Biological Filtering and Substrate Promiscuity Prediction for Annotating Untargeted Metabolomics

Neda Hassanpour[✉], Nicholas Alden^Δ, Rani Menon[✉], Arul Jayaraman[✉], Kyonbum Lee^Δ, and Soha Hassoun^{✉Δ*}

Department of Computer Science, Tufts University, Medford, MA

Department of Chemical and Biological Engineering, Tufts University, Medford, MA

Department of Chemical Engineering, Texas A&M, College Station, TX

KEYWORDS *Metabolomics, metabolite annotation, enzyme promiscuity, extended metabolic models*

ABSTRACT: Mass spectrometry coupled with chromatography separation techniques provides a powerful platform for untargeted metabolomics. Determining the chemical identities of detected compounds however remains a major challenge. Here, we present a novel computational workflow, termed Expanded Metabolic Model Annotation (EMMA), that aims to strike a balance between discovering previously uncharacterized metabolites and the computational burden of annotation. EMMA engineers a candidate set, a listing of putative chemical identities to be used during annotation, through an expanded metabolic model (EMM). An EMM includes not only canonical substrates and products of enzymes already cataloged in a database through a reference metabolic model, but also metabolites that can form due to substrate promiscuity. EMMA was applied to untargeted LC-MS data collected from cultures of Chinese hamster ovary (CHO) cells and murine cecal microbiota. EMM metabolites matched, on average, to 23.92% of measured masses, providing a > 7-fold increase in the candidate set size when compared to a reference metabolic model. Many metabolites suggested by EMMA are not catalogued in PubChem. For the CHO cell, we experimentally confirmed the presence of 4-hydroxyphenyllactate, a metabolite predicted by EMMA that has not been previously identified as part of CHO cell metabolism.

Metabolomics is an expanding field of research that involves the characterization of small molecules in cells, tissues and other biological systems. Metabolites are direct products of enzymatic reactions and provide a functional readout of cellular state¹⁻². Compared to genes and proteins that are regulated and post-translationally modified, respectively, metabolites are most predictive of the phenotype³. Metabolomics now plays a critical role in many fields including drug discovery and precision medicine, nutritional analysis, and examining environmental responses. Importantly, the ability to collect thousands of measurements on the metabolome using untargeted metabolomics, where thousands of features are measured within the sample under study, and to annotate these measurements with chemical identities, promises to broadly profile the metabolome and revolutionize phenotyping and biological discoveries.

Mass Spectrometry (MS) techniques coupled with liquid or gas chromatography separation techniques, LC-MS or GC-MS, have become standard analytical platforms for untargeted metabolomics⁴. The LC or GC step aims to separate compounds within the sample, while the MS step ionizes, fragments and detects a fragmentation pattern. There are now techniques for data processing (e.g., peak picking, missing value imputation, and adduct and degenerative feature removal). These tools convert raw MS data into features. Each feature corresponds to an ionized chemical compound, and is characterized using a *spectral signature*, comprising a chromatographic retention time (RT) paired with mass-to-charge ratio (m/z) and relative intensities for the parent compound and its fragments.

Interpretation of metabolomics data is facilitated by assigning putative chemical identities to the features. Relying on the mass of the ionized parent compound for annotation is problematic as a particular mass may be associated with many chemical formulas (e.g., there are 10,132 known molecular structures in PubChem⁵ that are associated with $C_{20}H_{22}N_2O_4$)⁶. The spectra of detected compounds can be matched against those within an in-house library generated using the same instrument and method as the samples. However, this is impractical due to the large number of compounds detected in an untargeted MS experiment. Instead, feature annotation typically relies on libraries in reference databases. The two largest spectral databases in terms of number of unique compounds, METLIN⁷ and NIST⁸, cover less than 17,000 compounds each, which is less than 0.002% of the ~96,000,000 compounds catalogued in PubChem. Due to these limitations, the annotation rate, which we define as the fraction of features annotated with a putative chemical identity, using in-house or spectral databases is typically low. The maximum annotation rate across several metabolomics studies that we surveyed was 16% but averaged only 7.26%⁹⁻¹⁵.

In recent years, computational tools have become available to recommend a ranked list of chemical structures that best explain a spectral signature. This ranked list is selected amongst a pre-specified *candidate set*, a listing of metabolites with

formula weights that match the measured masses of parent compounds in the sample. Earlier tools used rule-based approaches to generate fragmentation patterns of candidate metabolites¹⁶⁻¹⁸. Subsequent efforts introduced combinatorial enumeration methods¹⁹⁻²² and machine-learning algorithms. For example, CFM-ID²³ uses the candidate set to create a probabilistic model of collision-induced fragmentation process. The model is then used to predict a fragmentation pattern for a given compound. CSI:FingerID²⁴ first predicts a fragmentation tree based on a spectral signature²⁵. CSI:FingerID then uses multiple-kernel learning²⁶ and support vector machines to predict fragmentation tree properties, which are searched against fragmentation tree properties of compounds in a molecular structure database. CSI:FingerID, which outperforms other tools^{19, 23, 27-29} in terms of accuracy. Despite progress, however, the annotation runtimes, are problematic. For example, annotating 3,868 spectral signatures in GNPS, a public repository of mass spectra data³⁰, using CSI:FingerID against PubChem structures was estimated to take 12 days and 12 hours²⁴. While there are tradeoffs between accuracy and runtime^{24, 31}, the runtime for all tools is dependent on the size of the candidate set. Hence, evaluating candidate sets from large compound databases such as PubChem and ChemSpider remain problematic.

We propose a novel annotation workflow for untargeted metabolomics that addresses current limitations regarding spectral database coverage and computational cost of annotation. The key step is to *filter* the detected masses through a metabolic model that we call an Expanded Metabolic Model (EMM). An EMM includes not only the defined substrates and products of enzymes cataloged for the organism(s) associated with the sample, but also additional metabolites reflecting the potential for promiscuous enzymatic activities. The central premise is that an EMM can be used to define a candidate set that is more comprehensive than a standard genome-scale metabolic model, but still enforces a degree of specificity for the system of interest. Our workflow, termed EMMA (EMM-based Annotation), broadens the search space for annotation beyond compounds in a reference metabolic model assembled from catalog definitions of enzymatic reactions, thus enhancing discovery while avoiding the computational cost of analyzing every compound in large chemical structural databases. We demonstrate the utility of EMMA on untargeted LC-MS data from cultures of Chinese hamster ovary (CHO) cells and bacterial isolates from murine cecum. We compare the candidate sets from reference metabolic models, EMMs, and a large structural database (PubChem). EMMA suggests biologically relevant chemical identities for almost a quarter of measured features, providing a > 7-fold increase in the candidate set size when compared to using a reference metabolic model. Importantly, EMMA allows the discovery of novel relevant putative identities that are not currently catalogued in PubChem. Targeted LC-MS experiments confirmed the presence of a predicted CHO cell metabolite that has previously not been cataloged as a Chinese hamster enzyme substrate or product.

METHODS

To describe and evaluate the EMMA workflow, we present it alongside two other annotation workflows (**Figure 1**). In describing the workflows, “annotation” refers to the use of any computational annotation tools. A *model-based* annotation workflow (**Figure 1A**) consists of filtering masses of measured metabolites against those expected in the sample based on a metabolic model that is built from a reference genome (or set of reference genomes). Model metabolites with exact masses that match, within a small error, to measured masses are designated

Figure 1. Comparison between annotation workflows. The candidate set for annotation is derived by filtering the measured masses based on: (A) the metabolic model, (B) databases, and (C) extended metabolic model (EMM). The candidate sets in (A) and (C) are biologically relevant, while candidates in (B) prior to filtering may not all be biologically relevant.

as the candidate set. The candidate set is then annotated, where candidates that best explain the experimentally observed spectra are ranked. This workflow provides two advantages. Metabolites within the candidate set are all biologically relevant. Consequently, all computing time will be used to evaluate biologically relevant candidates. While there is now a growing collection of annotated genome sequences (e.g. KEGG database³², BioCyc³³, and BiGG³⁴) and tools for the reconstruction of genome-scale metabolic models (GEMs)³⁵⁻³⁶, the completeness of these models is not guaranteed. GEMs models are typically constructed using sequencing and annotation³⁷⁻³⁸. Significant experimental and computational efforts are required to augment the models based on gene expression, proteomics, and metabolomics data³⁹. Current models do not account for enzyme promiscuity, where an enzyme transforms alternate substrates in addition to its natural substrate, as defined by a reference metabolic model and as catalogued in organism databases⁴⁰⁻⁴³. As a result, defining the candidate set based only on metabolites within the metabolic models naturally limits annotation.

Selecting the candidate set from a large database can potentially enhance annotation by increasing the number of measured masses that have a match in the candidate set (**Figure 1B**). This *database-based* workflow first identifies candidate metabolites by querying one or more specified compound databases for all molecules whose exact masses match the experimentally measured masses detected in the sample. The resulting candidate set is then annotated and ranked as in the workflow shown in **Figure 1A**. As the size of the candidate set is large in comparison to the one in the model-based workflow, the annotation runtime increases, and so does the chance of annotation. As annotation accuracy is low, some measurements, however, may be annotated with biologically irrelevant identities, e.g., phytochemicals or drug compounds that cannot possibly accumulate in a mammalian cell culture. The end user could then sift through the ranked candidate metabolites to select biologically relevant candidates. This manual curation is time-consuming and relies on the user’s judgment. Using a metabolic model to filter the ranked candidate metabolites can facilitate this process. However, this results in the same discovery-related limitations as the workflow shown in **Figure 1A**, while also incurring a large computational cost. Importantly, not all the computational cost is necessary. It is highly unlikely that every compound in candidate sets derived from large database is biologically relevant. Using manually curated metabolite databases such as KEGG to derive the candidate set is an attractive option, as the size of the

Table 1. Size of experimental data sets and models

	(A) Experimental data			(B) Metabolic model			(C) Expanded metabolic model using PROXIMAL			(D) Fold change for EMM relative to metabolic model	
Biological sample	Dataset	MS mode	Number of measured masses	Number of reactions	Number of metabolites	Number of unique masses	Number of unique operators	Number of unique derivatives	Number of unique derivative masses	Number of metabolites	Number of unique masses
CHO cell	HiNeg	negative	2,502	1,619	1,353	775	2,392	76,745	17,930	56.72	23.14
	HiPos	positive	3,856								
	SynNeg	negative	5,336								
gut microbiota	Neg	negative	1,651	1,381	1,307	779	2,756	94,186	23,356	72.06	29.98
	Pos	positive	1,657								

(A) Three experimental datasets under different conditions were collected for the CHO cell, and two for the gut microbiota sample. (B) The size of the metabolic model: number of reactions, metabolites, and unique masses. (C) The size of the expanded metabolic model: number of operators derived using *PROXIMAL*, unique derivatives generated by *PROXIMAL*, unique derivative masses due to *PROXIMAL*. For comparison purposes, the numbers of derivatives and derivative masses exclude those in the metabolic model. (D) Fold increase in number of metabolites and masses when comparing the size of these sets for EMM against the metabolic model.

candidate set is reduced when compared to using a large structural database. However, not all biologically relevant compounds are catalogued in such databases.

Our novel annotation workflow (**Figure 1C**), EMMA, applies an EMM-based filter to identify the candidate set. To create this model, we adopt a previously described method, *PROXIMAL*⁴⁴ (**Supplementary Methods**). From reactant-product pair(s) (RPAIR) of an enzymatic reaction⁴⁵, *PROXIMAL* identifies a molecular pattern that transforms the reactant into product. Each pattern is associated with a reaction center and its first and second-level neighboring atoms. If a substrate of interest matches a pattern, then the corresponding operator is applied to generate a product, which we call a “derivative” metabolite. The EMM for a system of interest is generated using *PROXIMAL* by applying the operators generated from the enzymatic reactions encoded in the system’s genome(s) to all of metabolites already associated with the system based on the enzymes’ reaction definitions. This generates a set of *derivative* metabolites. The calculated exact masses of derivative metabolites are then used to filter the measured masses. If a derivative has a mass that matches a measured mass, then the KCF or SMILES string of this derivative is searched against a chemical structure database (PubChem) to determine if it has been cataloged with a chemical name and identifier. The masses of metabolites in the reference metabolic model are also matched against the measured masses (as in **Fig 1A**). The union of matched derivatives and reference model metabolites constitute a biologically relevant candidate set. This candidate set is then used for annotation and the candidates are ranked, as in prior workflows. Pseudo-code for the EMMA workflow is provided in **Supplementary Methods**.

RESULTS

Datasets, reference metabolic modes, and EMMs

We compared the annotation workflows in **Figure 1** by analyzing untargeted LC-MS data collected on samples from two different biological systems (**Table 1, column group A**). One set of LC-MS experiments were performed on samples from Chinese hamster ovary (CHO) cell cultures grown in a chemically defined medium. The second set of experiments was performed on samples from anaerobic cultures of bacteria collected from murine cecum. Each set of LC-MS experiments comprised two or more different methods. By treating the datasets

independently, we were able to explore the influence of sample source and instrument method on EMMA’s performance. Details for the culture and LC-MS experiments are provided in **Supplementary Methods**. The processed data were arranged into feature tables, where each feature was specified by a chromatographic retention time (RT), measured mass (*m/z*), and a set of associated product ion (fragment) masses and their relative intensities, i.e., MS/MS signature. The reference metabolic models for CHO cell and murine cecal microbiota were derived from genomes in the KEGG database. For the CHO cell, we obtained lists of metabolites and reactions cataloged in KEGG that are associated with the organism code *cge*. The cecal culture is a consortium of many species. We assembled a community-level model based on the taxonomic groups detected in the culture using a previously described procedure⁴⁶. The numbers of reactions, metabolites and unique masses in the two reference models are listed in **Table 1 (column groups B)**.

The EMM for each sample was generated using biotransformation operators derived from each model (**Table 1, column group C**). EMMs augment a metabolic model to include molecules that are not originally part of the metabolic model. This augmentation increases the number of unique masses within the model. The number of biologically relevant molecules in the candidate set thus significantly increases (**Table 1, column group D**) when compared to the number of metabolites in the reference metabolic model (57x and 72x for CHO cell and the gut microbiota, respectively). Similarly, the number of unique masses in EMM is increased over the number of unique masses in the reference metabolic model (23x and 30x for the CHO cell and the gut microbiota, respectively). EMMs thus promise to provide a large annotation space when compared to the reference metabolic model.

Novel Annotation opportunities

Compared to using the reference metabolic model for a biological sample, using an EMM as the search space during metabolite annotation increases the size of the candidate set for annotation in terms of: (a) matching to a larger number of measured masses, and (b) suggesting a larger set of putative chemical identities. Using these two metrics, we compare the size of the “biologically relevant candidate sets” in the model- and EMMA-based workflows and compare that with the size of the candidate set using PubChem (**Table 2**). A small percentage of

Table 2. Candidate set size using different workflows

Biological sample	(A) Metabolites in metabolic model			(B) All EMM derivatives			(C) EMM derivatives with previously known chemical IDs			(D) Using PubChem-based filtering	
	Number of measured masses matched to those in metabolic model	Percentage of measured masses matched to those in metabolic model	Number of chemical IDs associated with measured masses	Number of masses matched to those in EMM	Percentage of masses matched to those in EMM	Number of unique mass-matched derivatives in EMM but not in the model	Number of masses matched to those with previously known chemical IDs	Percentage of masses matched to those with previously known chemical IDs	Number of previously known chemical IDs for EMM derivatives that mass-match to measurements	Number of unique mass matches in PubChem	Number of corresponding chemical IDs associated with measured masses
CHO cell	118	4.72%	178	678	27.10%	2,725	174	6.95%	386	3,951,635	7,657,564
	75	1.95%	93	715	18.54%	2,729	132	3.42%	226	3,362,305	6,406,877
	198	3.71%	229	1,490	27.92%	4,944	293	5.49%	527	7,058,696	14,133,885
gut	51	3.09%	131	445	26.95%	2,470	77	4.66%	207	2,448,238	5,192,205
microbiota	36	2.17%	43	316	19.07%	1,236	84	5.07%	149	2,774,074	5,572,587
Averages	96	3.13%	135	729	23.92%	2,821	152	5.12%	299	3,918,990	7,792,624

(A) Candidate set size when using the model: number of measured masses that match to metabolites in the model, the equivalent percentage of the number of measured masses reported for experimental data in **Table 1**, and corresponding number of chemical identities. (B) Candidate set size when using EMMA: number of measured masses that match to metabolites in the EMM, equivalent percentage in reference to the number of measured masses reported for experimental data in **Table 1**, and corresponding number of chemical identities. (C) Further filtering of the EMM derivatives reported in column group (B) to include only mass measurements that match to previously known chemical IDs as reported in PubChem, and reporting the number of matched masses, the relative percentage of these masses to the number of measured masses reported for experimental data in **Table 1**, and the corresponding number of chemical IDs. (D) Size of the candidate set when filtering using PubChem.

the measured masses are matched to masses of metabolites in the metabolic model. On average, 3.31% of measured masses can be potentially annotated using the metabolic model only. When using the EMMs, this number increases to 23.92%, a 7.6-fold increase. When restricting the EMM derivatives to those that have a catalog entry in PubChem, the annotation rate drops to 5.12% as there are many compounds that are not yet catalogued in PubChem, currently the largest structural database. Using PubChem, the number of mass matches are in the millions. Not all such metabolites are biologically relevant. The use of reference metabolic models or large database such as PubChem therefore provide some limitations in annotation when compared to using EMMs. Using EMMA allows for novel biological discovery by suggesting biologically relevant compounds not in PubChem and reduces the annotation space considerably.

As the size of the KEGG database is significantly smaller than PubChem, KEGG may not be as computationally

prohibitive to use for annotation. Further, using only a biological database such as KEGG for annotation guarantees biological relevance of candidate metabolites. A question that often arises is regarding the benefits of utilizing a general database for annotation compared to when employing a database that mostly comprises biomolecules. Using the EMMA workflow as a reference and restricting derivatives to those with chemical identities in PubChem, we are able to explore and quantify the benefits. Specifically, we utilized the EMMA workflow to identify candidate sets for our datasets. We then compared the EMM candidates against those obtained using the database-based workflows using KEGG and PubChem (**Table 3**). Many candidate molecules identified by EMMA that are catalogued in KEGG (e.g. For CHO cell HiNeg data, 93 out of 174 candidate compounds). However, there were also EMMA candidate compounds found in PubChem that are not catalogued in KEGG. For our datasets and using EMM metabolites as a reference,

Table 3. Using EMMs to compare the annotation opportunities of PubChem against KEGG

Biological sample	(A) Experimental Data		(B) In EMM and in KEGG		(C) In EMM and PubChem, and not in KEGG		(D) Lower-bound fold increase of PubChem over KEGG	
	Dataset	Number of measured masses	Number of matched masses	Number of candidate chemical IDs	Number of matched masses	Number of candidate chemical IDs	Number of matched masses	Number of candidate chemical IDs
CHO cell	HiNeg	2,502	56	93	118	200	2.11	2.15
	HiPos	3,856	26	39	106	148	4.08	3.79
	SynNeg	5,336	88	122	205	283	2.33	2.32
gut microbiota	Neg	1,651	25	47	52	113	2.08	2.40
	Pos	1,657	23	28	61	93	2.65	3.32
Average							2.65	2.80

(A) Experimental data for different data sets (repeated for convenience). (B) Number of matched masses and candidate chemicals found using EMMA that are reported in KEGG. (C) Number of matched masses and candidate chemicals found using EMMA reported in PubChem but not in KEGG. (D) Lower-bounds on discovery of biologically relevant matched masses and candidate chemicals when using PubChem over KEGG.

Table 4. Right candidate metabolites identified by EMMA used for experimental validation

(A) Candidate metabolites		(B) KEGG		(C) PubChem		(D) PROXIMAL		(E)
Mass measurement (Daltons)	Candidate metabolite identified by EMMA	Rank	Matches	Rank	Matches	Number of Reactions used to derive operator	Number of E.C.s. associated with reactions	Experimentally validated?
122.04	Salicylaldehyde	1	1	1	1	1	1	No
182.06	4-Hydroxyphenyllactate	1	2	1	4	12	15	Yes
101.05	Acetoacetamide	1	1	2	3	1	1	No
117.79	5-Aminopentanoate	1	2	1	5	4	4	No
132.04	Glutarate	1	1	3	6	12	11	No
167.06	3-Methoxyanthranilate	1	1	2	3	8	2	No
152.05	2-Hydroxyphenylacetic acid	NA	1	1	4	1	1	No
183.05	4-Pyridoxate	NA	0	1	1	1	1	No

(A) Candidate mass and name. (B) Ranking of metabolite and number of candidates that matched mass measurement using KEGG. (C) Ranking of metabolite and number of candidates that matched mass measurement using PubChem. (D) The number of reactions that yielded the *PROXIMAL* operator that yielded each candidate metabolite and the associated number of enzymes that catalyze these reactions. (E) The status of experimental validation

there is at least 2x or more additional biologically relevant candidates in PubChem for each candidate identified in KEGG. The 2x-fold increase over KEGG is a lower bound on the number of biologically relevant metabolites in PubChem. Using a large database such as PubChem thus significantly increases biologically relevant annotation opportunities when compared to KEGG. Relying only on small biological databases limits annotation. EMMA provides an alternative candidate set that provides different tradeoffs between annotation opportunities and speed.

Computational time required for annotation

To generate the candidate set as the input to *in silico* annotation analysis in database-based workflow (Figure 1B), we identified metabolites in the KEGG and PubChem databases that mass-matched within 10 ppm to the masses in our experimental data for each dataset. We investigate the computational time required to annotate the candidate sets from EMMs and from the combined PubChem and KEGG databases. Annotation of each candidate set identified by EMMA required a handful of hours, averaging 2.5 hours per data set (Table S1, group A). The number of candidate metabolites from databases PubChem and KEGG for each of our datasets exceeded 5 million candidates, with an average dataset size of 7.8 million candidates (Table S1, group B). It was computationally prohibitive to annotate all mass-matched metabolites from the databases. To calculate the required runtime, we estimate it using annotation runtimes based on the EMMA workflow (Table S1, group A). Dividing the runtime by number of metabolites in the candidate set, on average, annotation requires 0.0085 hours per match. Using this average, the estimated runtime for annotation of database-based workflow is computed for each dataset. The average required runtime per dataset is over 65,000 hours (Table S1, group B).

Experimental validation of EMMA

We next investigated whether any of the derivatives predicted by EMMA and matched to a detected MS feature based on mass and MS/MS signature could be experimentally confirmed with a chemical standard. We selected eight predicted derivatives that matched an LC-MS feature for CHO cell samples (Table 4, group A). The selection was based on two factors: the rank assigned by the *in silico* annotation tool and availability from a vendor. The selected derivatives are: salicylaldehyde, one of the three isomers of hydroxybenzaldehyde; 4-hydroxyphenyllactate, a tyrosine metabolite; acetoacetamide, a monocarboxylic acid amide of acetoacetic acid; 5-aminopentanoate, a lysine degradation product; glutarate, produced in lysine and

tryptophan metabolism; 3-methoxyanthranilate, an ester of anthranilic acid; 2-hydroxyphenylacetic acid, associated with styrene degradation pathway; and 4-pyridoxate, a product of vitamin B₆. When using KEGG as a database for annotation, CFM-ID ranked 6 of the 8 derivatives as the highest ranked candidates, while two of the derivatives were not in KEGG (Table 4, group B). Further, a small number of candidate matches were found for each mass measurement. When using PubChem as a database for annotation, all derivatives ranked among the three top candidates (Table 4, group C). As expected, the number of putative matches increased when compared to the number of matches using KEGG. We analyzed the number of reactions in CHO that contributed an operator that was used to generate each derivative and the number of E.C. numbers that are associated with each set of reactions (Table 4, group D). The number of reactions and enzymes varied for each derivative. For example, 12 different reactions catalyzed by 15 enzymes correspond to the operator that generated 4-hydroxyphenyllactate, whereas only one reaction and enzyme corresponds to the operator that generated acetoacetamide.

Comparing the RTs and MS/MS spectra of standards for these chemicals against the corresponding CHO cell culture sample features (Supplementary Methods), we were able to confirm correct annotation of 4-hydroxyphenyllactate (Figure 2). This demonstrates that the EMMA can indeed identify a novel metabolite that is not found among metabolites cataloged

Figure 2. Mirror plot for 4-hydroxyphenyllactate, KEGG compound C03672. (A) experimental data collected using untargeted metabolomics from the CHO cell culture. (B) data from high-purity chemical standard. This is considered a match by retention time (RT difference < 3 minutes) and by MS/MS (spearman rank correlation p-value < 0.05 and r-value > 0.6)

for the organism of interest, in this case the Chinese hamster. In addition to KEGG, we searched for 4-hydroxyphenyllactate in BioCyc and MetaCyc. Neither database associated this metabolite with the Chinese hamster. In KEGG, 4-hydroxyphenyllactate is associated with three enzymatic reactions. Reactions 4-hydroxyphenyllactate:NAD⁺ oxidoreductase (R03336) and 4-Hydroxyphenyllactate:NADP⁺ oxidoreductase (R03338) are both catalyzed by D-hydrogenase (E.C. 1.1.1.222, which was recently deleted and transferred to E.C. 1.1.1.110) and hydroxyphenylpyruvate reductase (E.C. 1.1.1.237). Reaction 3-(4-hydroxyphenyl)lactate hydro-lyase (4-coumarate-forming) (R08766) is associated with an enzyme that have yet to be characterized (E.C. 4.2.1.-). It is unlikely that the source of 4-hydroxyphenyllactate in our sample is exogenous, as our cell culture medium was chemically defined and did not include this metabolite. Further evidence that the metabolite is endogenously derived is provided by a recently updated GEM for the CHO cell in the BiGG database⁴⁷. This GEM reconstruction (iCHOv1) included 4-hydroxyphenyllactate as a “universal” metabolite that could be formed enzymatically, but for which a specific gene encoding the enzyme remains unknown.

DISCUSSION AND CONCLUSION

Our EMMA workflow addresses the challenge of creating an annotation candidate set that is enzymatically relevant to the sample under study and that includes metabolites beyond what is already catalogued in reference metabolic models. One important contribution of the work is conceptually separating the engineering of the candidate set from annotation. Only a limited number of prior works engineered candidate sets and in limited ways. These works focused on exclusions of particular elements, substructures or compounds³¹ or on inclusion sets⁴⁸. Filtering PubChem compounds using PubMed Medical Subject Heading (MeSH) labels⁴⁹ can reduce the candidate set size by including only naturally occurring compounds that are biologically relevant (e.g. carbohydrates, lipids, etc.). However, only a tiny fraction (124,049 compounds) of PubChem compounds has MeSH labels. Filtering using current MeSH labels was reported to reduce a candidate set of 62,782 structures that match in mass to 3,868 compounds in GNPS dataset to only 36 compounds⁵⁰.

Results from comparing the three workflows emphasize the need for optimizing the engineering of candidate sets. We demonstrated for our two case studies that using candidate sets from large databases is computationally prohibitive, as others have also noted²⁴. We also demonstrated that using a biological database such as KEGG yields a small candidate set when compared to using a large structural database. We further demonstrated that using a reference metabolic model is inadequate as only a very small percentage (3.31% on average) of measurements can be annotated. In this regard, EMMA contributes two key advances. First, filtering candidate chemicals using an EMM allows for the identification of novel metabolites that are missing from a GEM reconstruction. This advance addresses the need to enable discovery, which is inherently limited in the simpler approach of using a model comprising only known metabolites to filter the candidate chemicals or when using a small biological database, without incurring a prohibitive computational cost. Second, filtering the measurements through an EMM specific for the system of interest provides a biologically relevant and computationally feasible candidate set. This advance eliminates unnecessary and time-consuming computations on chemicals from large databases that are likely irrelevant to the system of interest. Not all biologically relevant candidates from a large database are in the EMM. This issue could be addressed by further expansion of the EMM candidate set by the

repeated application of the biotransformation operators derived from the reference model to derived promiscuous products. While we demonstrated the EMMA workflow using specific tools and databases (e.g., CFM-ID for annotation; PubChem and KEGG databases) the overall workflow is generic and can be readily modified to use other *in silico* annotation tools.

EMMA relies on a reference metabolic model for annotation. Other recent studies have also exploited the metabolic network to enhance annotation. One method, iMet, suggests that neighboring metabolites within a metabolic network have similar MS/MS spectra and trains a classifier to predict if two spectra belong to neighboring metabolites⁵¹. The classifier is trained using MS/MS spectra from spectral databases and mass differences between reactant pairs from KEGG that are not specific to the biological sample. Another method, BioCAN creates a network based on measured features and assigns aggregate annotation scores based on spectral lookups and annotation tools⁵². Mummichog maps measured features to metabolic models, and performs statistical pathway and module enrichment⁵³. There are also other studies that exploit putative biotransformation for annotation. In one method, the mass difference between a pair of features is matched against mass differences between substrate-product pairs of common metabolic conversions (e.g., oxygenation, acetylation, etc.), with a match indicating a potential biochemical transformation between the pair of detected feature masses⁵⁴. These transformations can be used to propagate metabolite annotation from an identified metabolite to its potential reactants and products. In contrast to this method, EMMA does not require any MS/MS training data and utilizes biological context that is specific to the sample. There is a common limitation when using metabolic models to improve annotation. Genome-scale metabolic reconstructions can be incomplete, especially for non-model organisms. Our biotransformation prediction by PROXIMAL suggests that the metabolite 4-hydroxyphenyllactate may result from the promiscuous activity of one or more carboxylic acid dehydrogenases expressed in the CHO cell on the substrate molecule 4-hydroxyphenylpyruvate.

This work presents the first *in vivo* experimental evidence for a computationally predicted metabolite derived through promiscuous action of an enzyme. Using a chemical standard, we confirm the presence of 4-hydroxyphenyllactate in a CHO cell culture, even though there is no documented gene associated with the CHO cell metabolism that can catalyze the reaction with 4-hydroxyphenyllactate as product. We were however able to confirm only one out of the eight predicted metabolites. This could be due to inaccuracies in the rankings by the annotation tool. For example, CSI:FingerID reports an accuracy of only 39.5% when annotating a data set from MassBank by searching PubChem. The low confirmation rate can also be due to the assumption that all enzymes are promiscuous. As an enhancement, we are currently investigating methods to improve PROXIMAL to rank predicted derivatives based on enzyme designations as generalists or specialists⁵⁵ and participation in primary or secondary metabolism⁵⁶.

Despite limitations due to the underlying potentially incomplete metabolic models and to the accuracy of current annotation tools, EMMA demonstrates utility in creating an expanded, biologically relevant candidate set and in utilizing it to enhance annotation. Importantly, EMMA promises to offer annotation opportunities beyond those possible with metabolic models without the high computational cost of searching large structural databases that contain many non-biological compounds.

ASSOCIATED CONTENT

Supporting Information

Supplementary Methods is a PDF file that provides a detailed description of PROXIMAL and EMMA.

AUTHOR INFORMATION

Corresponding Author. * soha@cs.tufts.edu

Notes. The authors declare no competing financial interest.

Author Contributions

S.H. conceived the EMMA workflow. N.H. performed all the computational work. K.L. conceived the experimental validation. K.L.'s lab provided the experimental data for CHO. N.A. conducted the experimental validation for the CHO Cell. A.J.'s lab provided the experimental data. R.M. conducted the experimental work on the murine cecal cultures. N.H., K.L. and S.H. edited the manuscript with input from N.A.

ACKNOWLEDGMENT

This work was supported by the National Institute of Health under project number 1R03CA211839-01 and NSF grant #1421972.

REFERENCES

1. Fiehn, O., Metabolomics—the link between genotypes and phenotypes. In *Functional Genomics*, Springer: 2002; pp 155-171.
2. Patti, G. J.; Yanes, O.; Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology* 2012, 13 (4), 263.
3. Raamsdonk, L. M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M. C.; Berden, J. A.; Brindle, K. M.; Kell, D. B.; Rowland, J. J., A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature biotechnology* 2001, 19 (1), 45.
4. Alonso, A.; Marsal, S.; Julià, A., Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology* 2015, 3, 23.
5. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem substance and compound databases. *Nucleic Acids Research* 2016, 44, D1202-D1213.
6. May, J. C.; McLean, J. A., Advanced Multidimensional Separations in Mass Spectrometry: Navigating the Big Data Deluge. *Annual Review of Analytical Chemistry* 2016, 9, 387-409.
7. Smith, C. A.; O'maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G., METLIN A Metabolite Mass Spectral Database. *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring & Clinical Toxicology* 2005, 27, 747-751.
8. Yang, X. New Features of the 2017 NIST Tandem Mass Spectral Library. http://chemdata.nist.gov/dokuwiki/lib/exe/fetch.php?media=chemdata:nist17:new_features_nist2017_tandem_mslibrary.pdf.
9. Lynn, K.-S.; Cheng, M.-L.; Chen, Y.-R.; Hsu, C.; Chen, A.; Lih, T. M.; Chang, H.-Y.; Huang, C.-j.; Shiao, M.-S.; Pan, W.-H., Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Analytical chemistry* 2015, 87 (4), 2143-2151.
10. Han, T.-L.; Yang, Y.; Zhang, H.; Law, K. P., Analytical challenges of untargeted GC-MS-based metabolomics and the critical issues in selecting the data processing strategy. *F1000Research* 2017, 6.
11. Wang, H.; Muehlbauer, M. J.; O'Neal, S. K.; Newgard, C. B.; Hauser, E. R.; Bain, J. R.; Shah, S. H., Recommendations for Improving Identification and Quantification in Non-Targeted, GC-MS-Based Metabolomic Profiling of Human Plasma. *Metabolites* 2017, 7 (3), 45.
12. Denihan, N. M.; Kirwan, J. A.; Walsh, B. H.; Dunn, W. B.; Broadhurst, D. I.; Boylan, G. B.; Murray, D. M., Untargeted metabolomic analysis and pathway discovery in perinatal asphyxia and hypoxic-ischaemic encephalopathy. *Journal of Cerebral Blood Flow & Metabolism* 2017, 0271678X17726502.
13. Romo-Hualde, A.; Huerta, A. E.; González-Navarro, C. J.; Ramos-López, O.; Moreno-Aliaga, M. J.; Martínez, J. A., Untargeted metabolomic on urine samples after α -lipoic acid and/or eicosapentaenoic acid supplementation in healthy overweight/obese women. *Lipids in health and disease* 2018, 17 (1), 103.
14. French, K. E.; Harvey, J.; McCullagh, J. S., Targeted and Untargeted Metabolic Profiling of Wild Grassland Plants identifies Antibiotic and Anthelmintic Compounds Targeting Pathogen Physiology, Metabolism and Reproduction. *Scientific reports* 2018, 8 (1), 1695.
15. Kera, K.; Fine, D. D.; Wherrett, D. J.; Nagashima, Y.; Shimada, N.; Ara, T.; Ogata, Y.; Sumner, L. W.; Suzuki, H., Pathway-specific metabolome analysis with 18 O 2-labeled Medicago truncatula via a mass spectrometry-based approach. *Metabolomics* 2018, 14 (5), 71.
16. Mass Frontier software, version 7.0. <http://highchem.com/index.php/support/>.
17. ACD/MS Fragmenter, version 12. http://www.acdlabs.com/products/adh/ms/ms_frag/.
18. Zhou, J.; Weber, R. J.; Allwood, J. W.; Mistrik, R.; Zhu, Z.; Ji, Z.; Chen, S.; Dunn, W. B.; He, S.; Viant, M. R., HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. *Bioinformatics* 2014, 30 (4), 581-3.
19. Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S., In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 2010, 11, 148.
20. Heinonen, M.; Rantanen, A.; Mielikäinen, T.; Pitkänen, E.; Kokkonen, J.; Rousu, J., FiD: new software for de novo identification of metabolite fragments from tandem mass spectrometry data. 2008.
21. Wegner, A.; Weindl, D.; Jager, C.; Sapcaru, S. C.; Dong, X.; Stephanopoulos, G.; Hiller, K., Fragment formula calculator (FFC): determination of chemical formulas for fragment ions in mass spectrometric data. *Anal Chem* 2014, 86 (4), 2221-8.
22. NIST Mass Spectrometry Data Center. Mass Spectrum Interpreter, ver. 2. <http://www.chemdata.nist.gov/mass-spc/interpreter/>.
23. Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D., CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research* 2014, 42.
24. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America* 2015, 112, 12580-5.
25. Rasche, F.; Svatos, A.; Maddula, R. K.; Bottcher, C.; Bocker, S., Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem* 2011, 83 (4), 1243-51.

26. Shen, H.; Dührkop, K.; Böcker, S.; Rousu, J., Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics* 2014, 30 (12), i157-i164.
27. Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J., Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* 2012, 28 (18), 2333-41.
28. Ridder, L.; van der Hoof, J. J.; Verhoeven, S.; de Vos, R. C.; Bino, R. J.; Vervoort, J., Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea. *Anal Chem* 2013, 85 (12), 6033-40.
29. Wang, Y.; Kora, G.; Bowen, B. P.; Pan, C., MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Anal Chem* 2014, 86 (19), 9496-503.
30. Global Natural Products Social Molecular Networking (GNPS) Database. <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp>.
31. Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics* 2016, 8 (1), 1.
32. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M., The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, 32 (Database issue), D277-80.
33. Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; Paley, S. M.; Subhraveti, P., The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2017.
34. King, Z. A.; Lu, J.; Dräger, A.; Miller, P.; Federowicz, S.; Lerman, J. A.; Ebrahim, A.; Palsson, B. O.; Lewis, N. E., BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 2016, 44 (D1), D515-22.
35. Feist, A. M.; Herrgård, M. J.; Thiele, I.; Reed, J. L.; Palsson, B. Ø., Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* 2009, 7 (2), 129-143.
36. Schellenberger, J.; Park, J. O.; Conrad, T. M.; Palsson, B. Ø., BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* 2010, 11 (1), 213.
37. Kim, T. Y.; Sohn, S. B.; Kim, Y. B.; Kim, W. J.; Lee, S. Y., Recent advances in reconstruction and applications of genome-scale metabolic models. *Current opinion in biotechnology* 2012, 23 (4), 617-623.
38. Saha, R.; Chowdhury, A.; Maranas, C. D., Recent advances in the reconstruction of metabolic models and integration of omics data. *Current opinion in biotechnology* 2014, 29, 39-45.
39. Baker, M., *Metabolomics: from small molecules to big ideas*. Nature Publishing Group: 2011.
40. D'Ari, R.; Casades, J., Underground metabolism. *Bioessays* 1998, 20 (2), 181-6.
41. Nobeli, I.; Favia, A. D.; Thornton, J. M., Protein promiscuity and its implications for biotechnology. *Nat Biotechnol* 2009, 27 (2), 157-67.
42. Tawfik, O. K.; Dan, S., Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry* 2010.
43. Khersonsky, O.; Malitsky, S.; Rogachev, I.; Tawfik, D. S., Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry* 2011, 50 (13), 2683-2690.
44. Yousofshahi, M. S.; Wu, C. L. K.; Hassoun, S., PROXIMAL: a method for Prediction of Xenobiotic Metabolism. *BMC systems biology* 2015, 9, 94.
45. Oh, M.; Yamada, T.; Hattori, M.; Goto, S.; Kanehisa, M., Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *Journal of chemical information and modeling* 2007, 47 (4), 1702-12.
46. Sridharan, G. V.; Choi, K.; Klemashevich, C.; Wu, C.; Prabakaran, D.; Pan, L. B.; Steinmeyer, S.; Mueller, C.; Yousofshahi, M.; Alaniz, R. C.; Lee, K.; Jayaraman, A., Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat Commun* 2014, 5, 5492.
47. King, Z. A.; Lu, J.; Dräger, A.; Miller, P.; Federowicz, S.; Lerman, J. A.; Ebrahim, A.; Palsson, B. O.; Lewis, N. E., BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research* 2015, 44 (D1), D515-D522.
48. NORMAN (Network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances) Association, NORMAN Suspect List Exchange. <http://www.norman-network.com/?q=node/236>.
49. Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S., CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* 2012, 84 (1), 283-9.
50. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences* 2015, 112 (41), 12580-12585.
51. Aguilar-Mogas, A.; Sales-Pardo, M.; Navarro, M.; Tautenhahn, R.; Guimerà, R.; Yanes, O., iMet: A computational tool for structural annotation of unknown metabolites from tandem mass spectra. *arXiv preprint arXiv:1607.04122* 2016.
52. Alden, N.; Krishnan, S.; Porokhin, V.; Raju, R.; McElearney, K.; Gilbert, A.; Lee, K., Biologically Consistent Annotation of Metabolomics Data. *Anal Chem* 2017, 89 (24), 13097-13104.
53. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F. H.; Khan, N.; Soltow, Q. A.; Jones, D. P.; Pulendran, B., Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* 2013, 9 (7), e1003123.
54. Morreel, K.; Saeys, Y.; Dima, O.; Lu, F.; Van de Peer, Y.; Vanholme, R.; Ralph, J.; Vanholme, B.; Boerjan, W., Systematic structural characterization of metabolites in Arabidopsis via candidate substrate-product pair networks. *The Plant Cell* 2014, 26 (3), 929-945.
55. Nam, H.; Lewis, N. E.; Lerman, J. A.; Lee, D.-H.; Chang, R. L.; Kim, D.; Palsson, B. O., Network context and selection in the evolution to enzyme specificity. *Science* 2012, 337 (6098), 1101-1104.
56. Bar-Even, A.; Tawfik, D. S., Engineering specialized metabolic pathways—is there a room for enzyme improvements? *Current opinion in biotechnology* 2013, 24 (2), 310-319.