jbt.18-2902-002.

35. NCI Thesaurus Available online: https://ncithesaurus.nci.nih.gov/ncitbrowser/ (accessed on 11 July 2023).

36. Research, C. for D.E. and Drugs@FDA Data Files. *FDA* **2023**.

37. Scholz, M.; Fiehn, O. SetupX--a Public Study Design Database for Metabolomic Projects. *Pac Symp Biocomput* **2007**, 169–180.

3.9 Supplemental Information



Main Term	Valid Term
Skin	Skin
	Dermis
Liver	Liver
	Hepta
etc.	etc.

	(s,k,i)	(k,i,n)	(d,e,r)	(e,r,m)	...	etc.
Skin	0.237	0.846	0	0		0
Dermis	0	0	0.9453	0.5435		0
Liver	0	0	0	0		0
Hepta	0	0	0	0		0
etc.						

AuthorID	StudyID
parkerbremer	001
	003
oliverfiehn	002
	004
	005

SampleID	StudyID	species	organ	...	Etc.
1	001				
2	001				
3	001				
4	001				
5	001				

Figure S1: Detailed Workflow of SMetaS
a) formal ontologies supply curated and relevant terms. Here we show a simplified version of the Medical Subject Headings ontology. b) ontologies are coerced into initial vocabularies for metadata categories. Here, the organ subgraph from MeSH becomes the organ vocabulary. c) each vocabulary generates a tf-idf vectorizer, which creates a numeric space based on triplets of letters. d) each vocabulary and vectorized space is used to create a nearest neighbors model, which is used to find similar terms to strings provided by users. e) in the first pass, described in the Results, the API connects the tf-idf/nearest-neighbors models to the frontend. f) in the event that new terms are added to vocabularies (like a new drug to the drug vocabulary), the API adds terms to the vocabularies, which triggers a retraining of the tf-idf vectorizer/nearest neighbors models based on the expanded vocabulary. g) completion of a curation process stores the resultant curated values into tables in the database. h) completion of a curation process generates an excel file that is immediately available for the user. This excel file contains the same information that is stored in g). i) the stored studies can be programmatically accessed. Details for practical usage are provided in the documentation website at https://metabolomics-us.github.io/metadatastandardizer/ .

Appendix: Additional Projects

Part 1: ClusterBase

ClusterBase was a project developed during a summer internship at the Chan-Zuckerberg Biohub with colleagues Dr. Brian DeFelice and Wasim Sandhu. It is an in-house software tool that ingests processed mass spectrometer data files into a database and then performs network analysis. This project had several benefits for the metabolomics team at the Biohub. First, it established modern infrastructure for their data so that any subsequent project will be significantly easier. Second, the network analysis automatically annotated metabolites. This sped the processing of

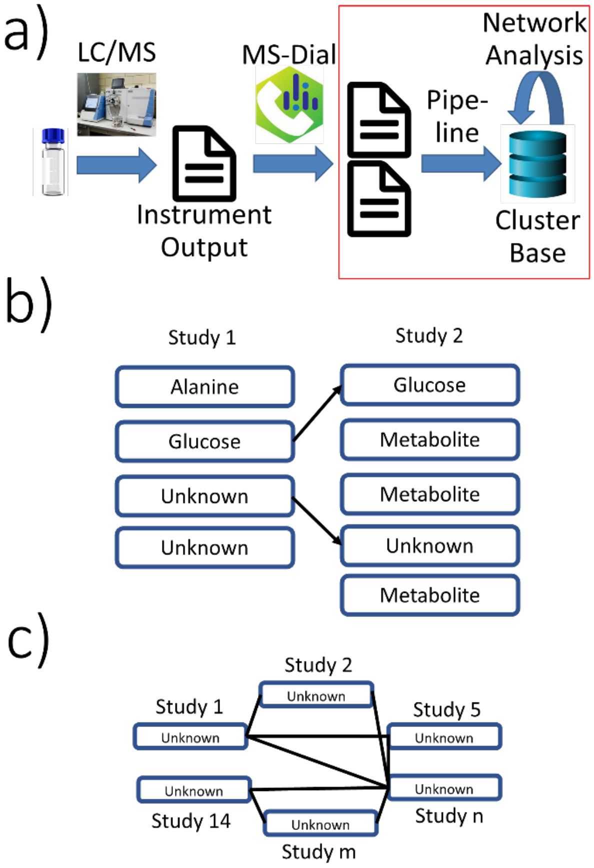
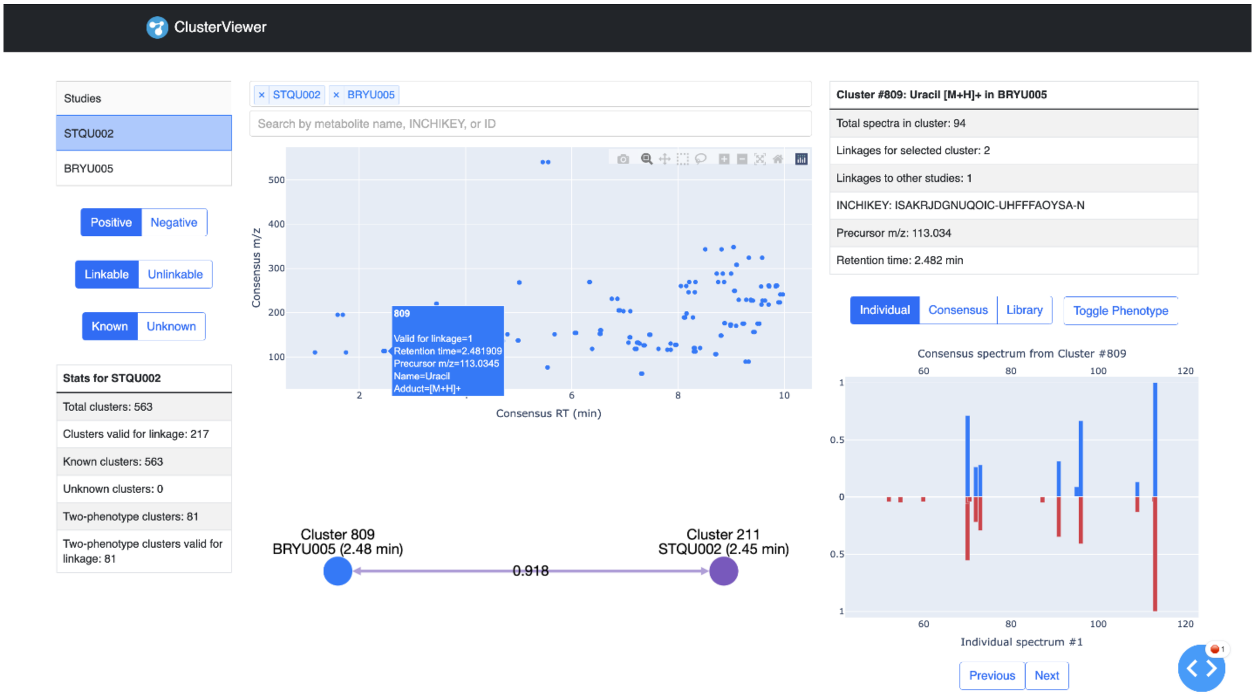


Figure 1: ClusterBase Outline.
a) the ClusterBase project starts files derived from MS-Dial, which in turn accepts raw instrument output. The Clusterbase project is outlined in red. These files are ingested into a database and then network analysis is performed. b) network analysis occurs between studies, where nodes are metabolites and edges between nodes are formed based on similarity (retention time, MS2, etc.) In this way, the nth study can be automatically annotated based on previous studies. c) After many studies, subgraphs form, and the sample metadata from studies can be associated in order to identify unknowns of interest.

data and reduced turn-around time for clients. Third, the networking allowed for the organized association of unidentified compounds that were observed in independent studies. Determining metabolites that were observed for samples with metadata of interest (certain diseases, ages, etc.), allowed the team to focus on those metabolites for the process of chemical identification. in This allows for the retrospective identification of trends among those samples from disparate studies but common metadata attributes, i.e., the appearance of certain metabolites in certain diseases. These benefits are outlined in **Figure 1**.

The process of ClusterBase can be outlined as follows. ClusterBase begins with the ingestion of files from MS-Dial that are in turn derived from raw mass spectrometer files. MS-Dial generates 1) an "individual file" for every sample, which contains a set of observed peaks, an MS/MS spectrum (MS2), adducts, intensities, etc. 2) an "alignment file" for every study (set of samples), which contains a sample-feature matrix, where each sample is one of the individuals and each feature is the union of all peaks sets among indiviuals. Magnitudes are derived by algorithmically shifting/stretching the features in each samples' mz-rt space, and 3) a mapping file containing key-value relationships between the summary alignment file and each of the individual files.

For the nth study, ClusterBase extracts all information from the alignment and individual files and coerces those data into a relational database. ClusterBase then performs network analysis among the aligned features from the nth study against the aligned features from all other studies that it has seen thus far. Each feature becomes a node, and edges between nodes and among are created if the nodes in the nth study have sufficiently close retention times, precursor masses, adducts, MS2 similarities, etc. to nodes from previous studies. Clusterbase then chooses those nth study nodes that have sufficiently confident identifications based on connections to earlier nodes, and returns predicted identities to the user for verification. Upon verification, CluserBase updates properties of node-groups, such as consensus spectra and consensus retention times. A user interface for ClusterBase was developed and is showcased in **Figure 2**.



Stats for STQU002

Total clusters: 563
Clusters valid for linkage: 217
Known clusters: 563
Unknown clusters: 0
Two-phenotype clusters: 81
Two-phenotype clusters valid for linkage: 81

Cluster #809: Uracil [M+H]+ in BRYU005

Total spectra in cluster: 94
Linkages for selected cluster: 2
Linkages to other studies: 1
INCHIKEY: ISAKRJDGNUQOIC-UHFFFAOYSA-N
Precursor m/z: 113.034
Retention time: 2.482 min

Figure 2: ClusterBase user interface. In this example, for visual simplicity, only two studies are loaded (STQU002 and BRY005). A study is selected and the aligned mz-rt space is displayed (center) along with study metadata (left). A peak can be selected from the mz-rt space, which generates the corresponding subgraph (bottom-center), cluster metadata (top-right), and spectra (bottom-right).

The entire process is heavily parameterized (retention time cutoffs, consensus spectrum binning parameters, etc.) and therefore we needed to determine the optimal parameters. To do this, we used a customized hyperparamter optimization process. This involved comparing the identities in the automatically generated subgraphs to identities that were manually generated over the past several years by the Biohub metabolomics team. For a single set of parameters, the entire network was generated. Then, we determined the precision and recall of each identity. For each identity, we examined the subgraph with the greatest percent population. The precision was treated as the fraction of nodes in that subgraph with that identity. The recall was treated as the fraction of nodes in the entire network that belonged to that subgraph. We could then take the average of those precisions and recalls for each identity and assign a value for the entire network for that hyperparameter combination. We chose a parameter set that offered good precision in order to have confidence in our identities.

Throughout this process, we challenged many assumptions in the software that we used. One of the assumptions worth mentioning is shown in **Figure 3**. Here, we show that MS-Dial's alignment process is not completely reliable.

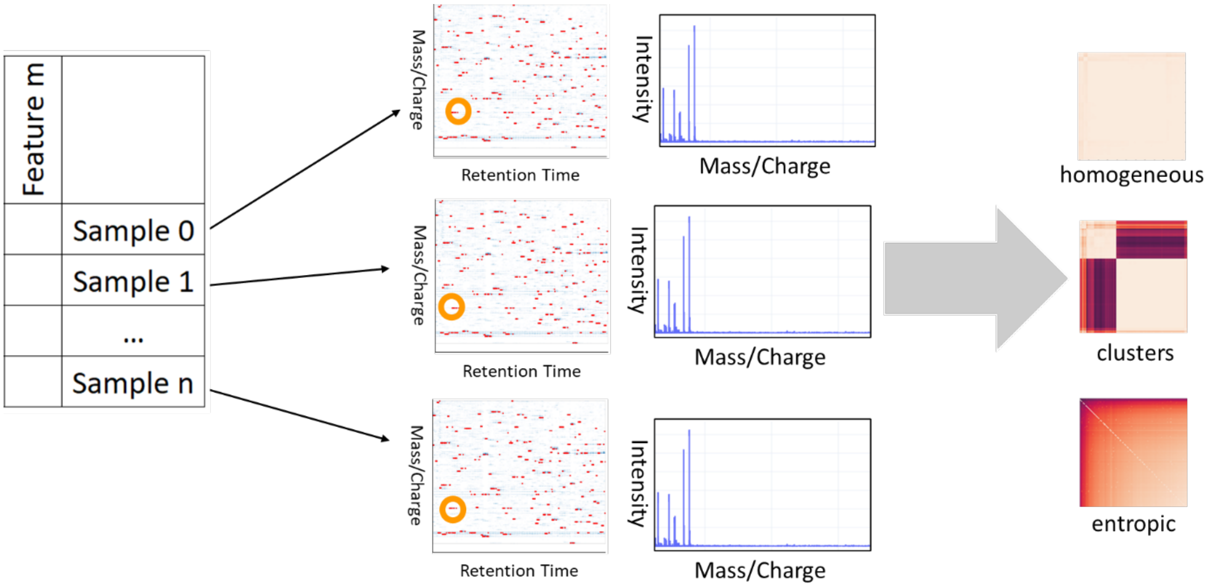


Figure 3: Explorations into MS-Dial's alignment. In MS-Dial, the mz-rt spaces among samples are aligned based on intensities in the space. This is challenging, especially when aligning matrices that are qualitatively different (e.g., plasma vs liver). We checked the alignment by generating similarity matrices for the MS2 associated with a particular feature. This revealed three cases: in the homogenous case, the alignment succeeded and all MS2 were relatively similar. In the subcluster case, where multiple distinct metabolites were identified as the same feature. In the entropic case, a noise element is assigned to be a metabolite.

Part 2: CFM-ID Webtool Automation

One of the outstanding challenges in metabolomics is to identify what compounds are present in complex mixtures from biological samples such as human plasma. The traditional method is an information retrieval problem – a database of (identity:observables) pairings is generated, and for each observable in the sample, the identity corresponding to the best-matched observables is chosen. There is a significant bottleneck in the generation of these pairings because each pair must be manually synthesized and tested. Therefore there is great interest in in silico tools to aid the generation of these libraries.

In this project, in conjunction with colleagues Dr. Arpana Vaniya and Fanzhou Kong, we compared in silico tools for compound identification. I was tasked with identifying 500 compounds using, CFM-ID, the tool extensively explored in Chapter 1. CFM-ID has an online component that extends its functionality beyond library generation into the realm of compound identification. Unfortunately, this tool does not have a "batch mode", so, in this project, I wrote a pipeline that automated the webtool. We identified those 500 compounds with this automation wrapper and then performed statistical analysis on the identification attempt. The workflow is outlined in **Figure 1**.

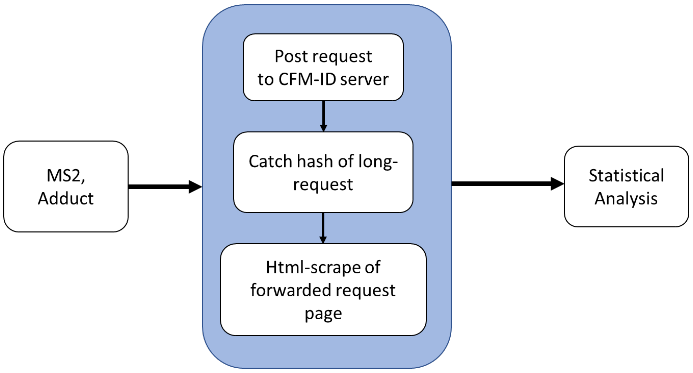

MS2,
Adduct
Post request
to CFM-ID server
Catch hash of long-request
Html-scrape of forwarded request page
Statistical
Analysis

Figure 1: Workflow of CFM-ID automation.

The challenging aspect of this work is that the CFM-ID webtool is not intended for automatic use. Therefore, we had to deconstruct the webtool from the outside and write custom requesters and scrapers. After doing this, the results were analyzed. For each compound attempted, we were interested in how far from the top suggestion was the genuine identity. Our results are shown in **Figure 2**. In this figure, we see that CFM-ID dramatically underperforms similar tools. We spoke with developers, and determined that the CFM-ID webtool had some kind of error. After testing, it was discovered that the CFM-ID webtool database had become corrupted and that they needed to regenerate most of their MS2 predictions.

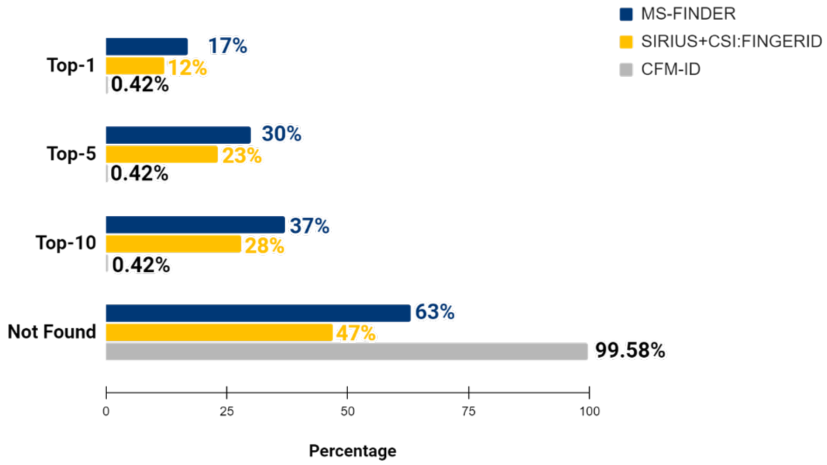


Figure 2: Comparison of CFM-ID to other top compound identification tools.

Part 3: MS2 Intensity Prediction

Throughout my Ph.D. I had the privilege to help others as well as explore my own scientific ideas. I am truly grateful for the excellent funding in the Fiehn lab that afforded me the time to pursue my scientific passions.

One of the avenues that I explored was the creation of an intensity predictor to augment quantum mechanical predictions of MS2 spectra. The quantum mechanical predictions modeled the fragmentation process well, and therefore had a good jaccard score when compared to their

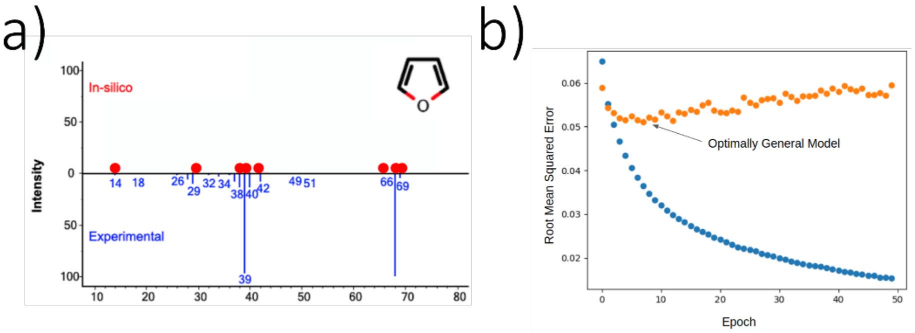


Figure 1: Outline of Spectral Prediciton problem. a) the goal is to predict an intensity for each peak in a predicted spectrum. Therefore, we have n predictions per compound, where n is the number of non-zero m/z values. b) performance vs. training iteraction shows that we quickly achieve optimized generalization.

experimental analogs, but did not capture the intensities well. We suspected this was because they did not model interactions with the detector. Therefore, we were motivated to create a tool to replace the quantum mechanically predicted intensities, as illustrated in **Figure 1a**. Here, we show that, we rely on quantum mechanics for the substructures, and therefore the m/z values, but we desire learn to predict an intensity for each peak. Therefore, every compound actually requires n prediction, where n is the number of peaks in the spectrum. We think of the relative intensities as a competition, and therefore, as feature input, utilize structural information via graph fingerprints as well as intensities of surrounding peaks.

The success is shown in **Figure 1b**, where we quickly arrive at a generalized model. Indeed, we encountered the problem often seen with machine learning in mass spectrometry – the complexity of the physics and statistical mechanics is barely encapsulated in the descriptor space. Moreover, interpolation is extremely difficult because of the complexity of these processes. Indeed, we quickly achieve an optimally generalized model, despite the fact that our model can continue to over-optimize on its training set.

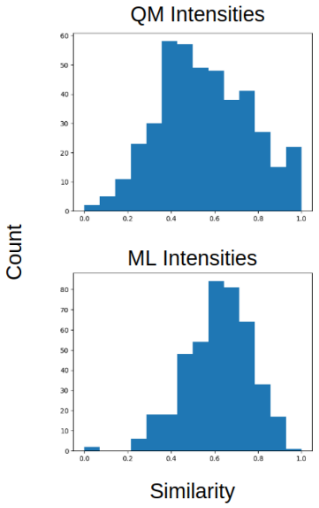


Figure 2: Similarity distributions with quantum mechanical and machine learning spectra.

Ultimately, we achieved moderate success in this approach, as illustrated in **Figure 2**. Here, we see that we have slightly increased the similarity of predicted spectra to their empirical spectra.

re directions would include transforming this regression problem into a classification problem, in order to remove the tendency for predictions to be a moderate intensity.