

Analyst

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: C. Broeckling, *Analyst*, 2026, DOI: 10.1039/D5AN00914F.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

**Creating taxonomically-informed metabolome libraries for any species using the
pubchem.bio R package.**

Corey D. Broeckling, PhD
Colorado State University, Analytical Resources Core – Bioanalysis and Omics Center,
Fort Collins, CO, USA, 80523.
cbroeckl@colostate.edu
970-491-2273

Analyst Accepted Manuscript

Open Access Article. Published on 15 July 2025. Downloaded on 12/18/2025 12:38:41 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.



Abstract:

Annotation remains a significant challenge in metabolomics, in large part due to the enormous structural diversity of small molecules. PubChem represents one of the largest curated chemical structure databases, with more than 122,000,000 structures, supplemented by extensive biological metadata provided by numerous external sources. While many of these structures are relevant to metabolomics, a majority are unlikely to be measured in a typical metabolomics experiment. This article describes the R package, pubchem.bio, which enables users to: (1) Download metabolomics-centric subset of PubChem onto their local computer, (2) Build a metabolomic structured library of biological compounds in PubChem, (3) Develop custom metabolite structure libraries for any species or collection of species using selected or all available taxonomic data in PubChem and (4) Define a core biological metabolome, comprising metabolites plausibly found in any species. Species-specific metabolomes are enabled through the adoption of a lowest-common-ancestor chemotaxonomy approach, which is implemented by associating PubChem CIDs into the NCBI Taxonomy database hierarchy, enabling extrapolation of taxonomic range beyond the species reported. This package is available via CRAN, and can be used to simplify the annotation process and embed biological metadata into the annotation process.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Introduction:

A critical step in any mass spectrometry-based metabolomics workflow is ‘annotation’ - the process of assigning chemical structures to chromatographically coupled mass spectrometry signals. The annotation task has been considered a one of the largest challenges in metabolomics since the field’s inception¹, and remains problematic today^{2–4}.

There have been innumerable approaches designed to improve annotation accuracy. One of the major challenges in metabolomics annotation is appropriately defining the chemical search space for any given untargeted metabolomics experiment. The full theoretical chemical search space has been estimated to be on the order of 10⁶⁰ potential structures at a molecular weight of 1000 or less⁵. PubChem is one of the largest publicly available repositories for small molecule structures, with more than 122,000,000. This size makes it one the most comprehensive structure databases available, but there are many structures in PubChem which are unlikely to ever be observed in most metabolomics experiments. PubChem Lite⁶ was designed to extract the compounds that are most likely to be observed, with a focus toward support for NTA or exposomics experiments.

Numerous biological databases have been developed for metabolomics research as a way of restricting chemical structure search space, each with different focus. If a study is focused on clinical samples, the Human Metabolome Database⁷ is an invaluable resource. For natural products, Coconut⁸, Lotus⁹, NPASS¹⁰, and others are available. These databases are extremely valuable in cataloging the existing known chemical space, particularly for specialized metabolites. However, there are relatively few species for which a comprehensive database exists, due to the effort necessary to compile databases from the literature and the incredible diversity of life. Some of the natural product databases explicitly link chemical structures to taxonomy, enabling taxonomy filtering, but these databases are biased toward natural products and therefore not inclusive of more highly conserved metabolic pathways.

Taxonomy has been shown to be a useful piece of metadata to include in annotation approaches^{11,12}. An ideal metabolomics library for a given biological sample would include all the potential small molecules which would plausibly be found in a given sample, with as few extraneous compounds as possible. This is feasible through manual effort, but remains a cumbersome task¹³. PubChem¹⁴ has become not only a vast repository of chemical structure data, but also a vast repository of associated biological metadata, incorporating metadata from metabolomics, metabolic pathway, and natural product databases. The pubchem.bio R packaged described in this manuscript is an informatic resource which streamlines the building of biological and taxonomically informed metabolite libraries from PubChem in support of metabolomics,

Analyst Accepted Manuscript

and other applications which can benefit from a comprehensive list of plausible small molecules found in a given species.

Approach:

In this article, the term 'metabolite' is used to be inclusive of all small molecule structures which are linked to biology, including traditional metabolites, but also lipids, polysaccharides, and even exogenous compounds such as pesticides, which are frequently found in biological metabolome libraries and pathways.

PubChem is an accessible and freely available data source, which incorporates a wealth of metadata. The pubchem.bio functions were executed Aug 22, 2025, to generate summary statistics presented here. There are several biological classes of metadata built into PubChem, including:

1. Data source: PubChem can be subset based on which organizations deposited structures. 885 different data sources have contributed. Suitable biological data sources include HMDB, ChEBI, Metabolomics Workbench, Lipid Maps, and others. Any data source can be selected, but the default values are those considered to clearly fit the category of 'biological.'
2. Pathways: 10 organizations have deposited pathway data. The presence of a chemical in a pathway means that there is biological transformation which either produces or consumes it. These chemicals will be either native biological metabolites, or metabolic products of enzymatic processes acting on well-characterized exogenous compounds.
3. Taxonomy: 11 organizations have submitted taxonomy/structure relationship datasets to PubChem in the form of 'Annotations.' For example, the Natural Product Activity and Species Sources (NPASS) has 540,494 annotations in PubChem. The Lotus database has submitted 434,081 annotations, where an 'annotation' is defined as a relationship between a taxonomy identifier and a pubchem structure.

The pubchem.bio package is designed to utilize all these data sources to enable efficient and comprehensive custom metabolome library creation through PubChem centralization. The vast majority of the functionality of pubchem.bio is arranged as five R functions. These functions retrieve, organize, and subset PubChem data to enable generation of custom taxonomically informed libraries.

get.pubchem.ftp: The puchem.bio R package accesses NCBI's PubChem and Taxonomy databases programatically, downloading data primarily through the FTP interface. The downloaded files are stored in a temporary directory, unzipped, and parsed, retaining only the portions of the data needed. Several derivative data sets are generated, each indexed by PubChem CID or taxonomy ID, and stored internally as a

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

data.table¹⁵ formatted data frames, enabling fast searching and filtering. These files are saved and retained to a local drive determined by the user. This function only downloads and parses data into smaller, more easily managed chunks, to enable further downstream handling.

build.cid.lca: This function utilizes all selected sources of taxonomy data, which can include both pathway data such as WikiPathways¹⁶, which is sometimes built for a specific species, or the taxonomic annotations from PubChem's data sources, such as the Lotus database⁹. To make full use of this taxonomy data, pubchem.bio organizes each PubChem CID into the nested NCBI Taxonomy¹⁷ hierarchy. This network structure (also stored as a data.table) enables extrapolation of metabolite (CID) presence even in species for which the metabolite has not been reported, through the notion of lowest common ancestor (LCA). The LCA approach has been adopted for metaproteomics studies as a mechanism to deal with redundancy in peptide sequence within a given taxonomic clade¹⁸. The pubchem.bio package adapts this logic for small molecule structures, enabling inference on the plausible metabolites for any given species.

Table 1: Example demonstrating inference of LCA from taxonomy data.

cid	metabolite	Latin binomial	species	genus	tribe	subfamily	family	phylum	kingdom	domain	root
174174	atropine	<i>Atropa belladonna</i>	33113	24609	424566	424551	4070	35493	33090	2759	1
174174	atropine	<i>Datura metel</i>	35625	4074	424565	424551	4070	35493	33090	2759	1
1548943	capsaicin	<i>Capsicum frutescens</i>	4073	4071	424564	424551	4070	35493	33090	2759	1
1548943	capsaicin	<i>Capsicum annuum</i>	4072	4071	424564	424551	4070	35493	33090	2759	1
44254980	Ceratodictyol B	<i>Haliclona cymaeiformis</i>	1385788	6057	NA	NA	6056	6040	33208	2759	1
44254980	Ceratodictyol B	<i>Ceratodictyon spongiosum</i>	38331	38330	NA	NA	31498	2763	NA	2759	1

Table 1 provides an example of how LCA is assigned. Note that not all taxonomic levels are displayed, for simplicity. If one wished to determine the lowest common ancestor for capsaicin (in blue, Table 1), the metabolite in peppers which provide their spicy heat, all biological species which are known to contain capsaicin are catalogued (only a small subset is shown in Table 1). The full taxonomic hierarchy is then generated for each species, and the lowest taxonomic level which contains all examples of capsaicin is assigned at the LCA of 4071, the genus *Capsicum*. Alternatively, consider atropine, in red. Atropine is found in both *Atropa* and *Datura* species, and the lowest common ancestor is Taxonomy ID 424551, the subfamily Solanoideae, since each of the two species have a distinct genus and tribe. The Ceratodictyol B example is more complex, and will be discussed in the results.

This output cid.lca dataset generated by the build.cid.lca function is both saved with results from get.pubchem.ftp and returned to the R console, and is used for downstream

Analyst Accepted Manuscript

library creation. This function is separated from `get.pubchem.ftp` only for practical reasons, as the function takes a bit of time to run.

build.pubchem.bio: With all of the data now organized from the first two functions, the `build.pubchem.bio` function will generate a `data.table` containing only metabolites which are found from the selected data sources, which may include structure databases, pathway databases, and/or taxonomy-structure databases. The returned dataset will include CID, compound name, molecular formula, monoisotopic mass, smiles structure, InChIKey, etc. Additionally, this function enables calculation of all physicochemical properties available through the `rdck` package¹⁸. The data set is both saved in the local directory and returned as an R `data.table`.

build.taxon.metabolome: The `pubchem.bio` dataset created in the step above contains ALL biological compounds from PubChem. The `build.taxon.metabolome` function uses all selected taxonomy data to filter and/or score all biological compounds that have an assigned LCA. The user supplies one or more target taxa, using a NCBI Taxonomy number. For each taxon identifier, the full taxonomic hierarchy is extracted, and all CIDs which have a lowest common ancestor that matches any of the taxon's hierarchy levels are assigned a score of '1'. All CIDs which have taxonomy data assigned, but do not fall within the selected taxonomic hierarchy are assigned a score between '0' and '1', based on how many taxonomic levels separate the metabolite lowest common ancestor from the 'root' of the taxonomic tree. Put another way, parent taxa up to the lowest common ancestor inherit the metabolome of child taxa, and then traverse the tree using the proposed scoring system. In this way, highly specialized metabolites from *Streptomyces*, for example, are assigned a low taxonomy score if the user is building a *Solanum* metabolome library. All CIDs which have no taxonomic data are left with a taxonomy score of NA.

Table 1 demonstrates the `cid.lca` theory. If we build a taxon metabolome for *Datura metel*, our exported metabolome will contain both atropine and capsaicin. However Atropine will have a taxonomy score of '1', while capsaicin will have a taxonomy score between 0 and 1. Since capsaicin is found in a genus (*Capsicum*) that is relatively closely related to *Datura metel* – they share a common subfamily – the taxonomy score will be larger, but less than 1. In practice the assigned score for capsaicin, when building a library for *Datura metel*, is 0.76, reflecting the fractional number of taxonomic levels which separate the two, relative to the total number of taxonomic levels.

Note also that if one were to build a taxon metabolome library for the genus *Capsicum*, the metabolome would contain atropine with a taxonomy score of 1, since the lowest common ancestor for atropine – Taxonomy ID 424551 – is a direct taxonomic ancestor of the *Capsicum* genus. This may not, at first glance, appear to be a desirable result. However, this captures evolutionary concepts. Atropine is found in multiple genera



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

within the subfamily Solanoideae. We must either assume that (1) the only species atropine is found in are those listed or (2) that the species that have been found to contain atropine is some subset of all species that contain atropine, and that there is missing data in PubChem. This missing data is due to some combination of (a) unpublished or uncatalogued studies which *demonstrate the absence* of atropine from other *Datura* species, and (b) studies looking at atropine in other species which have yet to be performed. If no one has published on the presence or absence of atropine in *Datura ferox*, for example, we have no evidence that atropine is absent from that species, which is not evidence that atropine is absent from *Datura ferox*.

The pubchem.bio package uses the LCA approach to infer metabolome content for any species, whereby one can assume that the data listed in PubChem is a subset of the species in which Atropine can be found. Rephrased, we do not currently have complete data, so must infer the taxonomic true range of each given metabolite. Given that all occurrences of Atropine are found within the subfamily Solanoideae, we infer that any member of this subfamily may plausibly have evolved the capacity to synthesize atropine. There is of course also a practical reason for wanting to be inclusive in our metabolome search space – if we are measuring the metabolome of peppers, and atropine is present in some samples, we want to identify it as such, as it can have toxic properties.

build.primary.metabolome: In theory, the PubChem metabolites which have been assigned an LCA of ‘1’ – the root of all life – can be considered a list of primary metabolites contained within PubChem. This function enables users to extract all highly conserved metabolites from the full pubchem.bio output.

Additionally, there are currently three ‘export’ functions, for exporting the full dataset in .csv format, or in formats compatible with Sirius or MSFinder.

Results:

The pubchem.bio package was demonstrated Aug 22nd, 2025. All processing reflects performance on a Windows 11 PC, with 64 GB RAM, a 1.0 TB SSD hard drive, and 13th Gen Intel Core i9-13900K 3.00 GHz processor. The get.pubchem.bio function was run to retrieve and store data for later use. The downloaded content occupies approximately 18 GB on disk, though users should ensure there is at least triple that to accommodate temporary files during processing. The get.pubchem.ftp function took approximately 2.5 hours to run.

There are 885 total sources of PubChem data. For this example, five were considered as biological databases: Metabolomics Workbench, Human Metabolome Database (HMDB)⁷, ChEBI¹⁹, LIPID MAPS²⁰, and MassBank of North America (MoNA)²¹. Users can select any of the 885 databases as source databases. There are ten sources of

Analyst Accepted Manuscript

pathway data, with PathBank²², PlantCyc²³, Plant Reactome²⁴, BioCyc²⁵, Reactome²⁶, and WikiPathways¹⁶ each listing more than 10,000 records. All pathway sources were used, and any taxonomy-assigned pathway was incorporated into the CID-LCA assignment. Eleven sources provide taxonomic-structure (CID) pairs, with FooDB²⁷, NPASS¹⁰, LOTUS⁹, HMDB⁷, KNApSAcK²⁸, and NPA²⁹ each supplying more than 10,000 records. For this example, only LOTUS was used, but the user is free to use any or all of the sources. CID taxonomy associations within PubChem amount to 3,857,536 records, including 434,081 from LOTUS. Adding taxonomic associations from pathway sources added an additional 375,375 CID-taxonomy pairs, bringing the total to 809,456. The build.cid.lca function took approximately 1.1 hours to run, resulting in 247,050 cid-lca pairs. Some pathways and biological data sources do include non-biological compounds, such as pesticide degradation pathways in plants. It should be noted that this package does not remove those compounds from the resulting structure library.

There are numerous 'NA' values in the taxonomy hierarchy – not all species have assignments at each level. In fact, the only two taxonomic levels that have no missing values are species and domain. This can be seen in Table 1, specifically for the Ceratodictyol B example.

In preliminary CID-LCA assignments, there were nearly 10,000 metabolites which were demonstrated to have a listed lowest common ancestor of '1' – the root of all biology. These could be interpreted as being the most highly conserved, and therefore classified as 'primary' metabolites. It was, however, observed that sparse and spurious CID-Taxonomy associations can generate false positive assignments with unexpectedly high LCA taxonomic assignments. For example, Ceratodictyol B was assigned an LCA of '2759,' as depicted in Table 1. Taxonomy ID 2759 is 'Eukaryota,' or eukaryotes. Assignment to this level would result in assignment of Ceratodictyol B to every eukaryote metabolome library. This occurred due to Ceratodictyol B being reported from precisely two species, each listed by two sources: a marine sponge, *Haliclona cymaeformis*, and a red algae, *Ceratodictyon spongiosum*. *Ceratodictyon spongiosum* can form symbioses with sponges, calling into question whether *Haliclona* is producing Ceratodictyol B, or harboring *Ceratodictyon*, which is producing Ceratodictyol B. In fact, the paper describing these results does not distinguish between the two³⁰. This result can therefore be interpreted as a false positive assignment of LCA at a much higher taxonomic level than is warranted. The LCA approach is built on a premise that each taxon-CID association reflects phylogeny, while the LCA approach implemented is dependent on taxonomy as a surrogate, and the case in question reflects complex cross-taxa symbioses which taxonomy doesn't capture well.

To enable customization in the assignment of LCA, the build.cid.lca function was provided an option, 'min.taxid.table.length', which can prevent such spurious LCA

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

assignments. In the case of Ceratodictyol B, it can be seen that there are 2 taxonomic identifiers at each level species, genus, family, phylum, and kingdom, and only when we arrive at domain do we find that the number of unique taxonomy identifiers drops to 1. There are very few levels of unique taxonomy ID count across taxonomy levels, which is used as an indicator that there are few records which are taxonomically (and presumably phylogenetically) isolated from each other. For each taxonomic level, the number of unique taxonomy identifiers which map to the CID are calculated, and the frequency examined. All instances with frequency of '0' (all NA) are removed. If the length of the resulting tables is less than or equal to min.taxid.table.length, then the LCA is assigned within the lowest taxonomic level with the most frequently observed n.taxa, where n.taxa is the number of taxa within that taxonomic level. The default value for min.taxid.table.length is '3'. In this way, metabolites which are observed in very few taxa which are very disparate in their taxonomic distance from each other can be assigned to two (or rarely more) LCAs.

In the example depicted in table 1, Ceratodictyol B is found in two species, genera, families, and phyla. As such, the most frequently observed n.taxa is '2'. The lowest common taxonomic level for all taxonomic levels with n.taxa = 2 is species, so LCA is assigned within the level 'species.' In this case, there are two species, and Ceratodictyol B is assigned two LCA values, one for each species. Hypothetically, if there were three *Ceratodictyon spp* linked to Ceratodictyol B, the LCA would be assigned within the level 'genus' instead. Since there is still only one species within the genus *Haliclona*, the first LCA would be still be assigned to the species *Haliclona cymaeformis* (taxonomy ID = 1385788). For the second taxa set, there are three species of *Ceratodictyon*, the LCA for which would then be assigned at the genus level (taxonomy ID = 38330).

After running the build.cid.lca function, min.taxid.table.length=3, 2859 metabolites were assigned an LCA equal to 1. When considering construction of in-house libraries of metabolites, this list can be used to ensure that the investment in metabolite standards is spent on metabolites most likely to be observed across all sample types. The full table of assigned primary metabolites is provided as supplementary material, but can also be regenerated using the R package to ensure that as the data in PubChem grows, so can the list of primary metabolites.

The build.pubchem.bio function was used to create a biological metabolome libraries, using default values, which include the use of the biologically associated datasource including Metabolomics Workbench, Human Metabolome Database (HMDB), ChEBI, LIPID MAPS, and MassBank of North America (MoNA). XLogP, AcidicGroupCount, BasicGroupCount, and TPSA for each metabolite were predicted using rcdk¹⁸. This biological subset of PubChem took approximately 1.1 hours to build, and resulted in

1,268,778 metabolites. The molecular weight distribution of metabolites is clearly altered between PubChem (Fig 1a) and the pubchem.bio dataset (Fig 1b).

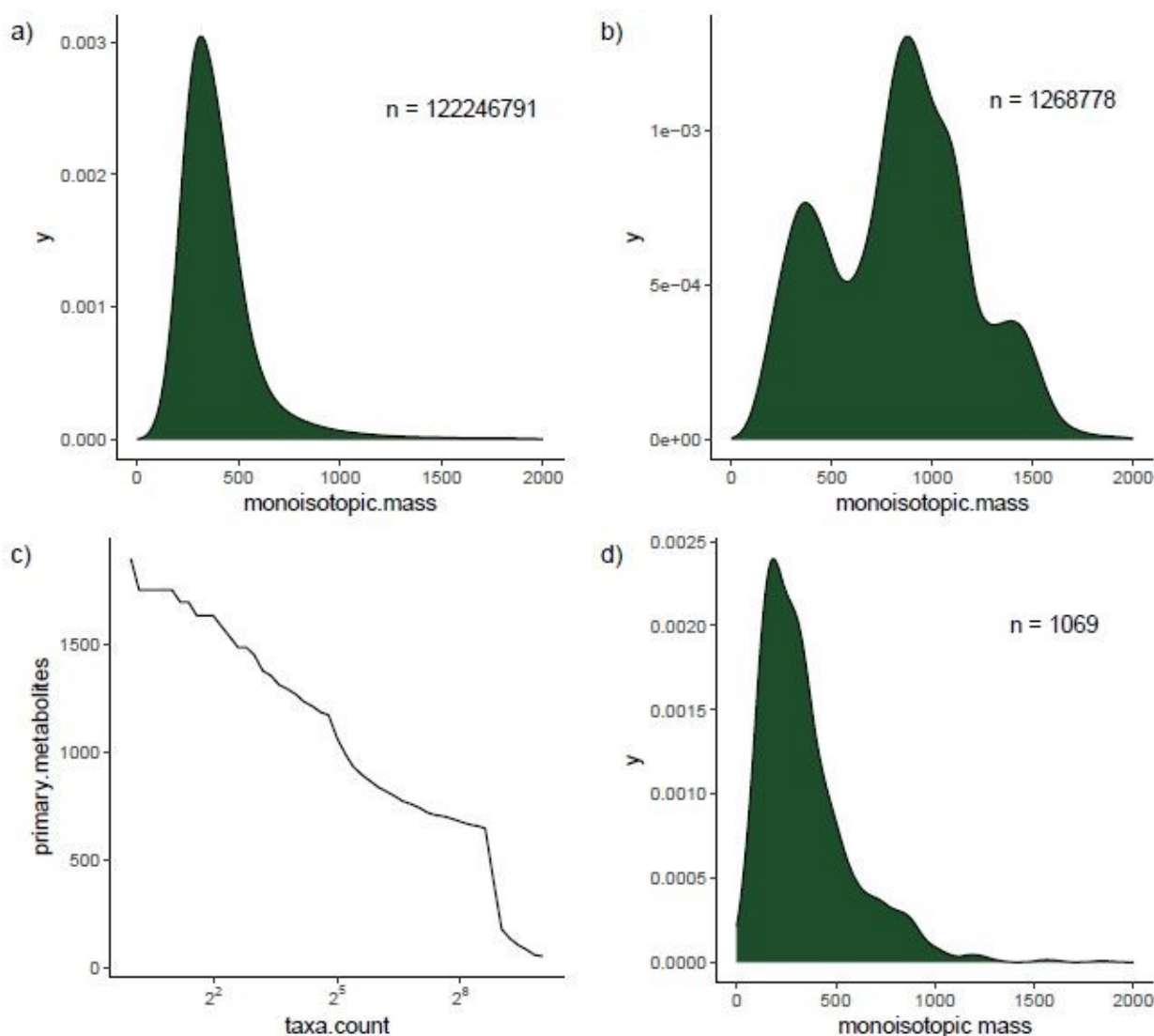


Figure 1. pubchem.bio filtering changes the molecular weight distribution of pubchem structures. A. the molecular weight distribution for all structures in PubChem is dominated by structures with monoisotopic weight < 500 Dalton. B. After filtering for biological compounds, a new peak around 1000 emerges, derived from larger lipids and natural products. C. The relationship between the number of conserved 'primary' metabolites – those with an assigned LCA of '1', the root of all cellular life – and the number of assigned taxa. D. The distribution of monoisotopic masses for all metabolites assigned an LCA of '1' and which have at least 2^5 taxonomic assignments, representing a conservative estimate of the conserved 'primary' metabolome size.

Figure 1c plots the relationship between the number of taxa which map to a given metabolite and the number of metabolites which also have an LCA of '1' (root). A

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

strong log-linear relationship is observed which begins to plateau at taxa counts above 2⁵, before falling rapidly at 2⁸. These inflection points in the curve represent some combination of undersampling of taxa-structure relationships, and the loss of metabolic conservation across taxa – we cannot disentangle the two from the available data. The default value for assignment of a metabolite as ‘primary’ in this function is therefore conservatively set to 2⁵, though this value is a variable in the function that can be further refined as more and more data are incorporated into PubChem. At a minimum frequency of 2⁵ taxonomic occurrences, there are 1069 metabolites considered as primary. This value falls between the ~200 reported recently in a survey of mammalian metabolomics data³¹ and the ~6000 reported in a genome-informed approach³². As can be seen from Figure 1c, even within the pubchem.bio package, the value can range from several hundred to over 2000. This package can help to inform on core metabolic functionalities, but is not going to provide an unambiguous answer to this important evolutionary question.

To demonstrate the utility of taxonomy filtering and scoring, a food example is considered. Salsa is comprised largely of tomato, pepper, cilantro, onion and garlic. These foods map to Taxonomy database IDs of 4081, 4072, 4047, 4679, 4682. Building the salsa metabolome from these five ingredients took 6.5 minutes. For the ingredients listed above, there were 17162, 17058, 16895, 17062, and 17027 metabolites mapped, respectively. In total, 18116 metabolites are mapped to these five species, from four different genera. Tomato and pepper are closely related, and 98% of all tomato metabolites are also in the pepper metabolome. Garlic and onion are also closely related to each other, sharing the same genus, *Allium*. 99.7% of garlic metabolites are also present in the onion metabolome. Tomato and garlic are more distantly related, with a common taxonomic ancestor at the kingdom level, Viridiplantae. 97% of all tomato metabolites are also present in the garlic metabolome. A total of 243,792 metabolites are assigned an aggregate taxonomy score, with scores ranging from 0 (extremely unlikely to be found given the five species listed) to 1 (highly plausible, as they are metabolites which map to at least one of the species). Intermediate scores represent increasing probability that a metabolite may be present, given imperfect knowledge of species-specific metabolism (Figure 2). Of course, the small molecule composition of salsa will ultimately be additionally impacted by cooking and preservation, so the reported library should be viewed as the potential small molecules components of salsa which are derived directly from the ingredients, not comprehensive of, for example, maillard products that may form.

Analyst Accepted Manuscript

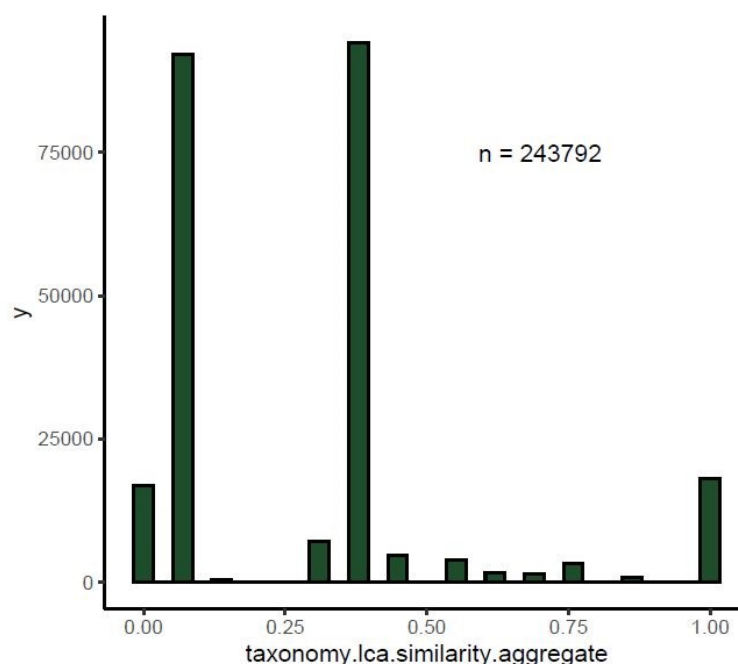


Figure 2: The metabolome for the five species mixture for major salsa ingredients contains approximately 18,116 metabolites which map to those species. These metabolites are assigned a score of '1'. All other metabolites which have taxonomic mapping are assigned values between 0 and some value less than 1, reflecting the taxonomic distance to the target species. Biological metabolites with no mapped taxonomy are not assigned a score. A total of 243,792 metabolites have been assigned a taxonomy score.

Usage Discussion:

The pubchem.bio package is performing tasks that would take humans much longer to do. The process of assigning lowest common ancestor to PubChem structures is relatively novel, but based on the same principles as chemotaxonomy – a field with decades of applications³³, and highly similar to that used in metaproteomics studies³⁴. The LCA approach turns a limitation of chemotaxonomy – that even specialized metabolites are often not unique to a given species – into a useful trait, enabling generalization of metabolite distributions for creating more comprehensive metabolome libraries for any species, even those poorly studied in the past. However, it must be noted that the LCA approach is difficult to validate and optimize, given the incomplete knowledge of any given biological sample's metabolic components. Default values are a reasonable starting point for users, and should provide relatively conservative results – a database which utilizes all available data to generate a comprehensive list of analytes which may plausibly be observed.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Taxonomic and biological data in PubChem is both incomplete (false negative) and can contain errors (false positive). Each of these types of errors can result in inaccurate taxon-specific metabolome data and assigned LCA. For example, Capsaicin is widely considered to be a specialized metabolite of peppers, genus *Capsicum*. The ‘taxonomy’ section of the PubChem webpage for Capsaicin lists numerous *Capsicum* species, but also the prokaryote *Streptomyces roseofulvus*, submitted by the NPASS database. The original reference supporting this association is not apparent, so it is difficult to evaluate this claim. That said, if the user selects both Lotus and NPASS databases when building LCA, the LCA for Capsaicin may be ‘1’ – the root of all biology, depending on the assigned ‘min.taxid.table.length’ value used. However, if the users opt to use only the more selective Lotus database (the current default value), the LCA is returned as ‘4071’, the Taxonomy ID for the *Capsicum* genus.

Biological databases and pathway databases may each incorporate exogenous compounds into their data in a taxonomy informed manner. Capsaicin can be degraded by *E. coli*, for example, and is therefore can be reported as an *E. coli* metabolite, despite the fact that *E. coli* is not known to be able to synthesize capsaicin. In theory, if all reaction data for a given species were available in computer readable format, one could remove metabolites which are not considered metabolic products of a reaction – this is an opportunity for additional development in the future.

The taxonomy scoring applied here is based on taxonomic relationships as described by decades of taxonomic research. It should be noted that the taxonomic approaches not necessarily consistent across all clades. While taxonomy is a representation of phylogeny, scoring algorithm of pubchem.bio is based on what must be considered an imperfect taxonomic representation of phylogenetic time and distance between species.

The work here demonstrates that the pubchem.bio package is able to return meaningful results to users, enabling users to create taxonomy-informed libraries in an automated manner. This does not mean that the process is completely objective – the user will need to select which sources to use and to assign an appropriate ‘min.taxid.table.length’ value, for example. The ‘correct’ metabolome library is one which serves the need of the user, whether that be the minimal most confident metabolome, the maximum plausible metabolome, or some intermediate level. The package is designed to enable users to build the library they need, depending on the circumstance. As the resources in PubChem grow, so to does the ability for pubchem.bio to convert that collected knowledge into metabolome libraries.

The pubchem.bio package fills a functional void. There are no other resources to enable rapid generation of customizable metabolome libraries for any species. While individual researchers have done this in the past manually, pubchem.bio enables what would have taken hours, at best, to be done in minutes. The most similar tool

Analyst Accepted Manuscript

available, an R package called tima, available via zenodo³⁵, performs comparable tasks to pubchem.bio, but does so using more limited resources, and does not appear to utilize a lowest common ancestor extrapolation approach. There are also tools for retrieving and handling existing structure libraries. The CompoundDB R package is useful for retrieving and storing structure data from a few select well curated database sources, such as HMDB, and provides an extensible data structure to integrate into other Rformassspectrometry packages³⁶.

Conclusion:

Using freely available data from PubChem, derived through many invaluable sources that have been contributed individual database efforts, the pubchem.bio package enables the creation of metabolomics-ready libraries fully informed by biological and taxonomic digital resources. After an initial setup phase requiring several hours of computer time, each new species metabolome library can be generated in minutes.

Future Directions:

Additional metadata is downloaded as part of the get.pubchem.ftp function, notably MeSH data, full CID-pathway membership details, CID-substance relationships, synonyms, and CAS numbers. These metadata have potential utility in, for example, pathway analysis³⁷. The pubchem.bio package makes no effort to incorporate any additional exogenous compounds beyond what is already incorporated into biological databases such as CheBI or HMDB. Future work may incorporate the data from PubChem Lite, which is focused on these exogenous analytes. Additionally, the use of reactions rather than pathways may enable removal of exogenous catabolic reactants from the species specific metabolome output, when desired. Metabolite libraries will also be integrated into compounds libraries via the CompoundDB R package, a component of the Rformassspectrometry ecosystem, to enable broader interoperability.

Software Availability:

The pubchem.bio package is available via CRAN under a GPL-3 license.

Acknowledgements:

I would like to acknowledge the scientists who compiled datasets and submitted to PubChem^{7–10,14,16,17,22–29}, without which there would be no pubchem.bio. I would also like to acknowledge Paul Theissen with PubChem, who patiently provided valuable guidance for accessing PubChem programmatically. CDB works in the Colorado State University Analytical Resources Core, Research Resource ID (RRID: SCR_021758).

References:



- (1) Bino, R. J.; Hall, R. D.; Fiehn, O.; Kopka, J.; Saito, K.; Draper, J.; Nikolau, B. J.; Mendes, P.; Roessner-Tunali, U.; Beale, M. H.; Trethewey, R. N.; Lange, B. M.; Wurtele, E. S.; Sumner, L. W. Potential of Metabolomics as a Functional Genomics Tool. *Trends in Plant Science* **2004**, *9* (9), 418–425. <https://doi.org/10.1016/j.tplants.2004.07.004>.
- (2) de Jonge, N. F.; Mildau, K.; Meijer, D.; Louwen, J. J. R.; Bueschl, C.; Huber, F.; van der Hooft, J. J. J. Good Practices and Recommendations for Using and Benchmarking Computational Metabolomics Metabolite Annotation Tools. *Metabolomics* **2022**, *18* (12), 103. <https://doi.org/10.1007/s11306-022-01963-y>.
- (3) Delabrière, A.; Gianfrotta, C.; Dechaumet, S.; Damont, A.; Hautbergue, T.; Roger, P.; Jamin, E. L.; Puel, O.; Junot, C.; Fenaille, F.; Thévenot, E. A. mineMS2: Annotation of Spectral Libraries with Exact Fragmentation Patterns. *Journal of Cheminformatics* **2025**, *17* (1), 111. <https://doi.org/10.1186/s13321-025-01051-y>.
- (4) Chang, C. H.; Schwartz, S. C.; Im, A. K.; Bloodsworth, K. J.; Webb-Robertson, B.-J. M.; Ewing, R. G.; Metz, T. O.; Ross, D. H. Assessing the Impact of Measurement Precision on Metabolite Identification Probability in Multidimensional Mass Spectrometry-Based, Reference-Free Metabolomics. *Anal. Chem.* **2025**, *97* (26), 13861–13871. <https://doi.org/10.1021/acs.analchem.5c01067>.
- (5) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Medicinal Research Reviews* **1996**, *16* (1), 3–50. [https://doi.org/10.1002/\(SICI\)1098-1128\(199601\)16:1<3::AID-MED1>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6).
- (6) Elapavalore, A.; Ross, D. H.; Grouès, V.; Aurich, D.; Krinsky, A. M.; Kim, S.; Thiessen, P. A.; Zhang, J.; Dodds, J. N.; Baker, E. S.; Bolton, E. E.; Xu, L.; Schymanski, E. L. PubChemLite Plus Collision Cross Section (CCS) Values for Enhanced Interpretation of Nontarget Environmental Data. *Environ. Sci. Technol. Lett.* **2025**, *12* (2), 166–174. <https://doi.org/10.1021/acs.estlett.4c01003>.
- (7) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res* **2022**, *50* (D1), D622–D631. <https://doi.org/10.1093/nar/gkab1062>.
- (8) Chandrasekhar, V.; Rajan, K.; Kanakam, S. R. S.; Sharma, N.; Weißenborn, V.; Schaub, J.; Steinbeck, C. COCONUT 2.0: A Comprehensive Overhaul and Curation of the Collection of Open Natural Products Database. *Nucleic Acids Res* **2025**, *53* (D1), D634–D643. <https://doi.org/10.1093/nar/gkae1063>.
- (9) Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J. G.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; Pauli, G. F.; Wolfender, J.-L.; Bisson, J.; Allard, P.-M. The LOTUS Initiative for Open Knowledge Management in Natural Products Research. *eLife* **2022**, *11*, e70780. <https://doi.org/10.7554/eLife.70780>.
- (10) Zhao, H.; Yang, Y.; Wang, S.; Yang, X.; Zhou, K.; Xu, C.; Zhang, X.; Fan, J.; Hou, D.; Li, X.; Lin, H.; Tan, Y.; Wang, S.; Chu, X.-Y.; Zhuoma, D.; Zhang, F.; Ju, D.;


- Zeng, X.; Chen, Y. Z. NPASS Database Update 2023: Quantitative Natural Product Activity and Species Source Database for Biomedical Research. *Nucleic Acids Res* **2023**, *51* (D1), D621–D628. <https://doi.org/10.1093/nar/gkac1069>.
- (11) Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; Bisson, J.; Bagheri, M.; Saesong, T.; Ebrahimi, S. N.; Ingkaninan, K.; Wolfender, J.-L.; Allard, P.-M. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Front. Plant Sci.* **2019**, *10*. <https://doi.org/10.3389/fpls.2019.01329>.
- (12) Zuffa, S.; Schmid, R.; Bauermeister, A.; P Gomes, P. W.; Caraballo-Rodriguez, A. M.; El Abiead, Y.; Aron, A. T.; Gentry, E. C.; Zemlin, J.; Meehan, M. J.; Avalon, N. E.; Cichewicz, R. H.; Buzun, E.; Terrazas, M. C.; Hsu, C.-Y.; Oles, R.; Ayala, A. V.; Zhao, J.; Chu, H.; Kuijpers, M. C. M.; Jackrel, S. L.; Tugizimana, F.; Nepthali, L. P.; Dubery, I. A.; Madala, N. E.; Moreira, E. A.; Costa-Lotufo, L. V.; Lopes, N. P.; Rezende-Teixeira, P.; Jimenez, P. C.; Rimal, B.; Patterson, A. D.; Traxler, M. F.; Pessotti, R. de C.; Alvarado-Villalobos, D.; Tamayo-Castillo, G.; Chaverri, P.; Escudero-Leyva, E.; Quiros-Guerrero, L.-M.; Bory, A. J.; Joubert, J.; Rutz, A.; Wolfender, J.-L.; Allard, P.-M.; Sichert, A.; Pontrelli, S.; Pullman, B. S.; Bandeira, N.; Gerwick, W. H.; Gindro, K.; Massana-Codina, J.; Wagner, B. C.; Forchhammer, K.; Petras, D.; Aiosa, N.; Garg, N.; Liebeke, M.; Bourceau, P.; Kang, K. B.; Gadhavi, H.; de Carvalho, L. P. S.; Silva Dos Santos, M.; Pérez-Lorente, A. I.; Molina-Santiago, C.; Romero, D.; Franke, R.; Brönstrup, M.; Vera Ponce de León, A.; Pope, P. B.; La Rosa, S. L.; La Barbera, G.; Roager, H. M.; Laursen, M. F.; Hammerle, F.; Siewert, B.; Peintner, U.; Licon-Cassani, C.; Rodriguez-Orduña, L.; Rampler, E.; Hildebrand, F.; Koellensperger, G.; Schoeny, H.; Hohenwallner, K.; Panzenboeck, L.; Gregor, R.; O'Neill, E. C.; Roxborough, E. T.; Odoi, J.; Bale, N. J.; Ding, S.; Sinninghe Damsté, J. S.; Guan, X. L.; Cui, J. J.; Ju, K.-S.; Silva, D. B.; Silva, F. M. R.; da Silva, G. F.; Koolen, H. H. F.; Grundmann, C.; Clement, J. A.; Mohimani, H.; Broders, K.; McPhail, K. L.; Ober-Singleton, S. E.; Rath, C. M.; McDonald, D.; Knight, R.; Wang, M.; Dorrestein, P. C. microbeMASST: A Taxonomically Informed Mass Spectrometry Search Tool for Microbial Metabolomics Data. *Nat Microbiol* **2024**, *9* (2), 336–345. <https://doi.org/10.1038/s41564-023-01575-9>.
- (13) Bragagnolo, F. S.; Ibáñez, E.; Cifuentes, A.; Rostagno, M. A. Building Your Own Database for Foodomics: Tips and Tricks. *Food Chemistry International n/a* (n/a). <https://doi.org/10.1002/fci2.70009>.
- (14) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res* **2023**, *51* (D1), D1373–D1380. <https://doi.org/10.1093/nar/gkac956>.
- (15) Barrett, T.; Dowle, M.; Srinivasan, A.; Gorecki, J.; Chirico, M.; Hocking, T.; Schwendinger, B.; Krylov, I. *Data.Table: Extension of `data.Frame`*; 2025. <https://doi.org/10.32614/CRAN.package.data.table>.
- (16) Agrawal, A.; Balci, H.; Hanspers, K.; Coort, S. L.; Martens, M.; Slenter, D. N.; Ehrhart, F.; Digles, D.; Waagmeester, A.; Wassink, I.; Abbassi-Daloui, T.; Lopes, E. N.; Iyer, A.; Acosta, J. M.; Willighagen, L. G.; Nishida, K.; Riutta, A.; Basaric, H.; Evelo, C. T.; Willighagen, E. L.; Kutmon, M.; Pico, A. R. WikiPathways 2024: Next

- Generation Pathway Database. *Nucleic Acids Res* **2024**, 52 (D1), D679–D689. <https://doi.org/10.1093/nar/gkad960>.
- (17) Schoch, C. L.; Ciufu, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; McVeigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-Mizrachi, I. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database (Oxford)* **2020**, 2020, baaa062. <https://doi.org/10.1093/database/baaa062>.
- (18) Guha, R. Chemical Informatics Functionality in R. *Journal of Statistical Software* **2007**, 18 (6).
- (19) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res* **2016**, 44 (D1), D1214–D1219. <https://doi.org/10.1093/nar/gkv1031>.
- (20) O'Donnell, V. B.; Dennis, E. A.; Wakelam, M. J. O.; Subramaniam, S. LIPID MAPS: Serving the next Generation of Lipid Researchers with Tools, Resources, Data, and Training. *Science Signaling* **2019**, 12 (563), eaaw2964. <https://doi.org/10.1126/scisignal.aaw2964>.
- (21) Wohlgemuth, G.; Mehta, S. S.; Mejia, R. F.; Neumann, S.; Pedrosa, D.; Pluskal, T.; Schymanski, E. L.; Willighagen, E. L.; Wilson, M.; Wishart, D. S.; Arita, M.; Dorrestein, P. C.; Bandeira, N.; Wang, M.; Schulze, T.; Salek, R. M.; Steinbeck, C.; Nainala, V. C.; Mistrik, R.; Nishioka, T.; Fiehn, O. SPLASH, a Hashed Identifier for Mass Spectra. *Nat Biotechnol* **2016**, 34 (11), 1099–1101. <https://doi.org/10.1038/nbt.3689>.
- (22) Wishart, D. S.; Kruger, R.; Sivakumaran, A.; Harford, K.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Jackson, H.; Sykes, G.; Ramirez-Gaona, M.; Marcu, A.; Li, C.; Yee, K.; Garros, C.; Rayat, D. Y.; Coleongco, J.; Nandyala, T.; Gautam, V.; Oler, E. PathBank 2.0—the Pathway Database for Model Organism Metabolomics. *Nucleic Acids Res* **2024**, 52 (D1), D654–D662. <https://doi.org/10.1093/nar/gkad1041>.
- (23) Hawkins, C.; Xue, B.; Yasmin, F.; Wyatt, G.; Zerbe, P.; Rhee, S. Y. Plant Metabolic Network 16: Expansion of Underrepresented Plant Groups and Experimentally Supported Enzyme Data. *Nucleic Acids Res* **2025**, 53 (D1), D1606–D1613. <https://doi.org/10.1093/nar/gkae991>.
- (24) Gupta, P.; Naithani, S.; Preece, J.; Kim, S.; Cheng, T.; D'Eustachio, P.; Elser, J.; Bolton, E. E.; Jaiswal, P. Plant Reactome and PubChem: The Plant Pathway and (Bio)Chemical Entity Knowledgebases. In *Plant Bioinformatics: Methods and Protocols*; Edwards, D., Ed.; Springer US: New York, NY, 2022; pp 511–525. https://doi.org/10.1007/978-1-0716-2067-0_27.
- (25) 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* **2019**, 20 (4), 1085–1093. <https://doi.org/10.1093/bib/bbx085>.
- (26) Rothfels, K.; Milacic, M.; Matthews, L.; Haw, R.; Sevilla, C.; Gillespie, M.; Stephan, R.; Gong, C.; Ragueneau, E.; May, B.; Shamovsky, V.; Wright, A.; Weiser, J.; Beavers, D.; Conley, P.; Tiwari, K.; Jassal, B.; Griss, J.; Senff-Ribeiro, A.; Brunson,

- T.; Petryszak, R.; Hermjakob, H.; D'Eustachio, P.; Wu, G.; Stein, L. Using the Reactome Database. *Current Protocols* **2023**, 3 (4), e722. <https://doi.org/10.1002/cpz1.722>.
- (27) Wishart, D. S. Knowledge Translation and Knowledge Mobilization from the FoodBALL Project. *Appl. Physiol. Nutr. Metab.* **2024**, 49 (9), 1279–1285. <https://doi.org/10.1139/apnm-2023-0573>.
- (28) Nakamura, Y.; Mochamad Afendi, F.; Kawsar Parvin, A.; Ono, N.; Tanaka, K.; Hirai Morita, A.; Sato, T.; Sugiura, T.; Altaf-UI-Amin, Md.; Kanaya, S. KNApSACK Metabolite Activity Database for Retrieving the Relationships Between Metabolites and Biological Activities. *Plant Cell Physiol* **2014**, 55 (1), e7. <https://doi.org/10.1093/pcp/pct176>.
- (29) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M.-L.; Duncan, K. R.; Linington, R. G. The Natural Products Atlas 2.0: A Database of Microbially-Derived Natural Products. *Nucleic Acids Res* **2022**, 50 (D1), D1317–D1323. <https://doi.org/10.1093/nar/gkab941>.
- (30) Akiyama, T.; Ueoka, R.; van Soest, R. W. M.; Matsunaga, S. Ceratodictyols, 1-Glycerol Ethers from the Red Alga–Sponge Association Ceratodictyon Spongiosum/Haliclona Cymaeformis. *J. Nat. Prod.* **2009**, 72 (8), 1552–1554. <https://doi.org/10.1021/np900355m>.
- (31) Liska, O.; Boross, G.; Rocabert, C.; Szappanos, B.; Tengölics, R.; Papp, B. Principles of Metabolome Conservation in Animals. *Proceedings of the National Academy of Sciences* **2023**, 120 (35), e2302147120. <https://doi.org/10.1073/pnas.2302147120>.
- (32) Peregrín-Alvarez, J. M.; Sanford, C.; Parkinson, J. The Conservation and Evolutionary Modularity of Metabolism. *Genome Biol* **2009**, 10 (6), R63. <https://doi.org/10.1186/gb-2009-10-6-r63>.
- (33) Smith, P. M. *The Chemotaxonomy of Plants*; 1976.
- (34) Vande Moortele, T.; Devlaminck, B.; Van de Vyver, S.; Van Den Bossche, T.; Martens, L.; Dawyndt, P.; Mesuere, B.; Verschaffelt, P. Unipept in 2024: Expanding Metaproteomics Analysis with Support for Missed Cleavages and Semitryptic and Nontryptic Peptides. *J. Proteome Res.* **2025**, 24 (2), 949–954. <https://doi.org/10.1021/acs.jproteome.4c00848>.
- (35) Rutz, A. Tima: Taxonomically Informed Metabolite Annotation, 2024. <https://doi.org/10.5281/zenodo.14515116>.
- (36) Witting, M.; Rainer, J. Bio- and Chemoinformatic Approaches for Metabolomics Data Analysis. *Methods Mol Biol* **2025**, 2891, 67–89. https://doi.org/10.1007/978-1-0716-4334-1_4.
- (37) Wieder, C.; Frainay, C.; Poupin, N.; Rodríguez-Mier, P.; Vinson, F.; Cooke, J.; Lai, R. P.; Bundy, J. G.; Jourdan, F.; Ebbels, T. Pathway Analysis in Metabolomics: Recommendations for the Use of over-Representation Analysis. *PLOS Computational Biology* **2021**, 17 (9), e1009105. <https://doi.org/10.1371/journal.pcbi.1009105>.

Analyst Editorial Board,

This manuscript describes no new analytical data. Rather it describes a new informatic resource, the R pubchem.bio package. The package is freely available currently through my github repository: <https://github.com/cbroeckl/pubchem.bio> and has been submitted to CRAN to improve robustness and accessibility. I await a response from CRAN on any necessary edits, but am confident that it will be accepted to CRAN in the near future.



Corey Broeckling
Analytical Resources Core: Bioanalysis and Omics Center
Fort Collins, CO 80523
cbroeckl@colostate.edu
(1)-970-491-2273

