De novo molecular formula annotation and structure elucidation using SIRIUS 4

Marcus Ludwig, Markus Fleischauer, Kai Dührkop, Martin A. Hoffmann and Sebastian Böcker

Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany, sebastian.boecker@uni-jena.de

This is a preprint of an upcoming book chapter.

Abstract. SIRIUS 4 is the best-in-class computational tool for metabolite identification from high-resolution tandem mass spectrometry data. It offers *de novo* molecular formula annotation with outstanding accuracy. When searching fragmentation spectra in a structure database, it reaches over 70 % correct identifications. A predicted fingerprint, which indicates the presence or absence of thousands of molecular properties, helps to deduce information about the compound of interest even if it is not contained in any structure database. Here, we present best practices and describe how to leverage the full potential of SIRIUS 4, how to incorporate it into your own workflow and how it adds value to the analysis of mass spectrometry data beyond spectral library search.

1 Introduction

Comprehensive identification of small molecules is one of the most urgent needs in metabolomics, and related fields such as in natural products research, biomarker discovery and environmental science. Yet, this task remains highly challenging. Liquid-chromatography tandem mass spectrometry (LC-MS/MS) is one of the most prominent analytical techniques to identify biomolecules. The mere mass of a compound is not sufficient to determine the correct molecular formula, let alone its structure. Tandem mass spectrometry provides additional information but is non-trivial to interpret. Usually, metabolite identification is performed by searching fragmentation spectra in a spectral library [1–4]. However, spectral libraries are — and always will be — highly incomplete. This represents a major obstacle, particularly for secondary metabolism analysis. During the last years multiple tools were developed for searching in structure databases which are orders of magnitudes larger compared to spectral libraries; this includes CFM-ID [5], DEREPLICATOR+ [6], MAGMa [7], MetFrag [8, 9], MIDAS [10], MS-FINDER [11] and CSI:FingerID [12].

Currently, the best performing tool for this task is CSI:FingerID, successor of FingerID [13]. It is part of SIRIUS 4 [14], a software for metabolite identification from high-resolution fragmentation spectra. SIRIUS started off as a method for *de novo* molecular formula identification, but now integrates CSI:FingerID to offer combined molecular formula annotation and structure database search. SIRIUS performs metabolite identification in a two step approach: Firstly, the molecular formula of the query compound is determined via isotope pattern analysis and fragmentation trees. Second, SIRIUS uses CSI:FingerID to predict a molecular fingerprint from the given spectrum and fragmentation tree. This predicted fingerprint can be searched against a structure database to identify the most likely candidate. Searching CASMI 2016 [15] positive ion mode spectra in a database of 0.5 million structures of biological interest resulted in 74.0 % correct identifications [14]. When searching in PubChem [16], which contains many millions of structures, CSI:FingerID still achieves an identification rate of 39.4 % (74.8 % in the top 10). These rates were reached without using meta-information such as citation frequencies or production volumes; using such meta-information can be very harmful in practice [17].

Whereas spectral library search will only allow a "peek through the keyhole", SIRIUS enables untargeted identification to draw a more complete picture of a metabolic system [18]. It is understood that not every existing biomolecule is or will be contained in structure databases. But even for these instances SIRIUS offers valuable insight by providing a predicted molecular fingerprint to assist *de novo* structure elucidation and by searching in databases of hypothetical structures such as the *in silico* generated MINE databases [19]. Comprehensive compound identification is not a luxury but an indispensable step to answer biological questions. Compared to spectral library search SIRIUS offers highly increased coverage; compared to searching compounds only by mass it offers tremendously improved accuracy. Here, we present how to use SIRIUS to systematically annotate your compounds, and provide insight on common practices, judging the results and necessary prerequisites of your data.

2 What data can be processed by SIRIUS?

SIRIUS processes high-resolution, high mass accuracy fragmentation spectra, but also uses first stage of mass spectrometry (MS1) data. The statistical model of SIRIUS and the machine learning model of CSI:FingerID were trained on tandem mass spectra (MS/MS) created by collision-induced dissociation (CID), as commonly applied in LC-MS/MS experiments. Most of the training compounds were ionized by electrospray ionization (ESI). However, it has been reported that SIRIUS is also able to analyze compounds from GC-MS data which has been acquired using the soft ionization method dopant-assisted atmospheric pressure chemical ionization (dAPCI) and subsequently fragmenting ions by CID [20]. At present, SIRIUS only handles single-charged compounds.

3 Preprocessing

SIRIUS is specialized in metabolite identification and relies on other tools for proper preprocessing. Input spectra must be in centroid mode (peak picked). Besides, further preprocessing of the data is highly beneficial for good results. Open source software exists for feature finding, to group isotope peaks of each compound, estimate adducts and reject all MS/MS which cannot be assigned to a proper feature in the MS1. OpenMS [21] and MZmine 2 [22] both provide export functions tailored to the needs for SIRIUS.

It is beyond the scope of this paper to go into the details of the different preprocessing steps, but see [the Chapter on OpenMS in an upcomming book] for details on OpenMS processing. Unfortunately, we cannot propose optimal parameters, since these depend on the data. A metabolomics OpenMS workflow to preprocess data for SIRIUS may use the following OpenMS tools: `FeatureFinderMetabo`, `MetaboliteAdductDecharger` and `SiriusAdapter`. The `SiriusAdapter` can be used either to directly run SIRIUS or to export .ms-files for SIRIUS to import.

SIRIUS benefits from the following preprocessing steps:

- Reasonably averaged MS1 are more accurate than using a single MS1 spectrum. Determining the masses and intensities of the compound's isotope pattern using the chromatographic peaks can reduce errors.
- When measuring multiple MS/MS spectra of the same compound, in particular at different collision energies, it is beneficial to analyze a merged spectrum rather than the individual spectra. Fragmentation spectra can be grouped by their corresponding MS1 feature. SIRIUS will merge all grouped spectra. This is preferred over directly providing a merged spectrum as input for SIRIUS.
- MS/MS spectra which cannot be assigned to any MS1 feature should be rejected; these spectra are likely of bad quality.
- MS/MS spectra with low total intensity or very few signal peaks should be rejected. Usually it is difficult to confidently identify the corresponding compounds.

It is usually not necessary to preprocess fragmentation spectra by removing "noise peaks" or recalibrating masses; such preprocessing can substantially worsen results, as signal peaks may be removed or masses shifted into the wrong direction. SIRIUS can decide for itself which of the peaks in the spectrum are noise, but it cannot recover the masses of accidentally removed signal peaks. To this end, be cautious when using intensity thresholds. If the data is noisy and necessitates "noise peak" removal, use a low intensity threshold to remove as few signal peaks as possible. Furthermore, we propose to use a low MS1 intensity threshold and not-to-restrictive parameters for feature detection. A high number of spurious features might pose a problem for MS1-only analysis. But here, we concentrate on metabolite identification based on fragmentation spectra, and spurious features can easily be recognized because these will not produce significant signal peaks within the fragmentation spectrum. Using liberal parameters will help to detect more low intensity isotope peaks and include them into the compound's isotope pattern.

Instrumental setup has huge impact on spectrum quality and some setups might be more suitable for structure elucidation with computational tools. See Tip 3 for more information.

Tip 1

Spectra quality. High quality spectra are indispensable to obtain good compound annotations. Spectra of high quality possess many signal peaks with intensities considerably above the noise level and mass errors of less than 10 ppm. On the other hand, few high-intensity signal peaks and mass errors of over 15 ppm indicate a spectrum of bad quality. It is understood that some molecules produce few fragments. But the information content of a spectrum increases with the number of (non-noise) peaks; identifying a compound from one peak is mere guessing. A proper instrumental setup can facilitate peak-rich spectra. Instead of using a single collision energy, spectra should be measured at multiple energies and merged. Alternatively, a ramped collision energy can be used to cover a large range of energies. In both cases, we expect to see more fragmentation peaks and, hence, better results.
Broad isolation windows favor chimeric spectra, being composed of fragments from more than one compound. Such chimeric spectra will interfere with fragmentation tree computation and also complicate the identification of structures via CSI:FingerID. In addition, broad isolation windows will result in isotope patterns for all fragments. Selecting only the monoisotopic peak for fragmentation makes it easier to interpret the fragmentation spectrum. SIRIUS provides an option to account for isotopes in the fragmentation spectrum, but this assumes that the isolation window is broad and isotope patterns of fragments are undisturbed. Unfortunately, filtering is imperfect in practice: An isolation window of width, say, 3 Da may select 100 % of the monoisotopic peak, 80 % or the first and 50 % of the second isotope peak. This will distort the isotope patterns of fragments in a non-trivial way. At present, SIRIUS cannot deal with distorted fragment isotopes patterns.
Compound identification benefits from choosing an instrumental setup which minimizes chimeric spectra, and favors peak-rich and low noise fragmentation spectra.

4 Metabolite identification

SIRIUS identifies metabolites in two steps: namely, molecular formula annotation and searching in a structure database. Both steps can be performed on a complete dataset using a single command; but users are advised to manually validate all results, including intermediate results. Here, we will explain the usage of SIRIUS step-by-step. For the sake of a more vivid description we will refer to the graphical user interface (GUI) of SIRIUS. All computations can be performed via the command line interface (CLI), using the GUI as a mere visualization tool for final results (see Section 5).
An overview of the SIRIUS GUI is displayed in Figure 1. The analysis starts with importing the data; this is done via the import dialog or drag-and-drop. SIRIUS imports spectra from .csv, .ms or .mgf files. Imported compounds are displayed in the compound list located in the left panel. To find specific compounds, use the search field above the panel. Start computations by clicking the *Compute All* button or by selecting a set of compounds and using the context menu (right-click). If only a single compound is selected, additional parameters can be specified such as the known molecular formula.

4.1 Molecular formula annotation

SIRIUS finds the most likely molecular formula by considering all possible molecular formulas, and is able to annotate biomolecules with a molecular formula missing from any database. Necessary parameters for SIRIUS are:

Elements Set of considered elements. Some elements can be auto-detected if an isotope pattern is given (see Tip 4.1).

ppm Allowed mass deviation in ppm. This is the maximum value a molecular formula explanation is allowed to deviate from the peaks' measured mass. Molecular formulas with a higher mass error are ignored. Note that for all peaks below 200 Da an absolute error is assumed which corresponds to the specified deviation in ppm at 200 Da.

Considered ion types Set of considered ion types. For details see Tip 4.1.

Candidates Number of candidates to be displayed. Fragmentation trees are computed for all molecular formula candidates using the Critical Path[3] heuristic from [23]. The top k fragmentation trees are recomputed using an exact algorithm; here, k corresponds to the number of displayed candidates plus 10. Hence, a larger number of displayed candidates increases running times.

Depending on the dataset, anticipated elements and ion types can be selected. Select a reasonable set of elements. The mass deviation is the maximum allowed deviation. Spectra measured on an instrument with advertised sub-ppm mass accuracy might still have much larger mass deviation (e.g. if not properly calibrated or because of

bad peak picking). More restrictive parameters, in particular for the allowed elements, can make computations substantially faster. Never select all uncommon elements at once. This will lead to a combinatorial explosion of potential molecular formulas; running times will increase dramatically; the number of correct molecular formula annotations will decrease. SIRIUS provides scoring profiles for Q-TOF and Orbitrap, which mainly change some background parameters. In case you are unsure if your data really has the instrument's advertised accuracy, use the default profile and set your allowed mass deviation accordingly.
Fragmentation trees are computed from a merged spectrum combining all input fragmentation spectra. Isotope pattern analysis is performed on a merged MS1 spectrum or using the isotope pattern provided by a preprocessing tool. A fragmentation spectrum which possesses peaks broadly distributed across the whole mass range presents more information to SIRIUS than a spectrum composed of either low or high mass peaks only.

Judging results Molecular formula annotation results are displayed in the *Sirius Overview* tab (see Figure 1). Candidates are ranked by the sum of isotope pattern and fragmentation tree score (see Tip 4.1 on isotopes and Tip 4.1 on fragmentation trees). Colored bars for each score ease comparison between candidates. Each candidate molecular formula has an adduct. At this stage, this is an ion type; after structure database search with CSI:FingerID this adduct corresponds to an adduct type (compare Figures 1 and 3 and see Tip 4.1).
The displayed attributes are:

Score Overall score by which candidates are ranked. This is the sum of isotope and tree score.

Isotope score Similarity score comparing the measured isotope pattern with the theoretical pattern for each candidate molecular formula. Usually, a score close to zero or low in comparison to the remaining candidates indicates an incorrect molecular formula, or at least an annotation of low confidence. Besides being the incorrect candidate, this might indicate improper data quality such as high intensity deviation or a low number of detected isotope peaks. The scored isotope pattern is highlighted in the *merged MS1* and can be assessed via the *Spectrum view* tab.

Tree score Score of the computed fragmentation tree.

Explained peaks The number of peaks in the spectrum which can be explained by the fragmentation tree. A high number of unexplained peaks indicates an incorrect annotation, a noisy spectrum, or two compounds being fragmented simultaneously.

Total explained intensity Summed relative intensity of all explainable peaks. Values of 95 % or higher indicate good quality; for values below 80 %, results should be interpreted with care.

Median absolute mass deviation The median absolute mass deviation of explained peaks in ppm. Low deviations are clearly desirable.

Selecting a molecular formula candidate displays the corresponding fragmentation tree and spectrum in which explained peaks are highlighted. The merged MS1 spectrum displays the selected isotope pattern. Mass errors of each fragment are shown to spot unlikely explanations; the displayed fragmentation tree can be colored accordingly. The user can inspect fragmentation tree annotations in varying degree of detail; individual fragments may support or contradict a particular molecular formula candidate. The user may decide by manual validation how well a candidate is supported.

Tip 2

Isotope pattern and element detection. Isotope patterns offer valuable information about elemental composition. The presence of uncommon elements that result in characteristic isotope pattern changes can be automatically detected [24]. Detectable elements are sulfur, chlorine, bromine, boron and selenium. When detected, SIRIUS adds these elements to the default set of elements CHNOP to determine the molecular formula. A predictor for silicon is disabled by default, as it results in a relatively large number of false positive predictions; the silicon isotope pattern is not "special" enough to permit a reliable auto-detection. In contrast to [24], the current version of SIRIUS uses a deep neural network for auto-detection of elements. Automated detection can be enabled or disabled via the compute dialog. Not considering elements which are extremely unlikely, substantially improves running times and may slightly improve results [24]. SIRIUS may still choose a molecular formula which *does not contain* an element with positive auto-detection, just as it might choose a molecular formula which does not contain any other enabled element. The final score of each molecular formula candidate is a combination of the fragmentation tree score and the isotope pattern score.
CAUTION: If no isotope pattern is provided and compounds are expected to contains elements beside CHNOPS, we strongly recommend to restrict molecular formulas to those from a molecular structure database. Do not select all uncommon elements for molecular formula annotation with SIRIUS. This will lead to a combinatorial explosion of potential molecular formulas; running times will increase dramatically.

Fig. 1. The *SIRIUS Overview* tab displays the spectrum and fragmentation trees of the top molecular formula candidates. The best candidate $C_{24}H_{38}O_3$ is selected; the corresponding explained spectrum and fragmentation tree are shown. The left panel contains a searchable list of all compounds; selected compounds are highlighted. The data and results of the first selected compound are displayed in all the views to the right of the compound list. The upper panel provides functionalities to import spectra, save and load workspaces, export result tables, start computations and display their status in the jobs panel. The *SIRIUS overview* tab displays various scores for each molecular formula candidate and can be sorted accordingly.

Tip 3

Fragmentation trees. A fragmentation tree annotates peaks in the fragmentation spectrum with molecular formulas and identifies likely losses between the fragments — similar to "fragmentation diagrams" created by experts. The calculated tree must not be understood as ground truth but can be used to derive information about the measured compound's fragmentation [25]. Fragmentation trees are also used to identify the molecular formula of an unknown compound. For every molecular formula candidate of the precursor ion, a separate fragmentation tree is computed which best explains the spectrum, as evaluated by a Maximum A Posteriori estimator [26]. This estimation takes into account information such as mass deviations, intensities, common losses and loss sizes. The overall best-scoring fragmentation tree corresponds to the most likely molecular formula explanation. In addition, CSI:FingerID uses the fragmentation tree to predict the compound's molecular fingerprint.

A simplified example of a fragmentation tree is presented in Figure 2. A fragmentation tree is computed from the fragmentation spectrum given the (candidate) molecular formula of the precursor ion. Initially, a fragmentation graph is constructed in the following way: For every fragment peak, all possible molecular formula explanations are computed. These explanations must be subformulas of the precursor molecular formula — a fragment only looses, but never gains new atoms. Every such molecular formula is a node in the graph. Nodes are connected by an edge if one node is a subformula of another node — this represents a potential loss. Using combinatorial optimization, the best scoring fragmentation tree is computed which explains every peak at most once. Unexplained peaks are considered noise.

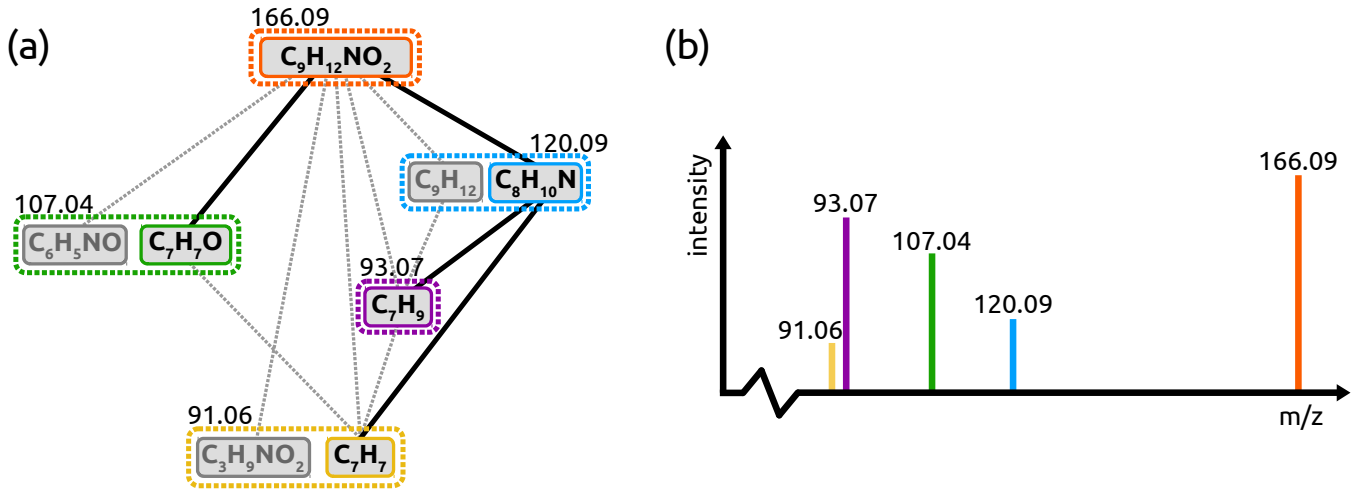

Fig. 2. Example of a fragmentation tree computed from a fragmentation graph in (a), given the spectrum in (b). The molecular formula of the neutral precursor is assumed to be $C_9H_{12}NO_2$. Molecular formulas are computed for all fragment peaks and serve as the nodes of the graph; nodes with the same color indicate molecular formulas corresponding to the same peaks. Nodes are connected by edges if one node is a subformula of another, thereby creating the fragmentation graph. A fragmentation tree is a connected subgraph which explains each color (peak) at most once and has no cycles. The best-scoring fragmentation tree, corresponding to a Maximum A Posteriori estimator, is computed by combinatorial optimization. The optimal fragmentation tree is indicated by solid lines; nodes which are not used are grayed out. These computations are repeated for each molecular formula candidate explaining the precursor mass, and the best such fragmentation tree is reported.

Tip 4

Ion and adduct types. SIRIUS differentiates between ion types and adduct types. Default ion types for positive ion mode spectra are protonation, sodium, and potassium; default ion types for negative ion mode spectra are deprotonation and chlorine. Adduct types can be seen as sub-types of an ion type. For example, the ion type protonation includes adduct types "intrinsically charged" ($[M]^+$), "protonated" ($[M + H]^+$), "protonated with water loss" ($[M - H_2O + H]^+$) and "ammonium group" ($[M + NH_4]^+$).
Adduct types cannot be determined from the fragmentation spectrum — the fragments $[C_4H_6O_2+NH_4]^+$ and $[C_4H_9NO_2+H]^+$ result in the exact same peak; and so will $[C_5H_7]^+$ and $[C_5H_8O - H_2O + H]^+$. That is why SIRIUS considers ion types, not adduct types, during the molecular formula annotation step. Multiple adduct types of the determined ion type can be considered for structure database search with CSI:FingerID (see Figure 3 and 4). When a specific ion type plus adduct type is provided by the user, it will be used during all computation steps. Users can specify additional ion and adduct types within the GUI or by modifying the config file.

Tip 5

Molecular fingerprint. A molecular fingerprint is a binary vector of fixed length where each position corresponds to a specific molecular property; for example, position #393 may encode the presence or absence of a benzene ring as a substructure. In general, a '1' indicates this specific substructure is present in the molecule, a '0' indicates it is not. There exist several types of fingerprints, such as PubChem CACTVS fingerprints[a], Klekota-Roth fingerprints [27], and MACCS fingerprints. Given a molecular structure, the corresponding fingerprint can be deterministically computed. Unfortunately, different structures can have the same molecular fingerprint.
Molecular fingerprints can be used to perform similarity search in a structure database. A common way to compare molecular structures using fingerprints is the Tanimoto similarity, also known as Jaccard index. Identical fingerprints produce a similarity of 1, whereas two structures not sharing a single molecular property have a Tanimoto of 0. Clearly, the similarity value depends on the choice of fingerprint type.
CSI:FingerID predicts a variety of molecular properties from several fingerprint types; only those molecular properties were selected which could also be predicted in evaluations. Given a spectrum and corresponding fragmentation tree, CSI:FingerID predicts a probabilistic fingerprint, see Sec. 4.3. This predicted fingerprint is compared to the deterministic fingerprints from a structure database to find the best match. The *CSI:FingerID Overview* tab also displays, for every structure candidate, the Tanimoto similarity against the predicted fingerprint. However, CSI:FingerID uses a different scoring function to rank candidates, which results in a larger number of correct identifications [12, 28].

[a] `ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf`

4.2 Searching in structure databases

After the molecular formula has been identified, the compound is searched in a structure database. Firstly, a molecular fingerprint of the query (see Tip 4.1) is predicted from the spectrum and fragmentation tree. Next, this predicted fingerprint is compared to (and scored against) fingerprints of structures in a database, to find the best matching structure. It must be understood that the molecular fingerprints of the candidate structures are fixed, known and independent of our tools.
To predict the molecular fingerprint, we have to know the molecular formula, ion type and adduct type of the query. By default, not only the top scoring molecular formula but multiple high-scoring molecular formula candidates are considered, applying a soft score threshold: All molecular formula candidates with a score above 0.75 of the optimal score are considered. To this end, we iterate over all possible combinations of molecular formula candidate and adduct type. The ion type of the query is determined by the molecular formula candidate; but various adducts types can be specified to search the database, see Tip 4.1 on ion and adduct types. When searching in the database, candidate structures must match the estimated molecular formula of the neutral molecule. Fragmentation trees of different adduct types differ as, say, a neutral loss is added to the top of the tree. These trees have exactly the same score. For each molecular formula and adduct type with candidate structures in the database, the resolved fragmentation tree is displayed in the *SIRIUS Overview* tab, see Figure 3. Scored structure candidates are displayed in the *CSI:FingerID Overview* tab. The *CSI:FingerID Details* tab allows to examine the scored structures in more detail for each molecular formula and adduct type separately (see Figure 4).

As a default, users should search compounds in the PubChem database, and filter results to the biocompound structure database or a subset thereof (see Tip 4.2). You may accept those query identifications for which there is a high-scoring structure candidate in the restricted database; potentially, this is even the highest-scoring candidate for all of PubChem. For those cases where no reasonable candidate was found in the biocompound structure database, and for cases where the best PubChem candidate scores substantially better than the best biocompound candidate, you can extend your search space to all of PubChem. Obviously, it makes much sense to *integrate biochemical background knowledge* at this point: This may be information about the organism the sample was taken from, or information about the biochemical preparation of the sample. Such meta information is not integrated into SIRIUS and CSI:FingerID, as this integration is highly non-trivial; but it is straightforward how to integrate the information manually.

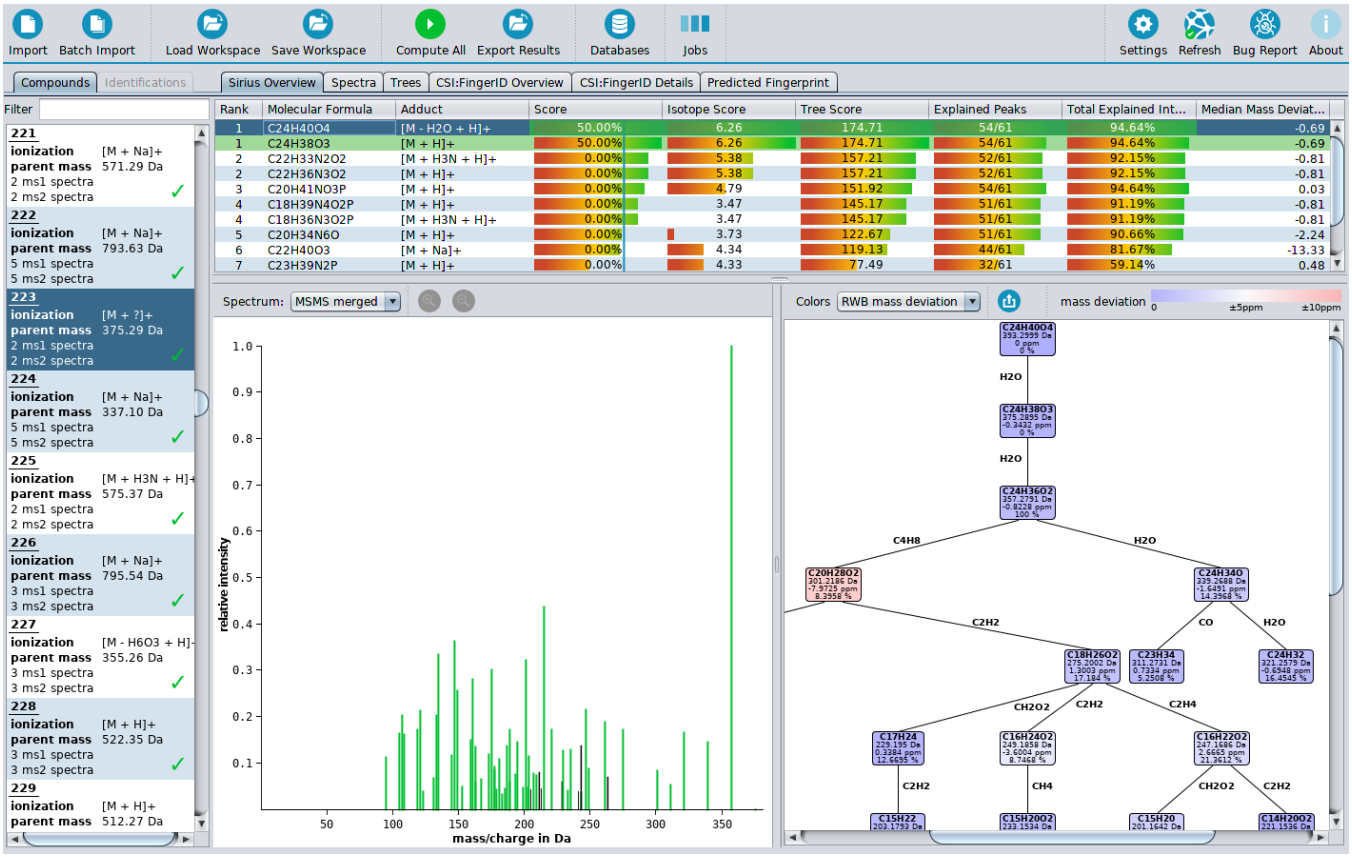

Fig. 3. Additional candidates are added to the *SIRIUS Overview* tab after searching with CSI:FingerID in a structure database considering adduct types $[M+H]^+$, $[M+NH_4]^+$ and $[M-H_2O+H]^+$. Molecular formulas $C_{24}H_{40}O_4$ and $C_{24}H_{38}O_3$ differ by an in-source loss of H_2O and are not distinguishable by MS/MS since in both cases, the ion $[C_{24}H_{38}O_3+H]^+$ is fragmented; hence, both have identical score. (The same holds for the pairs $C_{22}H_{33}N_2O_2$ vs. $C_{22}H_{36}N_3O_2$ and $C_{18}H_{39}N_4O_2P$ vs. $C_{18}H_{36}N_3O_2P$.) Displayed is the resolved fragmentation tree for $[C_{24}H_{40}O_4-H_2O+H]^+$, where an H_2O loss has been added to its top.

Judging results Users should check if the best structure candidate agrees with the best molecular formula candidate. Sometimes, CSI:FingerID decides that, based on its machine learning model and the given candidate structures, a structure with a different molecular formula better agrees with the data. Users should verify if the selected structure database does not contain any structures for the best-scoring molecular formula candidate; this can be an indication that the selected database is too restrictive. Besides, check if the correct adduct type has not been selected for database search.

CSI:FingerID ranks structure candidates by a logarithmic posterior probability [28], so that scores are negative numbers and zero is the optimum. Additionally, the predicted Tanimoto similarity is displayed. Since this is based on the predicted probabilistic fingerprint, this similarity usually underestimates the Tanimoto similarity

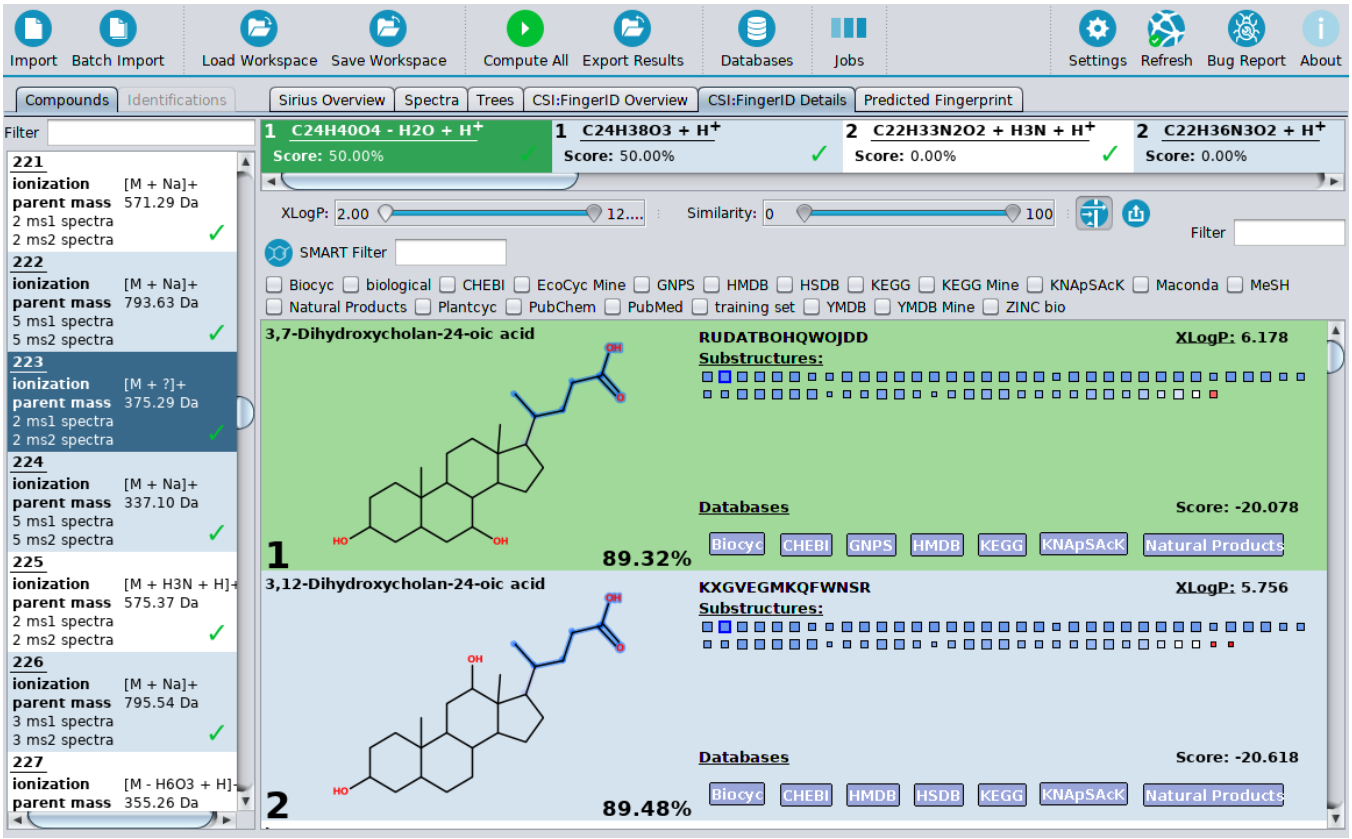

Fig. 4. The *CSI:FingerID Details* tab displays structure candidates for a selected molecular formula. The highlighted molecular property, which is predicted to be present in the query, is contained in the top 2 hits. Candidates are sorted by their score which is displayed on the right-hand side. Numbers in percent indicate the Tanimoto similarity between the predicted fingerprint and the fingerprint of each candidate. Candidates can be filtered by database, SMARTS string and XlogP value.

between the true fingerprints. Candidates can be filtered by database, XlogP values [29, 30] predicted using the Chemistry Development Kit [31, 32], or a specific SMARTS string. Structures are linked to database entries; clicking on the database icon opens the appropriate website. One CSI:FingerID candidate structure may link to several "3D structures" in a database, as CSI:FingerID ignores stereochemistry in its computations. The number of PubMed citations[1] is also displayed in the *CSI:FingerID Overview* tab. This value can contribute valuable information for the identification, for example as a sanity check. But on startup, these values must not be used to filter results: Doing so, we ignore the actual experimental data and potentially make our decisions based solely on prior knowledge [17].

The example in Figure 4 shows two top-scoring structure candidates. Both are structurally very similar and consequently, also have similar scores. The user may decide which structure is more likely, based on background knowledge about the sample. Comparing the, say, top 5 hits may also help to get an idea about a "core" structure which CSI:FingerID predicts to be present. Blue and red squares next to each candidate molecular structure represent its molecular properties. Blue properties are predicted to be present by CSI:FingerID and also present in the candidate; red properties are predicted to be absent but are present in the candidate. The size of the square represents the quality (F1 score, harmonic mean of precision and recall) of the predictor, as determined beforehand in cross validation; but a large F1 score does not guarantee that the prediction is correct for *this* query. In contrast, the saturation of the color indicates how sure CSI:FingerID is about the property, for this query. One specific property — a carboxyl group attached to a carbon chain — has been highlighted in Figure 4; it is present in the predicted fingerprint and in the first two candidates. A score close to zero and many blue squares usually indicates a confident identification — in this example, CSI:FingerID is very certain that the correct structure is at least very similar to the top hit. Even in case the best structure candidate is not correct, it is often structurally similar to the correct one and can help to elucidate the structure or answer the underlying biological question. Be warned that CSI:FingerID scores between different query compounds are usually not comparable; be cautious when using this score to differentiate between true and bogus identifications.

As explained in Section 4.1, users can also examine the fragmentation tree to decide how well a candidate is supported: For example, are specific side chains supported by fragments, losses or even fragmentation cascade in the fragmentation tree?

Tip 6

Some notes on database size. CSI:FingerID correctly identifies 39.4 % of CASMI 2016 positive ion mode spectra when searching in PubChem (in a structure-disjoint cross-validation setup). Searching in PubChem is difficult because it contains many millions of structures. If the search is performed in a database with 0.5 million structures of biological interest, correct identifications increase to 74.0 % [14]. To further increase identification rates, we might even be more restrictive and search in HMDB [2] or ChEBI [33]. Limiting CSI:FingerID search to the same structures which are contained in spectral libraries will even result in identification rates comparable to spectral library search! Does this mean it is advisable to search in a database with as few structures as possible? Clearly not! Results will look great in evaluation as long as all reference structures are contained in the restricted database. But in application, many compounds will be absent from the database, meaning you cannot find them at all.

Furthermore, there are — often ignored — side effects of searching in small databases. Firstly, the measured data becomes less important. You can easily identify a compound from one peak if you limit the candidate list to a few structures. Unfortunately, doing so does not increases the identification's confidence. It merely means that one candidate better matches the data compared with the other candidates, always assuming the correct structure is present in the candidate list. Second, incorrect identifications can be hard to spot, because they still "make sense": If all candidates in our database are frequently cited structures, then any identification (including the incorrect ones) will be a frequently cited structure and, hence, "reasonable".

Clearly, there is a trade-off between small and large databases. In a small database, many relevant biomolecules are missing. On the other hand, searching in PubChem decreases the number of correct identifications even though many PubChem structures are very unlikely to be actual biomolecules. CSI:FingerID provides a biocompound database with 0.5 million structures of biological interest, containing structures from ChEBI [33], KNApSAcK [34], HMDB [2], KEGG [35], HSDB [36], MaConDa [37], BioCyc [38], UNDP [39], a biological subset of ZINC [40], GNPS [1], MassBank [3] and MeSH-annotated PubChem compounds [16, 41]. In application, it is reasonable to search in this biocompound database, which is much smaller than PubChem, but still much more diverse than spectral libraries. For those queries where we find no reasonable explanation in the biocompound database, we can then consider the PubChem candidates.

[1] `https://www.ncbi.nlm.nih.gov/pubmed`

4.3 Beyond structure database search

It is understood that certain query biomolecules are not contained in any structure database. But even for such difficult instances, SIRIUS and CSI:FingerID can assist in structural elucidation. Recall that the SIRIUS molecular formula annotation step (Sec. 4.1) is done *de novo*. Hence, molecular formulas can be determined even for "novel compounds" absent from any structure database. Even if a structure is not contained in the structure databases, CSI:FingerID may find a very similar structure. Furthermore, CSI:FingerID allows the user to search in custom databases which may contain hypothetical structures, to identify "novel compounds".
But one key feature sets CSI:FingerID apart from other computational tools for structure elucidation: Predicting the molecular fingerprint of the query compound does not require any molecular structure database! The fingerprint is predicted from fragmentation spectrum and tree, and contains information about thousands of molecular properties. From that, we may draw conclusions what kind of substructures the query compound contains; and this information may be sufficient to decide if it is worth to further investigate the examined compound.

Judging results The predicted fingerprint is displayed in the *Predicted Fingerprint* tab, see Figure 5. Most molecular properties are described by SMARTS (SMiles ARbitrary Target Specification) strings[2]. SMARTS allows a flexible encoding of substructures; for example, a property might be described as "a methyl group bound to a hetero atom". Since SMARTS strings are usually hard to visualize, SIRIUS displays a set of example structures from the training data that have a particular molecular property.
A posterior probability is predicted for every molecular property. Estimates close to 1 indicate the property is likely being present in the query compound, whereas estimates close to 0 indicate it is not. But be careful: Since CSI:FingerID predicts thousands of properties, even some "rather certain predictions" must be wrong. A 98 % chance of being present also corresponds to 2 % chance of being absent; if 1000 molecular properties are predicted at this level of certainty, then 20 predictions are wrong. Also be reminded that these probabilities are *estimates*. To provide additional information on the quality of a prediction, the F1 score — a measure of the predictor quality — is displayed. The F1 score is the harmonic mean of precision (fraction of correct yes-predictions among all yes-predictions) and recall (fraction of correct yes-predictions among all yes-instances). A high F1 score indicates a good predictor, and 1.0 is the optimum. There is no general rule on what is a "good" F1 score; as a rule of thumb for this decision, one may assume that the F1 score equals precision and recall. Since many properties are rare and only present in few structures, the number of positive training examples is another indicator for the generalizability of the predictor. To help the user to concentrate on the most promising predictions, properties can be sorted by posterior probability, F1 score, or the number of atoms. The last option is useful to consider only larger, presumably more informative substructures.

5 Using SIRIUS in automated workflows

SIRIUS offers a powerful command-line interface (CLI) which allows for a flexible integration of SIRIUS into automated workflows. Technically speaking, the SIRIUS GUI is a visualization of the CLI functionality. Therefore, every task that can be done via the graphical user interface, can also be executed using the CLI. Corresponding to the two step approach in the GUI, the CLI provides self-contained sub tools for molecular formula identification (`sirius`) and structure elucidation (`fingerid`) with separate parameter sets.
Furthermore, CLI and GUI share the same input and output formats. Both, CLI and GUI store the computed results in the SIRIUS project-space (see Fig. 6) which in turn can also be an input for the GUI or the CLI. This allows the user to review results in the GUI that have been computed with an automated workflow using the CLI.

5.1 The SIRIUS project-space

The SIRIUS project-space is a standardized directory structure that is organized in a three hierarchy levels, namely, the *project level*, the *compound level* and the *method level* (see Fig. 6 for details).
On the *project level*, each compound corresponds to one sub-directory (*compound level*) storing the input data, parameters and results of the different analysis methods. These data is continuously written to the project-space, so that it represents the actual progress of a SIRIUS analysis. Further, the `.progress` file gives an overview about the progress of the ongoing analysis. On the *compound level*, each method provided by SIRIUS stores its results in its own sub-directory (*method level*). This allows the user to redo one analysis step without having

[2] `http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`

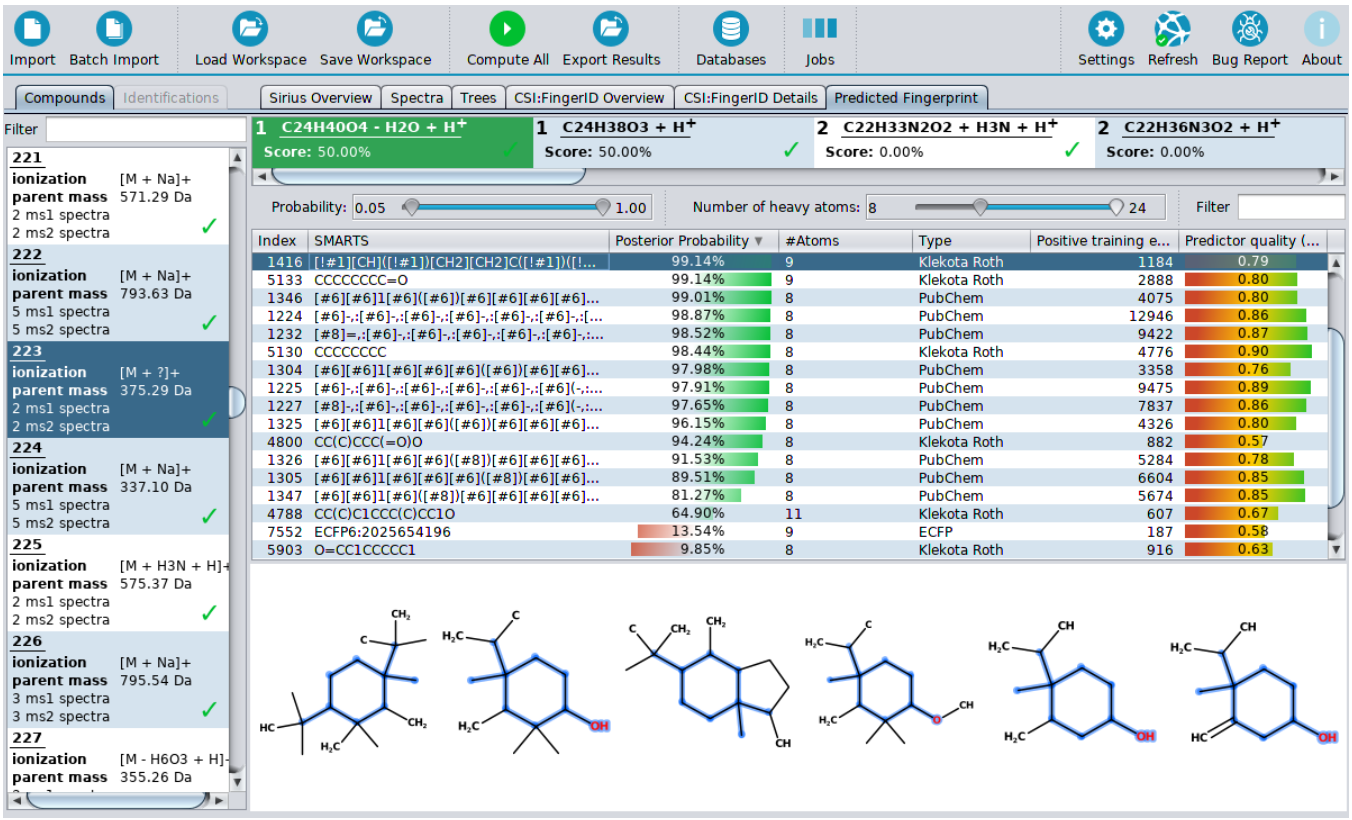

Fig. 5. The *Predicted Fingerprint* tab displays a predicted molecular fingerprint for each molecular formula candidate. The molecular fingerprint is predicted independently from any database. It can help deducing structural information on the compound even if the compound is not present in any structure database. Highlighted is a property mainly consisting of a ring. Training examples are displayed at the bottom. As shown, the oxygen is not a mandatory part of the substructure. The posterior probability of each property is also visualized as a color bar, to allow the user to swiftly distinguish properties predicted being present and absent. Green bars going to the right encode presence, red bars going the left encode absence.

to recompute the intermediate results it depends on. Further, SIRIUS is able to transfer intermediate results to a new project-space, so different parameters can easily be evaluated without having to recompute intermediate results. Since a project-space can be imported into the GUI, the user is able to judge intermediate results using the GUI before executing further analysis steps. Project-spaces can be read and written as an uncompressed directory or a compressed zip archive when using the `.sirius` file extension.
In addition to the *method level* results, the project-space contains summaries of these results on the *project level* and the *compound level*. These summaries are in *csv* format (`summary_<NAME>.csv`) to provide easy access to the results for further downstream analysis, data sharing and data visualization. The summaries are not imported into SIRIUS but are (re-)created based on the actual results every time a project-space is exported.

5.2 Standardized project-space summary with mzTab-M

The project-space is a SIRIUS-specific format that allows the user to access all results and analysis details, but may not be optimal for sharing this data with third party tools or data archives. For this purpose, SIRIUS provides an analysis report (`analysis_report.mztab`) in the standardized mzTab-M format [42]. All results summarized in this report are linked to the results in the corresponding SIRIUS project-space, allowing the user to share summarized results using mzTab-M without losing the connection to the detailed results provided in the project-space. Furthermore, SIRIUS passes meta information such as scan numbers and ids of the input data into this analysis report. This allows for an easy combination of the SIRIUS results with the results of other analyses such as MS1-based quantification.

6 Custom databases

Users may define their own structure databases to search in. These "custom databases" can be created via GUI and CLI. In the GUI, the *Databases* button opens a dialogue listing existing databases. New ones can be created with one click. Structures are imported by inserting structure descriptors (InChI or SMILES) into the import field; one structure per line. Custom databases are useful in case the users has a limited set of structures of interest. When screening for pollutants or drugs, a list of suspected structures can be collected in advance.
When searching with CSI:FingerID it does not matter if the structures in the database are known biomolecules or if these are hypothetical structures, which have not yet been discovered in any organism. Clearly, it is not reasonable to search in an arbitrarily large database. Databases of hypothetical structures have to be compiled with care to avoid combinatorial explosion. Available tools are BioTransformer [43] and the *in silico* generated MINE databases [19]. Currently, there exist MINE extensions for Ecocyc [44], YMDB[45] and KEGG [35]. But in principle, any existing structure database can be extended by such methods. Say, you are interested in finding new bile acids. A database of hypothetical bile acids can be created by applying biotransformations to known bile acids. This new database can then be searched with CSI:FingerID to find new bile acids synthesized by the investigated organisms.

7 Conclusion

To leverage the full potential of metabolomics, we need to overcome the limitations of spectral library search. This chapter presented concepts behind SIRIUS and CSI:FingerID, best-in-class computational tools for metabolite identification from high-resolution tandem mass spectra. We stress that computational tools currently cannot replace experts, but are meant to assist them. As a consequence, users must not accept identifications blindly but verify them properly. Here, we gave some advice on how this can be done.
SIRIUS ships with a command line tool which makes it easy to run computations on compute clusters and properly integrate it into automated workflows. Popular mass spectrometry data processing tools can create input files for SIRIUS, and SIRIUS outputs results in the standardized mzTab-M format to facilitate integration. The metabolomics community benefits from new computational tools, but tool development also benefits from the communities' input and more public training spectra. Finally, method development is an ongoing process, and SIRIUS is evolving to further improve metabolite identification.

References

1. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34,** 828–837 (2016).

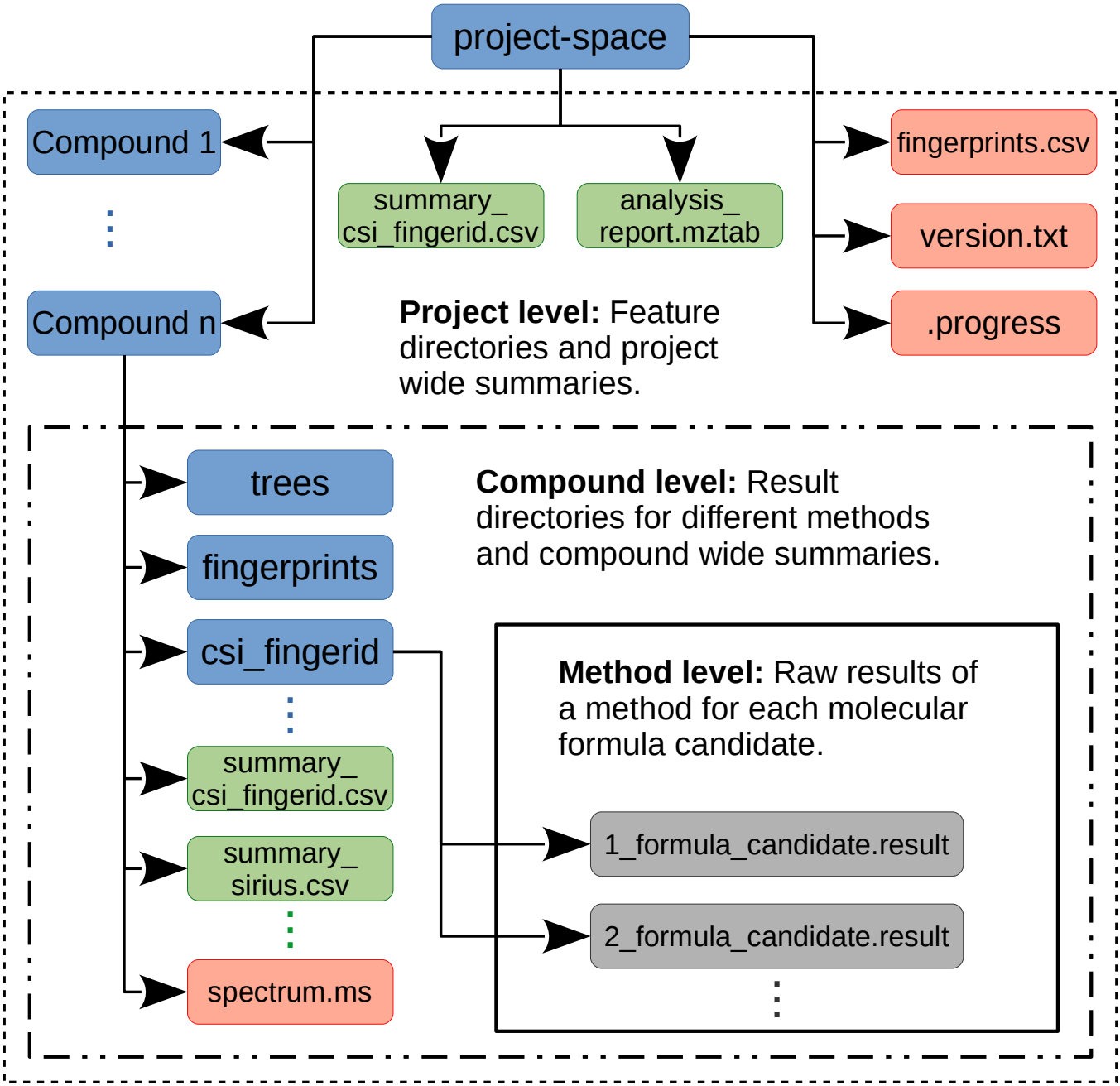

Fig. 6. The SIRIUS project-space is a standardized directory structure that stores results, summarized results, input data, parameters and version information of a SIRIUS analysis. It is organized on three levels, namely, the *project level* (dashed line), the *compound level* (dashed-and-dotted line) and the *method level* (solid line). The *compound level* contains sub-directories (blue) for each compound, summaries (green) about the whole dataset and additional information (red) about the version of SIRIUS that created the output. The *compound level* contains a sub-directory for each method that was applied to the compound as well as the summaries of these methods results. Further, it contains additional information, such as the input data and the parameters used for the computations. On the *method level*, SIRIUS stores the results of a specific method for a given compound (grey).

2. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46,** D608–D617 (2018).
3. Horai, H. *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **45,** 703–714 (2010).
4. Tautenhahn, R. *et al.* An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol* **30,** 826–828 (2012).
5. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11,** 98–110 (2015).
6. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **9,** 4035 (2018).
7. Ridder, L. *et al.* Automatic Chemical Structure Annotation of an LC-MS(n) Based Metabolic Profile from Green Tea. *Anal Chem* **85,** 6033–6040 (2013).
8. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf* **11,** 148 (2010).
9. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **8,** 3 (2016).
10. Wang, Y., Kora, G., Bowen, B. P. & Pan, C. MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Anal Chem* **86,** 9496–9503 (2014).
11. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* **88,** 7946–7958 (2016).
12. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* **112,** 12580–12585 (2015).
13. Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics* **28,** 2333–2341 (2012).
14. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* **16,** 299–302 (2019).
15. Schymanski, E. L. *et al.* Critical Assessment of Small Molecule Identification 2016: Automated Methods. *J Cheminf* **9,** 22 (2017).
16. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res* **44,** D1202–D1213 (2016).
17. Böcker, S. Searching molecular structure databases using tandem MS data: are we there yet? *Curr Opin Chem Biol* **36,** 1–6 (2017).
18. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* **112,** 12549–12550 (2015).
19. Jeffryes, J. G. *et al.* MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminform* **7,** 44 (2015).
20. Larson, E. A., Hutchinson, C. P. & Lee, Y. J. Gas Chromatography-Tandem Mass Spectrometry of Lignin Pyrolyzates with Dopant-Assisted Atmospheric Pressure Chemical Ionization and Molecular Structure Search with CSI:FingerID. *Journal of The American Society for Mass Spectrometry* **29,** 1908–1918 (2018).
21. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **13,** 741–748 (2016).
22. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf* **11,** 395 (2010).

23. Dührkop, K., Lataretu, M. A., White, W. T. J. & Böcker, S. *Heuristic algorithms for the Maximum Colorful Subtree problem* in *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2018)* **113** (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2018), 23:1–23:14.
24. Meusel, M. *et al.* Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal Chem* **88,** 7556–7566 (2016).
25. Rasche, F., Svatoš, A., Maddula, R. K., Böttcher, C. & Böcker, S. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem* **83,** 1243–1251 (2011).
26. Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *J Cheminform* **8,** 5 (2016).
27. Klekota, J. & Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24,** 2518–2525 (2008).
28. Ludwig, M., Dührkop, K. & Böcker, S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics* **34.** Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2018)., i333–i340 (2018).
29. Wang, R., Fu, Y. & Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J Chem Inf Comput Sci* **37,** 615–621. eprint: `http://dx.doi.org/10.1021/ci960169p` (1997).
30. Wang, R., Gao, Y. & Lai, L. Calculating partition coefficient by atom-additive method. *Perspect Drug Discovery Des* **19,** 47–66 (2000).
31. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **43,** 493–500 (2003).
32. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminf* **9,** 33 (2017).
33. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44,** D1214–9 (2016).
34. Shinbo, Y. *et al.* in *Plant Metabolomics* (eds Saito, K., Dixon, R. A. & Willmitzer, L.) 165–181 (Springer-Verlag, 2006).
35. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44,** D457–D462 (2016).
36. Fonger, G. C., Hakkinen, P., Jordan, S. & Publicker, S. The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): background, recent enhancements and future plans. *Toxicology* **325,** 209–216 (2014).
37. Weber, R. J. M., Li, E., Bruty, J., He, S. & Viant, M. R. MaConDa: A publicly accessible mass spectrometry contaminants database. *Bioinformatics* **28,** 2856–2857 (2012).
38. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42,** D459–D471. eprint: `http://nar.oxfordjournals.org/content/42/D1/D459.full.pdf+html` (2014).
39. Gu, J. *et al.* Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* **8,** 1–10 (2013).
40. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **52,** 1757–1768 (2012).
41. Nelson, S. J., Johnston, W. D. & Humphreys, B. L. in *Relationships in the organization of knowledge* (eds Bean, C. A. & Green, R.) 171–184 (Kluwer Academic Publishers, 2001).
42. Hoffmann, N. *et al.* mzTab-M: A data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal Chem* **91,** 3302–3310 (2019).

43. Djoumbou-Feunang, Y. *et al.* BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminf* **11,** 2 (2019).
44. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res* **45,** D543–D550 (2017).
45. Ramirez-Gaona, M. *et al.* YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res* **45,** D440–D445 (D1 2017).