analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 12799−12807

pubs.acs.org/ac

Integrated Probabilistic Annotation: A Bayesian-Based Annotation Method for Metabolomic Profiles Integrating Biochemical Connections, Isotope Patterns, and Adduct Relationships

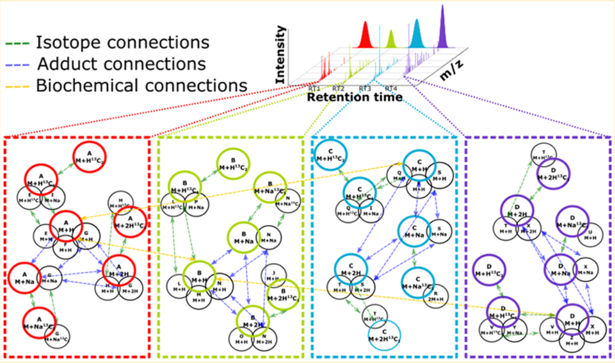
Francesco Del Carratore,[†] Kamila Schmidt,[†] Maria Vinaixa,[†] Katherine A. Hollywood,[†] Caitlin Greenland-Bews,[†] Eriko Takano,[†] Simon Rogers,[¶] and Rainer Breitling*[,†]

[†]Manchester Institute of Biotechnology, Faculty of Science and Engineering, University of Manchester, Manchester, M1 7DN, U.K.

[¶]School of Computing Science, University of Glasgow, Glasgow, G12 8RZ, U.K.

S *Supporting Information*

ABSTRACT: In a typical untargeted metabolomics experiment, the huge amount of complex data generated by mass spectrometry necessitates automated tools for the extraction of useful biological information. Each metabolite generates numerous mass spectrometry features. The association of these experimental features to the underlying metabolites still represents one of the major bottlenecks in metabolomics data processing. While certain identification (e.g., by comparison to authentic standards) is always desirable, it is usually achievable only for a limited number of compounds, and scientists often deal with a significant amount of putatively annotated metabolites. The confidence in a specific annotation is usually assessed by considering different sources of information (e.g., isotope patterns, adduct formation, chromatographic retention times, and fragmentation patterns). IPA (integrated probabilistic annotation) offers a rigorous and reproducible method to automatically annotate metabolite profiles and evaluate the resulting confidence of the putative annotations. It is able to provide a rigorous measure of our confidence in any putative annotation and is also able to update and refine our beliefs (i.e., background prior knowledge) by incorporating different sources of information in the annotation process, such as isotope patterns, adduct formation and biochemical relations. The IPA package is freely available on GitHub (https://github.com/francescodc87/IPA), together with the related extensive documentation.



When trying to convert raw mass spectrometry (MS) data into useful biological information, the association of the detected experimental features (or groups of features) to specific metabolites plays a pivotal role. This process, called annotation, is one of the major bottlenecks for untargeted metabolomics,[1] and the definition of what represents a valid "metabolite identification" is still under discussion.[2] The Metabolomics Standards Initiative (MSI, http://www.metabolomics-msi.org) defines 4 different levels of identification.[3] The highest confidence is achieved with level 1, which requires at least two orthogonal molecular properties to be confirmed (e.g., retention time and exact mass) with the aid of a pure standard analyzed in the same analytical conditions. Levels 2 and 3 (putatively annotated compounds or compound classes) are achieved with a simple comparison of molecular properties against literature and databases. Level 4 refers to unknown compounds. Confident and certain identification (level 1) requires a significant effort and is not always achievable[2] for a number of reasons: it is expensive, as in many cases, it is necessary to rebuild the database of standards when the analytical conditions change, and often there is no commercially available standard for the relevant metabolites. Therefore, in most cases only a level 2 identification is achieved, by comparing the detected *m/z*s with available databases, such as KEGG,[4,5] HMDB,[6] ECMDB,[7,8] Lipid Maps,[9] and PubChem.[10] Because of the nature of the technology, MS data show a high grade of redundancy, with several features corresponding to the same metabolite due to naturally occurring isotopes, adduct formation and in-source fragmentation.[11] Such redundancy negatively affects the annotation process by increasing the number of possible misidentifications. In the past few years, a number of different methods have been developed in order to improve the performance of the annotation process.[12] For example, Creek et al. successfully applied a retention time prediction approach. Through this approach, they removed 40% of the misannotated compounds, which showed inconsistency between *m/z* and retention time. Other methods tried to tackle the

Received: May 21, 2019
Accepted: September 11, 2019
Published: September 11, 2019

annotation by grouping together the features that appear to be related to the same compound.[14−17] Such an approach reduces the number of queries in the databases and the false annotation of adducts, fragments and isotopes. It is noteworthy that, regardless how sophisticated, the grouping algorithm is prone to errors, especially when coelution is observed. Additionally, this approach does not help in those cases where the same feature has more than one hit in the database. An interesting approach called CliqueMS was recently introduced by Senan et al.[18] This approach aims for the annotation of LC-MS features by considering the similarity between coelution profiles and a calculated natural frequency of adduct formation observed in real complex biological samples and pure compounds from the NIST database. Rogers et al. (2009) demonstrated that through a Bayesian approach it is possible to improve the annotation process by including different sources of information. Here we present IPA (integrated probabilistic annotation), a Bayesian annotation method building on this earlier work, which is able to provide a statistically rigorous assessment of our confidence in any putative annotation and is also able to update and refine our beliefs by incorporating different sources of information, such as isotope patterns, adduct formation, and biochemical relations. Compared to the original implementation of Bayesian metabolite annotation,[19,20] IPA provides substantial improvements. In fact, it is able to (a) integrate retention time information in the estimation of prior probabilities; (b) integrate the possibility that the peak under consideration is based on an "unknown" (i.e., not present in the database under consideration); (c) treat each source of information separately giving different weight to each of them; and (d) reject implausible connections for adducts and isotopes according to several criteria (e.g., mismatching retention time). The method has been validated both through a simulated data set and through two untargeted metabolomics experiments specifically designed to provide a reliable benchmark for annotation methods. Furthermore, the flexibility and the modular design of our method will allow the easy integration of additional sources of information (e.g., fragmentation patterns or ion mobility drift times).

■ MATERIALS AND METHODS

The typical data processing pipeline for untargeted LC/MS-based metabolomics usually ends up with a set of several thousands of redundant metabolic "features", generated by the presence of hundreds of metabolites in the biological sample analyzed.[21−25] Each of these features is defined by 3 quantities: its average mass-to-charge ratio (m/z), average retention time (RT), and maximum intensity. When fragmentation data is not available, the annotation of these features is usually based on the matching of all or some of the m/z values with different databases, which often leads to uncertain or incorrect assignments. When considering the association of a specific feature to one adduct of a specific compound our confidence could be simply based on how close the measured and theoretical m/z values are. However, our initial belief will change according to several observations: do we also observe the naturally occurring isotopes of the compound considered? Do the isotope signals show the expected relative intensity? Do we see the expected fragments or adducts that are usually detected in our system due to this compound? How plausible is the presence of this compound in the analyzed sample? Is this compound biochemically related to any of the other compounds found in the sample, for example as a commonly seen degradation product? The method described in this manuscript is based on the simple assumption that each annotation can be informed by all the others. Therefore, the performance of the annotation process can be significantly improved by incorporating additional information.

1. Adduct formation and in-source fragmentation: It is well-known that each compound generates several peaks due to in-source fragmentation and adduct formation. Even with the softest ionization method, electrospray ionization,[26,27] in-source fragmentation is always present and often generates fragments of identical m/z to common metabolites.[28] Ignoring this kind of information can easily cause misannotation, and it has been integrated into the IPA method as it can significantly help the annotation process.
2. Isotope patterns: Even with 1 ppm mass accuracy the information about the expected isotopes can significantly improve the annotation process.[29] Isotope peak intensity can also be informative; in fact, given the chemical formula, we are able to predict the isotopes' relative abundance[30,31] and use this information during annotation;
3. Biochemical connectivity: When analyzing a biological sample, it is safe to assume that almost all metabolites are members of the same metabolic network, that is, they are connected to other compounds in the same sample by simple (bio)chemical transformations. Considering this information during annotation helps to shift our beliefs toward compounds that are more likely to be part of the same metabolic network and away from more exotic compounds. For example, the presence of a particular compound becomes more likely if its biosynthetic precursors are also observed in the experiment.

All of these sources of evidence are routinely considered in the manual annotation of LC/MS metabolomics data, and IPA formalizes them in a unified statistical framework.

Formal IPA Description. *Database.* The IPA method relies on a structured database encoding all the information used during the annotation process. For every compound considered, the database must contain an unequivocal identifier, compound name, chemical formula, monoisotopic mass, positive and negative main adducts (i.e., the adduct known or thought to be the most intense) and positive and negative adducts and fragments (i.e., the complete list of adducts and/or fragments that are known or thought to be formed by the compound). Additionally, the database can contain a list of alternative names, a list of unequivocal identifiers for the reactions involving the compound, a list of alternative identifiers and a retention time range where the compound is considered more likely to be detected. The IPA package already includes a database containing all the compounds found in the KEGG database[4,5] with the addition of several compounds needed for the validation examples considered here (e.g., all the compounds involved in the mevalonate pathway and limonene biosynthesis[32]). This database has been initially populated assuming that all compounds show the same behavior: only a handful of adducts are generated by each compound ($[M + H]^+$, $[M + Na]^+$, $[M + 2H]^{2+}$, and $[2M + H]^+$ for positive mode and $[M - H]^-$, $[2M - H]^-$, $[M - 2H]^{2-}$, and $[3M - H]^-$ for negative mode) and the protonated ($[M + H]^+$) and deprotonated ([M

DOI: 10.1021/acs.analchem.9b02354

$-$ H]$^-$) ions are always considered as the main ions depending on the ionization mode. It should be noted that this assumption has only been used for the initial population of the database, and can be adjusted by the user as considered appropriate, given that adduct formation is different for different metabolites and for different experimental conditions. Moreover, the KEGG reaction database has been used to populate the list of reactions involving each compound. The IPA package provides the functionality for updating the database according to previous observations. For example in the course of this work, several standard mixes have been analyzed and used to update the database as described in the Supporting Information.

Prior Probabilities. Given the data set containing the measured m/z values, retention times, and intensities associated with all the M detected features, the function *find.hits()* is used to identify all the database hits given a user-defined accuracy window. With the help of the *enviPat* package,[31] this function also maps all the possible isotopes for each formula identified resulting in the collection of all the C adducts, fragments, and isotopes that could be associated with the measured mass-to-charge ratios. When trying to associate measured to theoretical mass-to-charge ratios, the conventional approach relies on selecting the theoretical mass-to-charge ratios that show an accuracy value better than an instrument-specific threshold. In the case of multiple hits and absence of additional information, it is reasonable to assume that the most likely annotation corresponds to the match with the best accuracy (closest mass match). Following this simple reasoning, the *compute.Priors()* function evaluates the prior probabilities using a Gaussian model based on the difference between measured and theoretical mass-to-charge ratio. While a Gaussian model appears to be the most reasonable choice, different noise models can be considered. For example, it would be easy to implement a model based on a uniform distribution where all assignments showing an accuracy better than a specific threshold have the same likelihood. Let $\mathbf{Z}$ be the ($C \times M$) binary matrix of assignments, where the single element $z_{c,m} = 1$ if the mass-to-charge ratio m is assigned to the formula c. The likelihood is defined as

$$p_{\text{prior}}(z_{c,m} = 1, x_m, y_c, \sigma^2) = \mathcal{N}(x_m - y_c | 0, \sigma^2) \quad (1)$$

where y_c represents the cth theoretical mass-to-charge ratio and x_m represents the mth measured mass-to-charge ratio. Describing the measurement error as a Gaussian distribution with zero mean requires the definition of the model standard deviation σ, which is related to the mass accuracy of the instrument used. A different standard deviation is computed for each mass-to-charge ratio according to the user-provided ppm value according to

$$\sigma = \frac{\text{ppm} \cdot m/z}{2 \times 10^6} \quad (2)$$

Using eq 2 to estimate the standard deviation, we are assuming that the instrument's accuracy is better than the user-defined threshold $\cong$ 95.45% of the times. It is reasonable to assume that the probability of any measured mass-to-charge ratio not being related to any of the chemical formulas contained in the database is always greater than zero, that is, there is always a finite probability that the observed mass-to-charge ratio corresponds to an unexpected or novel molecule. To take care of this, the function accepts an user-defined value for the accuracy expressed in ppm (ppm$_\text{u}$) associated with the "unknown" molecule. The estimation of the prior probabilities is then performed according to Algorithm 1. Moreover, the prior estimation has been designed in such a way that it is easy to include additional sources of information. For example, there may be evidence from previous experiments or from literature indicating that specific compounds are more or less likely to be detected in the sample analyzed. This prior knowledge can be easily added to the estimation of the prior as a multiplicative term (p_prior). Despite the relatively poor reproducibility of liquid chromatography retention times between different laboratories or experimental runs, IPA also offers the possibility of considering information regarding RT. In fact, for each compound contained in the database, it is possible to provide an RT range (as broad as reasonable) outside of which the presence of the compound is considered to be unlikely in light of previous experimental evidence or an RT prediction method.[13,33] The multiplicative term related to the retention time (p_RT) will be equal to a user-defined value ($\leq$1), when the RT is found outside this range, and will be equal to 1 otherwise. Additional multiplicative terms (e.g., regarding fragmentation patterns) could be easily implemented.

Algorithm 1 Prior Discrete Distribution

```
procedure PRIOR DISTRIBUTION
    $P \leftarrow matrix(0, M, C+1)$
    for i in 1:M do
        $di[1:C] \leftarrow exp[(-0.5 \cdot \frac{1}{\sigma^2}) \cdot (matrix(x[m], C, 1) - y)^2]$
        $di[C+1] \leftarrow exp[(-0.5 \cdot \frac{1}{\sigma^2}) \cdot (ppm_u \cdot x[m] \cdot 10^{-6})^2]$
        $di \leftarrow di \cdot p_{prior} \cdot p_{RT} \cdot \ldots$
        $P[i, 1:C] \leftarrow \frac{di}{sum(di)}$
    end for
    return $P$
end procedure
```

Creating Connectivity Matrices. As mentioned in the introduction, the IPA package improves the quality of the annotation process by considering the relationships between adducts, fragments, isotopes, and biochemically related compounds. Such information is encoded in three connectivity matrices generated by specific functions.

1. Adducts matrix ($\mathbf{W}_\text{add}$): The *build.add.connenctivity.matrix()* creates a ($C \times C$) binary matrix containing the relationships between adducts and fragments. By default, this matrix connects each monoisotopic adduct to its related main ion, typically the protonated or deprotonated version of the molecule. Alternatively, the user can select the *fully connected* parameter, and $\mathbf{W}_\text{add}$ will connect all adducts related to the same compound with each other.
2. Isotope matrix ($\mathbf{W}_\text{iso}$): The *build.iso.connectivity.matrix()* function creates a ($C \times C$) matrix connecting each m/z to its two (or one) isotopologues showing the highest predicted abundance. If the ith and jth m/z values are connected by an isotope relationship, the $w_{i,j}$ element contains the expected intensity ratio between the two. Thanks to this information, the package can filter out isotope connections showing inconsistent intensity ratios.
3. Biotransformation matrix ($\mathbf{W}_\text{bio}$): the *build.bio.connectivity.matrix()* creates a ($C \times C$) binary matrix connecting only the main adducts of those compounds thought to be involved in the same reaction. The function can achieve this by considering the list of reaction IDs provided in the database for each

DOI: 10.1021/acs.analchem.9b02354
Anal. Chem. 2019, 91, 12799−12807

compound or by considering a limited number of biochemical reaction classes that usually occur in a metabolic network.[34]

Posterior Probabilities. When annotating by simply querying a database, finding more than one possible hit is common. In such cases, the identification of the most reasonable assignment can be aided by the assignments of the other m/z values detected. For example, the presence of isotopes consistent with the predicted abundance ratios helps to reduce the number of plausible compounds.[35] This can be easily achieved when the assignments of the isotopes are considered as certain and independent. In reality, all assignments are potentially interdependent, which makes the incorporation of the information about possible connections between m/z values (i.e., computing the posterior probabilities) very challenging, as all the possible assignments have to be considered at once. It is possible, however, to define a prior distribution for one mass conditioned on the current assignments of all the other masses and this can be used to build an efficient Gibbs sampling scheme[19] which allows us to draw samples from the posterior distributions. Such conditional priors can be defined according to different sources of information (i.e., different networks of possible interactions). For example, the conditional prior probability of the mth m/z being assigned to the cth formula, given the network of all the possible biochemical connections defined in the $\mathbf{W}_{\text{bio}}$ binary matrix is defined as

$$p_{\text{bio}}(z_{c,m} = 1, \mathbf{Z}, \delta_{\text{B}}) = \frac{\beta_{cm} + \delta_{\text{B}}}{C \cdot \delta_{\text{B}} + \sum_{c'} \beta_{c'm}} \tag{3}$$

where C is the number of chemical formulas in the database, while the parameter δ_{B} can be thought of as the parameter for a Dirichlet prior on a multinomial distribution over β_{cm} and expresses our confidence on the information encoded by the connectivity matrix considered ($\mathbf{W}_{\text{bio}}$). In fact, the smaller is δ_{B}, the higher is the effect of the connections on the conditional prior and consequently on the posterior probabilities. β_{cm} represents the number of relationships the cth compound shows with all the other formulas that are currently assigned to all the other m/z values. It can be computed as

$$\beta_{cm} = \mathbf{W}_{c}\boldsymbol{\cdot} \cdot \mathbf{Z} \cdot \mathbf{1} - \mathbf{W}_{c}\boldsymbol{\cdot} \cdot \mathbf{Z}\boldsymbol{\cdot}_{m} \tag{4}$$

where $\mathbf{1}$ is the $(M \times 1)$ vector of ones and $\mathbf{W}_{c}\boldsymbol{\cdot}$ and $\mathbf{Z}\boldsymbol{\cdot}_{m}$ represent respectively the cth row of the connectivity matrix $\mathbf{W}$ and the mth column of the assignments matrix $\mathbf{Z}$. Similarly to eq 3, it is also possible to define the conditional prior related to the adducts and isotope relationships

$$p_{\text{add}}(z_{c,m} = 1, \mathbf{Z}, \delta_{\text{A}}) = \frac{\eta_{cm} + \delta_{\text{A}}}{C \cdot \delta_{\text{A}} + \sum_{c'} \eta_{c'm}} \tag{5}$$

$$p_{\text{iso}}(z_{c,m} = 1, \mathbf{Z}, \delta_{\text{I}}) = \frac{\omega_{cm} + \delta_{\text{I}}}{C \cdot \delta_{\text{I}} + \sum_{c'} \omega_{c'm}} \tag{6}$$

The computation of the number of connections for adducts and isotopes is slightly more complicated. In fact, both adduct and isotope relationships should be considered only between m/z values that might be generated from the same molecule. If two detected m/z values do not appear to be related to the same compound (e.g., they show a very different RT), not all possible adduct or isotope connections should be counted when computing the conditional prior. The IPA package can handle this problem in three different ways, according to the information available.

1. If the RT for each measured m/z is provided, the only connections considered are the ones between mass-to-charge ratios showing an RT difference lower than a user-defined threshold.
2. It is possible to group the detected mass-to-charge ratios that are more likely to be related to the same compound with a correlation based approach.[14] IPA is able to consider this grouping through an $(M \times 1)$ group label vector.
3. Alternatively, it is possible to provide an $(M \times M)$ correlation matrix. If two m/z values show a correlation lower than a user-defined threshold, all possible adduct of isotope connections between them are not considered.

Regarding the isotopes, it is also possible to consider information about peak intensity to filter out connections. As mentioned before, it is possible to estimate the expected intensity ratio between isotopes, and this information is stored in the $\mathbf{W}_{iso}$ matrix. If the measured intensities are provided, IPA is able to filter out isotope connections between m/z values if the observed intensity ratio is different from the theoretical one (given a user-defined tolerance). The Gibbs sampler has been implemented by considering the following distribution:

$$p(z_{c,m} = 1|\mathbf{Z}, x_m, \mathbf{y}, \delta_{\text{B}}, \delta_{\text{A}}, \delta_{I}, \sigma^2) \propto p_{prior} \cdot p_{bio} \cdot p_{add} \cdot p_{iso} \tag{7}$$

Normalizing over C, it is finally possible to obtain a proper discrete distribution. The functionality described in this section is available through the function *IPAposteriors()*, where the Gibbs sampler is implemented as described in Algorithm 2. Here, the user has to define both the number of iterations (N) and the number of initial iterations (burn_{in}) that will be ignored when computing the posterior probabilities.

Algorithm 2 Gibbs Sampler

```
procedure GIBBS SAMPLER
    initialize assignments vector $\mathbf{z}$ sampling from prior distributions
    for $s$ in $1 : N$ do
        order ← sample($1 : M$)
        for $m$ in order do
            $\mathbf{I} \leftarrow [\omega_{1,m}, \cdots, \omega_{C,m}]$
            $\mathbf{P} \leftarrow [\beta_{1,m}, \cdots, \beta_{C,m}]$
            $\mathbf{A} \leftarrow [\eta_{1,m}, \cdots, \eta_{C,m}]$
            $p_{iso} \leftarrow \frac{\mathbf{I}+\delta_{iso}}{sum(\mathbf{I}+\delta_{iso})}$
            $p_{bio} \leftarrow \frac{\mathbf{P}+\delta_{bio}}{sum(\mathbf{P}+\delta_{bio})}$
            $p_{add} \leftarrow \frac{\mathbf{A}+\delta_{add}}{sum(\mathbf{A}+\delta_{add})}$
            $p_0 \leftarrow p_{prior} \cdot p_{iso} \cdot p_{bio} \cdot p_{add}$
            $p_0 \leftarrow \frac{p_0}{sum(p_0)}$
            $z[m] \leftarrow sample(p_0)$
        end for
        $\mathbf{z}_s \leftarrow [\mathbf{z}_s; \mathbf{z}]$
    end for
    compute posteriors from $\mathbf{z}_s$ considering the last $N - burn\_in$ iterations
    return posteriors
end procedure
```

Spectral Acquisition and Data Availability. Two untargeted metabolomics experiments are considered in this study. The biological samples were analyzed on a QExactive Plus equipped with an Ultimate 3000 UHPLC (Thermo-Fisher, UK). The complete description of the procedure used for both experiments can be found in Supporting Information.

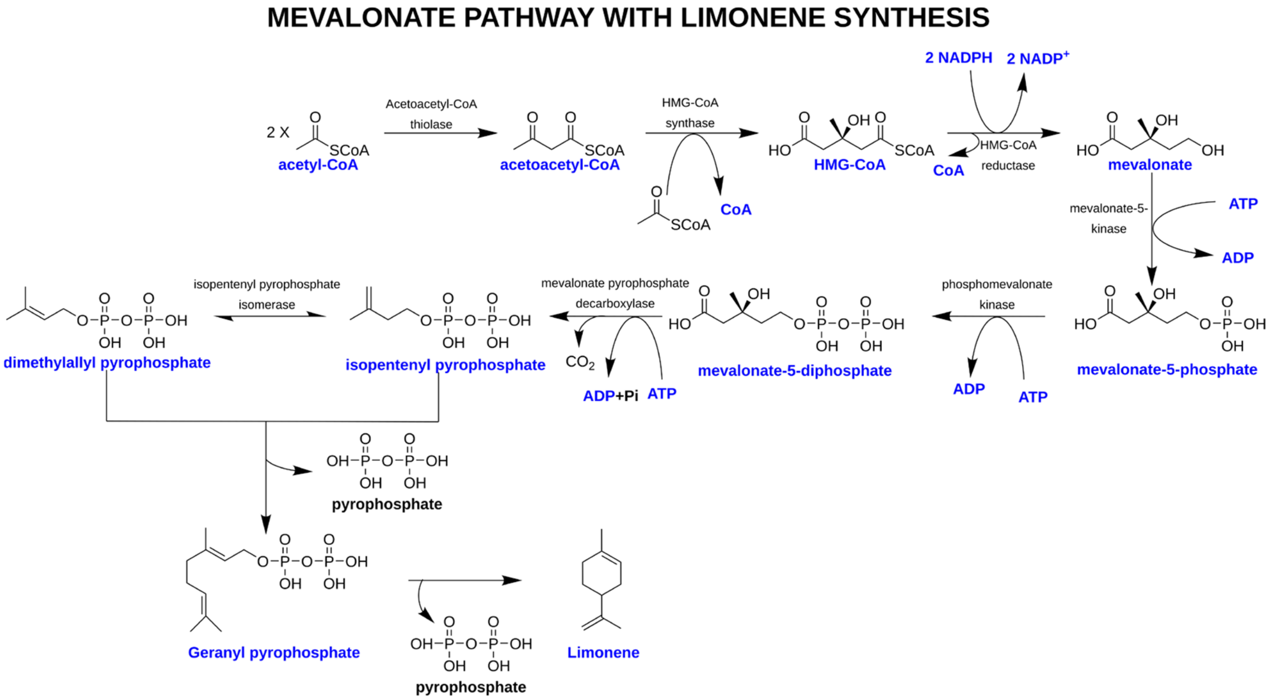


Figure 1. Mevalonate pathway with limonene synthesis. The compounds highlighted in blue have been included in the synthetic experiment.

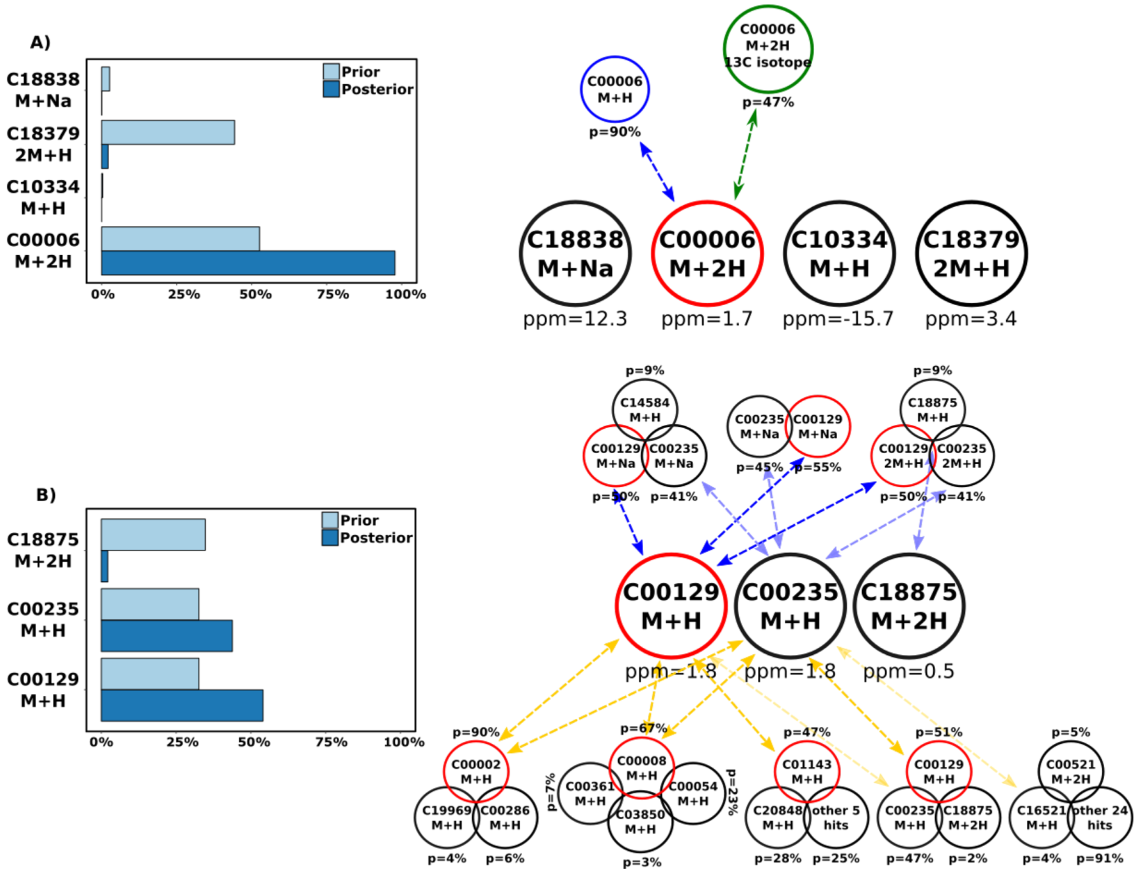


Figure 2. Estimation of the posterior distributions for the simulated feature for the $[M + 2H]^{2+}$ adduct of $NADP^+$ (A) and for the $[M + H]^+$ adducts of isopentenyl pyrophosphate (B). The annotation of the first feature is informed by an adduct connection (blue) and one isotope connection (green). The annotation of the second feature is instead informed by several adduct connections (blue) and biochemical connections (yellow). The circles clustered together represent the possible annotation associated with the same feature. The probabilities reported in the graph are the posterior probabilities estimated by the IPA method when considering all the sources of information.

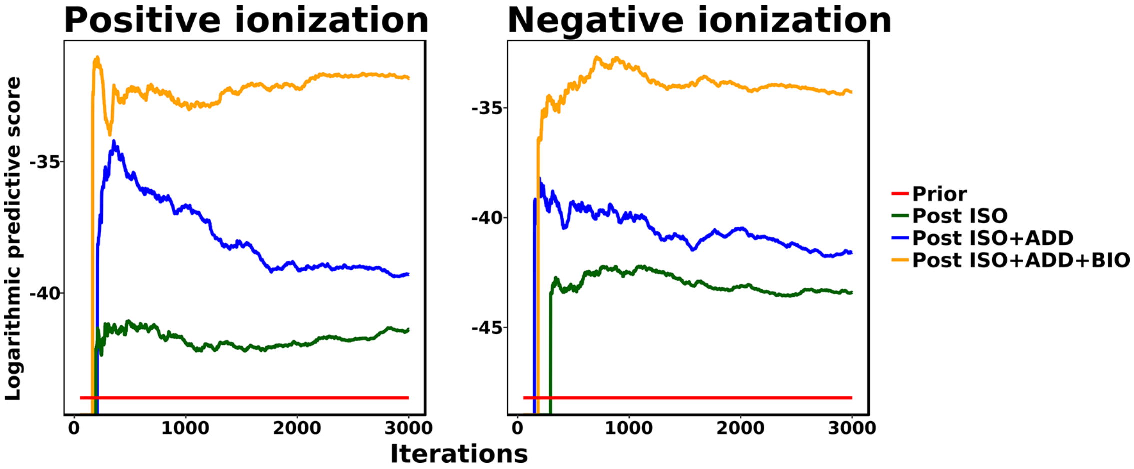


Figure 3. Logarithmic predictive score computed for the synthetic experiment. The number of *burn_in* iterations is always the 20% of the total.

■ RESULTS AND DISCUSSION

Synthetic Example. To evaluate the performance of our approach, a simulated metabolomics experiment has been considered.

This synthetic experiment contains 15 compounds involved in the mevalonate pathway and limonene synthesis[32] (specifically, the compounds highlighted in blue in Figure 1). Several adducts were simulated for each of the considered metabolites ($[M + H]^+$, $[M + Na]^+$, $[M + 2H]^{2+}$, and $[2M + H]^+$ for positive mode and $[M - H]^-$, $[2M - H]^-$, $[M - 2H]^{2-}$, and $[3M - H]^-$ for negative mode). Additionally, all the possible isotopes with a predicted relative abundance higher than 5% were included in the data set. Including all adducts and isotopes, a total of 83 m/z values were simulated for the positive mode, and 95 were simulated for the negative mode. A realistic experimental outcome was simulated by adding Gaussian noise to the m/z values ($\mathcal{N}(0, \sigma^2)$), where appropriate σ values were computed for each m/z assuming an instrument accuracy of 10 ppm (see eq 2). For each detected m/z, the measured intensities were chosen to be coherent with the theoretical abundance of the isotopes, and the same RT value (±2 s) was assigned to all the simulated m/z values related to the same compound. With the goal of showing that every source of information can positively affect the annotation process, the IPA method was applied in three settings: (1) only considering isotope relationships, (2) considering isotopes and adducts relationships, and (3) considering isotopes and adducts relationships and biochemical connections. The scripts used for the generation of the data sets and for the analysis with the IPA package are provided as Supporting Information. Even considering this simple toy example, several simulated features do not have a clear annotation. For example, the feature simulated for the $[M + 2H]^{2+}$ adduct of NADP$^+$ (C00006) shows several hits in the database when the annotation process is based on mass only. As shown in Figure 2A, the IPA method changes our initial belief toward the correct annotation by considering that it is the only one showing any kind of relationship with the other possible annotations.

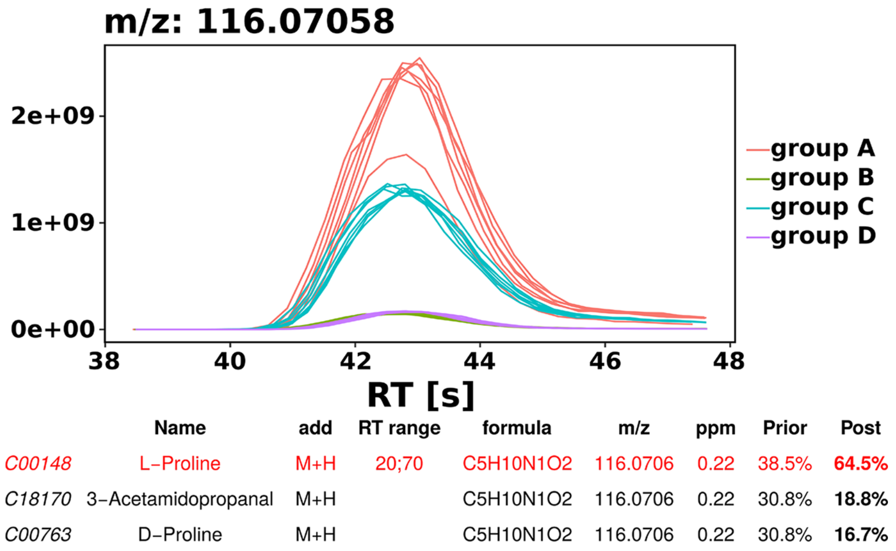
The example reported in Figure 2B is slightly more complicated. When considering the measured m/z values as the sole source of information, the feature considered (simulated for the $[M + H]^+$ adduct of isopentenyl pyrophosphate (C00129)) has three possible annotations in our database. Two of them (C00129 $[M + H]^+$, isopentenyl pyrophosphate and C00235 $[M + H]^+$, dimethylallyl pyrophosphate) are equally likely as they have the same chemical formula. It should be noted that both these compounds are actually present in the simulated data set. The third possible annotation (C18875 $[M + 2H]^{2+}$, novaluron) is initially the most likely having a m/z closer to the measured one. By considering all the possible connections with the other features present in the data set, this latter annotation becomes extremely unlikely despite being the one with the smallest difference between theoretical and measured m/z. Having one biochemical connection more, the C00129 $[M + H]^+$ annotation becomes slightly more likely than C00235 $[M + H]^+$. One should notice that all these assignment are interconnected, therefore this little advantage also increases the posterior probabilities associated with the other adducts associated with C00129. A couple of cherry picked examples cannot be used as an assessment of the annotation performance of the method.

For a comprehensive assessment, the logarithmic predictive score (LPS) was computed as

$$\mathrm{LPS} = \sum_{i=1}^{M} \log(p_i) \tag{8}$$

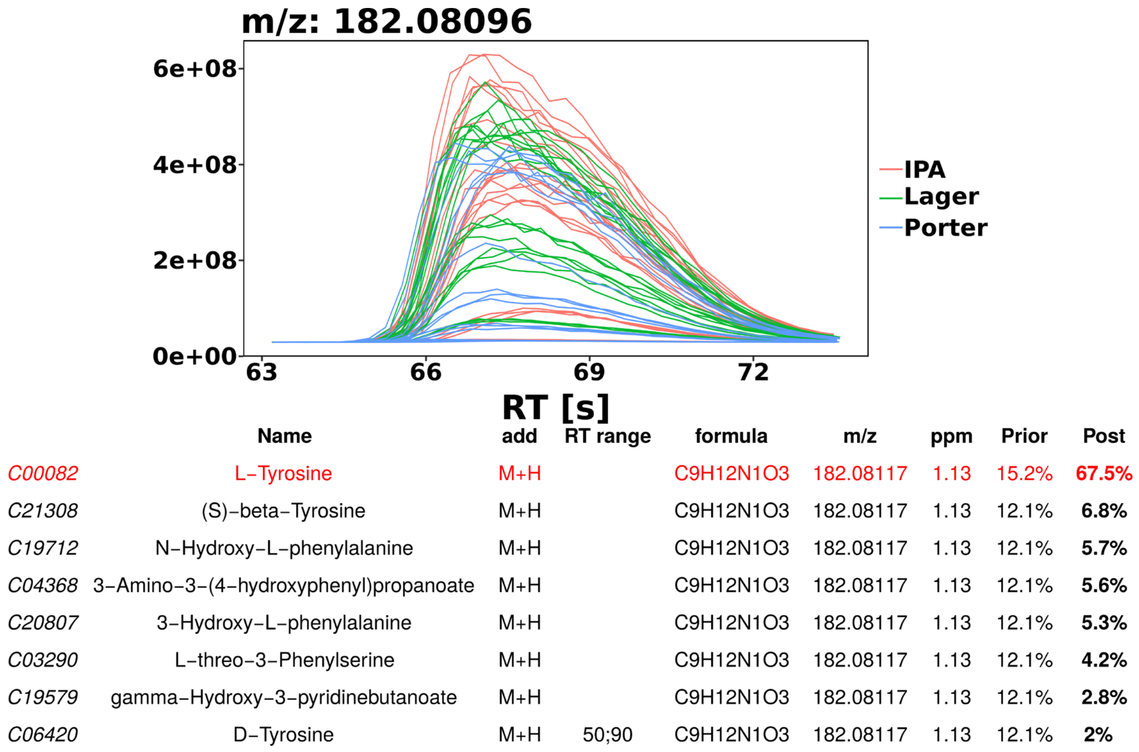
where p_i is the probability given to the correct assignment for the ith measured m/z. In the best case scenario, all the probabilities considered by eq 8 are equal to 1, that is, all features are correctly identified, therefore LPS = 0. In all other situations, LPS assumes a negative value. As shown in Figure 3, the integration of each of the additional sources of information drastically improves the predictive power of the annotation method for both simulated data sets.

***Escherichia coli* Data Set.** To show the value of our method in real-life conditions, we designed two different untargeted metabolomics experiments. In the first experiment, a cell extract of a culture of *Escherichia coli* was chosen as a biological matrix and divided into 4 groups. With the aim of providing a data set that could also be used easily as a benchmark for biomarker discovery methods, three of the groups were spiked with different amounts of 30 standards as described in Supporting Information. The data processing performed through xcms[36] and the mzMatch pipeline[21] led to a data set containing 1961 features. The IPA method was applied to the obtained data set with default parameters and



	Name	add	RT range	formula	m/z	ppm	Prior	Post
C00148	L−Proline	M+H	20;70	C5H10N1O2	116.0706	0.22	38.5%	**64.5%**
C18170	3−Acetamidopropanal	M+H		C5H10N1O2	116.0706	0.22	30.8%	**18.8%**
C00763	D−Proline	M+H		C5H10N1O2	116.0706	0.22	30.8%	**16.7%**

Figure 4. One of the features detected in this experiment has 3 possible annotations. According to the IPA method (and common sense), L-proline is the most likely annotation.



	Name	add	RT range	formula	m/z	ppm	Prior	Post
C00082	L−Tyrosine	M+H		C9H12N1O3	182.08117	1.13	15.2%	**67.5%**
C21308	(S)−beta−Tyrosine	M+H		C9H12N1O3	182.08117	1.13	12.1%	**6.8%**
C19712	N−Hydroxy−L−phenylalanine	M+H		C9H12N1O3	182.08117	1.13	12.1%	**5.7%**
C04368	3−Amino−3−(4−hydroxyphenyl)propanoate	M+H		C9H12N1O3	182.08117	1.13	12.1%	**5.6%**
C20807	3−Hydroxy−L−phenylalanine	M+H		C9H12N1O3	182.08117	1.13	12.1%	**5.3%**
C03290	L−threo−3−Phenylserine	M+H		C9H12N1O3	182.08117	1.13	12.1%	**4.2%**
C19579	gamma−Hydroxy−3−pyridinebutanoate	M+H		C9H12N1O3	182.08117	1.13	12.1%	**2.8%**
C06420	D−Tyrosine	M+H	50;90	C9H12N1O3	182.08117	1.13	12.1%	**2%**

Figure 5. One of the features detected in this experiment have 8 possible annotations. According to the IPA method, L-tyrosine is the most likely annotation.

using the updated database described in Supporting Information. Additionally, during the estimation of the prior probabilities, the multiplicative term p_{prior} was used to take into account our initial belief on the presence of the considered compounds in the biological samples here considered. Specifically, the p_{prior} is equal to 1 when the compound is also present in the *E. coli* Metabolome database,[7,8] and equal to 0.8 otherwise. After removing the features for which no hit was found considering a maximum accuracy of 15 ppm, the 1093 remaining features went through 5000 iterations of the Gibbs sampler (1000 *burn_in*). The Gibbs sampler represents the most computationally demanding step of the method described here, and with this data set, took ∼4 h (an average of 2.89 s per iteration) on a Linux desktop with 32 Gb memory and 8-core Intel Xeon E5-2620 2.1 GHz processor. As shown in Figure 4, one of the features detected in this experiment could be associate to the $[M + H]^+$ adducts of 3 different compounds: (1) L-proline (C00148), (2) D-proline (C00763), and (3) 3-acetamidopropanal (C00763). Having the same chemical formula, these three annotations should show the same initial probability when considering the mass as only source of information. The estimated prior probability estimated by the IPA package is slightly higher for L-proline (≃38.5%) since it is the only one, among the three possibilities, that is also present in the ECMDB database. Moreover, the estimation of the posterior probabilities further shifts our belief toward this annotation (posterior probability associated with L-proline ≃ 64.5%). As mentioned before, 30 different compounds were spiked in these samples. In positive mode, only 24 were detected and for 22 of them the IPA method increased the probability associated with the correct annotation of their main adduct (see Table S6). In negative mode, 20 of these compounds were detected and the IPA method shifted the probabilities toward the correct annotation for 18 of their main adducts (see Table S7). When dealing with real-life data, a comprehensive assessment of the method performance is not easy. In contrast to what is possible with the synthetic experiment, the correct annotation in real-life data will only be known for a very small fraction of the detected features. This makes the calculation of a meaningful LPS value impossible. Nevertheless, it is possible to quantify the impact of IPA on the annotation. Out of the 929 features showing more than one possible hit in the database, 268 (≃29%) showed a maximum difference between prior and posterior probabilities higher than 10%. More importantly, for 186 of these features (≃20%), the IPA method changed the most likely annotation or was a tie-breaker in the case of equal prior probabilities. The full results of this analysis together with the data set are available as Supporting Information.

Beer Data Set. In the second experiment, IPA was tested in another untargeted metabolomics experiment where 21 different beers (7 indian pale ales, 7 lagers, and 7 porters) were analyzed as previously described (see Supporting Information for details). After the data processing, a data set containing 3042 features was obtained. The IPA method was applied to this data set using the same parameters and the same database used in the previous example. During the estimation of the prior probabilities, the multiplicative term p_{prior} was again used to take into account our initial belief on the presence of the considered compounds in the samples here considered. Specifically, the p_{prior} is equal to 1, when the compound considered has been previously detected in published metabolomics studies involving beers,[37−42] and equal to 0.8 otherwise. After removing the features for which no hit was found considering a maximum accuracy of 15 ppm, the 2139 remaining features went through 5000 iterations of the Gibbs samples (1000 *burn_in*). The Gibbs sampler represents the most computationally demanding step of the method described here, and with this data set set took ∼9 h (an average of 6.4 s per iteration) on a Linux desktop with 32 Gb memory and 8-core Intel Xeon E5-2620 2.1 GHz processor. Also in this case, the IPA method is able to provide a probabilistic assessment of our confidence in each annotation. For example, Figure 5 shows one of the detected features, which m/z could be associated with 8 different compound all having the same chemical formula. The prior probability associated with L-tyrosine by IPA is slightly higher than the other possible annotations (≃15.2%) because it has previously been detected in a similar experiment.[38] After the estimation of the posterior probabilities, IPA makes this annotation extremely more likely (≃67.5%). Also in this case, it is possible to quantify the impact of the IPA method. Out of the 1846 features showing more than one possible hit in the database, 683 (≃37%) showed a maximum difference between prior and posterior probabilities higher than 10%. More importantly, for 618 of these features (≃33%) the IPA method changed the most likely annotation or was a tie-breaker in the case of equal prior probabilities. The full results of this analysis together with the data set are available as Supporting Information. The examples shown in Figures 4 and 5 might seem obvious to an expert in metabolomics, but they are not to the traditional automated annotation methods. They have been chosen to highlight how the IPA approach is able to replicate the reasoning of an expert. Two additional examples in the Supporting Information (Figures S1 and S2) further highlight the value of isotope and adduct connections in the automated annotation. To the best of our knowledge, no other method provides the same kind of functionalities introduced by IPA (i.e., a rigorous statistical integration of evidence, with real p-values rather than arbitrary scores). For this reason, a comparison with other methods is not straightforward. To validate the annotation precision, IPA has been compared with the xMSannotator package.[17] The results are reported in the Supporting Information.

■ CONCLUSION

The IPA method here presented implements a Bayesian-based approach able to incorporate several sources of information within the annotation process. This leads to a significant increase of the predictive power for assigning measured m/z values to putative formulas. Not only does IPA provide more reliable annotations of compounds, it is also able to quantify our confidence in such annotations and re-evaluate them when new information is provided. The IPA package provides a rigorous and comprehensive probabilistic assessment of the confidence in each annotation, which will be highly valuable for the downstream interpretation of the results. Moreover, IPA is also able to store and successfully utilize the additional information gained from previous experiments, thus leading to an iterative improvement of annotations, especially for data sets collected on the same experimental setup and similar biological samples. This "continuous learning" ability of the IPA approach mimics one of the most important features of human manual annotations and ensures that insights from earlier data sets are maintained for future exploitation.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b02354.

Description of standards analysis and database update (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rainer.breitling@manchester.ac.uk.

ORCID

Francesco Del Carratore: 0000-0003-1647-7818
Kamila Schmidt: 0000-0002-7763-1500
Maria Vinaixa: 0000-0001-9804-0171
Katherine A. Hollywood: 0000-0002-7028-047X
Eriko Takano: 0000-0002-6791-3256
Rainer Breitling: 0000-0001-7173-0922

Notes

The authors declare no competing financial interest.

Data and code for standards analysis and database update, data and code used for the generation of the synthetic experiment, data, code, and results of the *E. coli* data set analysis, data, code and results of the beer data set analysis, and data, code, and results of the comparison with xMSannotator can be found at http://doi.org/10.5281/zenodo.3414903.

■ ACKNOWLEDGMENTS

This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM) and acknowledges the Biotechnology and Biological Sciences Research Council (BBSRC) and Engineering and Physical Sciences Research Council (EPSRC) for financial support (Grant BB/M017702/1). This project has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant Agreements 720793 TOPCAPI−Thoroughly Optimised Production Chassis for Advanced Pharmaceutical Ingredients and 814408 SHIKIFACTORY100−Modular cell factories for the production of 100 compounds from the shikimate pathway.

■ REFERENCES

(1) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M. *Chemom. Intell. Lab. Syst.* **2011**, *108*, 23−32.

(2) Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.; Neumann, S.; Schymanski, E. L.; Sumner, L. W.; Trengove, R.; Wolfender, J.-L. *Metabolomics* **2014**, *10*, 350.

(3) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; et al. *Metabolomics* **2007**, *3*, 211−221.

(4) Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. *Nucleic Acids Res.* **2019**, *47*, D590−D595.

(5) Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. *Nucleic Acids Res.* **2017**, *45*, D353−D361.

(6) Wishart, D. S.; et al. *Nucleic Acids Res.* **2007**, *35*, D521−D526.

(7) Sajed, T.; Marcu, A.; Ramirez, M.; Pon, A.; Guo, A. C.; Knox, C.; Wilson, M.; Grant, J. R.; Djoumbou, Y.; Wishart, D. S. *Nucleic Acids Res.* **2016**, *44*, D495−D501.

(8) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. *Nucleic Acids Res.* **2012**, *41*, D625−D630.

(9) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H., Jr.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, D527−D532.

(10) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. *Nucleic Acids Res.* **2009**, *37*, W623−W633.

(11) Mahieu, N. G.; Patti, G. J. *Anal. Chem.* **2017**, *89*, 10397−10406.

(12) Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2018**, *90*, 480−489.

(13) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. *Anal. Chem.* **2011**, *83*, 8703−8710.

(14) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.

(15) Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K. E.; Breitling, R. *Bioinformatics* **2014**, *30*, 2764−2771.

(16) Broeckling, C. D.; Afsar, F.; Neumann, S.; Ben-Hur, A.; Prenni, J. *Anal. Chem.* **2014**, *86*, 6812−6817.

(17) Uppal, K.; Walker, D. I.; Jones, D. P. *Anal. Chem.* **2017**, *89*, 1063−1067.

(18) Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimerà, R.; Sales-Pardo, M. *Bioinformatics* **2019**, DOI: 10.1093/bioinformatics/btz207.

(19) Rogers, S.; Scheltema, R. A.; Girolami, M.; Breitling, R. *Bioinformatics* **2009**, *25*, 512−518.

(20) Silva, R. R.; Jourdan, F.; Salvanha, D. M.; Letisse, F.; Jamin, E. L.; Guidetti-Gonzalez, S.; Labate, C. A.; Vêncio, R. Z. *Bioinformatics* **2014**, *30*, 1336−1337.

(21) Scheltema, R. A.; Jankevics, A.; Jansen, R. C.; Swertz, M. A.; Breitling, R. *Anal. Chem.* **2011**, *83*, 2786−2793.

(22) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035−5039.

(23) Giacomoni, F.; Le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; Goulitquer, S.; Thevenot, E. A.; Caron, C. *Bioinformatics* **2015**, *31*, 1493−1495.

(24) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinf.* **2010**, *11*, 395.

(25) Franceschi, P.; Mylonas, R.; Shahaf, N.; Scholz, M.; Arapitsas, P.; Masuero, D.; Weingart, G.; Carlin, S.; Vrhovsek, U.; Mattivi, F.; Wehrens, R. *Front. Bioeng. Biotechnol.* **2014**, DOI: 10.3389/fbioe.2014.00072.

(26) Fenn, J.; Mann, M; Meng, C.; Wong, S.; Whitehouse, C. *Science* **1989**, *246*, 64−71.

(27) Banerjee, S.; Mazumdar, S. *Int. J. Anal. Chem.* **2012**, 282574.

(28) Xu, Y.-F.; Lu, W.; Rabinowitz, J. D. *Anal. Chem.* **2015**, *87*, 2273−2281.

(29) Kind, T.; Fiehn, O. *BMC Bioinf.* **2007**, *8*, 105.

(30) Valkenborg, D.; Mertens, I.; Lemiere, F.; Witters, E.; Burzykowski, T. *Mass Spectrom. Rev.* **2012**, *31*, 96−109.

(31) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. *Anal. Chem.* **2015**, *87*, 5738−5744.

(32) Alonso-Gutierrez, J.; Chan, R.; Batth, T. S.; Adams, P. D.; Keasling, J. D.; Petzold, C. J.; Lee, T. S. *Metab. Eng.* **2013**, *19*, 33−41.

(33) Bach, E.; Szedmak, S.; Brouard, C.; Böcker, S.; Rousu, J. *Bioinformatics* **2018**, *34*, i875−i883.

(34) Breitling, R.; Pitt, A. R.; Barrett, M. P. *Trends Biotechnol.* **2006**, *24*, 543−548.

(35) Kind, T.; Fiehn, O. *BMC Bioinf.* **2006**, *7*, 234.

(36) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(37) Heuberger, A. L.; Broeckling, C. D.; Lewis, M. R.; Salazar, L.; Bouckaert, P.; Prenni, J. E. *Food Chem.* **2012**, *135*, 1284−1289.

(38) Marova, I.; Parilova, K.; Friedl, Z.; Obruca, S.; Duronova, K. *Chromatographia* **2011**, *73*, 83.

(39) Andrés-Iglesias, C.; Blanco, C. A.; Blanco, J.; Montero, O. *Food Chem.* **2014**, *157*, 205−212.

(40) Intelmann, D.; Haseleu, G.; Hofmann, T. *J. Agric. Food Chem.* **2009**, *57*, 1172−1182.

(41) Schmidt, C.; Biendl, M. *Monatsschrift für Brauwissenschaft* **2017**, *70*, 198.

(42) Pai, T. V.; Sawant, S. Y.; Ghatak, A. A.; Chaturvedi, P. A.; Gupte, A. M.; Desai, N. S. *J. Food Sci. Technol.* **2015**, *52*, 1414−1423.