Searching molecular structure databases with tandem mass spectra using CSI:FingerID

Kai Dührkop[a], Huibin Shen[b], Marvin Meusel[a], Juho Rousu[b], and Sebastian Böcker[a,1]

[a]Chair for Bioinformatics, Friedrich Schiller University, 07743 Jena, Germany; and [b]Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, 02150 Espoo, Finland

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved August 11, 2015 (received for review May 19, 2015)

Metabolites provide a direct functional signature of cellular state. Untargeted metabolomics experiments usually rely on tandem MS to identify the thousands of compounds in a biological sample. Today, the vast majority of metabolites remain unknown. We present a method for searching molecular structure databases using tandem MS data of small molecules. Our method computes a fragmentation tree that best explains the fragmentation spectrum of an unknown molecule. We use the fragmentation tree to predict the molecular structure fingerprint of the unknown compound using machine learning. This fingerprint is then used to search a molecular structure database such as PubChem. Our method is shown to improve on the competing methods for computational metabolite identification by a considerable margin.

mass spectrometry | small compound identification | metabolomics | bioinformatics | machine learning

Metabolites, small molecules that are involved in cellular reactions, can provide detailed information about cellular state. Untargeted metabolomic studies may use NMR or MS technologies, but liquid chromatography followed by MS (LC/MS) can detect the highest number of metabolites from minimal amounts of sample (1, 2). Untargeted metabolomics comprehensively compares the mass spectral intensities of metabolite signals (peaks) between two or more samples (3, 4). Advances in MS instrumentation allow us to simultaneously detect thousands of metabolites in a biological sample. Identification of these compounds relies on tandem MS (MS/MS) data, produced by fragmenting the compound and recording the masses of the fragments. Structural elucidation remains a challenging problem, in particular for compounds that cannot be found in any spectral library (1): In total, all available spectral MS/MS libraries of pure chemical standards cover fewer than 20,000 compounds (5). Growth of spectral libraries is limited by the unavailability of pure reference standards for many compounds.

In contrast, molecular structure databases such as PubChem (6) and ChemSpider (7) contain millions of compounds, with PubChem alone having surpassed 50 million entries. Searching in molecular structure databases using MS/MS data is therefore considered a powerful tool for assisting an expert in the elucidation of a compound. This problem is considerably harder than the fundamental analysis step in the shotgun proteomics workflow, namely, searching peptide MS/MS data in a peptide sequence database (8): Unlike proteins and peptides, metabolites show a large structural variability and, consequently, also large variations in MS/MS fragmentation. Computational approaches for interpreting and predicting MS/MS data of small molecules date back to the 1960s (9): Due to the unavailability of molecular structure databases at that time, structure libraries were combinatorially generated and then "searched" using the experimental MS/MS data. "Modern" methods for this question have been developed since mid-2000. Particular progress has been made for restricted metabolite classes such as lipids (5), but as with peptides, results cannot be generalized to other metabolite classes. For the general case, several strategies have been proposed during recent years, including simulation of mass spectra from molecular structure (10, 11), combinatorial fragmentation (12–17), and prediction of molecular fingerprints (18, 19).

Searching in a molecular structure database is clearly limited to those compounds present in the database. Fragmentation trees have been introduced as a means of analyzing MS/MS data without the need of any (structural or spectral) database (20–22). In this paper, the term "fragmentation tree" is exclusively used to refer to the graph-theoretical concept introduced in ref. 20, not "spectral trees" that describe the dependencies of multiple MS measurements; see Vaniya and Fiehn (23) for a review. In more detail, our fragmentation trees are predicted from MS/MS data by an automated computational method such that peaks in the MS/MS spectrum are annotated with molecular formulas of the corresponding fragments, and fragments are connected via assumed losses. Clearly, there exist other approaches with the broad aim of identifying metabolites, such as network-based methods (24–26) and combined approaches (27); see Hufsky et al. (28) for a review of computational methods in MS-based metabolite identification.

It is undisputed that MS/MS data alone are insufficient for full structural elucidation of metabolites. We argue that elucidation of stereochemistry is currently beyond the power of automated search engines, so we try to recover the correct constitution (bond structure) of the query molecule, that is, the identity and connectivity (with bond multiplicities) of the atoms, but no stereochemistry information. Throughout this paper, we refer to the constitution of the molecule as its structure. In practice, orthogonal information is usually available, both analytical (retention time, ion mobility drift time, infrared and UV spectroscopy, and NMR data) and on the experimental setup (extraction procedure and organism) (29, 30).

Significance

Untargeted metabolomics experiments usually rely on tandem MS (MS/MS) to identify the thousands of compounds in a biological sample. Today, the vast majority of metabolites remain unknown. Recently, several computational approaches were presented for searching molecular structure databases using MS/MS data. Here, we present CSI:FingerID, which combines fragmentation tree computation and machine learning. An in-depth evaluation on two large-scale datasets shows that our method can find 150% more correct identifications than the second-best search method. In comparison with the two runner-up methods, CSI:FingerID reaches 5.4-fold more unique identifications. We also present evaluations indicating that the performance of our method will further improve when more training data become available. CSI:FingerID is publicly available at www.csi-fingerid.org.

Author contributions: J.R. and S.B. designed research; K.D., H.S., and M.M. performed research; K.D., H.S., J.R., and S.B. developed the method for fingerprint prediction; M.M. and S.B. developed fingerprint scoring functions; K.D., H.S., and M.M. analyzed data; and S.B. wrote the paper.

Conflict of interest statement: S.B. holds a patent on comparing fragmentation trees, whose value might be influenced by the manuscript.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 12549.

[1]To whom correspondence should be addressed. Email: sebastian.boecker@uni-jena.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1509788112/-/DCSupplemental.

We assume that this information is not presented to the search engines but rather used in a postprocessing step to manually select the best solution from the output list of the engine. This is comparable to the everyday use of search engines for the internet.

Here, we present CSI (Compound Structure Identification): FingerID for searching a molecular structure database using MS/MS data. Our method combines computation and comparison of fragmentation trees with machine learning techniques for the prediction of molecular properties of the unknown compound (19). Our method shows significantly increased identification rates compared with all existing state-of-the-art methods for the problem. CSI:FingerID is available at www.csi-fingerid.org/. Our method can expedite the identification of metabolites in an untargeted workflow for the numerous cases where no reference measurements are available in spectral libraries.

Results

Methods Overview. Recently, we used fragmentation trees to boost the performance of molecular fingerprint prediction using multiple kernel learning (19). Here, we further combine this method with a kernel encoding chemical elements, a kernel based on recalibrated MS/MS data, five additional kernels based on fragmentation tree similarity, and two pseudokernels based on fragmentation tree alignments (31). We then add PubChem (CACTVS) fingerprints (881 molecular properties) and Klekota–Roth fingerprints (32) (4,860 molecular properties) to the pool of predictable fingerprints. This results in 1,415 molecular properties that can be learned from the data; we will refer to these molecular properties as the fingerprint of a molecular structure. Finally, we use maximum likelihood considerations and Platt probabilities to refine the fingerprint similarity scoring.

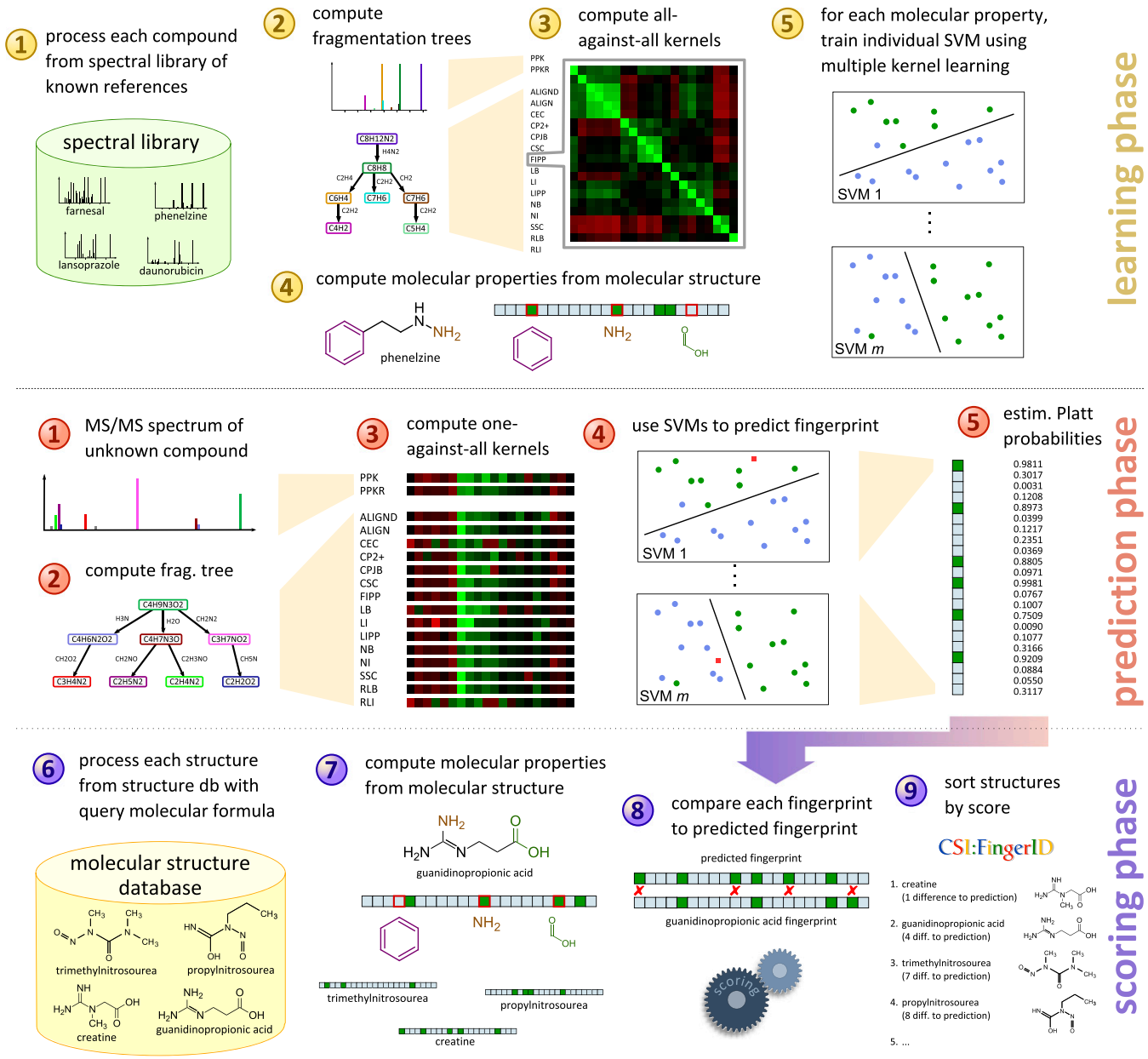

Fig. 1. Workflow of our method CSI:FingerID. During the learning phase, we use MS/MS reference data to train a set of predictors for molecular properties (the fingerprint). In the prediction phase we use MS/MS data of an unknown compound to find a fragmentation tree and to predict the fingerprint of the unknown. In the scoring phase we compare the predicted fingerprint of the unknown to fingerprints of molecular structures in a structure database, searching for a best match. See *Materials and Methods* for details.

Our method can roughly be divided into three phases (Fig. 1): one phase for training the method on some reference set of known compounds and two phases for identifying unknown compounds. In all cases, MS/MS spectra of each compound are first transformed into a fragmentation tree by the automated method described in refs. 22 and 33, and both the MS/MS spectra and the fragmentation tree of the compound are used as input for the subsequent analysis. In the learning phase, we use a database of reference compounds with known molecular structure. The used machine learning method falls into the class of kernel methods (34), where a kernel denotes a similarity measure for either MS/MS spectra or fragmentation trees. We compute several such similarity measures for each pair of compounds in the reference data and also determine weights to combine all similarity measures into one (multiple kernel learning) (19). In addition, we compute the molecular fingerprint of each reference compound using its known structure. For each molecular property in the fingerprint, we then train a support vector machine (35) (SVM) that, using the kernel similarities, tries to separate compounds into those that exhibit the molecular property and those that do not. Platt probabilities (36) allow for a more fine-grained prediction, replacing the 0/1 predictions of classical SVM by some (posterior) probability for the presence of the molecular property.

The second part, where we want to find an unknown compound in a database of molecular structures, consists of two phases. In the prediction phase, we are given the MS/MS spectra of an unknown compound. We compute kernel similarities of the unknown compound against all compounds in the reference dataset, based on MS/MS spectra and fragmentation trees. We then use SVMs trained above to predict the (probability of the) presence or absence of each molecular property for the unknown compound. This results in a predicted fingerprint of the unknown compound. In the scoring phase, we compare the predicted fingerprint of the unknown compound against fingerprints of compounds in a molecular structure database such as PubChem. For each candidate molecular structure, its fingerprint is scored against the predicted fingerprint; candidate structures are sorted with respect to this score and reported back to the user. We stress that the unknown compound is usually not part of the training data; in our evaluation below, we make sure that this is never the case, using cross-validation.

Identification Quality. We first evaluate each method using compounds from the combined Agilent and GNPS (Global Natural Products Social) dataset. Our method strongly outperforms all other available tools for searching MS/MS data in a molecular structure database (Fig. 2). Compared with the runner-up, FingerID, the number of correct identifications is 2.5-fold higher (34.4% vs. 13.8%) when searching PubChem. CFM-ID reaches third place with 13.2% identification. We achieve 63.5% correct identifications in the top five output; next come FingerID with 36.1% and CFM-ID with 36.0%. Our method reaches an identification rate of 50% at the fractional rank 2.23, far ahead of CFM-ID (fractional rank 13.5) and MAGMa (13.7). It reaches an identification rate of 66.7% for fractional rank 6.38, again far ahead of CFM-ID (50.0) and MAGMa (51.0). See Fig. S1 for identification rates for all ranks. Searching biocompounds in the biodatabase, we achieve 68.5% correct identifications, compared with 59.5% and 57.4% for the two next-best methods, MAGMa and CFM-ID, respectively (Fig. 2). For 92.3% of the query compounds, the correct answer is contained in the top five for our method, compared with runners-up FingerID (86.1%) and CFM-ID (84.2%). Searching the complete combined dataset in the biodatabase, this corresponds to 55.2% correct identifications for our method (Fig. S1).

The Agilent dataset is proprietary and, hence, cannot be used to evaluate future methods. We therefore repeated our analysis, this time searching with query instances from the two datasets individually (Figs. S2–S4). For the Agilent dataset, we reach 39.3% correct identifications, compared with 19.6% for FingerID and 15.3% for MAGMa. For the GNPS dataset, all methods suffer a slight loss in identification quality, but trends are highly similar to those reported above. For example, identification rates when searching PubChem decrease to 31.8% for our method, 12.1% for CFM-ID, and 11.8% for MAGMa, making the identification rate for our method 2.6-fold higher than for the runner-up. Our method achieves an identification rate of 50% for fractional rank 2.59, with the next-best being MAGMa (fractional rank 15.5) and CFM-ID (17.5). Our method reaches 66.7% identifications for fractional rank 7.62, compared with MetFrag (58.5) and MAGMa (63.1). When searching biocompounds in the biodatabase we reach 64.3% correct identifications, compared with 56.5% for MAGMa; the correct answer is in the top five for 90.2%, with runner-up FingerID (81.1%).

We also evaluated against the baseline method of randomly ordering candidates with the correct molecular formula. Random ordering performs well for searching biocompounds in the biodatabase, with 30.7% correct identifications and 64.3% in the top five for the combined dataset. This demonstrates the power of knowing the correct molecular formula for structure elucidation when searching in a restricted structure database.

Next, we compare methods for each individual query instance. See Fig. 3 for the overlap in identifications of our method, CFM-ID, and MAGMa; our method reaches 5.4-fold more unique identifications (correctly identified compounds not identified by one of the other two methods) than CFM-ID and MAGMa for the combined dataset. A method outperforms another for some query instance if it places the correct structure on a better rank. On the combined dataset, our method outperforms CFM-ID for 65.6% of the instances, whereas CFM-ID outperforms our method for 23.8% of them. Our method outperforms MAGMa for 66.1% and is outperformed for 24.1% of the instances. For MIDAS, MetFrag, and FingerID, our method outperforms each of these methods for more than 68.9% of the instances and is outperformed for at most 20.6%. In all cases, the significance (sign test P value) is below 10^{-127}. See Table S1 for an all-against-all comparison of methods. In Fig. S5 we

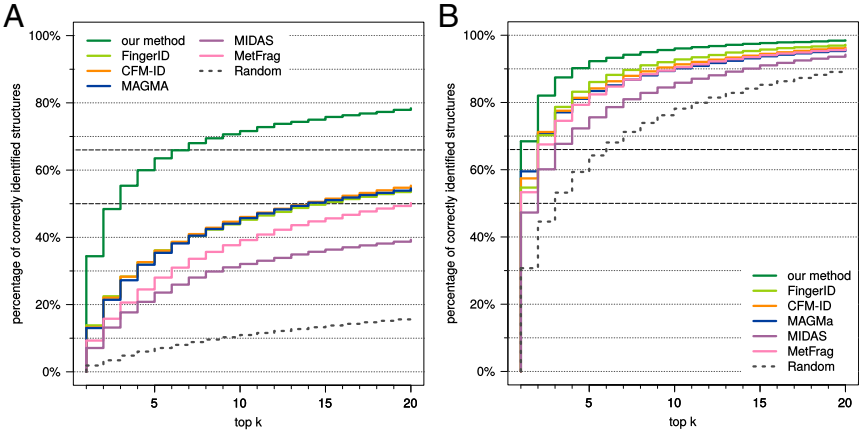

Fig. 2. Methods evaluation: percentage of correctly identified structures found in the top k output of the different methods, for maximum rank $k = 1, \ldots, 20$. Searching $N = 5{,}923$ compounds from the combined Agilent and GNPS dataset in PubChem (*A*) and the $N = 4{,}773$ biocompounds from the combined dataset in the biodatabase (*B*). Identification rates 50% and 66.7% marked by dashed lines.

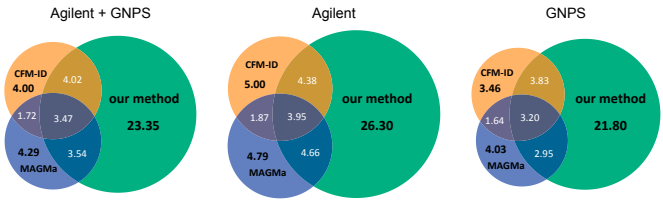

Fig. 3. Venn diagram, percentages of correct identifications of our method, CFM-ID, and MAGMa. (*Left*) Searching $N = 5{,}923$ compounds from Agilent and GNPS in PubChem. Here, 44.4% of the compounds were identified by at least one of the three methods. (*Middle*) Searching with $N = 2{,}055$ compounds from Agilent in PubChem, with 51.0% identified by at least one method. (*Right*) Searching with $N = 3{,}868$ compounds from GNPS in PubChem, with 40.9% identified by at least one method.

show nine exemplar compounds that were correctly identified by our method, but not by any other method in this evaluation.

We evaluated several new scoring functions for CSI:FingerID and found three that perform better than the function proposed in refs. 18 and 19 (Fig. 4). Among these, "modified Platt" achieves the best identification rates and was therefore selected as the default scoring function for CSI:FingerID. Compared with the original scoring function based on predictor accuracy (18, 19), we reach 17.0% (4.99 percentage points) more correct identifications.

Because our method employs machine learning to predict molecular properties of the query compound, additional data can improve the performance of the predictors. To estimate the scale of this effect, we repeated the learning step for all predictors but this time presented the method with only a fraction of the data for learning. Average accuracy and F1 score of the predictors, as well as their performance for searching PubChem, are shown in Fig. 5. Varying the amount of training data, a monotonic increase is observed in accuracy and F1 score of the molecular property predictions. This growth does not saturate with the amount of data available in our experiments. For the resulting identification rates, we observe an almost linear increase when varying relative training data size between 40% and 90%, whereby 400 additional compounds in the training data result in an increase of roughly one percentage point in the identification rate.

We found that the performance of our method is influenced by more than just the amount of available training data: To a great extent, our method depends on "useful" molecular properties it can predict. Whether or not a particular molecular property is "useful" is determined by a multitude of parameters, such as discriminating power in PubChem (a molecular property that is inherited by the vast majority of compounds in PubChem is of little use in filtering out wrong candidates) or availability of training data for both the presence and absence of the property. To estimate how additional molecular properties can further improve the power of our method in the future, we artificially restricted the properties available for prediction and evaluated the method for the reduced sets of molecular properties (Fig. 5). We find that increasing the available molecular properties causes a monotonic, logarithm-like increase in the identification rate.

Finally, we evaluated all methods on an independent dataset from MassBank (Fig. S6). For this dataset, our method reaches 39.5% correct identifications searching PubChem, compared with 19.0 and 5.77% for the runners-up FingerID and MIDAS. For 267 compounds, we find corresponding structures in the training data; our method correctly identifies more than two-thirds (68.8%) of these compounds. We observe a major drop in identification accuracy for all methods but ours and FingerID on the complete MassBank dataset, and a similar behavior for all methods on the "novel" compounds: Searching for the 358 "novel" compounds in PubChem, our method reaches 17.7% correct identifications, followed by FingerID (8.34%) and MetFrag (5.70%).

Discussion

When searching a molecular structure database using MS/MS data, CSI:FingerID achieves significantly better results than existing state-of-the-art methods. We observe a 2.5-fold increase of correct identifications compared with the runner-up method when searching PubChem and a 6.0- to 7.8-fold fractional rank decrease when trying to recover the correct solution for 50% or 66.7% of the instances, respectively.

It must be understood that finding the correct molecular structure in a molecular structure database as enormous as PubChem, being four orders of magnitude larger than existing MS/MS libraries, is highly challenging and will never be possible without a certain fraction of bogus identifications. We have deliberately left it to the expertise of the user to select the best molecular structure from the suggested candidates. Additional information such as citation frequencies or "number of PubChem substances" (16) can further assist the user in identifying the most promising candidates. We did not use such information in our evaluation to avoid overestimating the method's power: Spectral libraries mostly contain well-described compounds where pure reference standards are available, and such compounds also receive many citations and have many PubChem substance entries.

For the independent dataset, our method shows good identification performance, but for the 358 "novel" compounds where no corresponding structure is present in the training data, we observe a severe drop in identification rates for all methods. We manually inspected the MS/MS data but detected no peculiarities. Currently, we cannot convincingly explain the drop of identification rates, despite testing numerous possible explanations such as number of candidates of an instance, or structural similarity of candidates to the true solution. The only peculiarity we found is distinctively reduced identification rates for flavonoid compounds in the "Washington" subdataset.

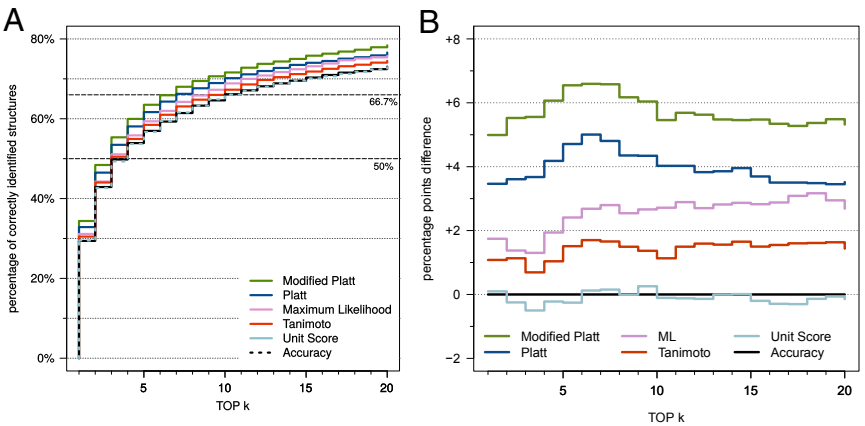

Fig. 4. (*A*) Evaluation of different scoring functions for our method. Searching with $N = 5{,}923$ compounds from the combined Agilent and GNPS dataset. See *Materials and Methods* for a description of the different scoring functions. (*B*) Difference in percentage points to the "accuracy" scoring function from refs. 18 and 19.

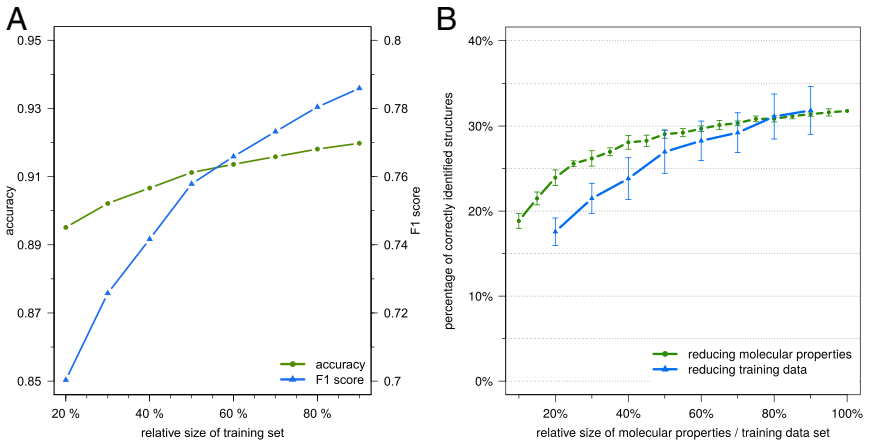

Fig. 5. Effects of training dataset size and of reducing the available molecular properties. (*A*) Average accuracy (green) and F1 score (blue) when training predictors on 20–90% of the combined training dataset (6,258 compounds), where 90% corresponds to the original 10-fold cross-validation. Note the different scales. From the 2,765 molecular properties that can be learned from the data, we use only those 712 for which the smaller class in the complete training dataset contains at least 5% of the structures. (*B*) Percentage of correctly identified structures when only a part of the training data (20–90%, blue) or only a part of the 2,765 molecular properties (10–100%, green) are available for prediction. Searching with $N = 3{,}868$ compounds from GNPS in PubChem. SD calculated from 10 replicates (molecular property plot) or 10-fold cross-validation (training data plot).

Running times of FingerID are fastest, whereas our approach, MetFrag, and MAGMa are roughly on par; in contrast, those of CFM-ID and MIDAS are two orders of magnitude higher (Table S2). This is not a problem for CFM-ID because spectrum simulation is done only once for each molecular structure during preprocessing, whereas comparison of spectra is very fast, but it can severely hinder the use of MIDAS in practice.

We found that choosing the correct cross-validation setup has a huge impact on our evaluation: If we choose cross-validation batches solely based on the individual measurements of the compounds, ignoring that two batches may contain the same structure, then our identification rate for searching in PubChem increases to a staggering 58.5% (Fig. S7). Manual inspection confirmed that different compounds with identical structure (constitution) often show highly similar MS/MS data.

Molecular structure databases keep growing at a pace beyond synthesizing capacities. To this end, CSI:FingerID and other methods for searching in molecular structure databases represent a paradigm shift in the metabolomics field. Clearly, CSI:FingerID can and should be combined with other search engines. In particular, it should be accompanied by a search in spectral libraries (37) such as GNPS itself, and alternative methods for structural elucidation (24–26). In cases where the class of the query compound is known, more specialized approaches may be available, such as LipidBlast for lipids (5), or database searching and de novo sequencing for small peptides (38).

Compared with the original FingerID method of 2012 (18), identification rates of our method are 2.5-fold higher. With this and our experiments on limiting training data and molecular properties (Fig. 5), we predict that CSI:FingerID will reach even better identification rates in the near future. Our method can open up new paths beyond searching in structure databases such as PubChem: An obvious next step will be to search in structure databases containing hypothetical compounds (39, 40), potentially allowing us to overcome the limits of molecular structure databases.

Materials and Methods

For training the method, we use a set of 4,138 small compounds from the public GNPS Public Spectral Libraries (https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp) and 2,120 compounds from the MassHunter Forensics/Toxicology PCDL library (Agilent Technologies, Inc.). Evaluation is carried out by 10-fold cross-validation, such that no two batches may contain the same structure. We evaluate all methods using 3,868 compounds from GNPS and 2,055 compounds from the Agilent dataset. We present results for the combined Agilent and GNPS dataset, and for the two datasets individually. As an independent dataset, we use MS/MS data of 625 compounds from MassBank (41). We search a version of PubChem (downloaded on September 15, 2014) containing 52,926,405 compounds and 40,805,940 structures, and a filtered version of PubChem (biodatabase; see Table S3 and Fig. S8) containing 268,633 structures of biological interest (about 300,000 compounds); 1,010 compounds from the GNPS dataset and 140 compounds from the Agilent dataset cannot be found in the biodatabase. We refer to the remaining $2{,}858 + 1{,}915 = 4{,}773$ compounds as biocompounds. Searching the biodatabase was performed using both the complete datasets and the subsets of biocompounds.

We assume that we are able to identify up front the molecular formula of the unknown compound. For this, several approaches have been developed that analyze the MS/MS and isotope pattern data of the compound (42); for example, CFM-ID (10) identified the correct molecular formula for more than 90% of 1,491 nonpeptide metabolites using MS/MS data, and SIRIUS (43) was able to find the correct molecular for 10 out of 12 instances of the CASMI (Critical Assessment of Small Molecule Identification) 2013 contest. For all evaluated tools, molecular structure candidates are extracted from PubChem using the known molecular formula of the query.

We evaluate our method against the original FingerID method (18), CFM-ID (10), MAGMa (16), MIDAS (15), and MetFrag (14). FingerID was retrained on the combined training data, to enable a sensible evaluation against its successor presented here. CFM-ID also uses machine learning techniques but was not retrained on the new dataset due to computational limitations, resulting in an overlap (972 structures) between training and evaluation set. Identification rates reported here are slightly better than to those reported in ref. 10 using cross-validation. Average running times per query range from 2 s (FingerID) to more than 1 d (MIDAS) (Table S2). To avoid proliferating running times, MIDAS was stopped after 24 h of computation (more than 10 times the estimated average running time of any other program) for any instance. If the output of a tool did not contain the correct candidate, then all candidates not in the output were added to the end of the output list with identical, minimal score. Similarly, if a tool was unable to process an instance, then all candidates received identical score.

Lists are sorted with respect to scores provided by each tool. Ties in the score of a method are broken randomly, comparable to adding weak random noise to the scores. A given query instance is correctly identified if the correct structure is at the top position of the output list; it is in the top k if its rank in the output list is at most k.

See *SI Materials and Methods* for details.

ACKNOWLEDGMENTS. We thank Agilent Technologies, Inc. for providing uncorrected peak lists of their spectral library and Lars Ridder (Wageningen University) for the MAGMa software. We are particularly grateful to Pieter Dorrestein and Nuno Bandeira (University of California) and the GNPS (Global Natural Products Social) community for making their data publicly accessible. This work was funded in part by Deutsche Forschungsgemeinschaft Grant BO 1910/16 (to K.D. and M.M.) and Academy of Finland Grant 268874/MIDAS (to H.S. and J.R.).

1. Patti GJ, Yanes O, Siuzdak G (2012) Innovation: Metabolomics: The apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 13(4):263–269.
2. Baker M (2011) Metabolomics: From small molecules to big ideas. *Nat Methods* 8(2): 117–121.
3. Rinehart D, et al. (2014) Metabolomic data streaming for biology-dependent data acquisition. *Nat Biotechnol* 32(6):524–527.
4. Kenar E, et al. (2014) Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol Cell Proteomics* 13(1):348–359.
5. Kind T, et al. (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat Methods* 10(8):755–758.
6. NCBI Resource Coordinators (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 41(Database issue):D8–D20.

7. Williams A, Tkachenko V (2014) The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. *J Comput Aided Mol Des* 28(10): 1023–1030.
8. Altelaar AFM, Munoz J, Heck AJR (2013) Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nat Rev Genet* 14(1):35–48.
9. Lindsay R, Buchanan B, Feigenbaum E, Lederberg J (1980) *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project* (McGraw-Hill, New York).
10. Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11(1):98–110.
11. Allen F, Pon A, Wilson M, Greiner R, Wishart D (2014) CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* 42(Web Server issue):W94-9.
12. Hill AW, Mortishire-Smith RJ (2005) Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun Mass Spectrom* 19(21):3111–3118.
13. Heinonen M, et al. (2008) FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom* 22(19):3043–3052.
14. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 11:148.
15. Wang Y, Kora G, Bowen BP, Pan C (2014) MIDAS: A database-searching algorithm for metabolite identification in metabolomics. *Anal Chem* 86(19):9496–9503.
16. Ridder L, et al. (2013) Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea. *Anal Chem* 85(12):6033–6040.
17. Ridder L, et al. (2012) Substructure-based annotation of high-resolution multistage $MS^{(n)}$ spectral trees. *Rapid Commun Mass Spectrom* 26(20):2461–2471.
18. Heinonen M, Shen H, Zamboni N, Rousu J (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* 28(18): 2333–2341.
19. Shen H, Dührkop K, Böcker S, Rousu J (2014) Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics* 30(12): i157–i164.
20. Böcker S, Rasche F (2008) Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* 24(16):i49–i55.
21. Rasche F, Svatoš A, Maddula RK, Böttcher C, Böcker S (2011) Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem* 83(4):1243–1251.
22. Dührkop K, Böcker S (2015) Fragmentation trees reloaded. *Research in Computational Molecular Biology*, Lecture Notes in Computer Science (Springer, Berlin), Vol 9029, pp 65–79.
23. Vaniya A, Fiehn O (2015) Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Analyt Chem* 69:52–61.
24. Watrous J, et al. (2012) Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA* 109(26):E1743–E1752.
25. Nguyen DD, et al. (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proc Natl Acad Sci USA* 110(28):E2611–E2620.
26. Morreel K, et al. (2014) Systematic structural characterization of metabolites in Arabidopsis via candidate substrate-product pair networks. *Plant Cell* 26(3):929–945.
27. Gerlich M, Neumann S (2013) MetFusion: Integration of compound identification strategies. *J Mass Spectrom* 48(3):291–298.
28. Hufsky F, Scheubert K, Böcker S (2014) Computational mass spectrometry for small molecule fragmentation. *Trends Analyt Chem* 53:41–48.
29. Sumner LW, et al. (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3(3):211–221.
30. Dunn WB, et al. (2013) Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9(1):44–66.
31. Rasche F, et al. (2012) Identifying the unknowns by aligning fragmentation trees. *Anal Chem* 84(7):3417–3426.
32. Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. *Bioinformatics* 24(21):2518–2525.
33. Rauf I, Rasche F, Nicolas F, Böcker S (2013) Finding maximum colorful subtrees in practice. *J Comput Biol* 20(4):311–321.
34. Shawe-Taylor J, Cristianini N (2004) *Kernel Methods for Pattern Analysis* (Cambridge Univ Press, New York).
35. Scholkopf B, Smola AJ (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA).
36. Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, eds Smola AJ, Schölkopf B (MIT Press, Cambridge, MA), Chap 5.
37. Stein S (2012) Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Anal Chem* 84(17):7274–7282.
38. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR (2008) Automated de novo protein sequencing of monoclonal antibodies. *Nat Biotechnol* 26(12):1336–1338.
39. Ridder L, et al. (2014) In silico prediction and automatic LC-MS(n) annotation of green tea metabolites in urine. *Anal Chem* 86(10):4767–4774.
40. Menikarachchi LC, Hill DW, Hamdalla MA, Mandoiu II, Grant DF (2013) In silico enzymatic synthesis of a 400,000 compound biochemical database for nontargeted metabolomics. *J Chem Inf Model* 53(9):2483–2492.
41. Horai H, et al. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45(7):703–714.
42. Scheubert K, Hufsky F, Böcker S (2013) Computational mass spectrometry for small molecules. *J Cheminform* 5(1):12.
43. Dührkop K, Hufsky F, Böcker S (2014) Molecular formula identification using isotope pattern analysis and calculation of fragmentation trees. *Mass Spectrom (Tokyo)* 3:S0037.
44. Shinbo Y, et al. (2006) KNApSAcK: A comprehensive species-metabolite relationship database. *Plant Metabolomics*, Biotechnology in Agriculture and Forestry, eds Saito K, Dixon RA, Willmitzer L (Springer, Berlin), Vol 57, pp 165–181.
45. Wishart DS, et al. (2013) HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res* 41(Database issue):D801–D807.
46. Hastings J, et al. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res* 41(Database issue):D456–D463.
47. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* 11(23-24):1046–1053.
48. O'Boyle NM, et al. (2011) Open Babel: An open chemical toolbox. *J Cheminform* 3:33.
49. May JW, Steinbeck C (2014) Efficient ring perception for the Chemistry Development Kit. *J Cheminform* 6(1):3.
50. Steinbeck C, et al. (2003) The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J Chem Inf Comput Sci* 43(2):493–500.
51. Rebentrost F, Ben-Shaul A (1981) On the fragmentation of benzene by multiphotoionization. *J Chem Phys* 74(6):3255–3264.
52. Jebara T, Kondor R, Howard A (2004) Probability product kernels. *J Mach Learn Res* 5:819–844.
53. Hufsky F, Dührkop K, Rasche F, Chimani M, Böcker S (2012) Fast alignment of fragmentation trees. *Bioinformatics* 28:i265–i273.
54. Cortes C, Mohri M, Rostamizadeh A (2012) Algorithms for learning kernels based on centered alignment. *J Mach Learn Res* 13:795–828.
55. Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):27-1–27-27.
56. Lin H-T, Lin C-J, Weng RC (2007) A note on Platt's probabilistic outputs for Support Vector Machines. *Mach Learn* 68:267–276.