Confident Metabolite Structure Annotation with COSMIC

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena
von M. Sc. Martin André Hoffmann
geboren am 23.02.1991 **in** Achim, Deutschland

Gutachter:

1. **Prof. Dr. Sebastian Böcker, Friedrich-Schiller-Universität Jena**
2. **Prof. Dr. Lennart Martens, Ghent University and VIB, Ghent, Belgium**
3. **Dr. Michael Witting, Helmholtz Zentrum München und Technische Universität München**

Tag der öffentlichen Verteidigung: 06.12.2022

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706

Abstract

Small molecules are key to biomarker discovery, drug development, toxicity screenings of ecosystems like rivers and lakes, and many more important research areas in multiple life sciences. Elucidating the exact structure of these metabolites is often crucial in determining their functionality, however, confident annotation of these structures remains a major challenge. To analyse and measure samples of small molecules occurring in nature, mass spectrometry is the currently predominant technique. While mass spectrometry is used to measure the mass of a compound, tandem mass spectrometry can be used to additionally measure the mass of its fragments. The resulting spectral data however is highly non-trivial to interpret, especially translating it back into a molecular structure. Comparing measured spectra to previously recorded spectra of reference compounds in a spectral library is a popular approach to structure annotation, but is naturally limited to those well-known structures. This bottleneck accelerates the development of computational tools to annotate metabolite structures from mass spectrometry data, which enables rapid, large-scale structure annotation independent from spectral libraries. As to be expected, these tools return some proportion of incorrect annotations, which, depending on the experimental setting, can vastly outnumber correct annotations. As such, scientists using these tools need to be able to differentiate correct from incorrect annotations.

In this thesis, we establish that hit scores of currently available structure annotation tools for mass spectrometry data can not be used to separate correct from incorrect annotations. CSI:FingerID is the currently best-in-class tool for metabolite structure annotation, predicting a molecular fingerprint that is then used to search structure databases. We develop an E-value computation that is based on proxy decoys drawn from the PubChem database and show that this E-value score outperforms the current CSI:FingerID hit score for the task of separating correct from incorrect annotations. To further improve on this, we develop a Percolator inspired machine learning approach, where we train linear support vector machines for this separation task. While many machine learning methods are often used as uninterpretable black boxes, we take multiple steps to ensure that our confidence score model is not overfitting as best we can. We enforce directionality of features to adhere to "common sense" and evaluate our model on artificially noisified data as well as multiple different collision energies. Features used in the confidence score include scoring features from CSI:FingerID, quality assessment features of the predicted molecular fingerprint as well as the E-value score.

The confidence score outperforms the original CSI:FingerID hit score, the E-value score and all other tools that participated in the CASMI 2016 contest by a wide margin. Arguably, our confidence score enables confident structure annotation for a relevant portion of a dataset for the first time. We then integrate the confidence score into SIRIUS, a software suite combining *de novo* molecular formula annotation with CSI:FingerID structure annotation.

We then show the power of this COSMIC (Confidence Of Small Molecule IdentifiCations) workflow by annotating novel bile acid conjugate structures never reported before in a mouse fecal dataset. The annotation of nine structures was confirmed by manual evaluation and two structures using synthetic standards. In human samples, we annotated and manually validated 315 molecular structures currently absent from the Human Metabolome Database. Application of COSMIC to data from 17,400 metabolomics experiments led to 1,715 high-confidence structural annotations that were absent from spectral libraries.

Zusammenfassung

Kleine Moleküle sind der Schlüssel zur Entdeckung von Biomarkern, zur Entwicklung von Arzneimitteln, zur Untersuchung der Toxizität von Ökosystemen wie Flüssen und Seen und zu vielen anderen wichtigen Forschungsbereichen in den Biowissenschaften. Die Aufklärung der genauen Struktur dieser sogenannten Metaboliten ist oft entscheidend für die Bestimmung ihrer Funktionalität, doch die sichere Annotation von Metabolitenstrukturen bleibt eine große Herausforderung. Zur Messung und Analyse von Proben kleiner Moleküle, die in der Natur vorkommen, ist die Massenspektrometrie die derzeit vorherrschende Technik. Während mit der einfachen Massenspektrometrie die Masse einer Verbindung gemessen wird, kann mit der Tandem-Massenspektrometrie zusätzlich die Masse ihrer Fragmente gemessen werden. Die daraus resultierenden Spektren sind jedoch äußerst schwierig zu interpretieren, und schwer in eine Molekülstruktur zu übersetzen. Der Vergleich von gemessenen Spektren mit zuvor aufgezeichneten Spektren von Referenzverbindungen aus einer Spektrenbibliothek ist ein beliebter Ansatz zur Strukturannotation, ist aber auf bereits bekannte Strukturen beschränkt. Diese Einschränkung führte zur Entwicklung von computergestützten Methoden zur Annotation von Metabolitenstrukturen aus Massenspektrometriedaten, die eine schnelle, vollautomatische Strukturannotation unabhängig von Spektralbibliotheken ermöglichen. Erwartungsgemäß liefern diese Methoden einen gewissen Anteil an inkorrekten Annotationen, welcher je nach Experiment die Zahl der korrekten Annotationen bei weitem übersteigen kann. Wissenschaftler, die diese Methoden verwenden, müssen daher in der Lage sein, korrekte von inkorrekten Annotationen zu unterscheiden. In dieser Arbeit stellen wir fest, dass die Scoringfunktionen der derzeit verfügbaren Methoden zur Strukturannotation für Massenspektrometriedaten nicht verwendet werden können, um korrekte von inkorrekten Annotationen zu unterscheiden. CSI:FingerID ist das derzeit beste Tool für die Strukturannotation von Metaboliten. Es sagt hierbei einen molekularen Fingerabdruck vorher, der dann zur Suche in Strukturdatenbanken verwendet wird. In dieser Arbeit entwickeln wir ein Scoring basierend auf der Berechnung eines E-Values, welches auf Proxy-Decoys aus der PubChem-Datenbank basiert. Wir zeigen, dass dieser E-Value Score das aktuelle CSI:FingerID Scoring bei der Aufgabe, korrekte von inkorrekten Annotationen zu unterscheiden, übertrifft. Um dies weiter zu verbessern, nutzen wir einen von Percolator inspirierten Ansatz des maschinellen Lernens, bei dem wir lineare Support Vektor Maschinen für diese Separationsaufgabe trainieren. Während viele Methoden des maschinellen Lernens oft als uninterpretierbare Blackboxen verwendet werden, ergreifen wir mehrere Maßnahmen, um sicherzustellen, dass unser Konfidenzscore-Modell so wenig wie möglich überangepasst wird. Wir erzwingen bestimme Vorzeichen von Featuregewichten, um den „gesunden Menschenverstand“ abzubilden, und evaluieren unser Modell an künstlich verrauschten Daten sowie an mehreren unterschiedlichen Kollisionsenergien. Zu den Merkmalen, die in den Konfidenzscore einfließen, gehören die Bewertungsmerkmale von CSI:FingerID, die Qualitätsbewertungsmerkmale des vorhergesagten molekularen Fingerabdrucks sowie das E-Value Scoring.

Der Konfidenzscore übertrifft das ursprüngliche CSI:FingerID Scoring, den E-Value-

Score und die Scorings aller anderen Methoden, die am CASMI-Wettbewerb 2016 teilgenommen haben, bei weitem. Unser Konfidenzscore ermöglicht nun zum ersten Mal eine zuverlässige Strukturannotation für einen relevant großen Teil eines Datensatzes. Anschließend integrieren wir den Konfidenzscore in SIRIUS, eine Software-Suite, die *de novo* Molekülformelannotation mit CSI:FingerID Strukturannotation kombiniert.
Zuletzt zeigen wir die Leistungsfähigkeit dieses COSMIC (Confidence Of Small Molecule IdentifiCations) Workflows, indem wir bisher unbekannte, neuartige Strukturen von Gallensäurekonjugaten in einem Mausfäkaldatensatz annotieren. Hierbei wurden die Annotationen von neun Strukturen durch manuelle Auswertung und die von weiteren zwei Strukturen durch synthetische Standards bestätigt. In menschlichen Proben haben wir 315 molekulare Strukturen annotiert und manuell validiert, die derzeit in der Humanmetabolom-Datenbank fehlen. Die Anwendung von COSMIC auf Daten aus 17.400 Metabolom-Experimenten führte zu 1.715 strukturellen Annotationen mit hoher Verlässlichkeit, die in nicht bereits in Spektrendatenbanken vorhanden sind.

Acknowledgements

I had saved writing this chapter for last, as I was always nearly overwhelmed by the thought of actually finishing this thesis. The time as a PhD student has probably been one of the most forming years of my life, and I would first like to thank my supervisor, Sebastian Böcker. Not only is he obviously the creative source behind many of the projects I have worked on, but he always gave me the freedom to try on my own. I always felt like if I worked hard on my projects, he would try hard for me if need be. The same goes for my colleagues, especially Kai, Markus, Marcus and Fleming, with whom I have worked closely over these past six years. Being such a small group, it always felt like everybody had everybody's back and each social meeting inspired and excited me about the work we do. I always liked bioinformatics for its duality of sitting in the figurative basement, hatching ideas, but then also seeing those ideas applied in the real world. I would like to thank Louis-Felix Nothias and Michael Witting for taking what we developed in our basements into the real world with excitement and a critical eye.

Similarly, the International Max Planck Research School allowed me to meet many scientists who work on the applied side of the field, which helped get an alternative perspective on the problems we are dealing with. Over the years, I have met many friendly and helpful people, who answered my technical or academic questions and provided guidance along the way. For that, I would like to especially thank Ales Svatos, Georg Pohnert, Juho Rousu, Pieter Dorrestein and Justin van der Hooft. The work presented in this thesis only was possible to this degree, because of the many researchers that came before me. SIRIUS and CSI:FingerID, on which all of this builds onto, have been developed for over ten years by a multitude of researchers, some of which I unfortunately never even met. I would like to thank the people who made access to the many data sets used in this work possible: The GNPS community, MassBank community, Agilent and NIST. I gratefully acknowledge the funding I received through the Friedrich Schiller University Jena, the International Max Planck Research School and the Deutsche Forschungsgemeinschaft (DFG).

The latter part of this chapter is dedicated to the people in my personal life, that helped and supported me through the last six years. First and foremost, I am grateful to my wife Julia, who has been the source of the emotional happiness and stability needed to finish this chapter of my life. I am grateful to my parents, sister and grandparents, for paving the way in my earlier years and continuous support until now. Next to my family, I would like to thank my close friends, who provided distraction and perspective when I needed it, especially Linte, Mario, Lyc, Nico, Yannik and Sandro. Last but not least, I would like to thank our secretary Kathrin Schowtka, without whom probably nobody in our group would have actually finished their PhD.

Contents

1 Introduction **1**
- 1.1 Contributions of this work 2

2 Backgrounds in Organic Biochemistry and Mass Spectrometry **5**
- 2.1 Molecules and Ions 5
- 2.2 Bioinformatics-driven Molecule Representations 7
  - 2.2.1 Molecular Similarity and Molecular Fingerprints 8
- 2.3 Metabolites 9
- 2.4 Mass Spectrometry 9
  - 2.4.1 Ionisation Modes and Adducts 10
- 2.5 Tandem Mass Spectrometry 11
- 2.6 Chromatography 11

3 Machine Learning and Statistics **13**
- 3.1 Machine Learning 13
  - 3.1.1 Supervised Machine Leaning and Classification 13
  - 3.1.2 Linear Support Vector Machines 14
  - 3.1.3 Kernel SVMs and Deep Neural Networks 15
  - 3.1.4 Evaluating Classifier Performance and Interpretability 17
- 3.2 False Discovery Rates 20
  - 3.2.1 Estimating False Discovery Rates 20

4 Related Work: Computational Mass Spectrometry **23**
- 4.1 Annotation and Identification 23
- 4.2 Data Preprocessing 23
- 4.3 De Novo Molecular Formula Annotation 24
  - 4.3.1 Isotope Pattern Analysis 24
  - 4.3.2 Fragmentation Trees 25
- 4.4 Spectral Library Search 25
- 4.5 Structure Database Search 27
  - 4.5.1 MS/MS Spectra Simulation for Database Structures 28
  - 4.5.2 Combinatorial Fragmentation of Structures 28
  - 4.5.3 Structural Feature Prediction from MS/MS Spectra 29
- 4.6 CSI:FingerID 29
- 4.7 Compound Class Prediction 31
- 4.8 Metascores 33

5 Datasets, Databases and Evaluation Principles **35**
- 5.1 MS/MS Reference Datasets and Noise Addition 35
- 5.2 Biological Datasets 37

5.3 Structure Databases 38
5.4 Training CSI:FingerID and Structure–disjoint Evaluation 39

6 COSMIC 41
6.1 False Discovery Rates in Computational MS 41
6.1.1 FDR Estimation in Proteomics 41
6.1.2 FDR Estimation in Metabolomics 42
6.2 A New task - Score Separation 42
6.2.1 The Importance of a Score that Separates 42
6.2.2 Evaluating Score Separation 43
6.2.3 Evaluating Score Separation for Popular Structure Library Search Tools 47
6.2.4 Designing a Confidence Score for CSI:FingerID 48
6.3 E-values 49
6.3.1 E-value Estimation 49
6.3.2 Evaluation of E-value Separation 50
6.4 Support Vector Machine Approach - the COSMIC Confidence Score 52
6.4.1 Overfitting Analysis and Enforced Directionality 53
6.4.2 Confidence Score Evaluation 58
6.4.3 Evaluation Against other *in-silico* Tools 65
6.4.4 Evaluation against Spectral Library Search 66
6.4.5 FDR Estimation Using the COSMIC Confidence Score 68
6.4.6 The COSMIC Workflow 70

7 Practical Application 73
7.1 Data Preprocessing 73
7.2 Annotation of Novel Bile Acid Conjugates 73
7.3 Annotating Structures Missing from HMDB 74
7.4 Repository-scale Annotation 76

8 Conclusion 79

9 Data and Code Availability 83
9.1 Data Availability 83
9.2 Code Availability 83

A Appendix 103

1 Introduction

Nature is diverse. In fact, it is so diverse and complex, that whole scientific fields emerge, just to study and understand small parts of it. Some of those fields are called the “-omics” sciences, each analysing and trying to understand a specific pool of molecules. One of the oldest “-omics”, genomics, is the study of one of the fundamental parts of evolution and life as we know it - the genome. It is comprised of DNA, large molecules consisting of only few different building blocks; A (Adenin), C (Cythosin), G (Guanin) and T (Thymin). From here, the “workflow of life” begins: DNA is transcribed into RNA, short readouts of specific parts of the genome; the entirety of RNA created in this fashion is called the transcriptome. One step further and we arrive at proteins, even smaller molecules translated from RNA, and generated from a slightly larger building block pool: The 22 proteinogenic amino acids. Genomics, transcriptomics and proteomics are nowadays considered established fields, with multiple decades of research conducted on them. Technological advances have made DNA/RNA sequencing affordable and fast, and the development of computational tools allows for high-throughput analysis of large data quantities.

By analysing the different molecules mentioned above, one can only obtain a rather static picture of an organism. The average lifetime of a protein ranges from days to years, while DNA molecules usually don’t degrade for decades [57, 84]. One can easily imagine the impact that the availability of such a snapshot has had on the medical, environmental and agricultural fields, to name just a few. There is however a plethora of remaining problems, for which a long term, static snapshot is not sufficient. Short term effects of drugs and toxicology of food or water are just a few examples that require the analysis of a more short-lived group of molecules - *metabolites*. Sometimes just consisting of a few atoms, these molecules allow for a more dynamic insight into the inner workings of an organism. The set of all metabolites contained in an organism is called the *metabolome*, and is studied by the field of metabolomics.

In recent years, research has shown that studying the metabolome can help understand complex biological systems and interactions [29, 123]. In contrast to proteins or DNA, metabolites do not consist of a known, small pool of building blocks, but are very diverse and non-linear. In addition, the metabolome is only partially encoded in our DNA; a substantial amount of metabolites stems from external sources, like food or cosmetics [23]. These external metabolites however, are not limited to what we willingly and knowingly consume; metabolites produced by microorganisms inhabiting our digestive system, skin and lungs have been shown to affect the host’s metabolism as well [129].

This biochemical diversity leads to metabolites being of critical interest in many research fields: In biomarker discovery, metabolites are used as indicators for conditions of biological systems, like diseases [74, 191]. In the environmental sciences, water bodies are screened for toxic metabolites [4], while the food industry screens nutritional items for those metabolites connected to known diseases.

Metabolomic screenings are usually performed in one of two ways: targeted or untargeted. Targeted metabolomics describes the rediscovery of a set of specific, already known molecules, for example contaminants or hazards. If the indicative metabolites for a certain condition are unknown, untargeted experiments can be performed to either analyse quantitative differences between samples or try to elucidate the structure of previously unknown molecules. This structural elucidation however is very challenging, and many metabolites still remain unknown [14, 35]. A major reason for that is the aforementioned structural diversity and short-livedness of metabolites.

Currently, the analysis of metabolites is facilitated by two predominant techniques: Nuclear magnetic resonance (NMR) and mass spectrometry (MS). Full structural elucidation of a molecule is only possible with NMR. However, relatively large amounts of the purified compound are needed, obtaining which can be extremely time- and cost intensive. Mass spectrometry on the other hand is much more sensitive and better suited for high-throughput analysis, but interpreting the data remains highly challenging. In contrast to the full elucidation NMR provides, mass spectrometry only measures the mass of a molecule and its fragments. Human interpretation of the resulting mass spectra is highly non-trivial, and requires compound specific expert knowledge and large time investments.

To alleviate these constraints, a popular method for metabolite identification is to compare mass spectra to a spectral library, containing spectra of already known reference molecules from previous measurements. These spectral libraries however only contain spectra of a small fraction of existing metabolites. In contrast, structure libraries like PubChem [88] are larger by several orders of magnitude, but a mass spectrum and a structure are not intrinsically comparable. To this end, many different computational tools were developed to make this comparison possible, and take advantage of the larger chemical space structure databases cover.

It should be noted however, that by searching in a database of known compounds, be it spectral or structure, one can never elucidate a truly novel compound's structure. While there are approaches to push that boundary by generating theoretical compound databases [76], the structural elucidation of novel metabolites is still an ongoing, exciting challenge in the field of metabolomics [107, 174].

1.1 Contributions of this work

I wrote my Bachelor thesis on a topic very much related to this thesis - the creation of decoy databases for molecular fingerprints. Later, for my master thesis, I worked on the separation of chimeric spectra using fragmentation trees implemented in SIRIUS [44]. It were these first scientific experiences that led to the realisation, that for me as a method developer, a close bond to experimentalists in the lab is indispensable. For that reason, I started my time as a PhD student as a member of the International Max Planck Research school at the Max Planck Institute for Chemical Ecology.

The primary area of research in the group of my supervisor Sebastian Böcker was and is the *in-silico* annotation of molecular structures from tandem mass spectrometry data. When I joined the group, CSI:FingerID coupled with SIRIUS had already established itself as a major player in the CASMI contests of 2016 and 2017 [148], a blind challenge for metabolite annotation. Using the *de novo* molecular formula annotation implemented in SIRIUS, CSI:FingerID predicts a molecular fingerprint from the input data and then compares it to structure candidates from a structure database such as PubChem [88] or HMDB [192]. This workflow produces correct structure annotation rates of up to 74% [47] when querying biomolecule structure databases. In this thesis, I focus on a problem that sits at the end of this workflow, but is of critical importance to enable high-throughput confident structure annotation: The separation of correct and incorrect annotations and its relation to false discovery rate estimation. Tools like CSI:FingerID use a scoring method to rank structure candidates, and return the highest scoring one as the annotation result. Since a sizeable portion of these annotations are incorrect, users require some metric to assess their confidence in the annotation. This is of particular importance, because follow up experiments based on *in-silico* annotations can be very much time- and cost intensive. In proteomics, false discovery rate estimations using decoy databases are used to give that metric [51, 79], this however, cannot be easily transferred to metabolomics. As I continued to work on creating decoy databases for molecular fingerprints, I realised that the scoring functions used in all tools that participated in the CASMI contests were unable to separate correct from incorrect annotations at a level, that would be required to enable sensible false discovery estimation in the first place. I then started to develop what we call a *confidence score* for CSI:FingerID. This would be an additional scoring that is not used to re-rank structure candidates for a single query, but rank only the top-scoring candidate for each query based on an assessment of how likely they would be correct annotations. To avoid overfitting, a common problem in machine learning, I first developed the estimation of an E-value, using PubChem structures as proxy decoys. This scoring already showed better separation power than the original CSI:FingerID scoring, but by itself was not able to reach a satisfying quality level. I then used a very simple machine learning approach utilising support vector machines, integrated the E-value as one of multiple features and was able to reach separation levels for correct and incorrect annotations that enable automated, rapid and confidence structure annotation of metabolites for arguably the first time. To make sure the model does not overfit on the training data, I artificially introduced different noise level into evaluation data, and restricted the freedom of the classifier by enforcing feature directionality. All of this work was done in close collaboration with my supervisor Sebastian Böcker and my colleagues Kai Dührkop, Markus Fleischauer and Marcus Ludwig.

While this part of my work was focused on theoretical advances and evaluations, I then showed that the confidence score I developed can be used in practical application. In close collaboration with Louis-Félix Nothias, the confidence score was used to discover multiple, previously unknown bile acid conjugate structures. In collaboration with Michael Witting, I processed ten publicly available datasets where human samples were measured, and we were able to annotate 315 molecular structures currently missing from the human metabolome database (HMDB) with high confidence. To showcase the possibility of high-throughput, confident structure annotation with this workflow, I processed 123 publicly available datasets consisting of 17,414 LC-MS/MS runs and was able to annotate 1,715 molecular structures with high confidence that were not present in our training data. I then created a publicly available web-based interface, in which these structures can be evaluated by experts in the field, or used in future research.

This thesis mainly covers my work on the development of the COSMIC confidence score, which has been published in 2022. Additionally, I have been involved in the development of CANOPUS by Dührkop et al. and ZODIAC by Ludwig et al.:

- **Hoffmann, M.A., Nothias, LF., Ludwig, M. et al.** "High-confidence structural annotation of metabolites absent from spectral libraries." Nature Biotechnology 40, 411–421 (2022).

- **Ludwig, Marcus, et al.** "Database-independent molecular formula annotation using Gibbs sampling through ZODIAC." Nature Machine Intelligence 2.10 (2020): 629-641.

- **Dührkop, Kai, et al.** "Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra." Nature Biotechnology 39.4 (2021): 462-471.

I presented my work or parts of it in a talk at the annual international conference on Intelligent Systems for Molecular Biology (ISMB) 2021, at the annual conference of the Metabolomics Society (Metabolomics) 2021 as well as at Symposia of the International Max Planck Research School in Jena 2018 and 2019 (best poster award). Together with Sebastian Böcker, Kai Dührkop, Markus Fleischauer and Marcus Ludwig, I was awarded the Thuringian Research Prize 2022 for applied research.

Before describing methods, evaluations and conclusions, I'm introducing the broad area of research. Starting with chemical background knowledge and the analytical technique called mass spectrometry on which my research is based, I then give a short overview of the fundamentals of machine learning and statistics that are needed in this thesis. In chapter 4, I will then introduce related work in the field of computational mass spectrometry, with a focus on SIRIUS and CSI:FingerID, on which my work is build on. The following chapters are then dedicated to introducing the problem of score separation at length, the development of the E-value and SVM-based confidence score as well as evaluation and application. For the remainder of this thesis, I will use "we" as the first person pronoun, as it is common in scientific literature.

2 Backgrounds in Organic Biochemistry and Mass Spectrometry

Chemistry is a vast field, branching into many different smaller ones. In this work however, we are focusing on biochemistry, the study of processes that occur in living organisms. As we have talked about building blocks before, we are now focusing on a specific one, that all molecules share: the *atom.* Atoms are usually displayed as their respective element, with a specific letter notating it, e.g. C for carbon, H for hydrogen, N for nitrogen, O for oxygen, P for phosphorus and S for sulphur. Notably, the chemistry studying compounds that contain carbon is called *organic chemistry.* An atom consists of protons, neutrons and electrons, where the amount of protons determines the chemical element, the amount of neutrons determines the isotope and the amount of electrons determines the charge of the atom. The sum of protons and neutrons that a molecule consists of, is also called the *atomic mass number*, and is displayed on the upper left of the element. Carbon for example possesses six neutrons and six protons in its most abundant form, resulting in the notation ^{12}C. The mass of an atom is defined by the amount of protons and neutrons it possesses, called the *nominal mass*, and is given in Dalton or the unified atomic mass unit (u). Both are defined as $\frac{1}{12}$ of a ^{12}C atom's mass, which is approximately $1.660539067 \cdot 10^{-27}$. For all atoms except ^{12}C, the atomic mass number differs from its actual mass. This effect is called the *mass defect*, and is very important in mass spectrometry. Atoms of the same element that possess different amounts of neutrons, are called *isotopes.* The unstable isotope of carbon, consisting of eight neutrons for example, would be noted as ^{14}C. Isotopes of an element appear in different abundances, depending on the element itself, and even geographical location. Refer to Table 2 for isotopes and abundances of some of the most common elements in organic chemistry. Molecules with the same atom composition, but differing isotopic compositions are called *isotopologues.* An atom is called *neutral* if the same amount of protons and electrons are present, *positively charged* or *cation* if the number of protons exceeds the number of electrons and *negatively charged* or *anion* if vice versa. Atoms that are either positively or negatively charged are called *ions.*

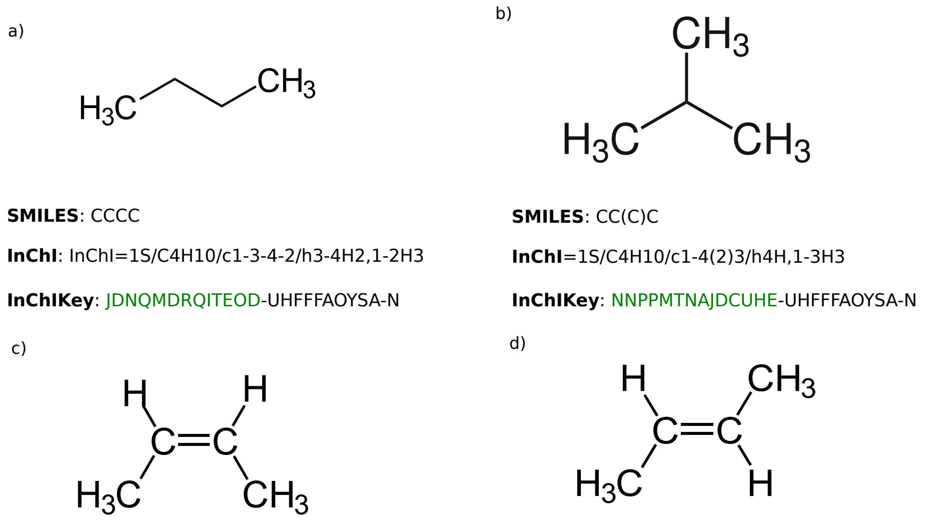
2.1 Molecules and Ions

Two or more atoms that are connected by a chemical bond are called a *molecule.* Bonds between molecules appear with different characteristics, regarding e.g strength and type of the bond. Covalent bonds are formed when two or multiple atoms share one or multiple electron pairs. Relevant examples of covalent bond types are *sigma bonds* and *pi bonds.* In this work, the most prevalent bonds are *single bonds*, consisting of one sigma bond, *double bonds*, consisting of one sigma and one pi bond and *triple bonds*, consisting of one sigma bond and two pi bonds. Chemical substances that consist of many identical molecules composed of at least two different elements are called *compounds.* Throughout this work, we use the terms “compound” and “molecule” interchangeably, as we are never

element	**symbol**	**isotopes**	**mass (Da)**	**abundance (%)**
carbon	C	^{12}C	12.0	98.93
		^{13}C	13.00335484	1.07
hydrogen	H	^{1}H	1.007825032	99.9885
		^{2}H	2.014101778	0.0115
nitrogen	N	^{14}N	14.003074	99.636
		^{15}N	15.000108898	0.364
oxygen	O	^{16}O	15.99491462	99.757
		^{17}O	16.99913176	0.038
		^{18}O	17.99915961	0.205
phosphorus	P	^{31}P	30.973761	100.00
sulphur	S	^{32}S	31.972071174	94.99
		^{33}S	32.9714589098	0.75
		^{34}S	33.967867	4.25
		^{36}S	34.96903231	0.01

Table 2.1: Common elements and their isotopes. Table showing the symbol letter, stable isotopes, isotopic masses and relative abundances of the most common elements in organic chemistry: carbon, hydrogen, nitrogen, oxygen, phosphorus and sulphur. Values taken from [176] (masses) and [12] (abundances).

interested in single molecules. The elemental composition of a molecule denotes of how many of each element it consists of, and is called the *molecular formula*. The most common notation of writing molecular formulas is the Hill notation. The amount of carbon atoms is written first, followed by the amount of hydrogen atoms. After that, all other elements follow in alphabetical order. The molecular formula of Taurin for example would be given as $C_2H_7NO_3S$. In the same fashion that atoms are called ions if they are positively or negatively charged, molecules can be ionic as well. The charge state of an ion is given at the upper right of the molecular formula, e.g. CH_3COO^- for the acetate ion. While molecular formulas denote the elemental composition of a molecule, they do not contain information about its *constitution*, that is the connectivity between atoms. If two molecules have the same molecular formula, but different constitutions, they are *structural isomers* of each other. Molecules with identical molecular formula and constitution can still differ in their orientation in three dimensional space, and are then called *stereoisomers*. To denote the constitution as well as stereochemistry of molecules, the *structural formula* can be used. See Fig. 2.1 for an example: butane and isobutane are structural isomers of each other (a,b), while cis-2-butene and trans-2-butene are stereoisomers of each other (c,d). Throughout this thesis we will use the term *structure* instead of constitution, as mass spectrometry is generally not able to differentiate between stereoisomers of a molecule.



SMILES: CCCC

InChI: InChI=1S/C4H10/c1-3-4-2/h3-4H2,1-2H3

InChIKey: JDNQMDRQITEOD-UHFFFAOYSA-N

SMILES: CC(C)C

InChI=1S/C4H10/c1-4(2)3/h4H,1-3H3

InChIKey: NNPPMTNAJDCUHE-UHFFFAOYSA-N

Figure 2.1: Examples for different forms of isomerism and text-based structure representations. Shown are the structural isomers butane (a) and isobutane (b) with their respective SMILES, InChI and InChIKey representation, as well as the stereoisomers cis-2-butene (c) and trans-2-butene (d). The first 14 characters of the InChIKey are marked in green.

2.2 Bioinformatics-driven Molecule Representations

While the structural formula discussed in the previous section is relatively easily readable for a human, computers require a different structural representation more optimised for machine reading. The two important representations used in this work, are the text-based *SMILES* [183] and *InChI* [68] formats. Both systems represent the molecule as a string, which is very favourable for database storing and reading. The SMILES of a molecule is a sequential string, where specific symbols are used to denote atoms, bonds and charge states. The elemental symbol is used to describe atoms, while different bond types have different symbols as well (e.g. "-" for single bonds, "=" for double bonds and "#" for triple bonds). Ring type structures are encoded by numerical labels, and side-chains or other branching structures by parenthesis. Implicit single bonds and hydrogen atoms are usually omitted in this representation. One specific problem with SMILES however is, that there are multiple SMILES expressions describing the same molecular structure. An example for this are CCCN and C(N)CC, both describing propylamine. Efforts have been made to develop a *canonical* SMILES, but failed to find a unique representation for every molecule [111, 119]. Even though no official canonical SMILES exists, implementations like the PubChem chemical structure standardisation [88] can be used. The IUPAC International Chemical Identifier, or InChI for short, creates a unique atom ordering by using graph isomorphism algorithms [68]. Contrary to the sequential SMILES representation, the InChI is composed of layers. The *main layer* contains the molecular formula as well as information about atom connectivity and hydrogen atom presence. Following layers then describe charge state and stereochemistry of the molecule. As the full InChI of a molecule can be quite long, the InChIKey was introduced as a compact, 27 character hashed representation of the InChI. It consists of three blocks separated by a hyphen. In this work only the first, 14 character long, block of the InChiKey is used, as that is where the molecules constitution is encoded. See Fig. 2.1 for examples of InChI and SMILES representations.

2.2.1 Molecular Similarity and Molecular Fingerprints

The text-based formats for molecular structure representations described above are a good way to quickly check a pair of molecules for identity. They are however unfit to infer how similar they are beyond this binary identity. Molecular fingerprints were developed for this very task, and are a widely used molecule representation in cheminformatics. Essentially, a molecular fingerprint is a vector of fixed size, in which each bit encodes a predefined molecular property. Molecular properties can range from the presence/absence of specific atoms or substructures, over chemical properties, to more complex information like atom neighbourhoods. Examples for fingerprint types encoding substructure information are PubChem CACTVS [180] and MACCS [49]. Extended connectivity fingerprints (ECFP) [134] are an example for neighbourhood-defining types. Many popular cheminformatics libraries, like CDK [157, 189], RDKit or OpenBabel [121], contain one or multiple methods for fingerprint computation from text-based structure representations like SMILES. Fingerprint performance is heavily dependent on how well the defined molecular properties of a fingerprint type fit to the data they are applied to [13, 42, 120, 188].

It should be noted that the fingerprint representation is not lossless, as the original molecule's structure can not be reconstructed from it. Furthermore, most fingerprints are binary, meaning that they don't encode for the frequency that a molecular property appears in a molecule. "Counting" fingerprints exist, but can still not encode the relative or absolute position of substructures. The vector representation that fingerprints offer allows for very fast comparisons of molecules. A common metric used to compare two fingerprints is the Tanimoto coefficient [186]:

$$Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.1}$$

where A and B are defined as a set of molecular properties.

This is of particular interest for the biochemical field, since similar structural features between two molecules can be a sign for similar biochemical activity. This structure/function relationship is called "structure activity relationship" [22, 77, 85, 185]. The prediction of a molecule's biochemical activity based on its (sub)-structure is done using $QSAR$ (quantitative structure activity relationship) models [43, 116, 188]. There are many biological applications for rapid molecule comparison, many of them making use of *virtual screening* [187]. Here, large databases of molecular fingerprints are compared to a query molecule and only those database molecules that pass a certain similarity threshold are retained for further analysis.

We want to point out here, that while the Tanimoto coefficient is widely adopted, its performance varies greatly based on the fingerprint type used, as well as size of the molecule. See Fig. 2.2 for an example of intuitively very similar molecules with a very low Tanimoto coefficient. Alternative similarity measures exist in the form of the cosine similarity, Dice coefficient and Soergel distance [8].

From a computer science perspective, it is intuitive to interpret a molecule's structure as a graph, where atoms represent vertices and bonds represent edges. The similarity of two molecules could then be measured as the minimum amount of edges that need to be removed from the two graphs to be isomorphic, with hydrogen atoms not being considered. This problem is called the "Maximum Common Edge Subgraph" (MCES) problem, which

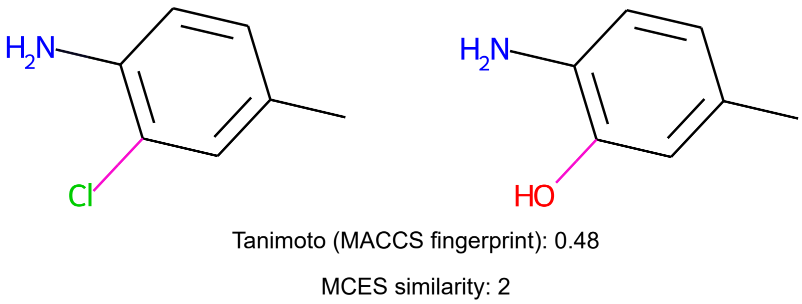


Figure 2.2: Example for a pair of molecules with low Tanimoto similarity but high MCES similarity. Edges that need to be removed to reach MCES distance of two are coloured in magenta, MACCS fingerprints were used to compute the Tanimoto.

is unfortunately NP-hard due to it being a generalisation of subgraph isomorphism [7]. For molecule pairs that are very similar to each other however, it can be used in a manual fashion to provide better similarities than the Tanimoto coefficient (Fig. 2.2).

2.3 Metabolites

Metabolites are a class of small molecules that are intermediate or end products of a living organism's metabolism. While there is no clear cut definition of what constitutes a metabolite, they are usually small molecules (below 1000 Da) that do not consist of a small pool of building blocks. The entirety of an organism's metabolites is called the *metabolome*, and provides a close snapshot of the phenotype [130]. As an exception, small peptides consisting of only few amino acids are also sometimes considered metabolites, even though they consist of well defined building blocks. In the same fashion, lipids, which are also part of the metabolome, are studied in a distinct field of study, *lipidomics*, as their very regular and linear structure sets them apart from most other metabolites. Metabolites vary greatly in structure, size and biochemical functionality and thus cover a large chemical feature space. We distinguish between *primary metabolites* and *secondary metabolites*. Primary metabolites are usually those involved in the organism's reproductive functions, growth and other "housekeeping" functions, while secondary metabolites are more specifically tailored to certain ecological niche functions or dealing with environmental stress. Primary metabolites are often shared between many different species, while secondary metabolites are more specific to an organism's environmental condition. A term often used synonymously with "metabolite" in the literature is *natural product*, however it usually refers to secondary metabolites. A natural product is defined as an organic compound that is produced by a living organism. Secondary metabolites are highly important for multiple scientific and economic fields [87, 113]. Many secondary metabolites remain unidentified [9, 123].

2.4 Mass Spectrometry

One of the two predominant techniques to analyse metabolites is mass spectrometry (MS). It requires only small amounts of sample and can measure different molecule species simultaneously. It is fit for high-throughput analysis, especially when compared to the other popular technique: nuclear magnetic resonance (NMR). Mass spectrometry is only able to measure ions, and thus all molecules need to be ionised before they can be detected by the mass spectrometer. Ions can be single- or multiply charged, as a mass spectrometer measures the mass-to-charge ratio (m/z) of an ion. Through this thesis, all ions will be single-charged, making it so the molecule's mass and its m/z are identical and used synonymously here. Even though it is highly sensitive, mass spectrometry does not measure singular ions, but all ions belonging to the same compound. Mass spectrometers have evolved to become very accurate measurement instruments, however they can naturally only provide accuracy to a certain degree. Since mass spectrometer's mass accuracy is relative to the measured ion's mass, we use *parts per million* (ppm) to specify an instrument's accuracy. Modern instruments reach accuracies of 1 ppm or even lower. The *resolution* of a mass spectrometer describes its ability to distinguish between signals of very similar mass, but produced from different compounds. Mass spectrometry measurements are usually categorised into low-resolution and high-resolution. The computational approaches described in this thesis all rely on high-resolution data, as overlapping signals produced by different compounds are in general hard to process.

A mass spectrometer consists of three major components: The *ion source*, *mass analyser* and *detector*. As mentioned above, all molecules need to be ionised before the measurement, which takes place at the ion source. Ionisation can either be *soft*, if the analyte molecule stays intact during the process, or *hard* if the analyte fragments. Commonly used techniques are *electron ionisation* (EI) [196] for hard, and *electrospray ionisation* (ESI) for soft ionisation. Following ionisation, ions enter the mass analyser, which is the central unit of a mass spectrometer. Ions are guided using electric fields, which is also why uncharged molecules cannot be measured. The general purpose of the mass analyser is to separate ions according to their masses. Many different mass analyser types exist, such as *time-of-flight* (TOF), *quadropole* or *Orbitraps*. An easy to understand example of how a mass analyser separates ions based on mass, can be given for a time-of-flight analyser: Ions are accelerated in a tube by an electric field through the same potential, resulting in identical kinetic energy for all ions with identical charge. As a result, the velocity of an ion is only determined by its mass; an ion with lower mass has a higher velocity than one with higher mass. After reaching the end of the tube, ions are detected by the mass detector. Intuitively, ions arriving at the mass detector at the same time should have the same mass, and the total flight time of an ion until it is detected by the mass detector can be used to infer its mass. The output of this measurement is called a *mass spectrum*. In principle, it consists of a list of m/z values with their corresponding ion signal intensities as measured by the mass detector. Mass spectra can be visualised as a 2-dimensional plot, where m/z values and corresponding intensities make up the axes.

2.4.1 Ionisation Modes and Adducts

Mass spectrometry experiments can be carried out in either *positive ion mode* or *negative ion mode*, which refers to the charge that the ions carry after ionisation. The ion mode to be used is mostly dependent on the analyte. In positive ion mode, molecules are *protonated*, that is a proton (H^+) is added to the molecule. In negative ion mode molecules are *deprotonated*, as they lose a proton. A common notation for protonated ions is ($[M + H]^+$), where "M" denotes the neutral molecule. Likewise, ($[M - H]^-$) is used for deprotonated ions. Depending on the analyte, other ionic species can form non-covalent bonds with the analyte ion and form an *adduct*. Common adducts include potassium ($[\mathrm{M}+\mathrm{K}]^+$) and sodium ($[\mathrm{M}+\mathrm{Na}]^+$) adducts in positive mode, as well as chlorine ($[\mathrm{M}+\mathrm{Cl}]^-$) adducts in negative mode. Depending on sample type as well as ionisation type and mode, the ratio of adducts in an experiment can vary greatly.

2.5 Tandem Mass Spectrometry

Even though mass spectrometers are capable of separating ion species in a sample based on their masses, there are certain limitations to it. Compounds with similar or identical nominal masses (*isobaric* compounds) or structural isomers are not differentiable by a standard mass spectrometry setup. To this end, *tandem mass spectrometry* (MS/MS) was introduced as an analytical technique. Its basic setup consists of two coupled mass spectrometers with a *collision cell* in between. Compounds are isolated based on mass by the first mass spectrometer, and then fragmented by introducing them to the collision cell. Fragment masses are then recorded by the second mass spectrometer and saved as the *tandem mass spectrum* (MS^2 spectrum or MS/MS spectrum). As spectra produced in this manner are highly reproducible, tandem mass spectrometry is widely used to aid in structural identification of unknown compounds. A commonly used fragmentation technique is called *collision induced dissociation* (CID). Here, the collision cell is filled with a noble gas, with which the analyte ions collide as they are passing through. This collision triggers a chemical reaction in which the analyte ion is fragmented. This fragmentation process is highly dependent on structure as well as *collision energy* (ce), given in electronvolt (eV). The collision energy is influenced by how fast analyte ions are moving through the collision cell and can be an important experimental parameter in MS/MS analysis. While certain compounds fragment very well even in lower collision energy settings, others might require higher collision energies to show sufficient fragmentation. The fragmentation process itself is highly complex and not fully understood. Fragmentation reactions can be cleavages of bonds, but also complex *rearrangements* of the structure [103]. In the common case of a single-bond cleavage, the ion fragment carrying the charge is called fragment or *product ion*, while the uncharged ion fragment is called *loss*. As only charged compounds can be detected by mass spectrometry, the loss is not recorded in the resulting MS^2 spectrum but can only be inferred. If an ion of the analyte ion species is not fragmented, it is called a *precursor ion*. The quality of a MS^2 spectrum is sometimes measured by the amount of fragment peaks recorded above a certain ion signal intensity threshold (commonly referred to as *noise threshold*). To increase the number of peaks in a MS^2 spectrum, spectra of multiple measurements with varying collision energies can be recorded and combined. See Fig. 2.3 for a depiction of a Tandem mass spectrometry setup. Instead of isolating one ion species by mass first, approaches exist in which all ions in a larger mass range are fragmented at once [21]. The resulting spectra contain fragment masses from multiple precursor ions, and are generally much harder to interpret, especially for *in-silico* methods.

2.6 Chromatography

As discussed in the previous section, isolating ion species can be tricky in real-world experimental conditions. This is especially true for more complex sample types containing

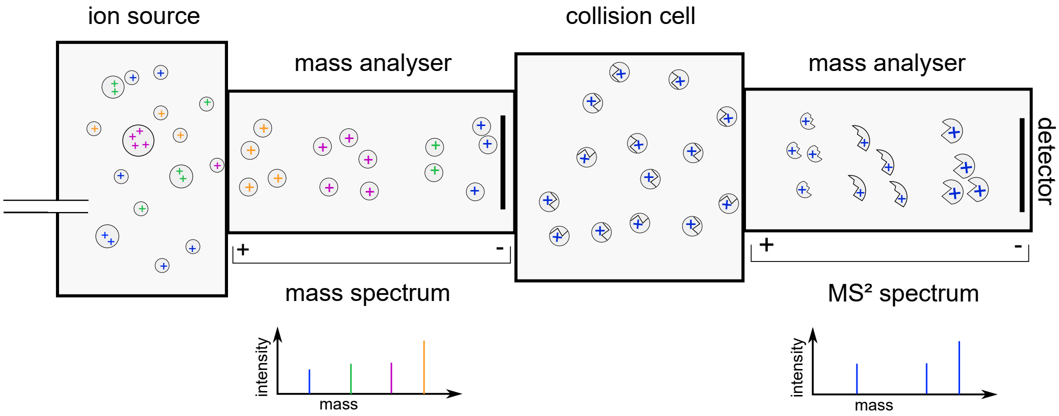


Figure 2.3: Schematic of a tandem mass spectrometry setup. Ions of different species are introduced and measured in the first mass analyser. Then, a mass spectrum is recorded. Ion species are filtered by their mass/charge values, and introduced into the collision cell, where they are fragmented. Fragments are then lead into the second mass analyser, where the MS/MS spectrum is recorded.

a high number of different metabolites (e.g. fecal or blood samples). Here, commonly occurring and highly abundant metabolites can drown out the signal of low abundant metabolites. Furthermore, if two metabolites with a sufficiently similar mass are measured, the mass spectrometer might not be able to isolate one from the other and fragment both ion species at the same time, resulting in a usually hard to interpret *chimeric spectrum*. Therefore, a *chromatography* step is often introduced prior to the mass spectrometry analysis. In essence, chromatography adds a second dimension to separating ion species from each other, which is most commonly polarity or structural volume. We differentiate between *gas chromatography* (GC) and *liquid chromatography* (LC). Coupled with a mass spectrometry setup, we talk about *gas chromatography-mass spectrometry* (GC-MS) and *liquid chromatography-mass spectrometry* (LC-MS). In metabolomics, liquid chromatography is often the chromatographic method of choice because it only requires the analyte to be soluble, while GC requires the analyte to be volatile [54]. Liquid chromatography consists of a mobile phase (the analyte in solution) and a stationary phase (also called *matrix*). As the mobile phase travels through the stationary phase, molecules interact with it based on structure, volume or other properties. This prolongs the time that it takes for a molecule to travel through the stationary phase completely. This travel time is called *retention time* and can be used as orthogonal information for the structural elucidation of unknown metabolites. Unfortunately, LC measurements of the same analyte solution can return varying results, mostly because of the high sensitivity of the system in regards to matrix type, containment from previous measurements or wear of use [193]. The retention time is still widely used as orthogonal information to aid in structural elucidation. Going back to the start of this section, liquid chromatography is helpful in separating metabolites with extremely similar masses, but very different structural properties (resulting in a large retention time difference). However, it is still not unusual for multiple ion species to have basically indistinguishable retention times and masses [125].

3 Machine Learning and Statistics

3.1 Machine Learning

Machine learning is a rapidly growing field in computer science, in which a system uses training data to “learn” and then predict properties of the input data supplied to it. In this chapter we are going to introduce and focus on the task of classification in supervised machine learning and introduce support vector machines as our method of choice.

3.1.1 Supervised Machine Leaning and Classification

Supervised machine learning describes a machine learning setup, in which the training data for a system is *labeled* and then used to calculate a mapping function for unlabeled input data. Data in this context describes a set of variables, called *features*, and is usually given in the form of vectors. Depending on the initial nature of the data, transforming it into a vector can be more or less challenging. To give an intuitive example, we can think of this problem: Given the colour of a tree leaf, predict the number of days until it falls of its branch. The features here could be the RGB values corresponding to the leaf’s primary colour, which are trivially easy to transform into a vector. Next, we are looking at the same problem, but instead of the colour of a leaf, we are given the structure of the leaf’s veins. Transforming this graph-like information into a vector can be a lot more difficult. The output of the machine learning system, the *prediction*, can take different ranges. Following our example from above, predicting the remaining lifespan of a leaf can return any numerical value in $[0, \infty)$. However if we reformulate our problem to just the question of if the leaf’s lifespan is more likely to be one, two or five days, our prediction will have to fall into one of these finite bins. This is called *classification*, and we differentiate between regular classification (two-class) and multi-class classification (three or more classes). In this thesis, we are going to focus on regular, two-class classification where class labels are in $\{-1, 1\}$.

Returning to the start of this section, finding a function that maps input training features to class labels is the core part of classification. This function needs to be optimised for the *empirical loss*, which denotes the divergence of predicted versus true labels on the training data. Given a training data set of size n, one can easily find a polynomial function of degree n that produces a perfect labelling and as such would be considered a perfect mapping function on the training data. However, if applied to a different data set, this function would likely severely underperform for data points not contained in the training data. This effect is called *overfitting* and is of central importance when evaluating classifier performance. A classifier function not only needs to perform well on its training data, but also on novel data points. This desired effect is called *generalisation.*

To achieve this, we can use a *regulariser*, which is essentially a penalty term. The regulariser is often a norm of the mapping function. Additionally, it might be beneficial or even necessary to restrict the space of functions to those with a lower degree.

In this thesis, we lean on the commonly used notation found in Friedman et al. [58]. We denote the training data matrix (also often called feature matrix) X as a $n \times m$ matrix, where n is the amount of feature vectors in the training data set and m is the length of the feature vectors. $X[i, j]$ then denotes the value of feature j in feature vector i. Similarly, Y denotes the matrix of all training labels y. In the case of binary classification, Y is $n \times 1$ dimensional and can be written as a vector. The matrix of predicted labels $\hat{y}$ is denoted as $\hat{Y}$ and the learned mapping function as f. The regulariser is the norm of the learned function f and is written as $\|f\|_r$. Our task of finding the predictor function f that minimises the sum of empirical loss L and the regulisation term $\|f\|_r$ can now be formalised as:

$$\min_{f \in F} \qquad L(Y, f(X)) + \|f\|_r \tag{3.1}$$

F here is the set of all functions that are permitted, for example all linear functions. Next, we are giving a short overview of popular machine learning strategies and formally introduce Support Vector machines.

3.1.2 Linear Support Vector Machines

Historically, separating data points of two classes has been described geometrically, rather than the loss and regularisation minimisation we introduced earlier. In three-dimensional space, one can imagine the classifier being represented by a hyperplane, where data points belonging to the first class lie on one side and data points belonging to the second class lie on the other side. The learning task would then be to find the hyperplane that best separates the data. Mathematically, a hyperplane is defined as a set of points that satisfy $\beta_0 + \beta \cdot x = 0$, where β_0 is the offset of that hyperplane from the origin (also called bias) and β is the normal vector to the hyperplane. β and β_0 are often denoted as w and b in literature, to better reflect their interpretation of "weight" and "bias". We are not using that notation to stay consistent with [58]. The first model developed to find such a hyperplane was the *Perceptron* [137], which iteratively updates feature weights to minimise the distance that misclassified data points have to the separating hyperplane. This iterative algorithm however only converges, if the training data can be fully linearly separated. Additionally, if multiple solutions exist to separate the data, one will be returned based on parameter initialisation, not optimal generalisation ability. To remedy these issues, the *Optimal Separating Hyperplane* [171, 172] was introduced, which maximises the distance between the closest data points of each class in the training data. The slab-like space between the separating hyperplane and these closest data points is called the *margin*, and by maximising it a unique solution is returned. The corresponding optimisation problem for maximising the margin M over n training data points is:

$$\begin{aligned} &\max_{\beta_0, \beta} M \\ &\text{subject to } \frac{y_i(x_i\beta + \beta_0)}{\|\beta\|} \geq M, i = 1, ..., n \end{aligned} \tag{3.2}$$

Since the distance of any given point x_i from the separating hyperplane is defined by $\frac{\beta \cdot x - \beta_0}{\|\beta\|}$, the constraints ensure that every data point is at least distance M from the hyperplane.

Because of β_0 and β being scaling invariant in eq. 3.2, we arbitrarily define $\|\beta\| = \frac{1}{M}$, to arrive at the equivalent optimisation problem of

$$\begin{aligned}&\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 \\ &\text{subject to } y_i(x_i\beta + \beta_0) \geq 1, i = 1, ..., n\end{aligned} \tag{3.3}$$

With that, we have solved one of the problems of the original perceptron, receiving a unique solution when solving this convex optimisation problem. The second problem, which is non-convergence if data is not fully linearly separateable however still exists. For that reason, the concept of *soft margin* [33] was introduced. The fundamental idea is to allow training data points to lie within in the margin, or on the wrong side of it, which essentially means to allow misclassification. Intuitively, these misclassified data points need to be penalised, for which *slack variables* ξ_i are defined. We integrate the sum over all slack variables into the optimisation problem, and arrive at:

$$\begin{aligned}&\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } y_i(x_i\beta + \beta_0) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, ..., n\end{aligned} \tag{3.4}$$

One can easily see, that the constant C here controls the scaling of misclassification penalisation, a smaller C results in a larger margin, but more misclassifications, while a larger C results in less misclassifications but a smaller margin. The empirical loss function of a SVM is called the *hinge loss* L_H, defined as:

$$L_H(y, \hat{y}) = max(0, 1 - \hat{y} \cdot y) \qquad \hat{y} \in \mathbb{R}, y \in \{-1, 1\} \tag{3.5}$$

The hinge loss for a point in the training data is equal to its slack variable ξ_i. From eq. 3.5 we can see, that the hinge loss is zero for training data points that lie outside of the margin but on the correct side of the hyperplane. See Fig. 3.1 for a visualisation of the concepts explained here. A common variation of the standard SVM model is to use quadratic hinge loss instead, which, analogous to eq. 3.4, can be expressed as:

$$\begin{aligned}&\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i^2 \\ &\text{subject to } y_i(x_i\beta + \beta_0) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, ..., n\end{aligned} \tag{3.6}$$

Training data points lying either within the margin or lying outside the margin but on the incorrect side of the hyperplane are called *support vectors*, and fully describe the prediction function. This becomes more clear if the optimisation problem is transferred to its Lagrangian dual form [144], that can also be used to calculate the distance of a new data point to the hyperplane without explicitly computing it.

3.1.3 Kernel SVMs and Deep Neural Networks

Kernel SVMs

In the previous paragraph, we introduced the *soft margin* SVM as a means to find a separating hyperplane even if the training data is not fully linearly separable. In practice, there might be cases where training data points are intuitively impossible to sensibly

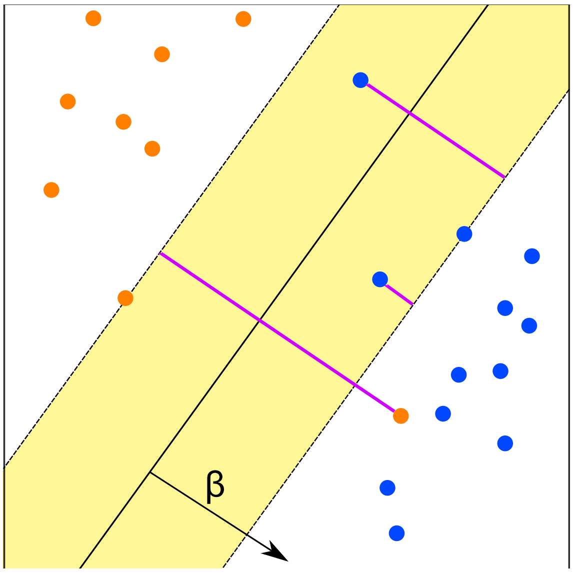


Figure 3.1: Schematic of a linear support vector machine. The separating hyperplane (solid black line) divides the space into two sides. Dashed black lines span the slab-like margin, coloured yellow. Orange and blue circles mark training data points belonging to either class 1 or -1. Data points within the margin or outside of the margin but on the incorrect side of the hyperplane have a non-zero hinge loss equivalent to their distance to the correct side of the margin (magenta lines). All points on the correct side of the margin have a hinge loss of zero. β is the normal vector orthogonal to the hyperplane.

separate by using a hyperplane at all. In these cases, it is common to transform the input space to a space of a higher dimension, where linear separability is possible again, see Fig. 3.2 for an example of such a transformation. Because of the high dimensionality of the resulting vectors, the computations performed on them in classifier training can lead to extreme computation costs. As a way to alleviate this, one can use kernels, which are functions that return the dot product of the high-dimensional, transformed vectors but take the lower-dimensional, original data points as input. This way, one does not have to actually apply the transformation into a higher-dimensional space. Since in this thesis we restrict ourselves to linear SVMs, we again point the interested reader to [58] for a much more exhaustive introduction to kernels.

Deep Neural Networks

As previously mentioned, in this thesis we focus on linear support vector machines. Nevertheless, we want to give a short introduction to arguably the most popular and well known machine learning method in use currently: *Deep Neural Networks (DNN)*. Historically, neural networks were created to be modelled after the human brain, where neurons are connected in a network. The earliest work on this was published by McCulloch and Pitts [102], who introduced the “McCulloch - Pitts Neuron”. Neurons are inspired by their biological counterparts, they take weighted input from different sources and compute an output via a (usually) non-linear function called the *activation function*. The most common activation function today is called the *rectified linear unit* (ReLU) [60], while the sigmoid function was most commonly used before that. Neurons are organised in layers, hence the name *neural network*. The first layer is called the *input layer* and encodes input

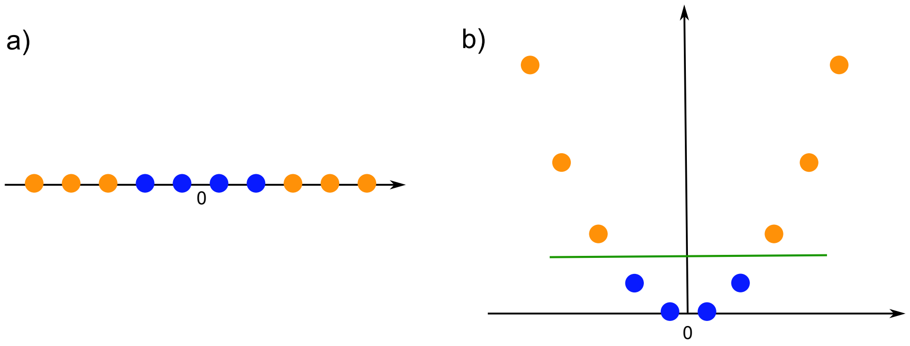


Figure 3.2: Example of data points that are not linearly separateable in one dimension (a), but easily separateable in two dimensions after applying a quadratic transformation (b). Green line depicts a possible separating hyperplane.

data in some form of vectorised representation. The last layer is called *output layer* and encodes the outcome (e.g. the predicted class in the case of binary classification). The layers in between the input and the output layer are called *hidden layer*, and signify some embedding of the input. A neural network is called *deep*, if at least two hidden layers exist. The way that layers and neurons are organised, is called the *architecture* of the network. A commonly used network architecture are *feedforward neural networks*, in which information flows one-directional from input layer to output layer. Over time, the idea of emulating the brain structure for neural networks diminished, as concepts that clearly differ from the way biological neural networks work, proved effective for deep neural network training. One of the most prominent examples is the backpropagation algorithm by Werbos et al. [184], which was later popularised by Rumelhart and Parker [122, 139]. The strength of the deep neural network model lies in its flexibility. Depending on the general architecture, and number of layers and neurons used, one can easily design a DNN for most problems. On the flipside, it is also very easy to design a DNN that just memorises training data and thus performs very poorly at generalising. Additionally, the internal representations of the input space used in the hidden layers can not be easily interpreted [159, 162, 195]. These properties are the main reason we are not using DNNs in this thesis. Obviously the field of DNN research contains much more than what we outlined here, we recommend the review by Schmidhuber [143] for a more extensive overview.

3.1.4 Evaluating Classifier Performance and Interpretability

As introduced earlier, the main goal of training a classifier is to create as much generalisation as possible and reduce overfitting to a minimum. In training, we minimise the empirical loss, in evaluation however that is not an intuitive or helpful metric to evaluate a classifier's performance. Using the metrics we now introduce directly as loss functions during training however is not possible, because they are not differentiable.

Evaluation metrics

As a short reminder, in two-class classification each instance belongs to one of two classes (we use "1" and "-1" as exemplary positive and negative class labels here). For each instance, the classifier then assigns a predicted label. This prediction can lead to the following four cases: An instance belonging to class "1", that had its class label predicted correctly, is

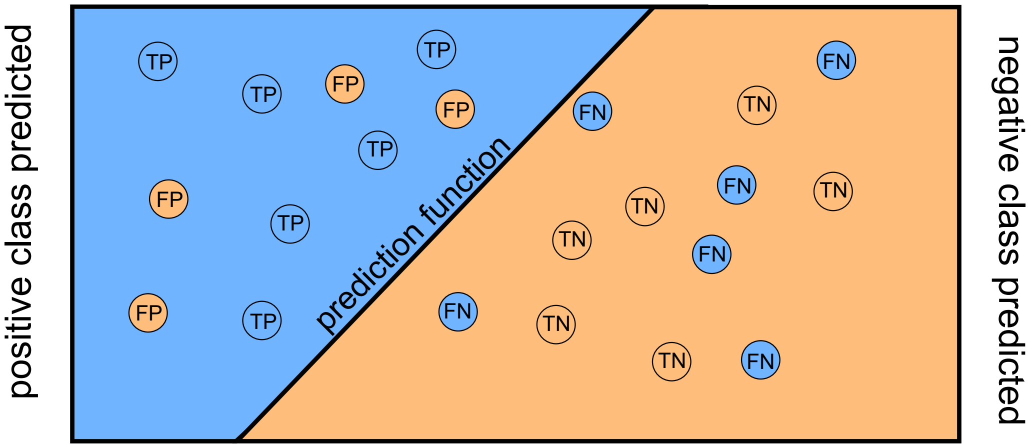


Figure 3.3: Depiction of the outcome of two-class classification. Data points belonging to the positive class (blue) or negative class (orange) can either be predicted correctly (TP or TN respectively) or incorrectly (FN or FP respectively).

called a *true positive* (TP), while one that had its class label predicted incorrectly is called a *false negative* (FN). An instance belonging to class "-1" that was predicted as "1" is called a *false positive* (FP), while one that was predicted as "-1" is called a *true negative* (TN). These four cases are commonly visualised as a 2x2 matrix, called the *confusion matrix*, see Fig. 3.3 for a depiction of this paragraph.

We can now use these cases to establish the following, frequently used classifier performance metrics: The *accuracy* is calculated as $(TP+TN)/(TP+FP+TN+FN)$ and describes the percentage of instances for which the label was predicted correctly. Its main drawback is being unfitting for cases in which class distribution is very lopsided. Here, prediction results for the majority class will drown out those of the minority class. In these scenarios we can measure *precision*, *recall* and *F1-value*. Precision is defined as $TP/(TP+FP)$ and describes the percentage of correct classifications between all instances assigned the positive class label. Its counterpart, recall, is defined as $TP/(TP+FN)$ and describes the percentage of instances belonging to the positive class that were correctly predicted. We can then calculate the harmonic mean of precision and recall to obtain the F1 score, which is defined as $\frac{2 \cdot precision \cdot recall}{precision + recall}$.

Receiver operating characteristic

The receiver operating characteristic (ROC) is a visualisation method widely used to evaluate classifier performance. It shows false positive rate (FPR), which is defined as $FP/(TN+FP)$, against recall (often labeled "true positive rate" (TPR)), for varying discriminative thresholds. The area under curve (AUC) can then be computed as a quantitative measure for classifier performance, where a perfect classifier would reach an area under curve of 1.0, and a classifier assigning labels randomly one of 0.5.

Independent and holdout datasets

Since we want to evaluate a classifier based on its ability to generalise, a common evaluation strategy is to simply evaluate its performance on a dataset that diverges from the training data in all relevant properties (called *evaluation data set* or *test dataset*). A dataset adhering to these restrictions is called an *independent* dataset. In real-world scientific appliance however, this is often extremely difficult or impossible to obtain. Given for example a training dataset consisting of mass spectrometry measurements of hospitalised patients. A truly independent dataset would likely need to be measured by a different scientist, on a different machine, at a different time of day, under different climatic conditions and so on. For that reason, datasets are usually called independent if they sufficiently diverge from the training dataset. A different strategy is the *holdout* method, in which part of the training data is omitted from training, and then used as an evaluation dataset. This is generally less preferred than independent datasets, but is still popular because of its simplicity in terms of implementation and data availability.

Cross-validation

Cross-validation is a method deployed to improve evaluation quality on imperfect independent and holdout datasets. Here, the data is partitioned into n folds of sufficiently similar size. One fold is then used as the test dataset, while the remaining ones are concatenated and used as the training dataset. This process is then repeated n times, so that every fold served as the testing set once. This reduces the risk of overestimated classifier performances being just the result of creating a partition containing only "easy" predictions. Usually, the number of folds is set to five or ten, which results in five or ten trained classifiers each with their own respective training and test partition. Depending on the training data, the partitioning into folds might require some additional constraints instead of just splitting randomly. This is especially the case if duplicate data points are part of the training set, as these need to always be grouped together in either the training or test partition.

As described, an independent dataset should not contain any data points also contained in the training dataset. In the case of overlapping data points, we can use the cross-validation setup described above. For every data point in the independent dataset that is also contained in the training dataset, we predict its label using a cross-validation classifier that contained this data point in its testing partition only.

Classifier interpretability

The evaluation metrics introduced in the above subsections are commonly applied to predictors of all popular machine learning approaches. Trained classifiers are given novel input data, and are then evaluated on their label-predicting performance. Here, we now want to introduce a concept that evaluates a classifier's quality data-free, which we call *classifier interpretability*. As discussed, the main goal is to train a classifier with high generalisation and low overfitting. Instead of using evaluation data to test for that property, it might also be helpful to evaluate the "knowledge system" that comprises the classifier directly. The feasibility of this approach varies greatly between machine learning models, and can only really be applied to relatively simple models. In the example of linear support vector machines, we can evaluate the trained feature weights. Since the prediction function of a linear SVM is just a linear combination of feature weights and the corresponding input feature values, we can perform simple checks to assess if the classifier contradicts with "common sense". Be reminded that a linear SVM assigns the positive class label if the linear combination of features and their corresponding weights is higher than some numeric threshold, which means that positive feature weights "push" for the positive class, while negative feature weights do the same for the negative class.

The higher the absolute value of a feature weight, the more impactful it is on the overall classification result. To give an example, let there be a classifier that is trained to predict if a person is going to be a billionaire in the next five years, and one of the features chosen is the person's monthly income. Intuitively, this monthly income should most likely be assigned a positive feature weight. If however, we see that the trained classifier assigned a negative weight (translating to "The higher your income, the less likely it is to become a billionaire"), we can use that information as a sanity check. In this particular example, the training data might have only contained low-earning lottery winners as examples for the positive class, resulting in high overfitting and low generalisation. It should be understood that this strategy needs to be applied with utmost care, as oftentimes correlations might be too complex to understand intuitively.

3.2 False Discovery Rates

The *false discovery rate* (FDR) is an important metric in life science research. It is defined as $FP/(FP+TP)$. Many large industries and academic fields focus on the search for something novel, be that new drug leads, novel biomarker for common diseases or previously unknown structures present in our very own metabolome. In most of these fields, the discovery of only one relevant novel compound is already considered a huge success, and the product of extremely high time and financial investment. It is also understood that the amount of currently unknown, but interesting biomolecules is very high, which is why preventing false positives is of much more relevance than preventing false negatives. A false positive biomarker candidate for example, can easily consume multiple years of research time before it is determined as such. As a result, detecting and discarding false positives is an extremely important task in many of these workflows, be they experimental or computational. In this thesis, we focus on false discovery rates for computational methods, where often thousands to millions of candidates are processed in some way at the same time. The output of such methods is usually a list of candidates (e.g. structures predicted for input data), that are ranked by an associated score. Since conceptually the "quality" of a candidate declines the lower its score, it is common practice to find a score threshold t at which the FDR for the sublist containing all candidates with score t or higher does not exceed a fixed threshold. The false discovery rate is always a property of a candidate list, rather than a singular candidate. To express the FDR-level of such a singular candidate, one can use the *q-value*, which is defined as the lowest FDR at which this candidate is still contained in the sublist defined by the score threshold t.

3.2.1 Estimating False Discovery Rates

To calculate the FDR of a candidate list as described above, one needs to know the true label of a candidate. In application this is usually not the case, the FDR then needs to be *estimated.* This estimated FDR must then be validated against exact FDR values, which is of course only possible in evaluation where the true labels are known. Over time, multiple approaches for estimating the FDR of candidate lists have been developed, many of them are based on the computation of the *p-value* as a measurement of statistical significance. In our context of scored candidates, the p-value is defined as the probability of obtaining a score, or a more extreme one, randomly. P-values are calculated for each candidate individually, and such must be corrected against multiple testing. By multiplying the p-value of a candidate with the total amount of candidates tested, we obtain the *E-value*, which denotes the amount of *expected* random candidates with a certain score or higher. Computing p-values or E-values for candidates is non-trivial, see Section 6.1 for an overview on FDR estimation in proteomics and metabolomics.

4 Related Work: Computational Mass Spectrometry

In this chapter we give a brief introduction on the advances in the field of computational mass spectrometry in recent years. In particular, we introduce methods to annotate molecular formulas and structures to tandem mass spectra. These methods are used as a foundation for the work presented later in this thesis.

4.1 Annotation and Identification

Before diving further into this chapter, we want to pay special attention to a specific part of the vocabulary used, which is the distinction between *molecule identification* and *molecule annotation*. Both expressions describe the process of trying to find the ground truth about a compound's molecular formula or structure, we are however deliberately using molecule annotation over identification. The reason is, that computer driven methods simply annotate the result of a computation to a query spectrum, which makes no promise about its correctness. Using the term "identification" might imply that the given result has been irrefutably proven against the ground truth.

In mass spectrometry, certain guidelines exist on how to judge the quality of an annotation. These are mostly based on the experimental level that was performed to ascertain the annotation [56].

4.2 Data Preprocessing

Data quality in mass spectrometry can greatly vary depending on instrument and setup used. As we are introducing computational methods to handle and interpret this data, it is important to understand that a high level of data quality is essential for a successful analysis. Since many different setups are used in application, it is important to preprocess and standardise incoming data. Different instruments for example produce varying levels of noise, which might need to be filtered prior [142, 190]. Other common steps of preprocessing LC-MS/MS data include baseline correction [11, 175], feature detection, integration and grouping [37, 86, 91, 110] as well as retention time alignment [71, 92]. While noise removal and baseline correction are primarily used to remove low quality parts of the data, feature grouping and retention time alignments can be seen as a way to enhance a molecule's measurement data.

Peaks are firstly output as a distribution over m/z values rather than a single value, which is called *profiled mode. Peak picking* describes the process of assigning a singular m/z value to a peak, which results in a *centroided* spectrum. Data preprocessing is often performed internally by tools like OpenMS [138], MZmine 2 [127] or SIRIUS 4 [47].

4.3 De Novo Molecular Formula Annotation

Molecular formula annotation is oftentimes the first step performed on the way to a molecule's structural elucidation. It can be used as guiding information in experimentally driven structure elucidation using NMR, but also aids in restricting the search space for computer driven structure library search [47]. Molecular formula annotation can be performed by searching a molecular formula database based on e.g. the molecule's mass, or *de novo*, which is independent of existing databases. While querying existing molecular formula databases can be done very fast, it restricts the search space to the size of the database. Even the largest molecular formula databases are considered vastly incomplete, which supports the need for *de novo* annotation. Here, dynamic programming can be used to enumerate all possible molecular formulas given a compound's mass and an alphabet of atoms that it might consist of [19, 45]. However, even for extremely small allowed mass deviations and an alphabet constrained to only the essential atoms biomolecules exist of, one can observe a combinatorial explosion of the search space [16]. This observation makes *de novo* molecular formula annotation a non-trivial problem. While heuristics exist to constrict the search space yet again, the true molecular formula might be omitted in the process [89]. Regardless of how the set of "allowed" molecular formulas is generated, candidates are then scored and ranked to determine the hopefully correct molecular formula candidate. This can either be done using MS1 data alone using the molecule's isotope pattern [3, 18, 20, 98] (see next section for details) or by incorporating available MS/MS data [16, 17, 47, 104]. Depending on the analysis, using the context between molecules can prove beneficial for molecular formula annotation. If molecular formulas are to be determined for multiple molecules contained in the same LC-MS/MS run, co-occurrences and common biotransformations can be taken into account to guide annotations [101, 136].

4.3.1 Isotope Pattern Analysis

As introduced earlier in this thesis, elements can occur as isotopes. The occurrence of isotopologues can also be measured in mass spectrometry, as isotopologues with sufficiently different masses produce different peaks. Isotopologues with extremely similar masses are commonly measured as one peak. A common notation is to name the isotopologue with the lowest mass the *monoisotopic peak* [20], the set of all isotopologues containing exactly one additional neutron the $M + 1$ peak, the set of all isotopologues containing exactly two additional neutrons the $M + 2$ peak and so on. Consecutive isotope peaks differ by about $m/z = 1$ for single charged compounds, making the *isotope pattern* easily visually distinguishable. The absence of an isotopic pattern for a measured peak can also indicate a noise peak or low data quality, and can help distinguish real metabolites from background noise or contamination. The measured isotope pattern of a compound can now be compared against the expected, simulated isotope pattern for a molecular formula candidate [20, 47]. Polynomial expansion methods [28] as well as Fourier transform methods [133] can be used to calculate masses and intensities of a molecular formula candidate's isotopic pattern. To ensure comparability to measured isotope patterns, isotopologues with very similar masses can be merged [20, 90]. Isotope pattern analysis as described, is integrated into SIRIUS 4, which considers relative and absolute errors when comparing masses and intensities between measured and simulated isotope patterns. Analysing the isotope pattern of a compound can also help to constraint the elemental alphabet and with that restrict the amount of molecular formula candidates that have to be considered for molecular formula annotation [47].

4.3.2 Fragmentation Trees

Fragmentation trees were introduced by Böcker et al. to provide a representation of a molecule's fragmentation based on its MS/MS spectrum. In this thesis, we are specifically interested in how fragmentation trees can be used to aid molecular formula annotations. For that, all possible molecular formula candidates that fit the precursor peaks mass are enumerated, and a fragmentation tree is constructed for each of them. Fragmentation trees are then scored by a maximum a posteriori estimator, which results in a ranked list of molecular formula candidates [16]. Fragmentation trees are directed, acyclic graphs, where a node represents a peak in the MS/MS spectrum, and an edge represents a loss between two peaks. Every peak in the underlying MS/MS spectrum can only be represented by one node. The root of a fragmentation tree is annotated with a molecular formula candidate representing ("explaining") the parent peak. Non-root peaks are annotated with molecular formulas explaining the specific peak's mass. A molecular formula annotated to a node of a fragmentation tree, usually has to be a subformula of the formula annotated to the parent node. As follows from previous sections, the mass of a peak can almost always be explained by multiple (and in most cases many) molecular formulas, which makes the process of finding the "best" annotations difficult.

To compute the fragmentation tree for a given molecular formula candidate, we first have to construct a fragmentation graph. A fragmentation graph is again a directed, acyclic graph where a node represents a peak in the MS/MS spectrum, and an edge represents a loss between two peaks. Different from a fragmentation tree, there is no limit as to how many nodes can represent the same peak. The nodes of the fragmentation graph can now easily be created as exactly the set of all possible molecular formulas for all peaks in the spectrum (except for the precursor peak, for which the molecular formula is fixed). Edges are now drawn between subformulas, completing the fragmentation graph. To now compute the fragmentation tree based on the fragmentation graph, requires solving an NP-hard problem, called the MAXIMUM COLOURFUL SUBTREE problem [17]. Luckily, metabolites measured on high resolution mass spectrometers produce fragmentation graph instances that can mostly be solved by Integer Linear Programming (ILP) [131] in a timely manner. If mass or elemental composition of a metabolite prevent the fast computation of its fragmentation trees, heuristics can be employed to solve the exact problem for only the high-ranked candidates.

Fragmentation trees do not try to represent the exact ground truth of the fragmentation process, as the exact chemical processes of fragmentation are not fully understood yet. See Fig. 4.1 for an example of a fragmentation tree representing a MS/MS spectrum.

4.4 Spectral Library Search

As introduced in previous sections, one of the fundamentals of what makes mass spectrometry such a widely used analysing technique is the reproducibility of the resulting spectra. Especially for GC-MS and EI measurements, spectra of the same compound measured on different setups usually are extremely similar [55]. The same is true for LC-MS spectra, although to a lesser degree. Fragmentation spectra produced by the Orbitrap's

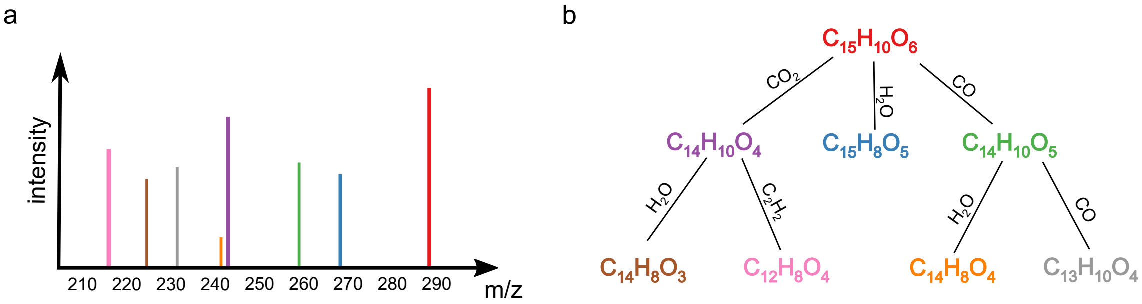


Figure 4.1: Fragmentation tree example. Example of a fragmentation spectrum (a) and the corresponding fragmentation tree (b). Each node in the tree corresponds to exactly one peak in the spectrum. An edge between two nodes in the fragmentation tree is labeled with the loss between the two molecular formulas. The mass of this loss corresponds to the mass difference between the two peaks in the spectrum.

"higher-energy C-trap dissociation" (HCD) can produce slightly different fragments and intensities than those produced by regular collision induced dissociation [75]. The same is true even for instruments that use very similar fragmentation methods, but from different vendors [118]. Additionally, collision energy settings are not perfectly accurate between machines of different vendors and can further lead to small dissimilarities [161]. Nevertheless, spectral library search has been a widely adopted technique for tandem mass spectrometry annotation for decades, as the approach is very simple technically. Here, the measured spectrum of interest is compared against a database of previously measured spectra, called the spectral library. To be efficient, only spectra of a close enough precursor mass are used for this. The similarity measure usually employed is called the *cosine similarity* or *cosine score*, and is calculated by first binning the spectra into vectors of mass and intensity tuples, then normalising intensities and calculating the dot product. As such, the cosine similarity is between zero and one for orthogonal and identical vectors respectively.

Over time, modifications to the cosine score have been adopted to combat a major problem: A spectrum often contains of only few high-intensity peaks, and many more low-intensity peaks. Cosine scores as calculated above would be dominated by only a few peaks in the spectrum, and fail to take smaller peaks into account. Transforming peak intensities by taking the square root alleviates this problem, and often shows better performance [73, 156]. Additionally, the precursor peak is often removed from both spectra.

While calculating the cosine score is generally simple, interpreting it is often not. As long as the spectral library contains spectra with a similar enough precursor mass, a best matching spectrum is always going to be returned. Choosing a cosine score threshold for which spectral matches are to be considered significant however, often involves extensive knowledge about the currently measured experimental data, as well as the spectral library. Scheubert et al. [141] recently developed a false discovery estimation approach for spectral library search to tackle the problem of false positive hits. One of the main ideas relies on the target-decoy approach, which is commonly used in proteomics. Here, a second spectral library is constructed *in-silico*, containing artificial spectra that cannot be produced by real-world compounds, but are still very similar to the ones contained in the target spectral library. Both databases are then searched for spectral matches to estimate the probability

that a hit in the target database is in fact a false positive (target-decoy).
Publicly accessible data repositories for MS/MS data include MassBank [73], HMDB [192] and GNPS [177]. GNPS in particular allows the upload of uncurated datasets by its users, combined with analysis workflows to aid in structure elucidation [117]. A visualisation technique that gained a lot of popularity in recent years involves the generation of molecular networks [109, 177, 182]. These networks show the relationship of compounds in one or multiple datasets, by creating a graph in which compounds are connected by edges denoting similarity (usually via cosine score). This helps to identify compound cluster in the dataset, and creates knowledge about biochemical reactions in the sample, as for example simple biotransformations can be hypothesised about using mass differences [168]. The process of using known spectral library hits in a molecular network to annotate unknowns or validate annotation results of *in-silico methods* is called *network propagation* [36, 94, 152]. Common spectral patterns between entries in a spectral library can be extracted by MS2LDA [169, 170], an unsupervised machine learning method, and manually annotated with structural properties. This can aid in partially annotating query spectra, that show such a known pattern.

While spectral library search is a powerful tool, it also comes with limitations. First, every compound contained in a spectral library needs to be measured in its pure, uncontaminated form to create a “reference standard” [149]. Depending on the compound, this process can be extremely cost- and labor intensive, or straight up impossible with current experimental methods. Despite spectral libraries growing over the past years [73, 163, 174, 177, 192], compared to the realm of all known molecular structures (PubChem for example contains over 110 million compounds), they are still vastly incomplete. As a result, the correct structure candidate for a query spectrum is not contained in the spectral library for a large majority of detected compounds [35]. In this thesis, we are mostly picturing structure library search as superior over spectral library search, however we want to stress that both approaches have their merits and should not be seen as mutually exclusive.

4.5 Structure Database Search

As mentioned in the last section, databases containing structural information of compounds are many orders of magnitude larger than spectral databases. PubChem [88], one of the largest publicly available structure databases, contains 110 million compounds of 277 million substances (as of January 2022). Even though the majority of these compounds are not relevant in the context of metabolomics (because they don’t naturally appear in biological samples), the subset of biologically relevant compounds still contains hundreds of thousands of structures [10, 47]. Obviously the limitation of not being able to elucidate truly novel structures not contained in the database also applies here. It can be tackled by generating *in-silico* databases containing artificially generated structures [27, 40, 62, 76, 128]. Since it is much easier to artificially create molecular structures than it is to generate sensible spectra, this approach is much more feasible for structure library search.
The main problem to overcome when attempting mass spectrometry driven structure library search is, that measured spectra and molecular structures are not inherently comparable, and thusly need to be transformed. We differentiate between the following thee approaches to this problem [14]:

4.5.1 MS/MS Spectra Simulation for Database Structures

Transforming structures into spectra or alternatively spectra into structures are two intuitive approaches to enable comparability. Here, we first describe the idea of predicting or simulating a MS/MS spectrum for a given molecular structure. This simulated spectrum can then be matched against a spectral library as described earlier. Since the fragmentation process of molecules in mass spectrometry is far from fully understood, creating models that accurately produce a theoretical MS/MS spectrum for a compound is very challenging. A relatively simple approach is called “rule-based”, and describes the process of specific, well-defined fragmentation rules to a given structure. These rules are generated from the literature, and are generally highly curated. This approach is severely limited by the amount of publicly observed and understood fragmentation processes and additionally can not produce relative peak intensities [83]. “Mass Frontier” and “ACD/MS Fragmenterm” are examples for commercial rule-based fragmentation tools, however the produced spectra are usually not suited for comparison to measured spectra because of the missing peak intensities [83]. Quantum-computing-based cheminformatics approaches which try to predict the MS/MS spectrum *ab initio* exist [179], but require tens to hundreds of CPU hours for the prediction of one spectrum. Additionally, the predicted spectrum showed no pronounced advantage in comparison to CFM-ID, which is also faster [154]. CFM-ID (competitive fragmentation modelling), a machine learning and stochastic-based method, models the fragmentation process using Markov-Chains [1, 2, 41]. Bonds or rings in the molecule’s structure can be broken or disconnected depending on chemical properties that describe the bonds strength. CFM-ID utilises a neural network for its predictions. Although the calculated MS/MS spectra are not extremely similar to the experimentally produced ones, they still perform well for the task of molecular structure elucidation.

4.5.2 Combinatorial Fragmentation of Structures

Combinatorial fragmentation describes the process of generating hypothetical fragments from a structure by bond breakage, hydrogen rearrangements or other rule sets. The resulting fragments are then used to annotate peaks of a measured query spectrum. While this method was initially used to annotate the fragmentation of known compounds [66, 69], MetFrag [140, 194] was the first method to use combinatorial fragmentation to annotate any given spectrum by searching in a structure database. Since version 2.2 [140], MetFrag also uses metascores to influence structure candidate rankings, see Section 4.8 for an introduction to metascores. The two main components of combinatorial fragmentation annotation methods is the efficient generation of fragments and the scoring function used to compare them to a measured query spectrum. The generation of fragments is usually done non-exhaustive and requires some thought as to speed up computation times. It is easy to understand that multiple different sets of successive fragmentations can lead to the same fragment, which is why MAGMa [132] hashes previously generated fragments to avoid duplicity. Scoring of generated fragments is usually done by assigning a cost to bond cleavages, depending on the bond’s type. MetFrag uses bond dissociation energies while MAGMa and MIDAS [181] use a simple predefined bond-cost model. MetFrag was later combined with MassBank [73] spectral library search by Gerlich et al. in the form of MetFusion [59], which aimed to combine the advantages of both database searches. MAGMa was further improved by Verdegem et al. by creating MAGMa+ [173], a wrapper script for MAGMa that optimises the parameters used in the scoring function based on the current query using machine learning. MS-FINDER [166], developed by Tsugawa et al., introduced rules for hydrogen rearrangement during fragmentation and also uses the amount of database occurrences of a structure candidate as a metascore. DEREPLICATOR+ [106], an extension of DEREPLICATOR [105], restricts the set of bonds that can be fragmented to C-C, N-C and O-C, and more importantly reports a false discovery rate estimation for annotations. This estimation is target-decoy-based, decoy spectra are generated with randomly sampled peaks from a reference spectra library. It is important to note however, that the estimated FDRs that are reported were not evaluated on their ability to reflect the true FDR. MolDiscovery [30] was introduced as a successor to DEREPLICATOR+, which included a more efficient algorithm for fragmentation graph construction.

4.5.3 Structural Feature Prediction from MS/MS Spectra

While the previous approaches were based on transforming structural information to a representation more comparable to a spectrum, here are others that try to predict the structure directly from the MS/MS spectrum. This would alleviate the need for even a structure database, and enable truly novel annotations. However, this task is extremely difficult, and so far no existing method is able to fully *de novo* predict a structure directly from an MS/MS spectrum [38]. Instead, predefined structural properties are being predicted, in the form of molecular fingerprints. These fingerprint vectors can be examined for the presence or absence of certain functional groups, structural backbones and other biochemical properties which can aid experts even if a full elucidation is impossible. Additionally, calculating the molecular fingerprint for structures in a structure database is trivial, allowing for a comparison of the predicted fingerprint with fingerprints of known structures. FingerID was developed in 2012 by Heinonen et al. [67], and uses kernel-based support vector machines (SVMs) to predict molecular fingerprints for MS/MS spectra. Later, Shen et al. combined fragmentation tree kernels with the probability product spectrum kernels used in FingerID [150]. Afterwards, Dührkop et al. improved the scoring function and extended the space of molecular properties that could be predicted, resulting in CSI:FingerID, which is widely used today [46]. SIMPLE [114] is also based on multiple kernel learning, and additionally promises higher classifier interpretability. Brouard et al. developed Input Output kernel regression (IOKR) [24–26], which is based on the same principle, but directly learns a mapping from fragmentation spectra to structure without explicitly predicting a fingerprint. Spectra and structures are mapped directly instead, based on a learned mapping function. In a similar fashion, ADAPTIVE [115] employs message passing neural networks to learn a mapping from structure to molecular vector, and then deploys IOKR to learn a mapping from spectra to molecular vector. ChemDistiller [93] combines combinatorial fragmentation and fingerprint-based predictions to increase annotation rates. MSNovelist [158] uses fingerprints predicted by CSI:FingerID as input for a recurrent neural network (RNN). It then predicts structures in the form of SMILES, this is done *de-novo* and shows promising results in evaluation.

4.6 CSI:FingerID

As introduced in the previous section, CSI:FingerID is a method that predicts molecular fingerprints from input MS/MS data, and compares them to molecular fingerprints of structure database candidates. Combined with SIRIUS (see Section 4.3.2), it uses combinatorial optimisation as well as machine learning to elucidate an unknown compound's structure. First, the MS/MS spectrum is transformed into a fragmentation tree. Then, kernel SVMs predict the presence or absence of predefined molecular properties of the molecular fingerprint. The calculated fragmentation tree serves two purposes: Information contained in it can be used as input for the SVMs predicting the fingerprint in the form of kernels. Multiple kernels are computed based on spectral information as well as information about edges, nodes and paths in the fragmentation tree. Then, these kernels containing (partially) orthogonal information about the spectrum are combined into a single kernel using multiple kernel learning. Dührkop et al. used ALIGNF [34] and ALIGNF+ [151] to compute kernel weights. An individual SVM is then trained for each molecular property of the molecular fingerprint, however, all SVMs use the same combined kernel. The second benefit of calculating the fragmentation trees for a spectrum, is to restrict the search space of the structure database as described below.

Structure database search

Without additional information about the query compound, every method comparing the input spectrum (or in this case the predicted molecular fingerprint) to candidates in a structure database, would need to perform this comparison for each candidate in the database. Depending on database size (PubChem for example contains over 100 million structures) and the complexity of the scoring function, this can lead to strenuous computation times. If the molecular formula of the query compound is known however, we can restrict our search to only the subset of structure candidates sharing this formula. Molecular formula candidates can be calculated using SIRIUS and ranked by score. We can then perform structure database search for each (sufficiently high-ranked) candidate individually. As mentioned earlier, computing the (binary) molecular fingerprint for a given structure is trivial.

Over time, different scoring functions that compare predicted fingerprints to database fingerprints have been proposed. In this thesis, we use the two latest scoring functions that were introduced by Dührkop et al. [46] and Ludwig et al. [100] respectively. The "Modified Platt scoring" from [46] combines Platt probabilities [126] with sensitivity and specificity of the predictors for the individual molecular properties.

Let $F = (x_1, ..., x_n) \in \{0,1\}^n$ be a binary candidate fingerprint present in the structure database and let $F' = (p_1, ..., p_n) \in [0,1]^n$ be the Platt probabilities of the predicted molecular fingerprint. Additionally, $sens_i$ shall denote the sensitivity and $spec_i$ shall denote the specificity of the predictor of the ith molecular property. Then, the Modified Platt score between the two fingerprints is given as:

$$\prod_{i=1}^{n} \begin{cases} (1-p_i)^{\frac{3}{4}} \cdot (1-spec_i)^{\frac{1}{4}} & \text{if } p_i < 0.5 \text{ and } x_i = 0 \\ p_i^{\frac{3}{4}} & \text{if } p_i < 0.5 \text{ and } x_i = 1 \\ (1-p_i)^{\frac{3}{4}} & \text{if } p_i \geq 0.5 \text{ and } x_i = 0 \\ p_i^{\frac{3}{4}} \cdot (1-sens_i)^{\frac{1}{4}} & \text{if } p_i \geq 0.5 \text{ and } x_i = 1 \end{cases} \tag{4.1}$$

Compared to the Platt scoring, the Modified Platt scoring offers no statistical interpretation, but outperformed previous scorings [46]. The Modified Platt scoring assumes independence between the molecular properties of a fingerprint, which is an assumption that does not hold. A trivial example to give would be that a structure encoded as a molecular property might be a substructure of a different molecular property. Thus, these positions in the fingerprint can not be independent.

The “Covariance score” introduced by Ludwig et al. in [100] is based on Bayesian networks, and assumes dependence between fingerprint positions. Nodes in the Bayesian network are molecular properties represented as binary random variables that encode presence and absence of the property. To keep the problem computationally tractable, the network topology is forced to be tree based. The tree topology is then derived based on mutual information between molecular properties, using deterministic “anchor” fingerprints from a structure database. Furthermore, covariances between nodes are computed as follows. Let X, Y be binary random variables denoting molecular properties, then the covariance is defined as:

$$\begin{aligned} cov(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{P}(X,Y) - (\mathbb{P}(X) \cdot \mathbb{P}(Y)) \end{aligned} \tag{4.2}$$

For a tree T, a root node r and a set of edges E, the joint distribution can be given as:

$$\mathbb{P}(X_1, ..., X_n) = \mathbb{P}(X_r) \cdot \prod_{i,j \in E} \frac{\mathbb{P}(X_i, X_j)}{\mathbb{P}(X_i)} \tag{4.3}$$

The marginal probabilities $\mathbb{P}(X)$ and $\mathbb{P}(Y)$ are known from the Platt probability estimate. In the scoring phase, joint probabilities $P(X,Y)$ are computed for each edge, using both the covariance as well as the marginal probabilities from the predicted fingerprint. Ludwig et al. suggest either using one global tree structure for all queries and fingerprints, or determining an individual tree structure for each different molecular formula. Both approaches outperform the Modified Platt scoring on cross-validated training data. Using individual trees, this new scoring outperforms the Modified Platt scoring on independent data. See Fig. 4.2 for an overview of the CSI:FingerID workflow.

4.7 Compound Class Prediction

While molecular structure databases are orders of magnitude larger than most spectral libraries, they still do not cover all biomolecules that may be contained in a biological sample. Complete coverage of the biochemical space by these libraries is not something that is expected within the intermediate future, if at all. Consequently, many compounds in a sample can not be fully structurally elucidated by spectral library search or structure library search alike. Instead, one can focus on determining the *compound classes* contained in a sample. Compound classes are chemical classifications that go beyond the simple presence or absence of certain substructures like in molecular fingerprints. While knowing the class of a compound entails less information than knowing its full structure, it still contains valuable information like general chemical properties or certain substructure information. Note that a compound can and in most cases will belong to multiple chemical classes. Databases like ChEBI [64] or the MeSH thesaurus [135] contain compound class annotations for a rather small fraction of structures. To assign classes to all structures independently of their presence in these databases, ClassyFire [39] deterministically assigns them based on various substructure features and logical expressions. In 2021, Dührkop et al. presented CANOPUS [48], an approach to compound class assignment directly from LC-MS/MS data that is implemented into SIRIUS. CANOPUS uses the molecular fingerprint predicted by CSI:FingerID as input for a Deep Neural Network, which then predicts the compound classes for a given query spectrum. Dührkop et al. show, that CANOPUS outperforms the following other methods for compound class assignment: *Direct prediction* [167], which describes the one-step prediction of a compound class from the MS/MS data without the intermediate step of fingerprint prediction and *k-Nearest Neighbour* for either spectral library [6, 99] or structure database search [52, 167].

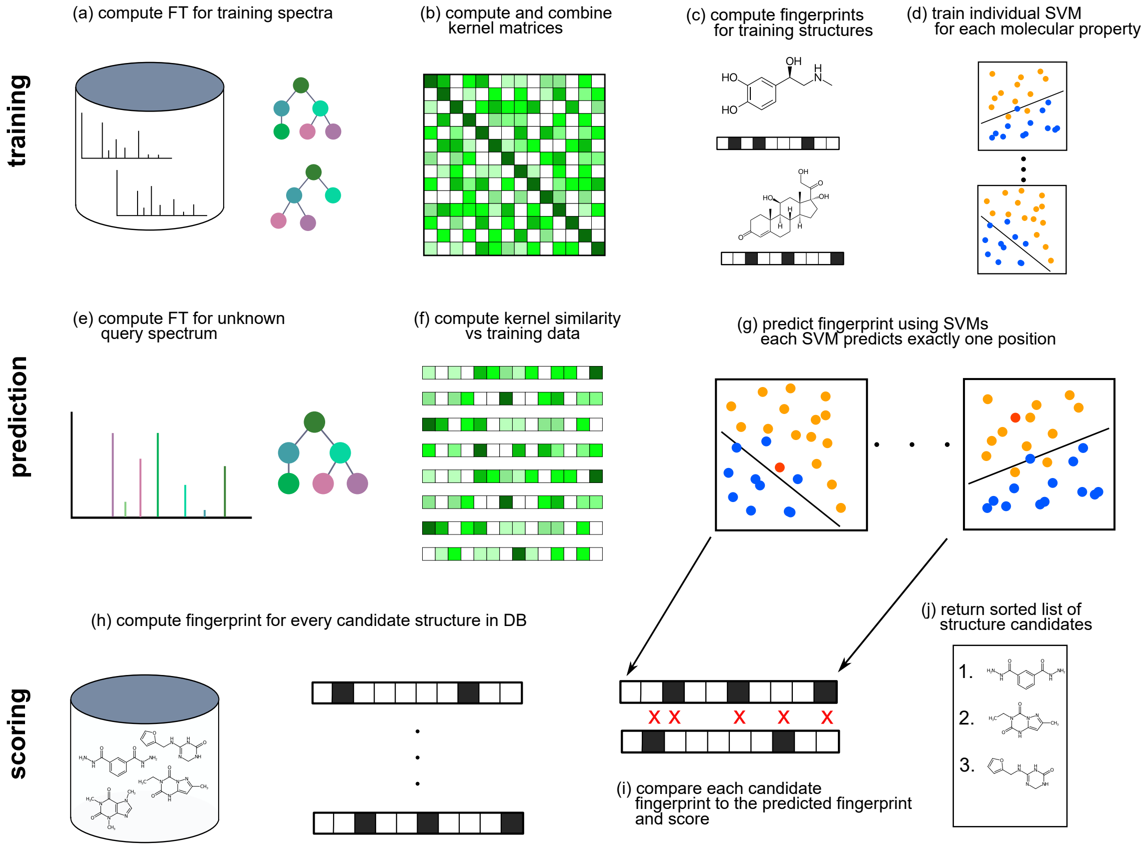

(a) compute FT for training spectra
(b) compute and combine kernel matrices
(c) compute fingerprints for training structures
(d) train individual SVM for each molecular property
training
(e) compute FT for unknown query spectrum
(f) compute kernel similarity vs training data
(g) predict fingerprint using SVMs each SVM predicts exactly one position
prediction
(h) compute fingerprint for every candidate structure in DB
(i) compare each candidate fingerprint to the predicted fingerprint and score
(j) return sorted list of structure candidates
1.
2.
3.
scoring

Figure 4.2: The CSI:FingerID workflow. In the training phase, fragmentation trees are constructed for each MS/MS spectrum in the training dataset (a). Then, all-against-all kernel matrices are computed and combined into a single kernel (b). Next, fingerprints are computes for the molecular structures corresponding to the the training spectra (c). For each position in the molecular fingerprint, an individual SVM is trained (d). In the prediction phase, the fragmentation tree for an unknown query spectrum is computed (e), as well as kernel similarities to the training data (f). The SVMs trained in the training phase now predict the molecular fingerprint of the query (g). Finally, in the scoring phase, fingerprints are computed for each structure candidate in a structure database (h). Every candidate is then compared and scored to the predicted fingerprint (i), and a sorted list of structure candidates is returned (j).

4.8 Metascores

In previous sections, we introduced various methods to compare spectral information to structural information, which use scoring methods to rank structure candidates based on this comparison. The term "Metascore" describes such a scoring function that not only takes into account the measured spectral data, but uses additional, non-experiment related information to rank structure candidates. Prominent examples of this are the citation frequency of a structure candidate or its production volume [96]. The underlying principle here, is the assumption that frequently cited compounds are more likely to be present in a given sample compared to lesser cited compounds. One should be aware of the implications that using such a metascore brings. First, by using a metascore on for example the citation frequency, a method is practically blind to novel or very rarely occurring structures if another candidate that is highly cited exists. Secondly, evaluation results can be misleading. In contests like CASMI [148], compounds which query spectra are presented to the different tools need to be somewhat easily obtainable and have their molecular structure known. As a consequence, these structures are often highly cited which inflates annotation rates for tools employing metascores.

We want to make the reader especially aware of the fact, that *metadata* is very different from metascores, and that these two should not be mistaken for one another. Metadata includes for example the experimental parameters used, like the MS instrument type or the column type used in chromatography. This kind of information is experiment related, and can be very helpful for annotating structures, for example by estimating the mass accuracy of an instrument by its model.

5 Datasets, Databases and Evaluation Principles

In this chapter we introduce the datasets and databases used in evaluations throughout this thesis. Additionally, we describe in details how the CSI:FingerID version used in this thesis was trained, as most of our work is based on it.

5.1 MS/MS Reference Datasets and Noise Addition

For evaluations, we limited ourselves to MS/MS spectra recorded in positive ion mode, as there are generally more such spectra available. This is not a restriction of COSMIC, and the publicly available version can also process negative ion mode data. Evaluations were carried out using reference measurements, as we do not know the correct answers for biological datasets.

For the CASMI 2016 evaluation, MS/MS spectra were downloaded from the CASMI web page (`http://casmi-contest.org/2016/`). MS/MS spectra were measured on a Q Exactive Plus Orbitrap (Thermo Fisher Scientific, Bremen, Germany) with 20/35/50 HCD nominal collision energies. Twenty-two mixes of synthetic standards were measured in one LC-MS run each, using data-dependent acquisition mode and inclusion lists. Each mix contained between 10 and 94 compounds. A reversed phase C18 column was used. See [148] for details. In full, MS/MS data of 127 compounds measured in positive ion mode were provided as part of the contest. Fragmentation spectra from different collision energies were merged.

To train CSI:FingerID, we used a combined dataset from MassBank [73], GNPS [177] and the NIST 2017 database (National Institute of Standards and Technology, v17). Reference MS/MS data were measured on different high-resolution instruments from multiple vendors. The *CSI training dataset* contains 16,703 structures with 23,965 independent MS/MS measurements. Some additional experiments experiments in this thesis were conducted using a newer version of CSI:FingerID. This version was trained on the *New CSI training dataset*, containing 26,432 independent MS/MS measurements of 18,970 structures. Experiments performed on this new dataset are labeled as such. As an independent dataset, we used the commercial MassHunter Forensics/Toxicology PCDL library (Agilent Technologies, Inc.) with 3,243 structures and 3,462 independent MS/MS measurements, all measured on an Agilent QTOF instrument. Unlike the commercially available library, these mass spectra were not curated. When discussing reference dataset evaluations, independent MS/MS measurements will be referred to as “compounds” for the sake of brevity.

Previous evaluations of CSI:FingerID [46, 47] were carried out using fragmentation spectra that merged all available collision energies. Here, we also want to evaluate COSMIC’s power if query spectra are recorded at a single collision energy, since LC-MS/MS datasets are often recorded in this way. To this end, we compiled fragmentation spectra sets for both training and independent data using single collision energies, namely, 10 eV, 20 eV and 40 eV. To ensure that COSMIC results are comparable between different collision energies, we only used those compounds for which all three collision energies are available. In the independent data, this is the case for all compounds; but in the training data, only NIST entries pass this criterion. Hence, the COSMIC training dataset exclusively contains spectra from NIST, all of which were measured on an Orbitrap instrument; and consequently, all cross-validation results on this dataset exclusively use MS/MS data from Orbitrap instruments. In case the NIST library did not contain fragmentation spectra for the exact collision energies 10 eV, 20 eV and 40 eV, we allowed for a deviation of up to 4 eV; in case fragmentation spectra for more than one collision energy were present in this interval, we used the one with collision energy closest to the desired one. Finally, *merged spectra* were generated by combining these three spectra (pseudo-ramp spectra).

Fragmentation spectra in reference libraries often have much better quality (more signal peaks, fewer noise peaks, better signal-to-noise) than fragmentation spectra from a biological LC-MS/MS run. To simulate this effect in our reference datasets, we "added noise" to each fragmentation spectrum. Distorting spectra followed comparable principles as the generation of decoy spectra [141]: We distorted spectra similar to what we expect for experimental spectra. For example, adding noise peaks with (uniform) random mass will result in spectra which are notably different from experimental ones [141]. We simulated two noise models, *medium noise* and *high noise.*

- We simulated a *global mass shift* (bias) by drawing a random number δ^* from $\mathcal{N}(0, \sigma_{\mathrm{mb}}^2)$, then shifting every peak mass m by $\delta^* m$. The standard deviation σ_{mb} was chosen as $\sigma_{\mathrm{mb}} = (10/3) \cdot 10^{-6}$ (medium noise) or $\sigma_{\mathrm{mb}} = (15/3) \cdot 10^{-6}$ (high noise), so that the $3\sigma_{\mathrm{mb}}$ interval represents a 10 ppm shift for medium noise and a 15 ppm shift for high noise.

- We simulated individual *mass deviations* by drawing, for each peak with mass m individually, a random number δ from $\mathcal{N}(0, \sigma_{\mathrm{md}}^2)$ and shifting the peak by $\delta\, m$. The standard deviation σ_{md} was chosen so that the $3\sigma_{\mathrm{md}}$ interval represents a 10 ppm shift for medium noise and a 20 ppm shift for high noise.

- We simulated intensity variations in the spectrum: Each peak intensity was multiplied by an individual random number ϵ drawn from $\mathcal{N}(1, \sigma_{\mathrm{id}}^2)$. Variance was chosen as $\sigma_{\mathrm{id}}^2 = 1$ for medium noise and $\sigma_{\mathrm{id}}^2 = 2$ for high noise. Furthermore, 0.03 times the maximum peak intensity of the spectrum was subtracted from each peak intensity. If a peak intensity fell below the threshold of one thousands of the maximum intensity in the spectrum, the peak was *discarded.*

- Finally, we added "noise peaks" to the spectrum. As uniformly choosing the mass of a noise peak would result in peaks which are too easy to spot and sort out by our subsequent analysis [141], we instead used peaks that appeared in other measured spectra. In preprocessing, a pool of "noise peaks" was gathered from the fragmentation spectra, using all peaks that did not have a molecular subformula decomposition of the known molecular formula of the precursor. For each spectrum, $\alpha\, n$ of these "noise peaks" were added to the spectrum, where n is the number of peaks in the spectrum and $\alpha = 0.2$ for medium noise, $\alpha = 0.4$ for high noise. Intensities of "noise peaks" were adjusted for maximum peak intensities in the contributing and receiving spectrum.

Parameters for medium noise and high noise were chosen in a way that the similarity between the original spectrum and the distorted spectrum reached a particular level, measured by the cosine score (dot product): For the cosine score, we allowed a mass deviation of 7 ppm when matching peaks. Precursor ion peaks were not considered for cosine score calculation, as their high intensities overshadow the lower intensity peaks. For medium noise, the cosine score between the original and the distorted spectrum had median value 0.880. For high noise, the median cosine score was 0.714. Datasets with different noise level were used for evaluations only, but not to train CSI:FingerID or individual confidence score SVMs.

Adding noise to the fragmentation spectra may result in an empty or almost empty spectrum, which would be regarded insufficient for structure annotation in applications. To this end, we removed fragmentation spectra with at most one peak. To ensure that evaluation results are comparable between collision energies and noise levels, we discarded the compound from all libraries if a fragmentation spectrum with at most one peak resulted for at least one collision energy and noise level. Doing so, 3,314 compounds were removed from the COSMIC training dataset and 171 compounds from the independent dataset. Substantially more compounds were removed from the COSMIC training dataset because many training dataset spectra have only few peaks, increasing chances that noisy spectra contain at most one peak. Here, 10 eV noisy spectra contain at most one peak for 75 % of the 3,314 removed compounds; 20 eV noisy spectra for 27 %; and 40 eV noisy spectra for 11 % (a compound can exhibit sparse spectra for more than one collision energy).

This resulted in eight libraries, four libraries with 4,046 compounds each for the *COSMIC training dataset*, and four libraries with 3,291 compounds each for the *independent dataset*. Notably, the COSMIC training dataset is a proper subset of the CSI training dataset; if we simply speak about "training data", this refers to the full CSI training dataset and includes the COSMIC training dataset. Recall that the COSMIC training dataset contains Orbitrap MS/MS data only, whereas the independent dataset contains QTOF MS/MS data only.

5.2 Biological Datasets

- For the *mice fecal dataset*, we analysed LC-MS/MS data of 278 samples from a public metabolomics dataset (MassIVE data repository, id no. MSV000082973). This dataset comes from a previously published study [165]; LC-MS/MS experiments were conducted on a Q Exactive Orbitrap instrument (Thermo Fisher Scientific, Bremen, Germany). See [70] for more technical dataset details.

- For the *human dataset*, we analysed ten MassIVE datasets from the MassIVE data repository (id nos. MSV000083559, MSV000079651, MSV000080167, MSV000080469, MSV000080533, MSV000080627, MSV000081351, MSV000082261, MSV000082629, MSV000082630). The dataset contains fecal, plasma, urine, lips, tongue and teeth samples from humans; all acquired on Q Exactive Orbitrap instruments (Thermo Fisher Scientific, Bremen, Germany) in positive ion mode. Runs were acquired using C_{18} RP Ultra High Performance Liquid Chromatography (UHPLC). Only files with extensions ".mzML" or ".mzXML" were considered, and LC-MS runs containing spectra in profiled mode were discarded. This resulted in 2,666 LC-MS/MS runs being processed.

- For the *Orbitrap* dataset, we followed the idea of “flipping the workflow” and reanalysing public data at a repository scale: We restricted ourselves to MassIVE datasets measured on a Q Exactive Orbitrap instrument (Thermo Fisher Scientific, Bremen, Germany), as this instrument had the largest number of MassIVE datasets. We applied no other constraints with regards to analysed organism, LC setup etc, resulting in 264 public MassIVE datasets (downloaded Feb 20, 2020). MassIVE datasets containing only spectra in profiled or negative ion mode were discarded, leaving us with 123 MassIVE datasets. Sample types range from environmental to natural products and include biological samples from at least 30 different species, covering diverse genera and phyla. Only files with extensions “.mzML” or “.mzXML” were considered, and LC-MS/MS runs containing spectra in profiled or negative ion mode were discarded, leading to 17,414 LC-MS/MS runs being processed. See Table A.1 for a list of all MassIVE datasets.

5.3 Structure Databases

Different from previous studies [46, 47] where structures were derived from InChI (International Chemical Identifier) strings, molecular structures were standardised using the PubChem standardisation procedure [88]. In particular, a canonical tautomeric form was chosen, as solvent, temperature, and pH in the sample influence the dominating tautomeric species. Standardisation of compounds not in PubChem was carried out using the web service at `https://pubchem.ncbi.nlm.nih.gov/rest/pug/`. Unfortunately, PubChem standardisation has changed multiple times over the last years without further noticing of users; to this end, it is possible that some non-PubChem compounds were standardised slightly differently than structures from the MS/MS training data.

We searched in the following structure databases with COSMIC:

- For the CASMI 2016 evaluation [148], we downloaded structures from the CASMI 2016 results web page (`http://casmi-contest.org/2016/`). Candidate structures were provided as part of the blind contest and originally retrieved from ChemSpider [124].

- The *biomolecule structure database* is a union of several public structure database including HMDB [192], ChEBI [64], KEGG [81, 82] and UNPD [61]. The resulting database contains 391,855 unique structures of biomolecules and compounds that can be expected to be present in biological samples.

- The *HMDB structure database* [192] was downloaded Aug 8, 2018 and contains 113,983 compounds, and 95,980 unique structures with mass up to 2000 Da.

- The *PubChem structure database* [88] was downloaded Jan 16, 2019, and contains 97,168,905 compounds, and 77,153,182 unique covalently-bonded structures with mass up to 2000 Da. We added all missing structures from the biomolecule structure database, which resulted in a total of 77,190,484 unique structures.

- A combinatorial database of 28,630 *bile acid conjugate structures* was generated with SmiLib v2.0 [145, 146], downloaded from `http://melolab.org/smilib/`. SmiLib generates chemical structures by combining scaffolds and building blocks provided as SMILES (Simplified Molecular Input Line Entry Specification). A list of initial bile acid "scaffolds" that represent common steroid cores (i.e. cholic acid, deoxycholic acid, hyocholic acid, chenodeoxycholic acid) was curated.These scaffolds were modified manually with common phase II metabolism reactions and resulted in 322 scaffolds. Scaffolds were combined with 91 building blocks, including proteinogenic and non-proteinogenic amino acids, along with their N-hydroxylated and N-methylated version, and acyls moieties. Stereochemical information was removed prior to the database generation with SmiLib. Notably, the bile acid conjugate structure database also contains unconjugated bile acids; for the sake of brevity, we will nevertheless speak about "bile acid conjugates" without explicitly mentioning this fact. The bile acid conjugate database was designed and generated by Louis-Felix Nothias

5.4 Training CSI:FingerID and Structure–disjoint Evaluation

We trained an array of Support Vector Machines (SVM) for fingerprint prediction from MS/MS data as described in [46, 47, 150]. Training of CSI:FingerID was carried out using merged spectra with all available collision energies from the CSI training dataset. In contrast, single collision energy and merged spectra libraries as well as noisified spectra were not used when training CSI:FingerID, but only in validation of COSMIC. We used PubChem-standardised structures [63] when computing the molecular fingerprint of a compound. In evaluations, we used the CSI:FingerID "Covariance score" from [100] to rank candidates, comparing the probabilistic query fingerprint and each structure candidate fingerprint. A hit was regarded as *correct* if the PubChem-standardised structures of query and top rank were identical.

As noted above, all evaluations were carried out structure-disjoint. For the tenfold cross-validation, we partitioned the training data into ten disjoint *batches* of almost identical size, ensuring that all fragmentation spectra from compounds with identical structure (such as L-threose and D-erythrose) *end up in the same batch*: Otherwise, L-threose could be part of the training data when evaluating on D-erythrose, and vice versa. For each batch, we trained the fingerprint SVM array using the remaining nine batches; we evaluated on the tenth batch. In this way, we ensured that all compounds are novel for CSI:FingerID: For each query, MS/MS training data for the corresponding structure, including independent MS/MS measurements, were not available for CSI:FingerID.

CSI:FingerID evaluations on the independent dataset were again executed structure-disjoint: We additionally trained an SVM array using the complete CSI training dataset. Given an MS/MS query from the independent data, we checked if the structure of the query is also part of the training data: If so, we used the appropriate SVM array from cross-validation for fingerprint prediction; otherwise, we used the SVM array trained on the complete training data. Again, this ensured all structures being novel in evaluation. Most evaluations in this thesis and in [70] were performed with CSI:FingerID version 1.2.0. Evaluations that were performed with a newer version of CSI:FingerID are labelled as such.

6 COSMIC

In this chapter, we introduce the COSMIC (Confidence Of Small Molecule IdentifiCations) confidence score for CSI:FingerID. Its purpose is to assign confidence to structure annotations of LC-MS/MS data given by CSI:FingerID. For that, we computed E-values, which were then integrated as features into a simple support vector machine model. Our approach is inspired by Percolator [78, 155], which is used in Proteomics.

In the next section, we are first going to introduce the role that score separation and false discovery rate estimation play in computational metabolomics. We then establish that the scoring models used in widely adopted structure elucidation tools are unfit to separate correct and incorrect annotations. Next, we show the development of the COSMIC confidence score, and evaluate it extensively. Our results show, that our model outperforms aforementioned scoring models by a wide margin for this specific task. We establish this for different noise levels and collision energy settings of the input data, to more accurately reflect real-world settings.

6.1 False Discovery Rates in Computational MS

False discovery rate (FDR) control plays an important role in computational mass spectrometry, where tens of thousands of spectra are annotated with structure candidates in a single analysis run. Naturally, a large chunk of these annotations are incorrect due to various reasons, leaving researchers in need of some metric by which they can decide on those annotations, that are trustworthy enough to further analyse. FDR control allows us to focus on sublists of annotations that contain a controlled ratio of incorrect annotations (false positives). Since a true FDR can only be calculated when the ground truth about the query spectrum's structure is known, in practice we are looking for a way to **estimate** this FDR. Estimating FDRs is a task that is present in many parts of computational mass spectrometry, in proteomics and metabolomics as well as in spectral and structure library search. Depending on the biological data that is being analysed (e.g. peptides vs metabolites), as well as data representations (e.g. spectra vs fingerprints), accurate estimation of FDRs can be of highly varying difficulty.

6.1.1 FDR Estimation in Proteomics

Since peptides are linear combinations of amino acids, which is a relatively simple building block structure, FDR estimation in proteomics is well established. A commonly used approach is called "target-decoy", and involves the construction of a decoy database which is then used to estimate the amount of false positives [32, 51, 80, 108]. A decoy database has to adhere to certain properties to be effective. In short, all candidates contained have to be incorrect candidates, but they still have to be sufficiently similar to potentially correct ones. Effectively, one has to create "convincing fake" peptides to populate the decoy database, which is commonly done by e.g. inverting the peptide sequence of entries in the target database. Searching a query in either both target and decoy database separately or concatenating them first allows for FDR estimation based on the number of decoy hits.

Percolator
Percolator [78, 155] is a machine learning-based approach, that works as a post-processor of the target-decoy approach. A support vector machine is trained to differentiate between correct and incorrect peptide-spectrum-matches (PSMs). Hereby, the PSMs derived from the decoy database search are used as negative training samples, and high-scoring PSMs from the target database as positive training samples. Notably, the classifier is re-trained for each individual dataset which removes the need for a well generalising model.

6.1.2 FDR Estimation in Metabolomics

In contrast to peptides, most metabolites do not possess a linear building block structure. In fact, they arguably don't consist of building blocks at all. This makes FDR estimation significantly harder, since there is no equivalent of reverting a peptides amino acid sequence for metabolites. As a result, creating decoy spectra or decoy structures to import the target-decoy approach from proteomics is highly non-trivial. For spectral library search, Scheubert et al. [141] developed a method to create high quality decoy spectra based on fragmentation trees. For structure library search however, the challenge of creating sensible decoy structures or fingerprints remains unsolved. In the following, we are focusing on FDR estimation for structure library search specifically.

6.2 A New task - Score Separation

In this section we are going to introduce the concept of *score separation*, and how it differs from the traditional task of annotation. Methods in computational mass spectrometry that annotate a structure to an MS/MS input spectrum, all require a scoring function, that is used to rank potential structure candidates for a singular query. The more often the correct candidate is ranked at the top of the list, the better the score's performance. This task is called the *annotation task*. We have introduced multiple examples in previous sections, such as the scoring functions for MetFrag, MAGMa+ and CSI:FingerID. These scoring functions are evaluated based on their ability to correctly annotate, for example in the CASMI contests [147, 148]. The annotation performance of a scoring function is designed to differentiate *locally* between structure candidates for a specific query spectrum. As such, the score's ranges can vary greatly between instances, and not be comparable *globally* between instances.
CSI:FingerID's performance for the annotation task lies between 45%-75% correct annotations, depending on input spectra and structure database used [46, 70, 100, 148]. While it is currently the best-in-class method for annotation, its performance is still far from a perfect, 100% annotation rate. This introduces the need for a second task, the *separation task*.

6.2.1 The Importance of a Score that Separates

An untargeted metabolomics MS/MS experiment usually contains hundreds to thousands of compounds that are in need of structural elucidation. When a researcher uses tools such as CSI:FingerID, MetFrag or MAGMa+, oftentimes more than half of annotations returned are incorrect. For experiments involving *in-silico* generated structure databases for the discovery of truly novel compounds, the portion of incorrectly annotated compounds can far exceed 90%, simply because these databases are non-exhaustive and the actual novel structure of the query compound is not contained. To give a practical example: When searching 10,000 query MS/MS spectra in a structure database, a method might return 10,000 structure annotations, of which only 10 are actually correct. The capability of separating these 10 correct annotations from the many thousand incorrect annotations is what we call *score separation.* Score separation is a global property of a scoring function: Instead of ranking candidate structures for a single query, we now want to rank the top-scoring candidates of **each** query. The top-ranked candidate together with its query fragmentation spectrum is called a **hit**; it can be either the correct candidate (correct hit) or an incorrect candidate (incorrect hit). Ideally, all the instances in which the annotation task produced a correct hit, would receive a high score, while the incorrect hits received a low score. We could then sort this list of hits by score, and choose a score threshold at which an acceptably low fraction of incorrect hits is contained in the resulting sublist (that is, a low FDR).

Separation before Estimation

Circling back to the beginning of this section, it is critical to understand that our ability to estimate the FDR is practically irrelevant if the underlying score does not sufficiently separate correct from incorrect annotations. To give an example: if scores of correct and incorrect hits were randomly ordered – by a poorly separating scoring function, with a correct annotations rate of 50% – then even if we had the perfect FDR estimation method, it would only ever return FDRs of 50% on average, which would be an accurate FDR estimate, but useless in practice. Therefore, we need to evaluate the separation ability of scoring functions used by existing structure annotation tools before we can move forward.

6.2.2 Evaluating Score Separation

When evaluating the separation performance of a score, it is helpful to first visualise the score distributions for correct and incorrect hits. For that we can plot the kernel density estimates of these score distributions. Naturally, a score that separates well would produce two distributions that overlap as little as possible. See Fig. 6.1 for exemplary score distributions of scores that do not separate well. We remind the reader that for each query we are only interested in the score of the hit (the highest scoring candidate). If one were to plot all candidates and their scores for all queries, they would naturally find themselves with a significantly larger amount of incorrect annotations than correct annotations (Fig. 6.2 (b,c)). Individually normalised, this distribution would give the idea that correct and incorrect annotations are somewhat well separated (Fig. 6.2 (a)), when in fact this is only because correct candidates receive a much higher score than randomly drawn incorrect candidates. In the context of separation, the score of a correct hit only competes with the scores of all other hits, not all candidates.

While this visualisation helps to get a initial feeling for how well a score separates, more sophisticated approaches are needed for proper evaluation. We are given a list of hits, one for each query, ordered by score. Each hit can either be *positive* (correct annotation) or *negative* (incorrect annotation). Varying a score threshold, we can modify the number of hits reported to the user; our goal is to report all positives and to reject all negatives. True positives (TP) and false negatives (FN) are positives (correct hits) which pass or do not pass the threshold; similarly, false positives (FP) and true negatives (TN) are

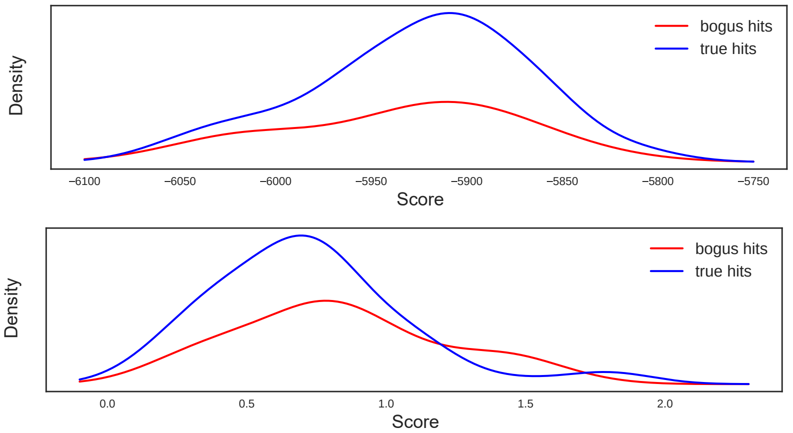


Figure 6.1: Exemplary score distributions with low separation. Positive ion mode. CASMI 2016 challenge data. Candidates retrieved by molecular formula. Searching the biomolecule structure database (n = 123 queries). Kernel density estimates of the score mixture distribution (correct and incorrect hits) for CFM-ID (a) and CSI:FingerID (b), ensuring structure–disjoint training data through cross-validation

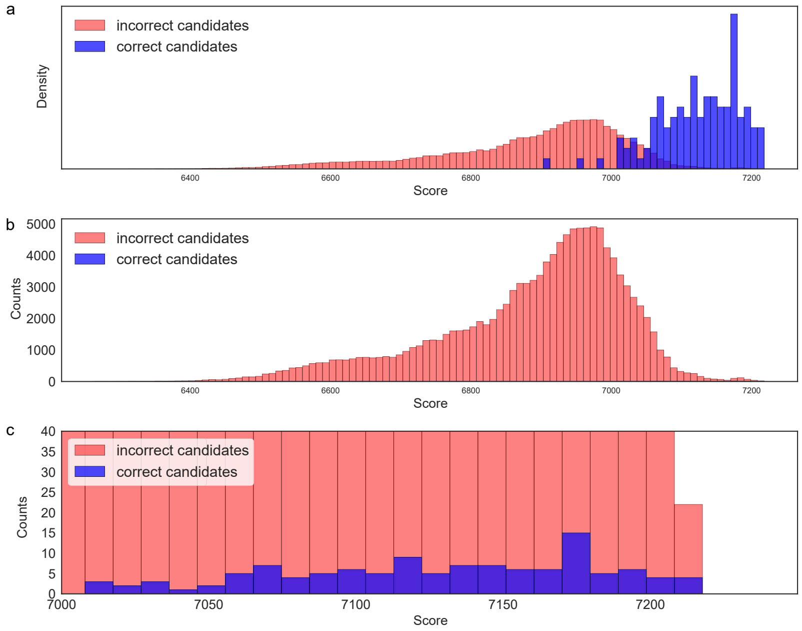


Figure 6.2: Score distribution of correct and incorrect molecular structure candidates, using CASMI 2016 contest results for CSI:FingerID. Histogram plots displaying all queries and *all candidates* of CASMI 2016 (positive ion mode) simultaneously. There are 120 correct candidates but 123 551 incorrect candidates, so incorrect candidates are three orders of magnitude more common. We plot scores of the original CSI:FingerID submission for CASMI 2016; since CASMI rules required that scores of all candidates are positive, an arbitrary constant value of 10,000 was added to each score. (a) Score distributions when both distributions have been normalised individually. (b) Score distributions without normalisation; correct candidates are practically invisible in this plot. (c) Zoom-in into (b): We observe numerous incorrect candidates with scores as high as correct candidates.

incorrect hits which pass or do not pass the threshold. For any score threshold, we plot the *true positive rate* $TP/(TP + FN)$ (ratio of reported correct hits among all correct hits) against the *false positive rate* $FP/(FP + TN)$ (ratio of reported incorrect hits among all incorrect hits), resulting in a Receiver Operating Characteristic (ROC) plot. The Area Under Curve (AUC) of the ROC curve is the integral of the ROC curve; the random score, corresponding to a random ordering of hits, reaches AUC 0.5. A method may reach AUC below 0.5, meaning that the hit score performs worse than random. A score that separates perfectly would reach an AUC of 1.0. Different from binary classification, we must not invert “predictions” to reach a better AUC: Logic dictates that the directionality of the hit score (such as, “high scores are good”) is fixed by the candidate annotation task.

HOP plots While ROC curves are a highly used and well established evaluation metric in binary classification, they fall short for this kind of evaluation. While it is certainly possible to evaluate annotation and separation task independently of each other with the metric introduced previously, it is of high importance to evaluate them in conjunction. A method that for $10,000$ query spectra only returns one correct hit can easily produce an AUC of 1.0, while the annotation performance of only 0.01% correct annotations would not be visible in the ROC plot. Conversely, a method can report an annotation rate of 75% or above, and produce an AUC of 0.6 or worse.

We introduce *hop plots* (inspired by the hop plant *Humulus lupulus* ranking to a supporting wire) to integrate this information: We again vary the score threshold but normalise reported correct hits and incorrect hits by the *total* number of hits (queries) $N = TP + FN + TN + FP$, plotting TP/N vs. FP/N. The resulting curve starts in the origin $(0, 0)$ and ends in some point $(x, y) \in [0, 1]^2$ with $x + y = 1$, where y is the ratio of correct hits for the complete list of queries. The hop curve lies in the lower-left triangle; random ordering of hits corresponds to a straight line from the origin to some point (x, y) with $x + y = 1$. For perfect results, the hop curve is a straight line between the origin and $(0, 1)$; in the worst case, it is a straight line from the origin to $(1, 0)$. Hop plots allow us to answer questions such as, “If I fix a certain false discovery rate, how many true discoveries will a method return?”. A zoom-in allows us to compare methods in the particularly interesting region close to the origin. Both ROC curves and hop plots allow us to visually compare the performance of a method for different datasets in one plot; here, the total number of hits N is different for each curve.

We can calculate the *area under curve* of a hop plot by mirroring the curve at the line $x + y = 1$ before taking the integral. A method with identification rate $y \in [0, 1]$ for the complete list of queries will have area under curve between y^2 and $y^2 + 2(1 - y)y = 1 - (1 - y)^2$, with random ordering reaching area $y^2 + (1 - y)y = y$. But much like the area-under-curve of a ROC curve, this number does not tell us whether a method performs well at the (highly relevant) lower-left or the (mostly irrelevant) upper-right of the curve; hence, we refrain from reporting hop plot area-under-curve.

Besides ROC curves, precision-recall curves are frequently used to asses the performance of a binary classifier. Similar to ROC curves, precision-recall curves are not appropriate for the identification task, since “recall” is normalised to the number of correct identifications, which is usually different for two methods. As “precision” equals one minus FDR, “precision” can directly be read from a hop plot, too. See Fig. 6.3 for a comprehensive example on hop plots in contrast to ROC and precision-recall curves.

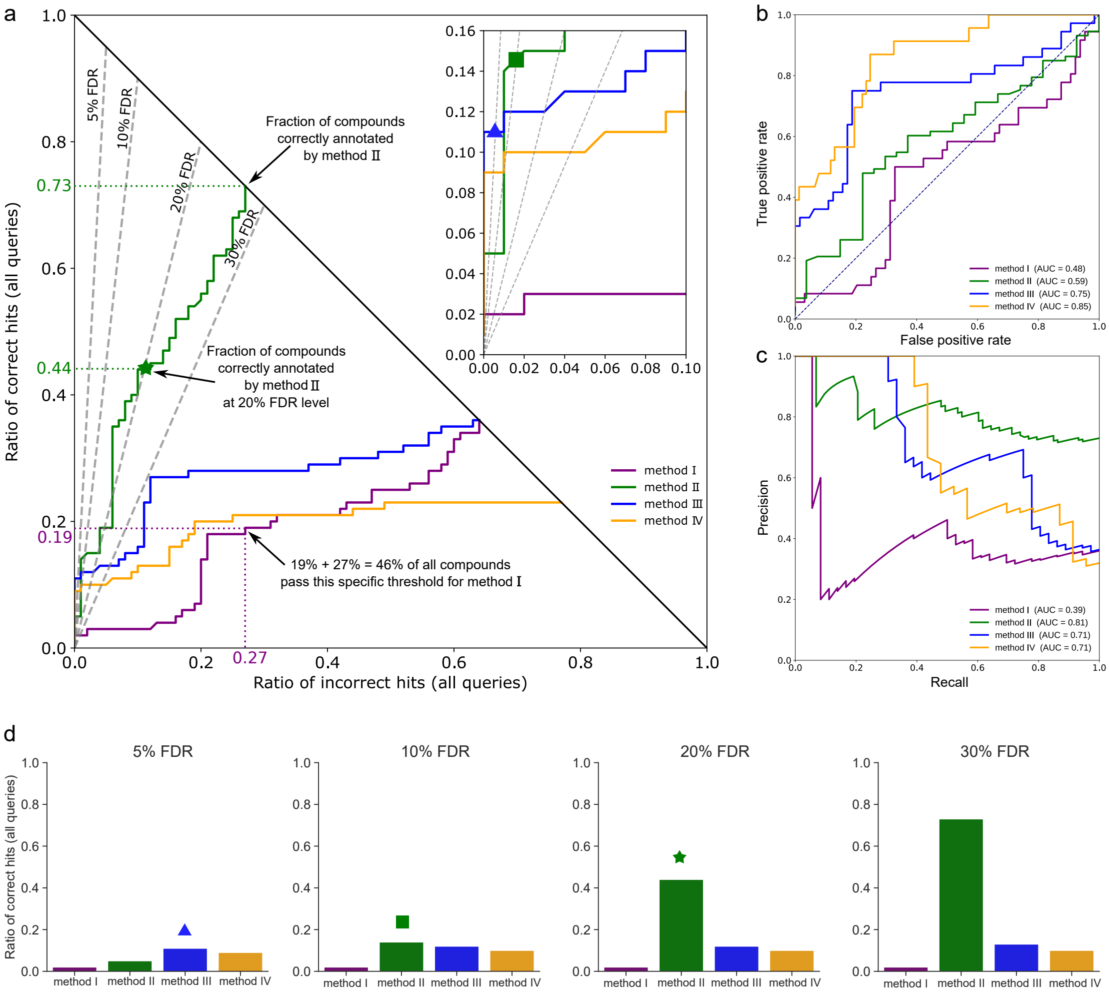


Figure 6.3: Introducing hop plots. (a) Hop plots allow us to simultaneously assess a methods annotation rate *and* its power to separate correct and incorrect hits. Two methods with identical annotation rate will end up in the same point (x, y) with $x + y = 1$, see methods I and III; these methods can differ substantially in their separation power. The plot shows which method performs best for a desired number of correct annotations (horizontal lines, not shown), incorrect annotations (vertical lines, not shown), or false discovery rate (FDR, dashed lines). For example, if we are willing to accept three incorrect annotations from a total of $N = 100$ queries, then method IV clearly outperforms method I; this ordering is reversed if we consider all queries $(x + y = 1)$. FDR levels correspond to lines through the origin; a hop curve may cross or touch some FDR line multiple times, or only in the origin. We report the maximum number of correct annotations among all crossing points. For example, method II returns 55 hits (44 correct, 11 incorrect) at FDR 20 % (star). We are usually interested in small FDR values such as FDR 10 %, so a zoom-in shows where different curves cross the corresponding FDR lines: For example, method III returns 11 hits (all correct) at FDR 5 % (triangle, zoom-in), and method II returns 15 hits (14 correct) at FDR 10 % (square, zoom-in). (b) ROC plot and (c) precision-recall curve for the data shown in (a). Both plots (b) and (c) hide the information that method II is by far the most powerful method. (d) Bar plots for four FDR levels. Notably, the information from the bar plot can directly be read from the hop plot: We mark the corresponding values by star, triangle and square, compare to the corresponding marks in (a).

6.2.3 Evaluating Score Separation for Popular Structure Library Search Tools

With the evaluation tools necessary to evaluate both annotation and separation task, we can do so for existing methods. As mentioned before, the CASMI 2016 contest was designed to compare only the annotation capabilities of methods [148]. We now extend this to include the separation task. Scores of MetFrag, MAGMa+ and CFM-ID were downloaded from the CASMI 2016 results web page (`http://casmi-contest.org/2016/`, category 2, automated methods). We only considered tools that scored all candidates. As some percentage of challenge compounds in CASMI 2016 were part of the CSI:FingerID training dataset that was used for the original submission, annotation results for CSI:FingerID are inflated. For that reason we recomputed scores for CSI:FingerID using a structure disjoint cross-validation setup. We computed scores for the structure-disjoint evaluation of CSI:FingerID using CSI:FingerID 1.2.0.

We used hit scores (score of the top-ranked candidate for each query) to order hits. We remind the reader, that we assume knowledge about the molecular formula of query compounds, and as such we restricted the set of candidate structures to those with the correct molecular formula for all tools. We want to evaluate the annotation and separation task independently of the task of correctly identifying a compound's molecular formula. In practice, molecular formulas can be established using SIRIUS 4 [47] and potentially ZODIAC [101].

We performed evaluation using either all ChemSpider [124] candidates, or restricting the search to those ChemSpider candidates that are simultaneously found in our biomolecule structure database. In four cases, this resulted in an empty list of candidates, and these queries were excluded from evaluation. In 13 cases, the set of candidates did no longer contain the correct structure; these queries were not excluded from evaluation. While these queries don't affect the comparison of annotation rates between tools and would usually be removed, they do affect the separation performance. As expected [14], MetFrag, MAGMa+ and CFM-ID profit more from restricting the set of candidates than CSI:FingerID [46]; hence, annotation rates varied less than those reported in the CASMI evaluation [148]. In fact, even randomly choosing one of the remaining candidates resulted in a decent annotation rate when searching the biomolecule structure database: In 38 cases, only a single candidate remained; and in 33 cases, the candidate list contained two or three structures. Even if there is only a single candidate, the score an *in-silico* tool assigns to this candidate is important information, as we use it to order hits. In practice, one would in most cases use the biomolecule structure database over ChemSpider or PubChem for annotation of metabolites. We show performances on ChemSpider to evaluate search tool performances on harder instances of the problem, as the candidate lists here are usually much larger. As can be seen in the hop plots of Figure 6.4, none of the hit scores of existing methods are capable of adequately separating correct from incorrect hits. When searching the biomolecule structure database, Fig.6.4 (a), no method evaluated is able to return a meaningful amount of hits for FDR levels 5%, 10% and 20%, see Fig. 6.5 for a bar plot visualisation. When searching ChemSpider, no tool is able to return any hits for any FDR level below 40%. Recall however, that none of the tools evaluated here were designed for the separation task, and as such it is not surprising that they perform poorly for it. To this end, *our findings must not be misunderstood as a critique* against these tools or their developers.

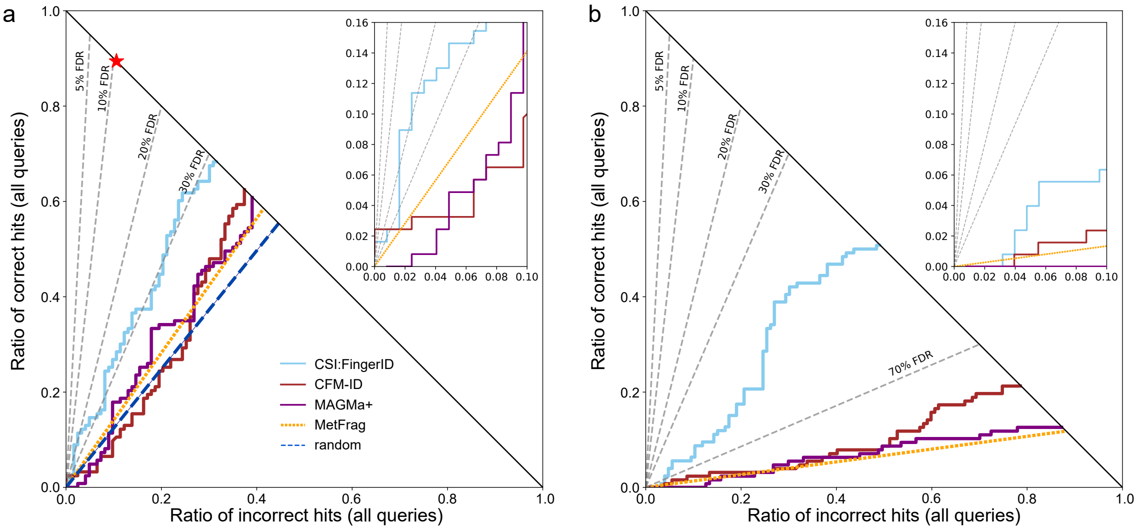


Figure 6.4: Hop plots showing separation by hit score for different *in-silico* tools, using the CASMI 2016 contest submissions. Positive ion mode, candidates retrieved by molecular formula. (a) searching the biomolecule structure database, $N = 123$ queries; (b) searching in ChemSpider, $N = 127$ queries. FDR levels shown as dashed lines; FDR levels are exact, not estimated. The blue dashed line in (a) indicates random scores, resulting in random ordering of candidates and hits; the red star in (a) is the best possible search result.

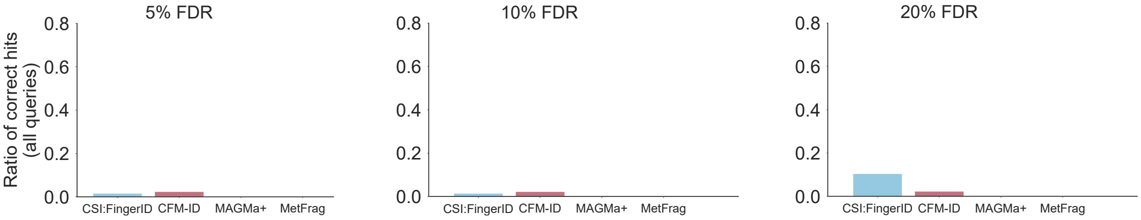


Figure 6.5: Bar plots showing the ratio of correct annotations for different *in-silico* tools and fixed FDR levels, using the CASMI 2016 contest submissions. Positive ion mode, candidates retrieved by molecular formula, searching the biomolecule structure database, $N = 123$ queries; FDR levels are exact.

6.2.4 Designing a Confidence Score for CSI:FingerID

From this section onward, we are going to focus on improving the score separation for CSI:FingerID, the current best-in-class tool for structure annotation from LC-MS/MS data. As we have established in the previous section, scoring functions developed and optimised for the annotation task, often perform poorly for the separation task. There is however no need for a score to perform well in annotation **and** separation, if we just separate the tasks and look at them individually. As such, we are not changing the hit scoring function currently used by CSI:FingerID, but instead develop a separate scoring that we name the *COSMIC confidence score* or just *confidence score* for the sake of brevity. This confidence score is decoupled from the annotation task, and can be seen as an additional scoring layer that is performed *after* the annotation task is completed. It should be interpreted as the answer to the question “Given the annotation output for a query (a ranked list of structure candidates), how sure are we that the top ranked structure candidate corresponds to a correct structure annotation?”. It is critical to understand that the confidence score does not re-rank structure candidates for a singular query. In the following sections, we introduce two designs for such a confidence score, one that is E-value-based, and one that is SVM-based inspired by Percolator.

6.3 E-values

The *p-value* of a score is the probability that a score this high or higher would be expected by chance; the *E-value* is the expected number of random hits with this score or higher. As mentioned earlier, since no methods for creating sensible decoy structures or fingerprints exist, we have to rely on proxy decoys as discussed below. However, there is still a large incentive to this approach over machine learning-based ones - the absence of overfitting. From here on out, we use “E-value score” and “calibrated score” synonymously.

6.3.1 E-value Estimation

We suggest to use the distribution of scores of PubChem [88] candidates as a proxy for the score distribution of incorrect hits. We empirically established that scores of an individual MS/MS query roughly followed a log-normal distribution; for other queries, the score distribution was multimodal (See Fig. 6.6). In particular, a small fraction of candidates had a much higher score than expected from the single log-normal distribution; ignoring this would result in inflated calibrated scores.

The log-normal distribution is a reasonable proxy if there are only few samples available. To model multimodal distributions as well as distributions that deviate from the log-normal distribution, we suggest to use a kernel density estimate of the probability density function. Clearly, we do not have to “compute” the kernel density; instead, we want to know the E-value under the resulting distribution. For the ease of presentation, we do not use log-normal kernel functions but instead, model the log-transform of the scores by normal kernel functions, which is mathematically equivalent. Let $y_i := \ln x_i$ for $i = 1, \ldots, n$ be the log-scores of the PubChem “proxy decoys” *excluding the hit score*, and let $y := \ln x$ be the log-score of the hit. We first determine the bandwidth of the kernel function; we use Silverman’s rule of thumb [153], first determining the standard deviation $\hat{\sigma}$ of the sample $y_1, \ldots, y_n$, then setting

$$h := 1.059223841 \cdot \hat{\sigma} n^{-1/5}.$$

We also tested a variation of Silverman’s rule of thumb, called “nrd0”, in which

$$h := 0.9 \cdot n^{-1/5} A.$$

with

$$A = min(\hat{\sigma}, \frac{\text{Interquartile range}}{1.34}).$$

Both variants showed the same level of separation power in evaluation (Fig. 6.7). For

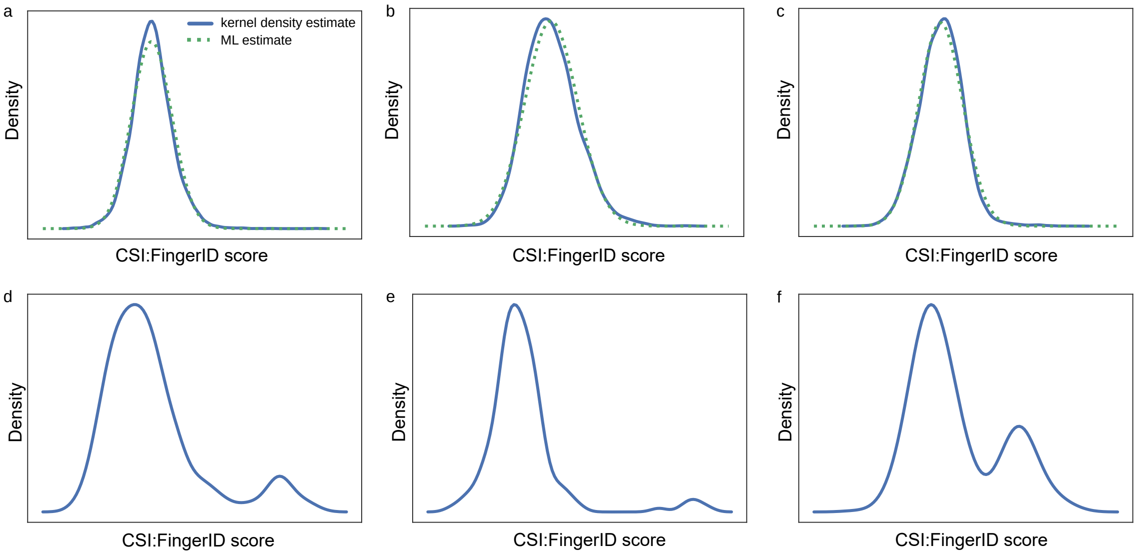


Figure 6.6: Examples of CSI:FingerID score distributions. Shown are kernel density estimates of candidate scores searching in PubChem. We find that unimodal score distributions (a–c) are often similar to a log-normal distribution ("kernel density estimate"); for comparison, we show the log-normal distribution with parameters fitted by Maximum Likelihood estimation ("ML estimate"). Other score distributions are clearly multimodal (d–f). (a) PyroGlu-Trp, $C_{16}H_{17}N_3O_4$, 4 862 candidates, NIST 1632483. (b) 3-Methyl-L-histidine, $C_7H_{11}N_3O_2$, 3 503 candidates, NIST 1346484. (c) 1,3-Benzodioxole-5-propanamine, $C_{12}H_{17}NO_2$, 15 786 candidates, NIST 1306465. (d) N-(2-Hydroxyethyl)-5(6)-epoxy-8Z,11Z,14Z-eicosatrienamide, $C_{22}H_{37}NO_3$, 471 candidates, NIST 1139175. (e) Methanone, $C_{24}H_{22}FNO_2$, 156 candidates, NIST 1300971. (f) Benzeneethanamine, $C_{18}H_{22}BrNO_3$, 483 candidates, NIST 1380115. Numbers of candidates from PubChem.

the Gaussian kernel $K(u) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$ we reach

$$K\left(\frac{y-y_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-y_i)^2}{2h^2}\right)$$

so this is just the usual probability density function of the normal distribution times h, which cancels out in the kernel estimator. We calculate

$$\text{E-value} = \frac{m}{n} \cdot \sum_{i=1,\dots,n} \left[\frac{1}{2} - \frac{1}{2}\,\text{erf}\left(\frac{y-y_i}{\sqrt{2}h}\right)\right] \tag{6.1}$$

where m is the number of candidates in the biomolecule structure database.

6.3.2 Evaluation of E-value Separation

Evaluation of score separation by the calculated E-values was carried out on the COSMIC training dataset as well as the independent dataset (see Chapter 5 for dataset details). For a query fragmentation spectrum, we again assume to know its molecular formula, and we obtained candidates from the structure databases using this molecular formula. For 325 compounds in the COSMIC training dataset and 278 compounds in the independent data, this resulted in an empty candidate list when querying the biomolecule structure

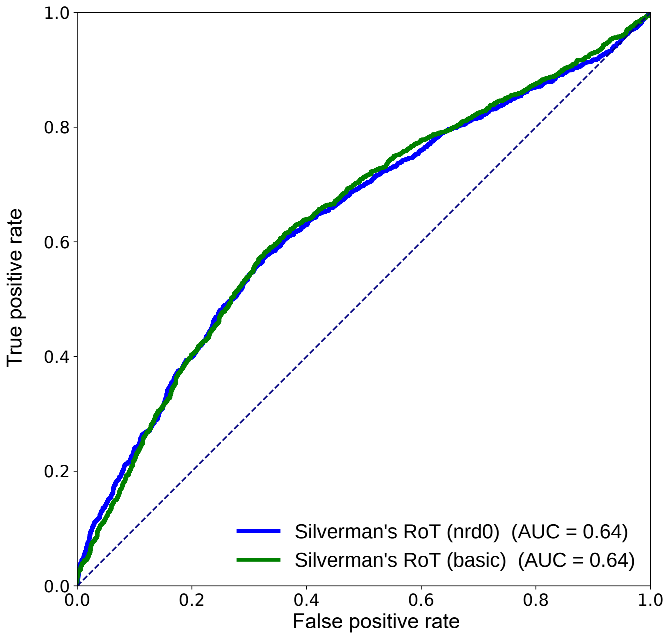


Figure 6.7: ROC plot of E-value estimations using either the basic Silverman's rule of thumb or the "nrd0" variant. Merged collision energies, new CSI training dataset, no artificial noise.

database; these compounds were excluded from evaluation, leaving us with 3,721 queries in cross-validation and 3,013 queries for independent data. Removing these compounds is reasonable, as the corresponding queries would not return a score at all, making it irrelevant for separation. For 845 compounds in the COSMIC training dataset and 521 compounds in the independent data, the correct structure is not present in the biomolecule structure database; *these compounds were not excluded*, as they are relevant for separation. We ensured structure-disjoint evaluation (all compounds novel) both for CSI:FingerID and the COSMIC confidence. As stated in Chapter 5, we evaluate score separation for different levels of noisified spectra, as well as for four different collision energies. For the sake of visual clarity and brevity, we show ROC plots for the independent dataset at medium noise here, additional data can be found in the Appendix.

ROC plots in Fig. 6.8 show, that E-values calculated from the proxy decoys drawn from PubChem separate correct from incorrect hits generally better than the CSI:FingerID score. This improvement however is moderate at best, as AUCs for the E-value score are reported between 0.71 and 0.74. In each plot, all curves end in the same number of correct hits (1,829 for (a), 1,901 for (b), 1,765 for (c), 1,948 for (d)), so a hop plot would not contain additional information. We note that E-values calculated in this fashion show improved separation power over the CSI:FingerID score, but still do not perform the separation task well enough for practical appliance. We conjecture that this is mostly due to the proxy decoys drawn from PubChem. Many structures contained in PubChem are not biologically relevant or even of organic nature, so our decoys might not be similar enough to structures in the biomolecule structure database. Additionally, for some percentage of queries PubChem only contains a low number of structures, making it much harder to estimate a sensible kernel density.

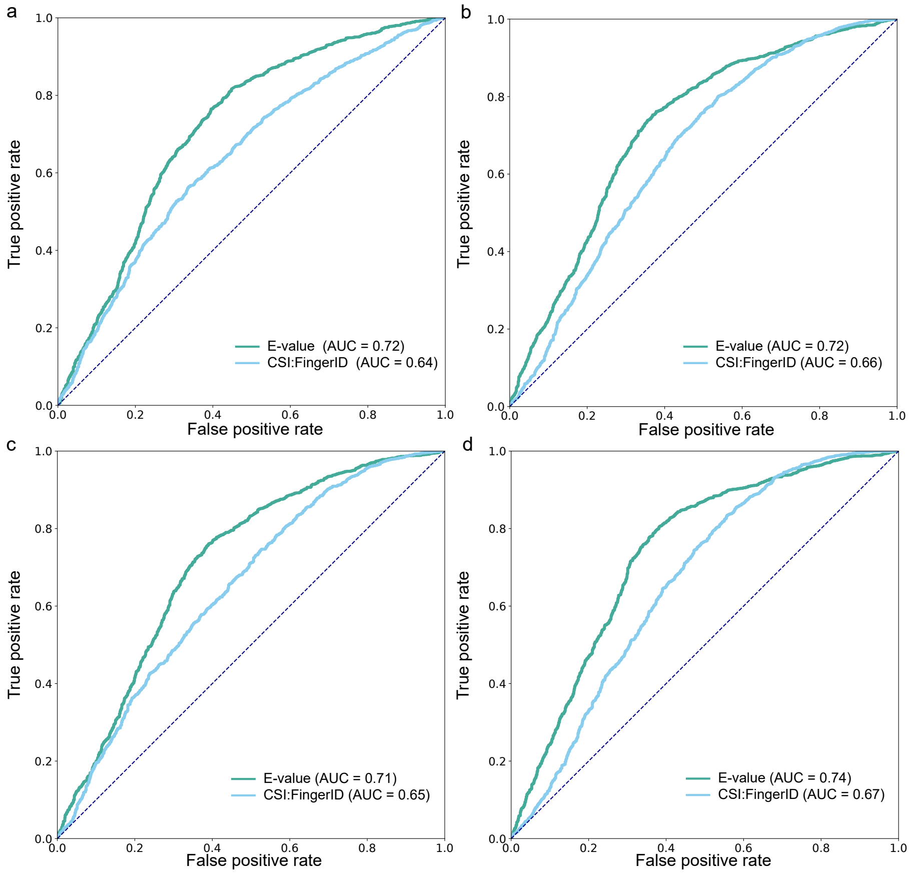


Figure 6.8: ROC plots showing E-value score performance. (a–d) Comparison of CSI:FingerID score and E-value score. ROC curves, structure-disjoint evaluation, independent data and medium noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra ("all collision energies")

6.4 Support Vector Machine Approach - the COSMIC Confidence Score

As the calculation of E-values alone is not sufficient for reasonable separation between correct and incorrect hits, we now introduce a Percolator [78, 155] inspired approach utilising SVMs. Different from there but similar to [5, 112], we do not train a classifier for an individual LC-MS run to "boost" annotation rates; instead, we train classifiers only once using the reference measurements, which are then applied to the biological data. As noted by Käll et al. [78], this approach is highly prone to overfitting: Characteristics of correct and incorrect hits may vary between experiments, instrument types, compounds present in the sample, and others. Here, we have taken extensive measures to counter overfitting, such as "noisifying" spectra and the restriction to *linear* Support Vector Machines. Using a simpler machine learning model such as a SVM allows us to spot and analyse overfitting effects to some degree, as described in Chapter 3.

We repeated the following for each collision energy (10 eV, 20 eV, 40 eV, and merged spectra), and trained individual SVMs using spectra without added noise from that energy as training data. Features of the linear SVMs are shown in Table 6.1. All features were individually standardised. Parameter $C \in \{10^{-5}, 10^{-4}, ..., 10^{5}\}$ of each SVM was chosen by a nested cross-validation. We used quadratic hinge loss and l_2 regularisation. SVMs were trained using LIBLINEAR [53].

For each collision energy, we trained three classifiers: (i) When searching PubChem, we used all appropriate features (all but features 20–22) from Table 6.1. Searching the biomolecule structure database, not all queries result in two or more candidates; but some features from Table 6.1 require a candidate list of at least size two, such as the difference between score of highest-scoring vs. runner-up candidate. To this end, we trained two classifiers for the biomolecule structure database: (ii) The regular SVM assumes that there are at least two candidates; it uses all features from Table 6.1 but is trained only on the appropriate subset of the training data. (iii) The single-candidate SVM uses only the appropriate sub-features (all but Features 1–4, 10, 13) but can be trained using all training data: For instances with two or more candidates, we uniformly selected one candidate.
When training the COSMIC confidence score SVMs, all CSI:FingerID fingerprint predictions of training spectra were carried out structure-disjoint using CSI:FingerID cross-validation models. The COSMIC training dataset was then partitioned for tenfold cross-validation in the same fashion as for CSI:FingerID training. Hence, cross-validation evaluation of the COSMIC confidence score is again structure-disjoint, and all compounds are novel. Similar to above, we also ensured structure-disjoint evaluations on the independent dataset, by choosing the appropriate SVM from cross-validation for computing the confidence score. We map decision values to posterior probability estimates using Platt probabilities [126]. Platt [126] proposed to use a sigmoid function as an approximation of posterior probabilities: $\mathbb{P}(y = \text{correct} \mid x) \approx P_{A,B}(f) \equiv \frac{1}{1+\exp(Af+B)}$, where $f = f(x) \in \mathbb{R}$ is the decision value for hit x, and $y \in \{\text{correct}, \text{incorrect}\}$ its label. We estimated parameters $A, B \in \mathbb{R}$ using maximum likelihood [95, 126] as implemented in LIBSVM [31]

Unlike Percolator, we do not learn a confidence score for individual LC-MS datasets. We do so because it is non-trivial to generate reasonable decoys for small molecules and, more importantly, since incorrect hits in the target database are often not random (Fig. 6.14) [15]. Also unlike Percolator, we do not use our scores to rerank candidates [78, 155]: All of our candidates share the same molecular formula, fragmentation tree and predicted fingerprint; these features are meaningless for reranking. To this end, curves of CSI:FingerID and the COSMIC confidence score in hop plots always end in the same point (x, y) with $x + y = 1$.

6.4.1 Overfitting Analysis and Enforced Directionality

The resulting linear classifiers showed clear signs of overfitting, when interpreting feature weights as introduced in Section 3.1.4. For example, some features received weights that were counter-intuitive, such as negative weight for the quality of the SIRIUS fragmentation tree or the CSI:FingerID score. Recall that the actual hit was chosen by CSI:FingerID as the candidate with the highest score; hence, logic dictates that the CSI:FingerID score of the hit must not receive a negative weight when deciding whether a hit is correct or incorrect. The same is true for selecting the best fragmentation tree by SIRIUS. To this end, we *enforced directionality* of the features: For each feature, we decided manually whether a high value of the feature would increase or decrease our confidence in an annotation. For example, a high CSI:FingerID score should clearly increase our confidence, and so should a small E-value. See Table 6.1 for enforced directions. Notably, enforcing directionality can

Table 6.1: Features of the COSMIC confidence score and classifier weights for merged spectra. Features 1–19 are used for PubChem and biomolecule structure confidence scores, features 20–22 are exclusive to the biomolecule structure confidence scores. For the linear SVM trained using merged spectra (all collision energies), we provide classifier weights for PubChem (“PC”), biomolecular structure database with one candidate (“bio_1”) and two or more candidates (“bio_{2+}”). Clearly, features that require at least two candidates cannot be used for classifier “bio_1”. *Unless explicitly stated otherwise*, we consider the candidate list from PubChem for the PubChem classifier, and the candidate list from the biomolecular structure database for the biomolecular structure classifier. CSI:FingerID scores are Modified Platt score and Covariance score. Multiple structures in the candidate list represented by the same fingerprint were treated as a single entry. Column ‘Δ’ shows if we enforced a feature to have positive (‘P’) or negative (’N’) weight in the classifier. Features are individually normalised.

		Classifier weights				
#	Δ	bio_{2+}	bio_1	PC	Name	Description
1	P	0.1452		0.1167	Log Score Diff. 1	Difference between log scores of highest-scoring vs. runner-up candidate, Modified Platt score
2	P	0.0000		0.0386	Score Diff. 1	Difference between scores of highest-scoring vs. runner-up candidate, Modified Platt score
3	P	0.1104		0.1167	Log Score Diff. 2	Difference between log scores of highest-scoring vs. runner-up candidate, Covariance score
4	P	0.0013		0.0316	Score Diff. 2	Difference between scores of highest-scoring vs. runner-up candidate, Covariance score
5	P	0.0861	0.0424	0.0340	Modified Platt Score	Modified Platt score of highest-scoring candidate
6	P	0.0139	0.0000	0.0000	Covariance Score	Covariance score of highest scoring candidate
7	N	−0.0701	−0.2975	−0.0096	Calibrated Mod. Platt	Calibrated score of highest scoring candidate using Modified Platt scores
8	N	−0.0237	−0.1039	0.0000	Calibrated Covariance	Calibrated score of highest scoring candidate using Covariance scores
9	P	0.0185	0.0137	0.0149	FT Explained Intensity	Sum of normalised peak intensities in the input spectrum which are “explained” by the SIRIUS fragmentation tree
10	N	−0.0753		−0.0900	Log No. Candidates	Logarithm of candidate list size
11	P	0.0000	0.0000	0.0000	FT Score	Score of the SIRIUS fragmentation tree
12	P	0.0000	0.0000	0.0000	Fingerprint Quality	“Quality” of the predicted fingerprint, measured as $\sum_i \max\{1 - p_i, p_i\}$ for predicted fingerprint $(p_1, \ldots, p_n)$
13	N	−0.0423		0.0000	Tanimoto Sim. runner-up	Tanimoto Similarity between highest-scoring and runner-up candidate
14	P	0.0012	0.0000	0.0000	Tanimoto Sim. predicted	Tanimoto Similarity between predicted fingerprint and highest-scoring candidate fingerprint
15	P	0.0178	0.0449	0.0234	FP Length Pred.	Cardinality of predicted fingerprint, only properties with posterior probability at least 0.5 are counted
16	P	0.0569	0.1178	0.0314	FP Length Hit	Cardinality of highest ranked candidate’s fingerprint
17	P	0.0907	0.0718	0.0480	Rescoring 1	Score of the highest-scoring covariance candidate when scored with the Modified Platt scoring method
18	P	0.0243	0.0000	0.0000	Rescoring 2	Score of the highest-scoring Modified Platt candidate when scored with the covariance scoring method
19	N	0.0000	0.0000	−0.0294	Rescoring Calibrated	Calibrated score of highest scoring Modified Platt scoring candidate, when scored with the covariance scoring
20	P	0.0119	0.0219		Score Diff Bio Pub	Score difference of the top hit in the biomolecular structure database and PubChem, Modified Platt scoring
21	P	0.0133	0.0257		Score Diff Bio Pub	Score difference of the top hit in the biomolecular structure database and PubChem, covariance scoring
22		−0.1119	−0.4195		Log No. Cand. PC	Logarithm of candidate list size in PubChem

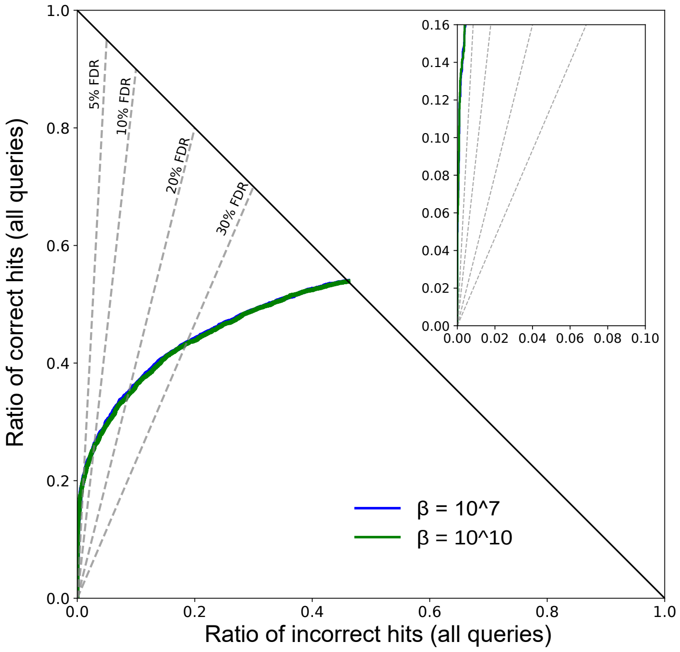


Figure 6.9: Comparison of SVMs trained with different β-values. Hop plot, structure-disjoint cross-validation, New CSI training dataset, merged spectra, no artificial noise, biomolecule structure database, CSI:FingerID version 2.0.0

be achieved by a regular SVM optimisation without additional constraints, allowing us to use established SVM solvers: For each feature with enforced directionality, we augmented one training sample where the corresponding feature was set to a large (positive or negative) value $\pm\beta$, whereas all other features were kept at zero; the sample received a positive label (correct hit). If the absolute feature value $\beta > 0$ is large enough, then an optimal solution must use the feature in the desired direction; the actual value β is of minor importance due to the hinge loss of SVM optimisation. To avoid potential numerical instabilities when finding the solution, β should not be chosen too large. Here, we used $\beta = 10^7$; using other absolute feature values, such as $\beta = 10^{10}$ resulted in basically identical models, and differences are of no practical consequence (Fig. 6.9). Notably, some features received non-zero weights for the classifier with enforced directionality, despite the fact that these features received "counter-intuitive" weights in the unrestricted optimisation: For example, feature "FP Length Hit" was repeatedly given negative weight in cross-validation but had high positive weight if we enforced directionality (unrestricted weight -0.00165, restricted weight 0.0568 in the same cross-validation fold).

The resulting classifier's feature weights as well as the enforced direction for each feature can be found in Table 6.1. We observed that the classifier using enforced directionality showed decreased evaluation performance compared to the unrestricted version (Fig. 6.10). We argue that this is a clear sign for our previous assumption that the unintuitive unrestricted feature weights are indeed contributing to overfitting.

To give a qualitative analysis of the features based on prior knowledge, in the following we give description and justification of directionality for the most impactful features grouped by category.

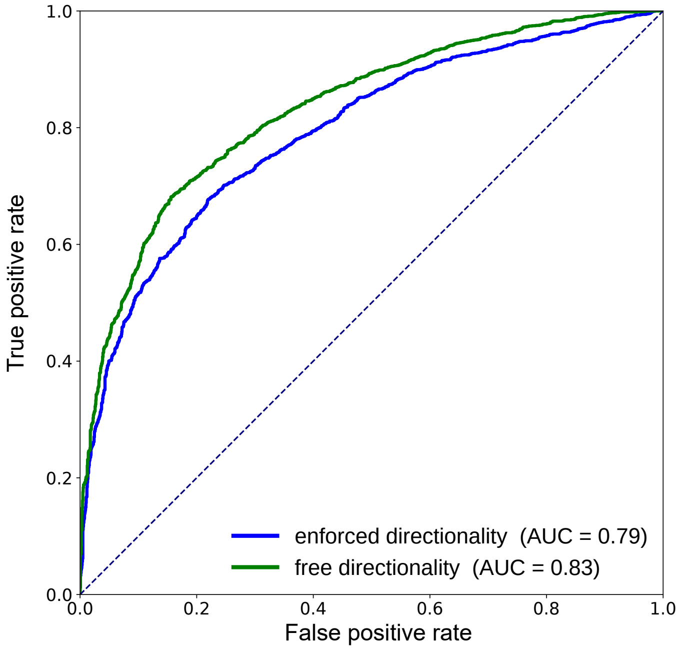


Figure 6.10: Comparison of SVMs trained with unrestricted feature directionality and restricted feature directionality. ROC curves, structure-disjoint cross-validation, merged spectra, no artificial noise, biomolecule structure database, $N = 3607$, CSI:FingerID version 2.0.0

(Log) Score difference

The score difference between the top-scoring candidate and the runner-up candidate is an intuitive and very effective feature (see its high weight in Table 6.1). Assuming the structure candidate list is comprehensive enough, a high score difference implies, that the predicted fingerprint has much higher similarity to the top-scoring structure candidate, than to any other. Vice versa, if this score difference is small, multiple candidate structures seem to be very similar in their fingerprint representation and the confidence that one of these candidates specifically is the correct one should be lower. For that reason, we enforced a positive feature weight.

Modified Platt score, Covariance score and Rescoring

The raw hit scores of CSI:FingerID using either the Modified Platt scoring or the Covariance score for the high-scoring structure candidate. We use both scorings because of instances in which the top-scoring structure candidate differs between the two. In these cases, we additionally score the top-scoring candidate of each scoring with the respective second scoring. Since a higher score should correlate to higher fingerprint similarity and as such higher structural similarity, we enforced a positive feature weight.

Calibrated Modified Platt score and Covariance score

This feature group contains the E-value estimates that we introduced earlier. Again, we use both scorings to spot potentially slightly orthogonal information. Since conceptually a small E-value corresponds to a higher probability of a top-scoring structure candidate to be correct, we enforced negative feature weights.

Fragmentation Tree score and Explained Intensity
The score of the fragmentation tree computed by SIRIUS as well as the sum of normalised peak intensities of the peaks explained by it. These features assesses the quality of the fragmentation tree and how well it represents the given input spectrum. As kernels computed on the fragmentation are used in fingerprint prediction, this has a direct influence on the predicted fingerprint. Since a higher fragmentation tree score and a higher amount of explained intensity correlates to higher quality of the predicted fingerprint, we enforced positive feature weights.

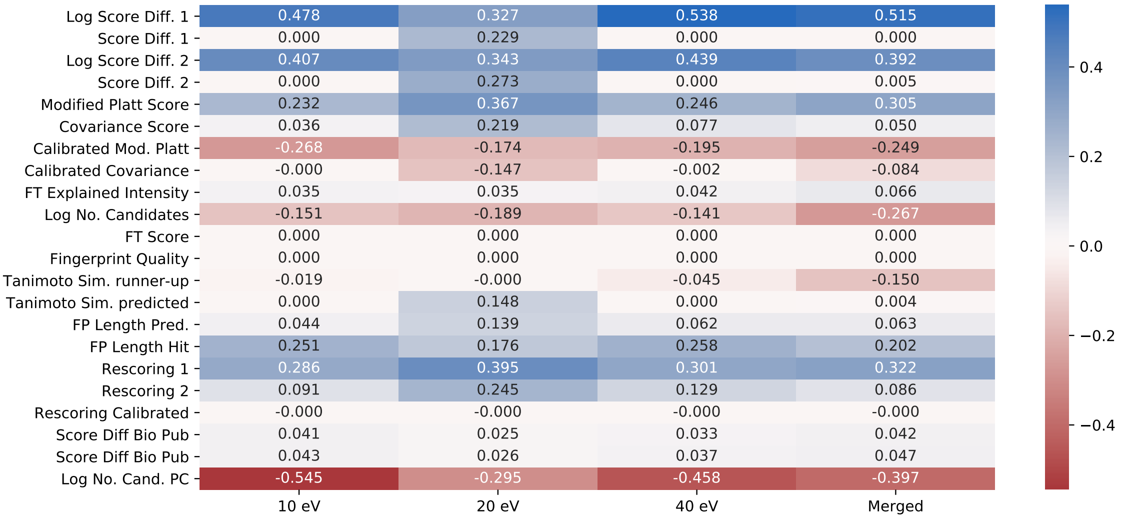
Fingerprint cardinality features
The cardinality of a fingerprint is defined as the number of positions with values greater or equal 0.5. This cardinality can be interpreted as a measure of how well a fingerprint model is able to represent a molecular structure. A lower cardinality corresponds to a weaker representation and forces the comparison of structures via their fingerprint representations to be more coarse-grained. Because of that, we enforced positive feature weights.

Predicted fingerprint quality
The predicted fingerprint is not binary, but contains posterior probabilities for each molecular property. The closer a posterior probability is to 0 or 1, the more confident the CSI:FingerID SVM predicting this position should be. Therefore, the more of these confident positions a predicted fingerprint contains, the higher its quality. We use $\sum_i \max\{1 - p_i, p_i\}$ for predicted fingerprint $(p_1, \ldots, p_n)$ as such a quality measurement, and as such enforced positive feature weight.

(Log) candidate list size
The larger the size of the structure candidate list for a given query, the harder the problem. This is true for both annotation and separation, as can be seen in the evaluation on PubChem, where candidate lists are usually much larger than in the biomolecular structure database. As a larger amount of candidates generally increases the chance of a random, incorrect structure receiving a very high score, we enforced negative feature weight.

Log PubChem candidate list size
This feature is exclusive to searching in the biomolecule structure database. The size of the PubChem candidate list likely influences the quality of the E-value estimation. We remind the reader, that we use PubChem structures as proxy-decoys. We did not enforce a directionality for this feature. One might expect that a larger number of decoys is beneficial for our E-value estimation. Because of the presumed low quality of our decoys however, we are not confident enough to assign a directionality.

We observe that the (calibrated) CSI:FingerID scores, score differences´between hit and runner-up, and the number of candidates turned out to be highly important features. Other features, such as a simple quality measure for the predicted molecular fingerprint or the score of the fragmentation tree, received weights close to zero. This is not unexpected, as much information is shared between these features, rendering some of them obsolete. We want to remind the reader that the feature weight directionality we enforced is derived from what we consider “closest to common sense” and is dependent on our interpretation of



	10 eV	20 eV	40 eV	Merged
Log Score Diff. 1	0.478	0.327	0.538	0.515
Score Diff. 1	0.000	0.229	0.000	0.000
Log Score Diff. 2	0.407	0.343	0.439	0.392
Score Diff. 2	0.000	0.273	0.000	0.005
Modified Platt Score	0.232	0.367	0.246	0.305
Covariance Score	0.036	0.219	0.077	0.050
Calibrated Mod. Platt	-0.268	-0.174	-0.195	-0.249
Calibrated Covariance	-0.000	-0.147	-0.002	-0.084
FT Explained Intensity	0.035	0.035	0.042	0.066
Log No. Candidates	-0.151	-0.189	-0.141	-0.267
FT Score	0.000	0.000	0.000	0.000
Fingerprint Quality	0.000	0.000	0.000	0.000
Tanimoto Sim. runner-up	-0.019	-0.000	-0.045	-0.150
Tanimoto Sim. predicted	0.000	0.148	0.000	0.004
FP Length Pred.	0.044	0.139	0.062	0.063
FP Length Hit	0.251	0.176	0.258	0.202
Rescoring 1	0.286	0.395	0.301	0.322
Rescoring 2	0.091	0.245	0.129	0.086
Rescoring Calibrated	-0.000	-0.000	-0.000	-0.000
Score Diff Bio Pub	0.041	0.025	0.033	0.042
Score Diff Bio Pub	0.043	0.026	0.037	0.047
Log No. Cand. PC	-0.545	-0.295	-0.458	-0.397

Figure 6.11: Classifier weights of the COSMIC confidence score. Shown are classifier weights for searching the biomolecular structure database with two or more candidates (“bio_{2+}” in Table 6.1). CSI:FingerID scores are Modified Platt score from Dührkop *et al.* (*Proc Natl Acad Sci USA* 112, 12580–12585, 2015) and Covariance score from Ludwig *et al.* (*Bioinformatics* 34, i333â€“i340, 2018). Shown are weights for 10 eV, 20 eV, 40 eV and pseudo-ramp spectra (“Merged”, all collision energies). Weights for each classifier are normalised to unit norm. We observe that classifiers for 10 eV, 40 eV and merged spectra have similar weights, whereas the classifier for 20 eV distributes weights more uniformly among similar features.

the data structures. Classifier weights are similarly distributed between different collision energies (Fig. 6.11), with an exception of the 20 eV classifier which shows a more uniform weight distribution amongst similar features.

6.4.2 Confidence Score Evaluation

When compared to the separation performance of the original CSI:FingerID score as well as the E-value score, the SVM-based COSMIC confidence score shows much improved separation (Fig. 6.12). We again evaluated on the structure disjoint COSMIC training dataset as well as the Agilent independent dataset on three noise levels and four collision energies. For the sake of visual clarity and brevity, we show ROC plots only for the independent dataset at medium noise here, additional data can be found in the Appendix.

As we have stressed throughout this thesis, our main goal is to detect and prevent overfitting of the machine learning classifier, to ensure that we trained a model that generalises well outside of our established reference data. To that end, we evaluated the COSMIC confidence score on our cross-validated training dataset as well as the Agilent independent dataset for three different artificial noise levels (Fig. 6.13, for details on the generation of the datasets, see Chapter 5). As a short reminder, reference datasets generally consist of very high-quality spectra that one would very rarely observe in real-world measured data. While there is no metric on how “similarly low-quality” the noisified spectra generated in this thesis are to those real-world measurements, we conjecture that evaluating on them still gives valuable insight on the generalisation ability of the classifier.

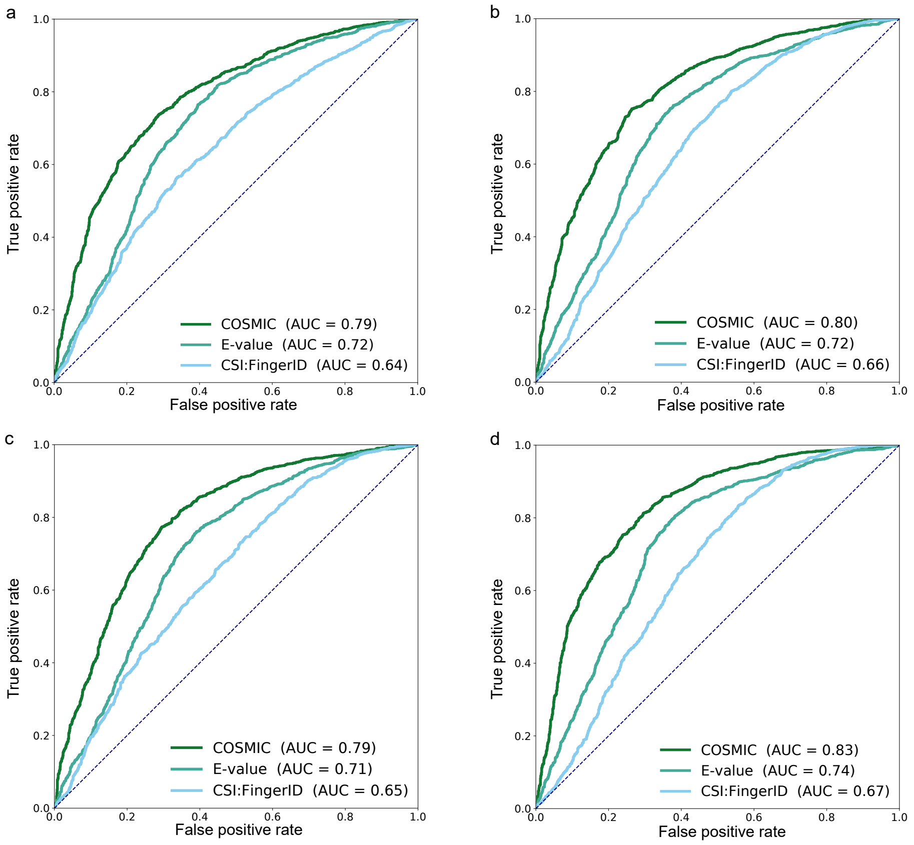


Figure 6.12: ROC plots evaluating confidence score performance. (a–d) Comparison of CSI:FingerID score, E-value score and SVM-based COSMIC confidence score. ROC curves, structure-disjoint evaluation, independent data and medium noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra (“all collision energies”)

For a query fragmentation spectrum, we again assume to know its molecular formula, and we obtained candidates from the structure databases using this molecular formula. For 325 compounds in the COSMIC training dataset and 278 compounds in the independent data, this resulted in an empty candidate list when querying the biomolecule structure database; these compounds were excluded from evaluation, leaving us with 3,721 queries in cross-validation and 3,013 queries for independent data. For 845 compounds in the COSMIC training dataset and 521 compounds in the independent data, the correct structure is not present in the biomolecule structure database; *these compounds were again not excluded.* We ensured structure-disjoint evaluation (all compounds novel) both for CSI:FingerID and COSMIC.

First we can observe that separation performance between training set (Fig. 6.13 (a-c)) and independent evaluation dataset (Fig. 6.13 (d-f)) is similar. Much improved performance on the training dataset compared to the independent dataset can often be understood as a sign of overfitting. In this case, the much better annotation performance of CSI:FingerID on the independent dataset over the training dataset, might just hint at the training dataset consisting of harder instances instead. Second, we can observe that separation performance of the COSMIC confidence score as well as annotation performance of CSI:FingerID are dependent on the collision energy settings on which the measurements

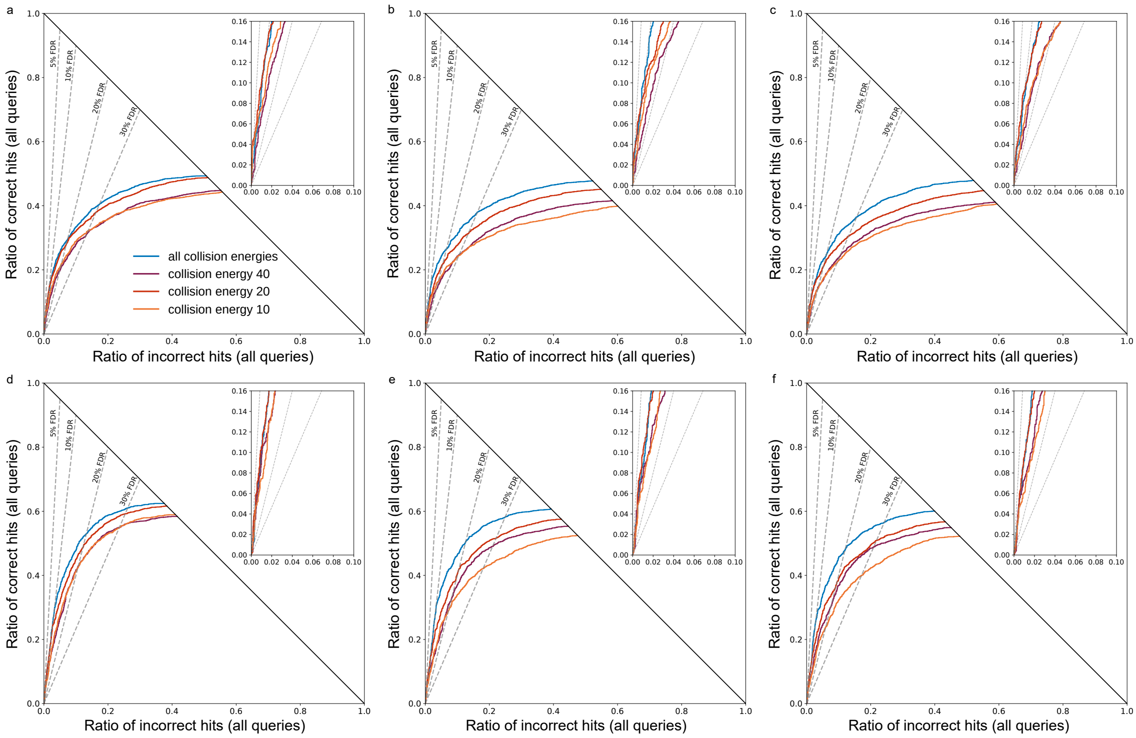


Figure 6.13: Hop plots evaluating confidence score performance. (a–f) Evaluation of COSMIC confidence score: Hop plots for different collision energies, biomolecule structure database. (a–c) Structure-disjoint cross-validation, queries are Orbitrap MS/MS data, $N = 3,721$. (d–f) Independent data with structure-disjoint evaluation, queries are QTOF MS/MS data, $N = 3,013$. (a,d) No added noise, (b,e) medium noise, (c,f) high noise. FDR levels shown as dashed lines; FDR levels are exact, not estimated

were recorded. Notably, the "all collision energies" spectra classifier outperformed the singular collision energy ones on every dataset and noise level. Third, the classifier trained on "all collision energies" seems to be most robust to increased noise in query spectra. Compare the classifier trained on 10 eV spectra only in Fig. 6.13 (d) and Fig. 6.13 (e) for an example of a classifier that is much more susceptible to noise. Notably, the "all collision energies" model returns close to 30% of queries with an FDR of 10% for high noise level independent data.

Inevitably, some incorrect hits received a high confidence score and, hence, would be wrongly regarded as "probably correct". Fig. 6.14 shows the nine incorrect hits with highest confidence scores when searching independent data with medium noise. In seven of nine cases, the true structure was not contained in the biomolecule structure database. In all nine cases, the true structure was highly similar to the corresponding hit. In practice, we conjecture that incorrect hits that receive a high confidence score often are very close to the desired true structure. In contrast, the bottom nine incorrect hits generally showed little structural similarity to the corresponding true structures (Fig. 6.15). Notably, the confidence score machine learning model has not been trained taking this structural similarity into account. Next, we evaluated the impact of precursor mass and number of structure candidates per query on the separation power of the confidence score. We showed

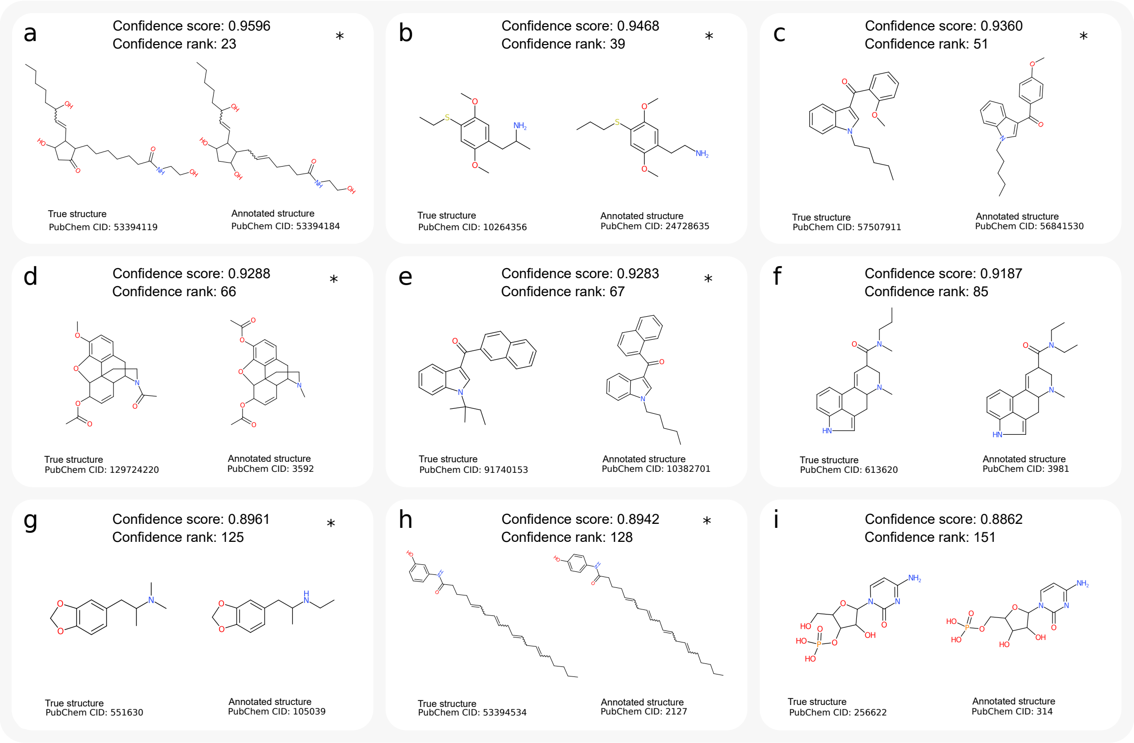


Figure 6.14: Examples of incorrect annotations with highest confidence scores. Queries are cross-validation data, merged spectra, medium noise, biomolecule structure database, structure-disjoint evaluation. Evaluations were carried out using reference spectra, so the true structure behind each query spectrum is known to us, but not known to CSI:FingerID or the confidence score. Each query spectrum is annotated with the structure top-ranked by CSI:FingerID; this pair is called “hit”, and can be either correct (annotation is identical to the true structure) or incorrect. All hits were then ordered by confidence score; it is inevitable that some incorrect hits will receive a high confidence score. Out of the 151 hits with confidence score above 0.8862, 142 were correct (not shown here) and only 9 were incorrect (a–i). Incorrect annotation (CSI:FingerID top-ranked structure) on the right and corresponding true structure on the left. Incorrect annotations may or may not be structurally similar to the true structure, compare to Fig. 6.15. Notably, the nine incorrect annotations with highest confidence score (a–i) show very high structural similarity to the corresponding true structures. This is particularly noteworthy as the confidence score machine learning model has not been trained taking into account this structural similarity. If incorrect hit i is at rank n, this implies that $n - i$ of the $n - 1$ top-ranked hits are correct and only $i - 1$ are incorrect, corresponding to exact FDR $(i-1)/(n-1)$. For example, only 8 out of 150 hits with highest confidence score were incorrect (exact FDR 5.33 %), for confidence score threshold 0.8863. “Confidence rank” is the rank of the (incorrect) hit in the complete ordered list of hits, and “PubChem CID” is the PubChem compound identifier number. Instances where the true structure was not contained in the biomolecule structure database are marked by an asterisk. For these instances, a correct annotation by CSI:FingerID is impossible; at the same time, it is highly challenging for the confidence score to identify these hits as “incorrect”. In seven cases, molecular graphs of the incorrect hit and true structure differ by the theoretical minimum of two edge deletions. Query spectra: (a) NIST 1210761/62/64, (b) NIST 1617825/29/34, (c) NIST 1320583/85/91, (d) NIST 1429464/65/71, (e) NIST 1483460/63/69, (f) NIST 1247455/57/63, (g) NIST 1480825/30/34, (h) NIST 1418771/73/80, (i) NIST 1276453/55/59.

that mass has no pronounced impact on the separation between correct and incorrect annotations (Fig. 6.16 (a)), while the amount of structure candidates has a strong impact (Fig. 6.16 (b)). Similar to this observation, CSI:FingerID annotation rates are also only minorly affected by mass, but majorly impacted by the amount of candidate structures (Fig. 6.16 (c,d)).

Obviously, performance of CSI:FingerID as well as the confidence score is heavily dependent on the quality of the input spectra, which is why we noisified evaluation data as introduced earlier. In addition to the lower quality of real-world biological data compared to reference data, which we tried to emulate using these noisified spectra, data is also often preprocessed using noise baselines that are too high, or were simply chosen for a

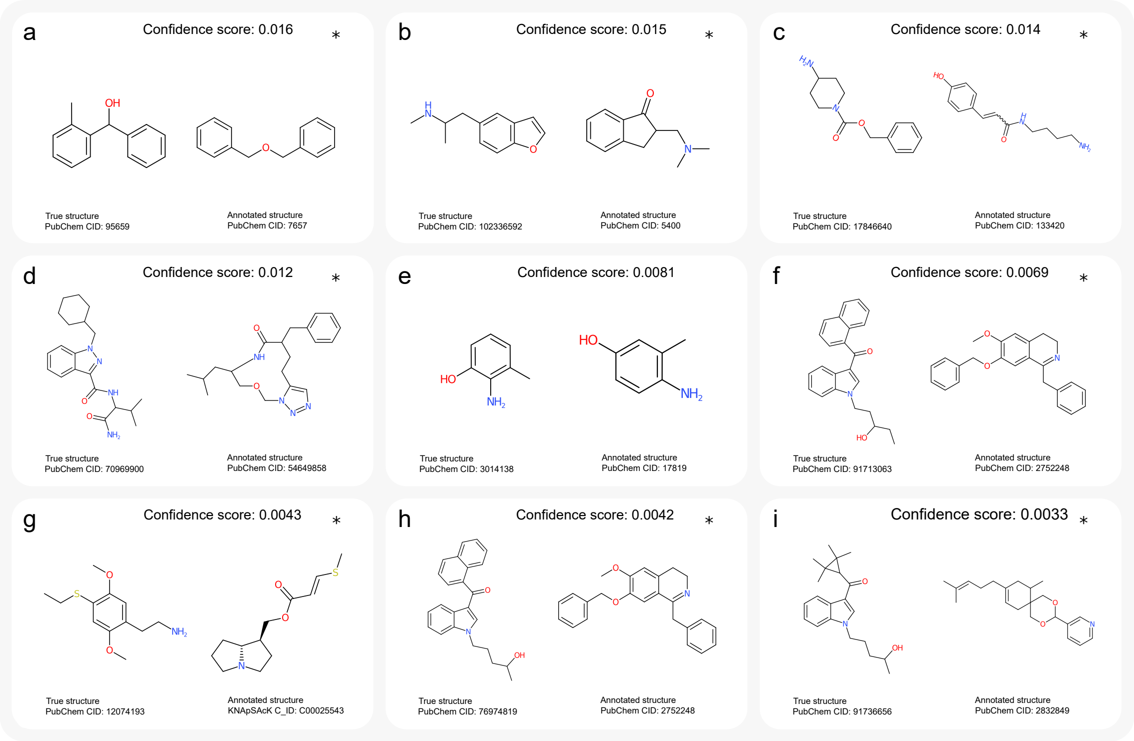


Figure 6.15: Examples of incorrect annotations with lowest confidence scores. Queries are cross-validation data, merged spectra, medium noise, biomolecule structure database, structure-disjoint evaluation. (a–i) Incorrect hits with lowest confidence scores. Top-ranked structure on the right and corresponding true structure on the left. "PubChem CID" is PubChem compound identifier number. Instances where the true structure was not contained in the biomolecule structure database are marked by an asterisk. For (g), the structure of the top hit is not contained in PubChem; we report the KNApSAcK compound identifier ("C_ID") instead. For (a) and (e), molecular graphs of incorrect hit and true structure differ by the theoretical minimum of two edge deletions. For (a), the query spectrum was heavily distorted, and only 8.6 % of peak intensities were explained by the fragmentation tree. For (e), the three top-ranked candidates — including the correct one — were structurally highly similar and received almost identical CSI:FingerID score. Hence, COSMIC rightfully showed little confidence in these (incorrect) hits. Query spectra: (a) NIST 1544714/19/23, (b) NIST 1322859/64/69, (c) NIST1627646/51/56, (d) NIST 1462584/87/93, (e) NIST 1340388/91/96, (f) NIST 1320854/56/62, (g) NIST 1386503/07/12, (h) NIST 1305770/72/78, (i) NIST 1325235/37/43.

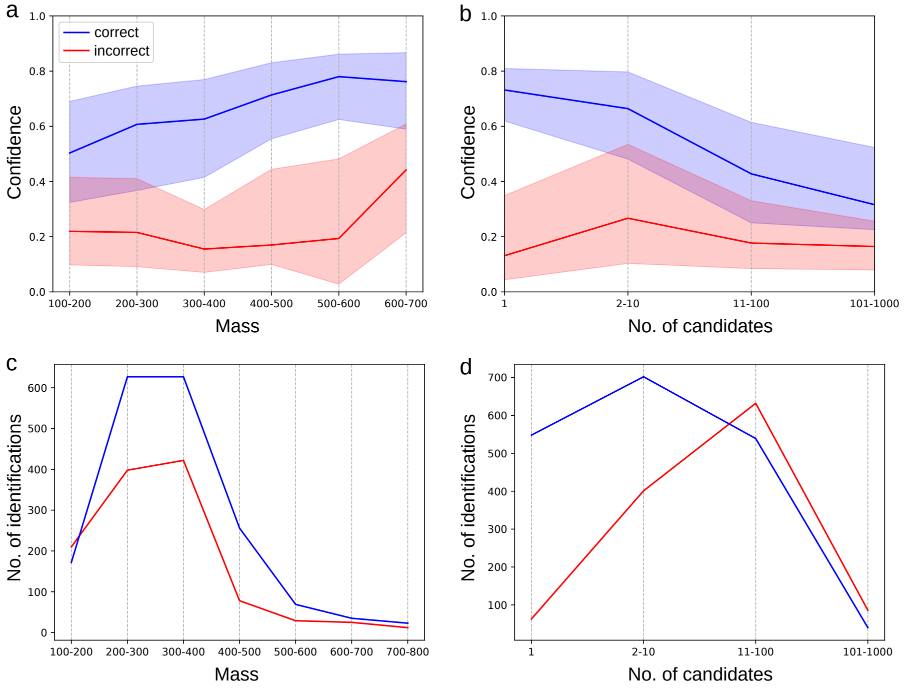


Figure 6.16: Effect of query compound mass and number of candidates on confidence scores. Independent data, merged spectra ($N = 3\,013$), structure-disjoint evaluation, medium noise, biomolecule structure database. (a,b) Confidence score of correct and incorrect annotations when varying query mass ranges (a) and number of candidates (b). Solid lines show median values, colored areas indicate first (25 %) and third (75 %) quartiles. (c,d) Number of correct and incorrect annotations for varying query mass ranges (c) and number of candidates (d). One compound with mass below 100 Da omitted from (a,c). Only few compounds exist above 500 Da and with more than 100 candidates, so curves (a,b) should be interpreted with care in these regions.

different purpose. This leads to spectra containing only very few peaks. To investigate how the separation performance of the confidence score is impacted by the number of peaks in a spectrum, we binned query spectra into three categories (up to 2 fragments, 3 to 5 fragments, 6 or more fragments), based on the number of peaks in the query spectrum with relative intensity at least 5 %. The number of intense peaks in a query spectrum has a clear impact on CSI:FingerID's annotation performance, but a weaker impact on the confidence score's separation performance (Fig. 6.17).

The evaluations carried out above all queried the biomolecule structure database, as that is the setting one would mostly use in practice. Nevertheless, we also evaluated the separation performance of the COSMIC confidence score when querying PubChem. We use the same four collision energy settings and the same three noise levels, and assume the molecular formula to be known. ROC curves are again only shown for the independent dataset, medium noise. While the COSMIC confidence score still outperforms the original CSI:FingerID score and the E-value score on separation (Fig. 6.18 (a-d)), the performance difference is less pronounced. Notably, while the E-value score still outperforms the CSI:FingerID scoring on the bottom left part of the ROC curve, its AUC is lower. In general, separation performance of the confidence score is (expectedly) much worse when querying PubChem over the biomolecule structure database (Fig. 6.18 (e-j)).

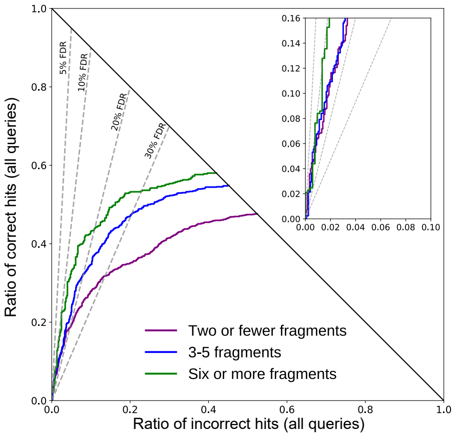


Figure 6.17: Evaluation of separation vs. number of intense peaks in the query spectrum. Independent data, 10 eV structure-disjoint evaluation, medium noise, searching the biomolecule structure database.

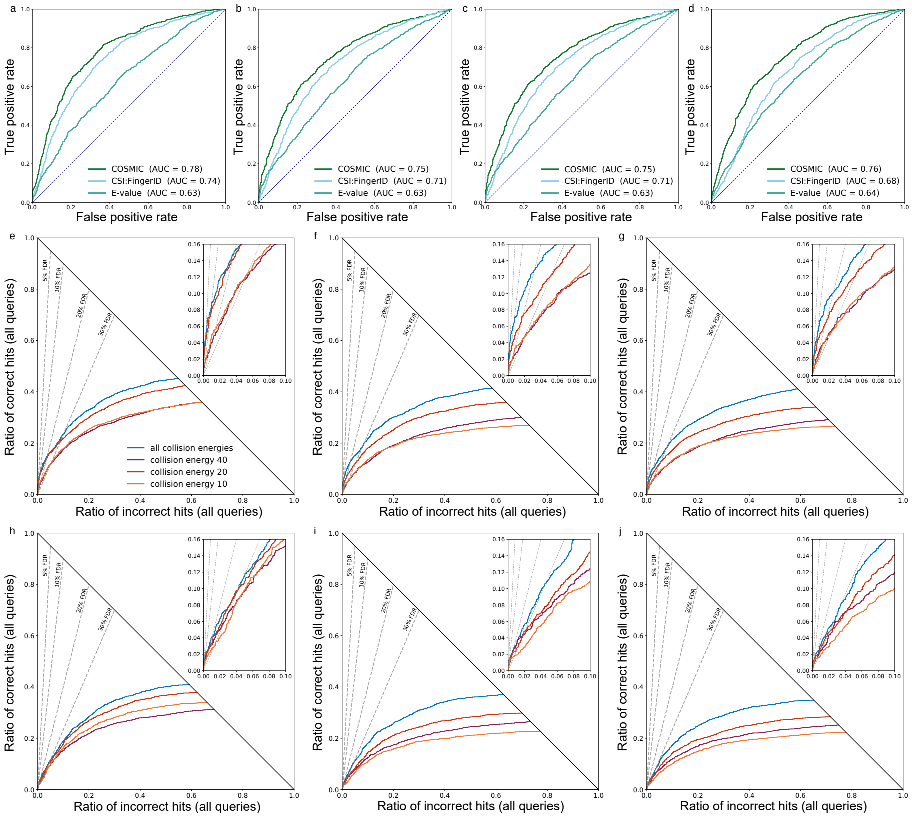


Figure 6.18: Evaluation of E-value and confidence score performance searching PubChem. (a–d) Comparison of CSI:FingerID score, calibrated score and COSMIC confidence score. ROC curves, structure-disjoint evaluation, independent data, medium noise, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra (“all collision energies”). Notably, E-values sometimes result in worse separation than the CSI:FingerID score. (e–j) Evaluation of the COSMIC confidence score: Hop plots for different collision energies; notably, these result in substantially different annotation rates. (e–g) Structure-disjoint cross-validation, $N = 3\,721$. (h–j) Independent data with structure-disjoint evaluation, $N = 3\,013$. (e,h) No added noise, (f,i) medium noise, (g,j) high noise.

6.4.3 Evaluation Against other *in-silico* Tools

We now compare the separation power of the SVM-based COSMIC confidence score against other *in-silico* tools using the CASMI 2016 contest data (Fig. 6.19, see Subsection 6.2.3 for evaluation setup details). Both for ChemSpider and the biomolecule structure database, we used the confidence score variant for searching the biomolecule structure database; this is reasonable as the number of ChemSpider candidates is often substantially smaller than the number of PubChem candidates. We use the confidence score model for "merged spectra". Using the COSMIC confidence score, we correctly annotated 57 hits (46.3% of queries) with FDR below 10 % (Fig. 6.19 (a),(c)) searching the biomolecule structure database (123 queries), and 16 (12.6% of queries) hits with FDR 0 % searching ChemSpider [124] (127 queries, Fig. 6.19 (b)). This is a substantial improvement over the separation power of previous hit scores. As a reminder, no other scoring we evaluated was able to annotate more than even 3% for 10% FDR on the biomolecule structure database.

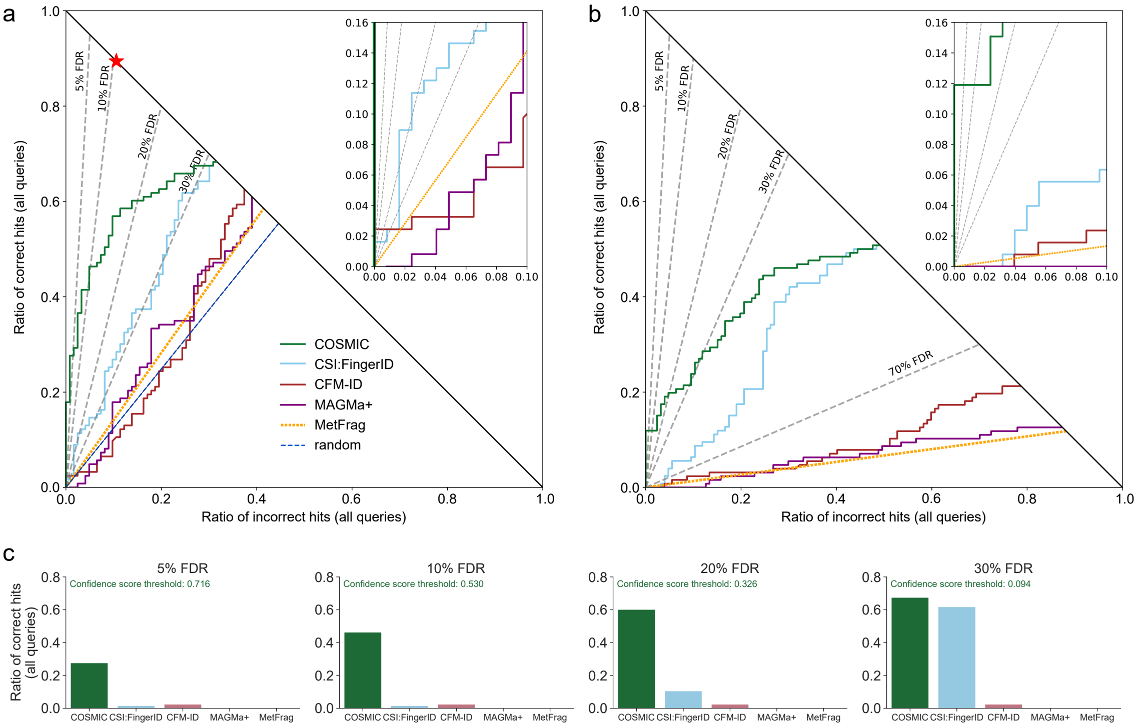


Figure 6.19: Hop and bar plots showing separation by hit score for different *in-silico* tools, using the CASMI 2016 contest submissions. Positive ion mode, candidates retrieved by molecular formula. (a) searching the biomolecule structure database, $N = 123$ queries; (b) searching in ChemSpider, $N = 127$ queries. FDR levels shown as dashed lines; FDR levels are exact, not estimated. The blue dashed line in (a) indicates random scores, resulting in random ordering of candidates and hits; the red star in (a) is the best possible search result. (c) Bar plots showing the ratio of correct annotations for different *in-silico* tools and fixed FDR levels, using the CASMI 2016 contest submissions. Positive ion mode, candidates retrieved by molecular formula, searching the biomolecule structure database, $N = 123$ queries; FDR levels are exact

6.4.4 Evaluation against Spectral Library Search

We also evaluated the dereplication power of COSMIC in comparison to spectral library search. Dereplication in our case describes the task of annotating already known structures to query spectra. Since we are not searching for novel structures, the smaller size of a spectral library is not an issue here. The true structure we are trying to annotate is in most cases contained in the spectral library. As introduced in Section 4.4, the cosine score used to measure similarity between spectra (the annotation task) is also frequently used as a metric of score separation for spectral library search. Here, we evaluate the separation power of the COSMIC confidence score in a setting that resembles a spectral library search setup. For this evaluation, CSI:FingerID and the confidence score were trained without cross-validation, and query spectra came from the independent dataset. Hence, this evaluation is not structure-disjoint, but still spectrum-disjoint: Not a single query spectrum is part of the training data. To evaluate against spectral library search, we generated two spectral libraries based on the CSI training dataset: One library with merged spectra, and one library with spectra at individual collision energies as well as merged spectra. We searched merged query spectra in the first library, and query spectra containing a single collision energy in the second library. Merged spectra are identical to those used for training CSI:FingerID, see above; this library contains 23,965 spectra. The second library contains all available fragmentation spectra at all available collision energies, plus the merged spectra, and contains 189,979 spectra. Notably, the spectral library contains MS/MS data from QTOF and Orbitrap instruments, whereas all query MS/MS spectra are QTOF data. We argue that this resembles how searching in a public or commercial spectral library is executed in practice. The situation is clearly different for an in-house spectral library, but such libraries are usually one to two orders of magnitude smaller. To ensure a fair comparison with COSMIC, spectral library search candidates were restricted to those with the correct molecular formula for each query; in practice, this information is usually not available, and spectral library search may perform worse than reported here. In case the spectral library did not contain at least one candidate with the correct molecular formula of the query, an incorrect annotation with score zero was assumed. We want to remind here, that an incorrect annotation with score zero is beneficial for separation evaluation. We evaluated both the standard cosine score, as well as a cosine score using the square root of intensities (see Section 4.4 for details). One may expect that targeting novel compounds (the true purpose of the COSMIC confidence score) instead of dereplication comes at a price: The biomolecule structure database is more than an order of magnitude larger than GNPS [177] and NIST spectral libraries, which make up the spectral libraries, and we cannot rely on direct spectral comparison. Somewhat unexpectedly, COSMIC annotated substantially more compounds for all reasonable FDR levels, see Fig. 6.20: At FDR 5 %, COSMIC outperformed library search 1,415 to 52 hits at 20 eV and 1701 to 1 hits using merged spectra, respectively. Notably, COSMIC correctly annotated compounds with high confidence although query spectrum and reference spectrum were (highly) dissimilar, with cosine scores between 0.06 and 0.63 (Fig. 6.21). We also observe that separation using the original CSI:FingerID score is much better than in structure-disjoint evaluations (Figs. 6.19 and 6.13). We attribute this increased separation power to the overlap in structures between training and evaluation data: Structures for which a fragmentation spectrum is present in the training data of CSI:FingerID often receive high CSI:FingerID hit scores, comparable to library search.

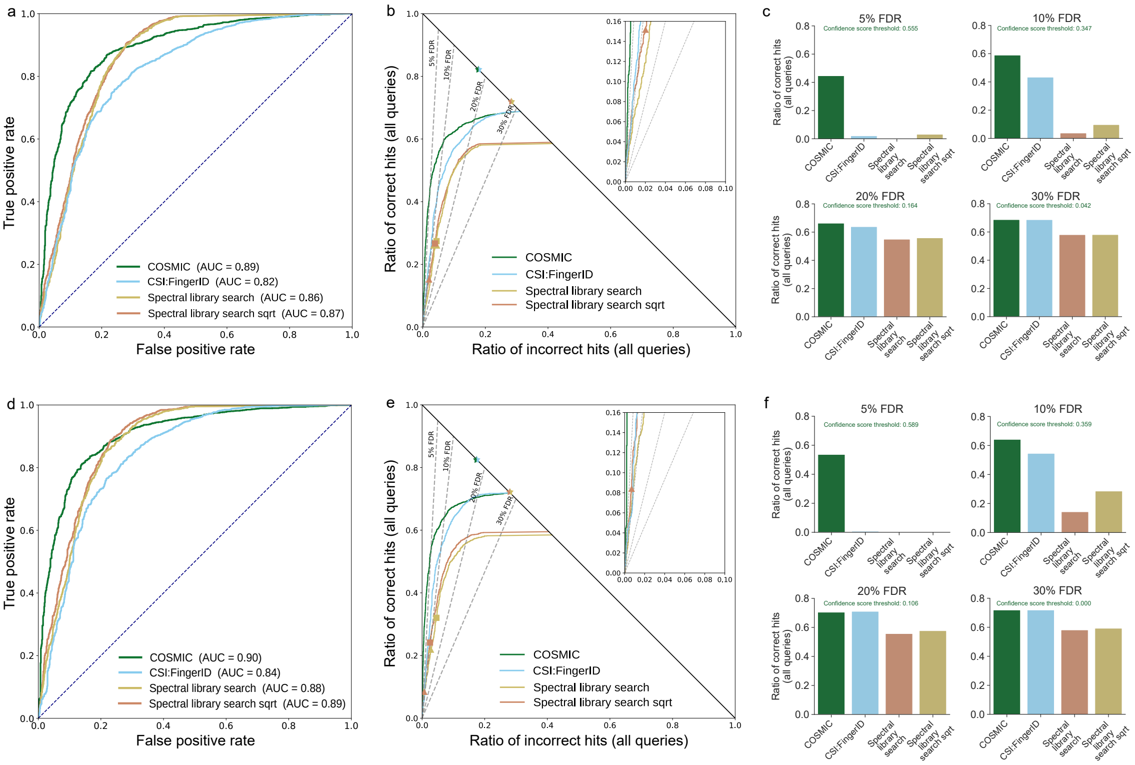


Figure 6.20: Comparison to spectral library search and separation without structure-disjoint evaluation. Query spectra (independent dataset) distorted with medium noise; COSMIC is searching the biomolecule structure database. ROC curves (a+d), hop plots (b+e) and bar plots (c+f) for collision energy 20 eV (a–c) and merged spectra (d–f). Bar plots (c+f) for FDR levels 5 %, 10 %, 20 %, and 30 %. There is no overlap in fragmentation spectra between training data and independent data, but we do not remove training data for which we find the same structure in the independent dataset. To this end, 2,192 of the $N = 3,013$ structures from the independent dataset (72.75 %) are also present in the spectral library. We compare search performance and separation of COSMIC, the CSI:FingerID score and spectral library search. All three methods utilise basically the same MS/MS data. For spectral library search, we compute the normalised dot product using either regular peak intensities, or the square root of peak intensities (“Spectral library search sqrt”) [156]. Spectral library search candidates were restricted to those with the correct molecular formula for each query. Query spectra are QTOF MS/MS data, whereas the spectral library contains a mixture of QTOF and Orbitrap MS/MS data. The spectral library is 16-fold smaller than the biomolecule structure database, giving library search a large competitive edge in evaluation. Notably, COSMIC results in substantially more correct annotations than library search for all reasonable FDR levels; FDR levels are exact, not estimated. For spectral library search, markers show commonly used cosine score thresholds 0.9 (triangle) and 0.8 (square), respectively. Finally, stars indicate best possible annotation results, for CSI:FingerID/COSMIC and library search.

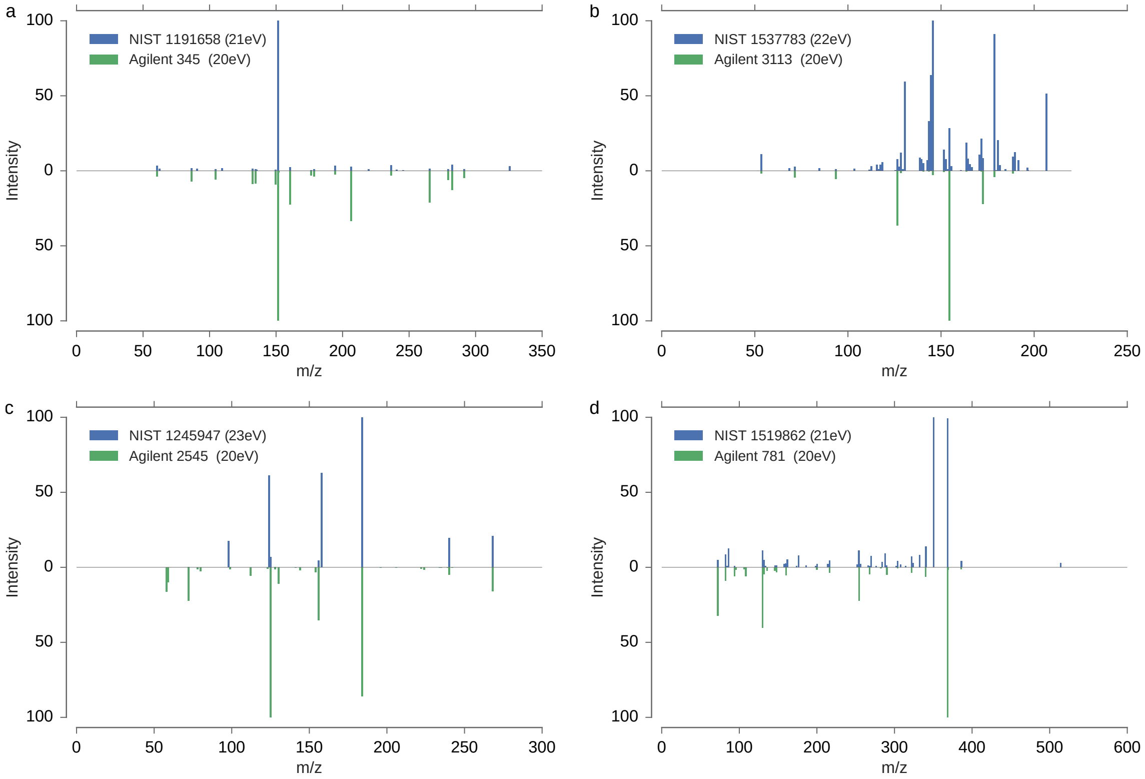


Figure 6.21: Mirror plots of low-scoring library hits that were correctly annotated with high confidence using COSMIC. Shown is the query spectrum (bottom) from the independent dataset, plus the top-scoring reference spectrum (top) from the spectral library, that is, the CSI:FingerID training dataset without merging spectra. Cosine scores were calculated using regular intensities (cosine) as well as square root of intensities (cosine-sqrt). All query spectra consist of a single 20 eV collision energy measurement with medium noise added. Reference spectra consist of a single collision energy measurement with no added noise; shown is the spectrum with the *highest* cosine, among *all* spectra in the spectral library for this compound. (a) Spectra of Thiophanate, PubChem CID 3032792, molecular formula $C_{14}H_{18}N_4O_4S_2$. Reference spectrum NIST 1191658, query spectrum Agilent PCDL 345. Correct COSMIC annotation with confidence 0.9092, cosine 0.0637, cosine-sqrt 0.3165. (b) Spectra of Chlorbufam, PubChem CID 16073, molecular formula $C_{11}H_{10}ClNO_2$. Reference spectrum NIST 1537783, query spectrum Agilent PCDL 3113. Correct COSMIC annotation with confidence 0.9347, cosine 0.1949, cosine-sqrt 0.3523. (c) Spectra of Duloxetine, PubChem CID 60835, molecular formula $C_{18}H_{19}NOS$. Reference spectrum NIST 1245947, query spectrum Agilent PCDL 2545. Correct COSMIC annotation with confidence 0.9283, cosine 0.5197, cosine-sqrt 0.4767. (d) Spectra of Proscillaridin, PubChem CID 5284613, molecular formula $C_{30}H_{42}O_8$. Reference spectrum NIST 1519862, query spectrum Agilent PCDL 781. Correct COSMIC annotation with confidence 0.9720, cosine 0.6312, cosine-sqrt 0.4852. Unlike the commercial Agilent library, the query spectra shown here are uncurated and artificial noise was added.

6.4.5 FDR Estimation Using the COSMIC Confidence Score

At the beginning of this chapter, we introduced the task of score separation as a requirement for useable FDR estimation. Now that the COSMIC confidence score allows us to decently separate correct from incorrect hits, we can aim to transform it into sensible FDR estimates. We remind the reader, that while the COSMIC confidence score lies in the interval between

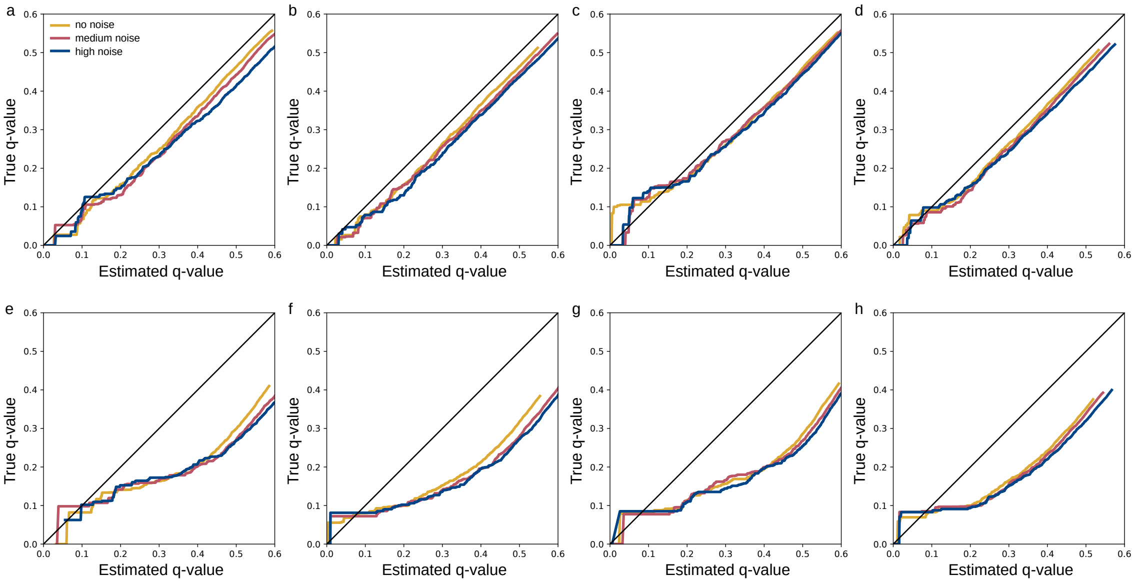


Figure 6.22: False discovery rate estimation. Q-Q plot of true vs. estimated q-values with no added noise, medium noise, and high noise. (a–d) cross-validation, $N = 3\,721$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra. (e–h) Independent data, $N = 3\,013$. (e) 10 eV, (f) 20 eV, (g) 40 eV, (h) merged spectra. The “step” at the beginning of most curves in (e–h) is not an issue of FDR estimation, but due to the fact that no non-zero (true) q-values below this exist in the dataset.

zero and one, they are not to be interpreted as probabilities and as such hold no confound statistical meaning. We now show how to transform COSMIC confidence scores to FDR estimates. The confidence score is an estimated posterior probability of the hit to be correct; to this end, it is one minus the posterior error probability for this hit. Hence, we can use the confidence score to estimate the FDR of the top k hits [50, 141]: Let p_j be the posterior error probability for hit j for $j = 1, \dots, n$, and assume that the hits are ordered by confidence score, so $p_j \leq p_{j+1}$. Viewing the annotations as (not necessarily independent) Bernoulli trials, the expected number of incorrect annotations for the top k hits is $\sum_{j=1}^{k} p_j$, and the expected false discovery rate is

$$\widehat{FDR}_k = \frac{1}{k} \cdot \sum_{j=1}^{k} p_j. \tag{6.2}$$

Since hits have been ordered by posterior error probability, FDR estimates $\widehat{FDR}_k$ are monotonically increasing, so $\widehat{FDR}_k$ is also the q-value estimate for hit k.

We evaluate the accuracy of our FDR estimates by plotting exact q-values against estimated q-values in a Q-Q-plot (Fig. 6.22); this has to be carried out using reference data where exact FDR values can be calculated. Unfortunately, estimated FDRs using this method are of mediocre quality. In particular, estimates for independent data were highly conservative: estimated q-values were much larger than true q-values. Consequently, confidence score values must still be treated as a score, but not as the probability that the annotation is correct.

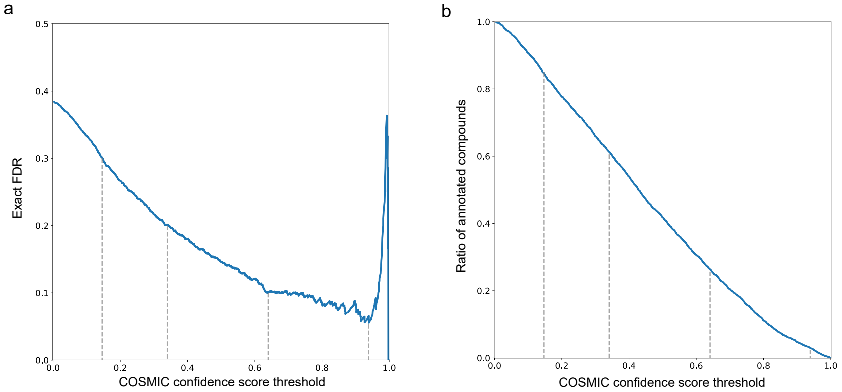


Figure 6.23: COSMIC confidence score vs. exact FDR and ratio of annotated compounds. Independent data (Agilent, QTOF), 20 eV, medium noise, $N = 3,013$. We vary the confidence score threshold and present the resulting exact FDR (a) and the ratio of annotated compounds (b). Dashed lines indicate COSMIC confidence score thresholds of 0.94, 0.64, 0.34, and 0.14, corresponding to exact FDR levels of roughly 5 %, 10 %, 20 %, and 30 %, respectively. The spike for high thresholds beyond 0.9 is an artifact of the small number of hits that pass this threshold; hence, a few incorrect hits with high confidence score can lead to high FDR. In practice, confidence scores depend on numerous factors such as the overall quality of the data and the identity of the query compounds. Hence, these thresholds come with no guarantee in either direction: For example, in the CASMI 2016 dataset, a smaller confidence score threshold of 0.53 corresponded to exact FDR 10 %, and using the abovementioned threshold of 0.64 would have returned fewer hits than possible. Nevertheless, these thresholds may serve as a starting point for practitioners.

To give practitioners an initial starting point for interpreting the confidence score in practice, we give a suggestion for rule of thumb thresholds in Fig. 6.23.

6.4.6 The COSMIC Workflow

In the previous chapters, we focused on developing the COSMIC confidence score, which can be used to better separate correct from incorrect hits. This confidence score is now integrated into the COSMIC workflow, that combines it with the selection or generation of a structure database and searching in that structure database with CSI:FingerID. COSMIC can process data at a repository scale, allowing us to repurpose the quickly-growing public metabolomics data. Doing so, COSMIC may allow us to flip the metabolomics workflow. (Fig. 6.24): We may concentrate on metabolites annotated with high confidence, without the need for intricate prior experiments, and try to develop a biological hypothesis from these annotations. Annotated fragmentation spectra can subsequently be searched in other datasets via “classical” spectral library search at the repository scale [178], allowing a more comprehensive annotation of public metabolomics datasets.

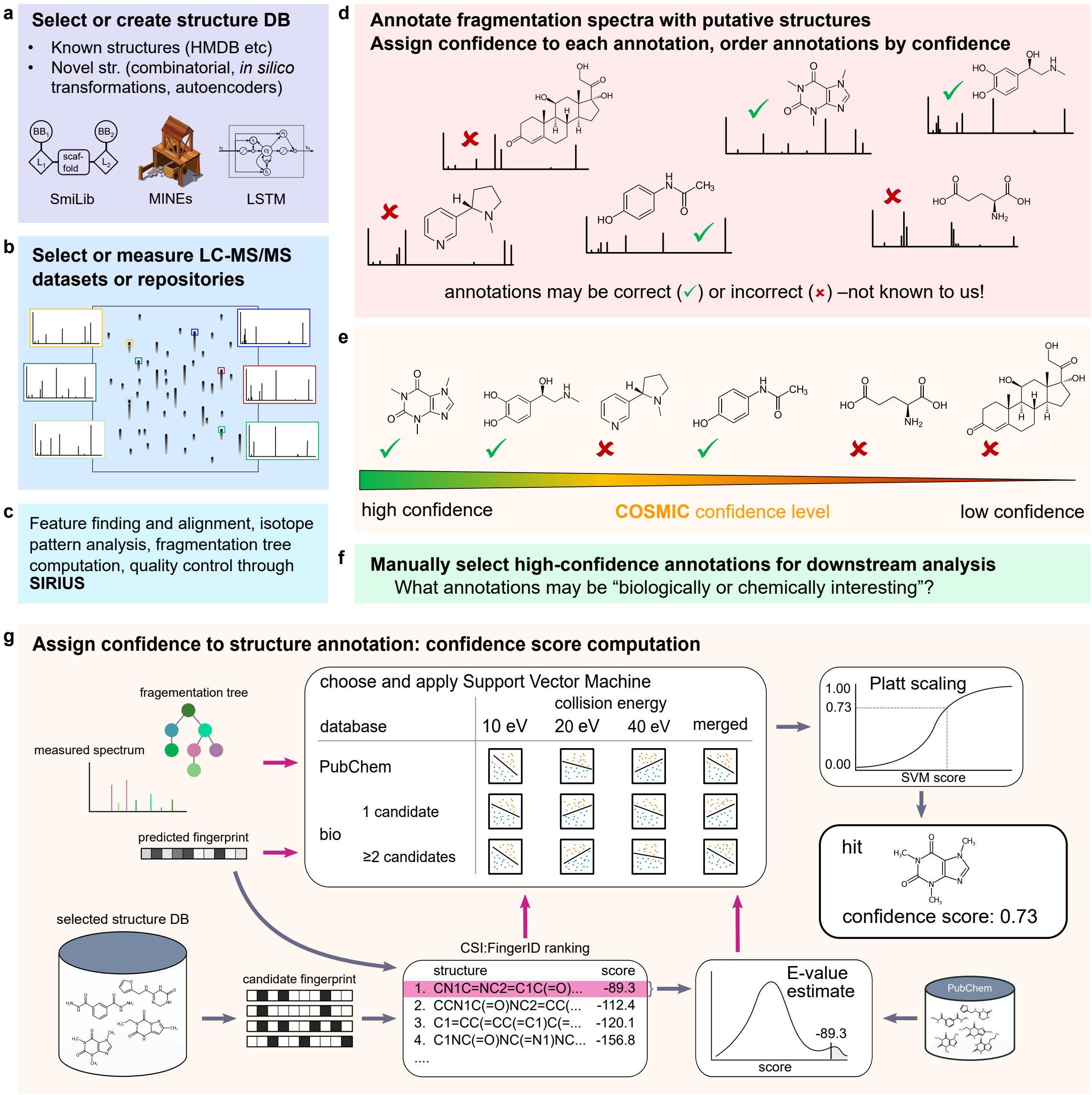

a
Select or create structure DB
• Known structures (HMDB etc)
• Novel str. (combinatorial, in silico transformations, autoencoders)
BB1
BB2
L1
scaf-fold
L2
SmiLib
MINEs
LSTM
b
Select or measure LC-MS/MS datasets or repositories
c
Feature finding and alignment, isotope pattern analysis, fragmentation tree computation, quality control through SIRIUS
d
Annotate fragmentation spectra with putative structures
Assign confidence to each annotation, order annotations by confidence
annotations may be correct (✓) or incorrect (✗) –not known to us!
e
high confidence
COSMIC confidence level
low confidence
f
Manually select high-confidence annotations for downstream analysis
What annotations may be "biologically or chemically interesting"?
g
Assign confidence to structure annotation: confidence score computation
fragementation tree
measured spectrum
predicted fingerprint
choose and apply Support Vector Machine
collision energy
database
10 eV
20 eV
40 eV
merged
PubChem
bio
1 candidate
≥2 candidates
Platt scaling
1.00
0.73
0.00
SVM score
hit
confidence score: 0.73
selected structure DB
candidate fingerprint
CSI:FingerID ranking
structure
score
1. CN1C=NC2=C1C(=O)... -89.3
2. CCN1C(=O)NC2=CC(... -112.4
3. C1=CC(=CC(=C1)C(=... -120.1
4. C1NC(=O)NC(=N1)NC... -156.8
....
E-value estimate
-89.3
score
PubChem

Figure 6.24: COSMIC workflow. (a) Select or create a structure database; this can be an existing structure database such as the Human Metabolome Database, or generated explicitly for this purpose. (b) Select or measure an LC-MS/MS dataset or select a complete data repository (data repurposing). (c) Data processing through SIRIUS. (d) Structure annotation of fragmentation spectra through CSI:FingerID; only the candidate top-ranked by CSI:FingerID is considered. We stress that at this point, there is no ordering of hits. (e) Each hit (structure annotation) is assigned a confidence score; annotations are then ordered by confidence, allowing users to concentrate on high-confidence annotations. (f) High-confidence annotations can be used to develop or test a biological hypothesis. (g) Detailed confidence score computation for the structure annotation of a spectrum (hit) applied in (e), including feature calculation (magenta arrows), E-value estimation, selection and application of the appropriate Support Vector Machine, and Platt scaling. Notably, COSMIC can annotate metabolites at an early stage of a biological analysis.

7 Practical Application

In this section we now apply the COSMIC workflow to real-world biological data to prove its capabilities of finding truly novel structures, as well as to demonstrate its potential in large-scale annotation studies. We present one use case where we search for novel bile acid conjugates in mouse fecal data, one use case where we process human-related public datasets to annotate structures missing from HMDB [192] and finally one where we process 123 public datasets in a repository-scale annotation study.

7.1 Data Preprocessing

SIRIUS 4 [47] was used to process LC-MS/MS runs and MassIVE datasets provided in mzML or mzXML format. Feature detection in SIRIUS 4 is similar in spirit to a targeted analysis: Instead of searching for all features in a run, SIRIUS first collects all fragmentation spectra and their precursor information, then searches for features that are associated with those fragmentation spectra (precursor ions, adduct ions, isotope peaks). Adducts and isotopes were detected using predefined lists of mass differences. Fragmentation spectra assigned to the same feature (precursor ion) are merged using an agglomerative clustering algorithm based on cosine distance. Compounds with mass beyond 700 Da were discarded to avoid high running time. MassIVE datasets that exceeded 600 LC-MS/MS runs were split to reduce memory consumption.

We use both isotope patterns and fragmentation patterns to determine the molecular formula *de novo* using SIRIUS 4 with default parameters and mass accuracy 10 ppm. CSI:FingerID with default parameters was used to rank structure candidates. We use SIRIUS default soft thresholding of molecular formulas when querying CSI:FingerID structure candidates. For confidence score computation, we restrict the candidate list to those candidates with the same molecular formula as the highest-scoring candidate (hit). We used the highest-scoring structure candidate and the corresponding fragmentation tree, isotope pattern and structure candidate list features for COSMIC.

For the mice fecal dataset, SIRIUS results were imported into GNPS, and data were further annotated and explored by performing feature-based molecular networking and spectral library search on GNPS.

7.2 Annotation of Novel Bile Acid Conjugates

Bile acids are amphipathic molecules that act as signalling molecules in many organisms [72], and facilitate the solubility of lipids in the small intestine. Bile acids and their conjugates are very structurally diverse, and their quantification profile is considered highly species-dependent [164]. Recently, three previously unknown bile acid conjugates were discovered [129], supporting the hypothesis that additional novel conjugations of amino acids to bile acid core structures exist.

We explored this hypothesis applying COSMIC to a public mice fecal metabolomics dataset. Plausible bile acid conjugate structures were computed by combinatorially adding amino acids to bile acid cores, yielding 28,630 plausible bile acid conjugates (see Chapter 5). The COSMIC workflow was then applied to search the combinatorial bile acid conjugate structure database. For the mice fecal dataset, MS/MS measurements were taken with a collision energy of 30 eV; we used the closest COSMIC version trained on 40 eV spectra. The output of this workflow is an ordered list of 1,456 COSMIC structure annotations ("MS features", see Hoffmann et al. [70]). In case multiple compounds were annotated with the same structure (for example, compounds being present in multiple runs, and different adducts of the same compound), entries in the COSMIC output were merged and represented by the hit with the highest confidence. This reduces the output to 626 unique structure annotations. Of these, 113 were present in PubChem. Here, we concentrated on the 513 "truly novel" bile acid conjugates.

The top 12 most confident bile acid conjugate annotations (Fig. 7.1) were manually inspected and their fragmentation spectra interpreted by Louis-Felix Nothias for validation. Nine of these annotations were found to be consistent, and one was found to be inconsistent with the fragmentation analysis. Annotations of two "truly novel" bile acid conjugates, tryptophan (Trp) and phenylalanine (Phe) conjugates of chenodeoxycholic acid (CDCA), were verified by comparing their fragmentation spectra and retention times with those of synthetic standards. Manual inspection and verification as well as the spectral comparison with synthetic standards was performed by Louis-Felix Nothias. MASST [178] was used to find samples of species that contain the novel bile acid conjugate structures in public mass spectrometry datasets, including MassIVE-GNPS [177], MetaboLights [65] and Metabolomics Workbench [160]. In addition to the MASST search, Louis-Felix Nothias carried out a statistical analysis of the novel bile acid conjugate's quantification profile in mice with a high fat diet, see Hoffmann et al. [70] for more technical details.

7.3 Annotating Structures Missing from HMDB

The Human Metabolome Database (HMDB) [192] contains the by far most comprehensive collection of molecular structures found in or on the human body, with version 4.0 embracing 114,265 structures. Yet, certain molecular structures connected to human metabolism may currently be missing from this database. To test this hypothesis, we searched the human dataset against the biomolecule structure database; this comprises ten MassIVE datasets [177] with 2,666 LC-MS/MS runs from different sources (serum, plasma, lips, tongue, teeth, fecal, urine). To estimate a reasonable COSMIC confidence score cut-off, we made use of our reference data evaluation results. In our evaluation using independent data, collision energy 20 eV and medium noise, a confidence score threshold of 0.64 corresponded to FDR 10 % (see Fig. 6.23). Our implicit assumption is that for the biological data, this threshold will correspond to a similar FDR. It must be understood that we cannot guarantee a similar FDR for structure annotations, given our inability to accurately estimate FDR. Clearly, numerous hits with confidence below this threshold will nevertheless be correct.

We searched the human dataset against the biomolecule structure database; this resulted in 114,012 hits. Multiple hits can annotate the same structure; for example, these hits may originate from different LC-MS/MS runs or different adducts. Hence, we report unique structures instead, where the hit with the highest confidence is used as a representative

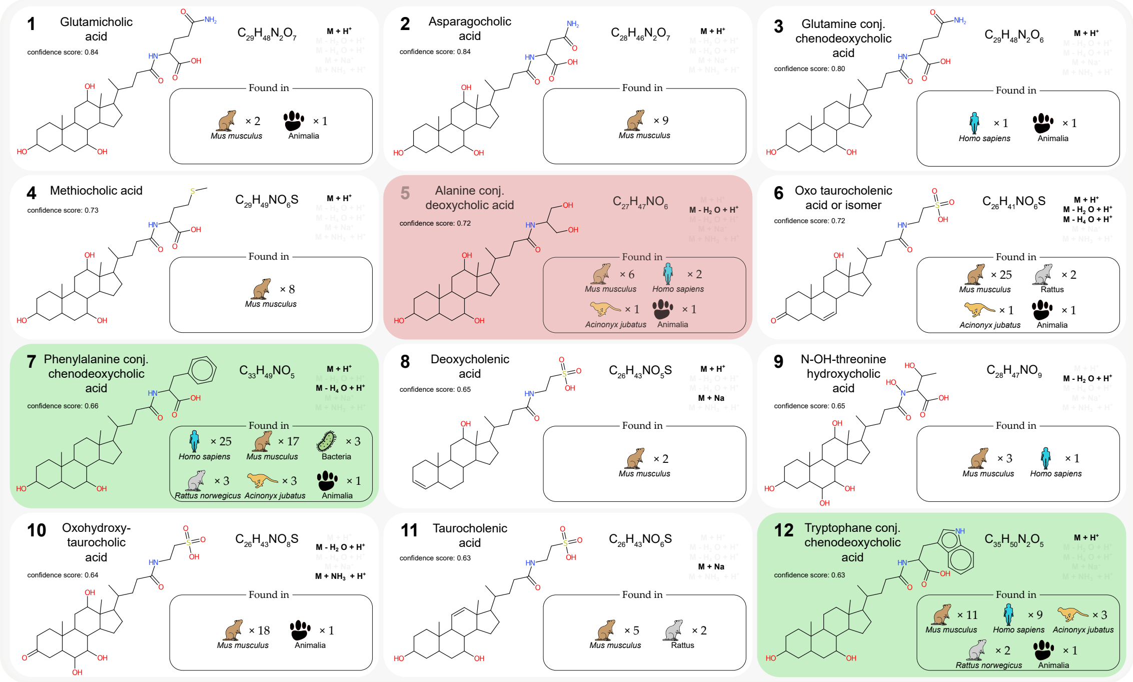


Figure 7.1: Applying COSMIC to discover novel bile acid conjugates in a mice fecal dataset. Top 12 highest-scoring COSMIC annotations of "truly novel" bile acid conjugates. Bile acid conjugates which are also present in PubChem are omitted from the list; see Hoffmann et al. [70] for the complete list. For each bile acid conjugate we report its chemical name, putative structure, molecular formula, and adducts of annotations for this structure. In addition, we report the confidence scores for each annotation. We also report species and number of datasets with spectral matches from a MASST search. Two annotations verified by authentic standards are highlighted in green, the single incorrect annotation in red.

for that structure. This resulted in 24,554 unique structures being annotated, of which 3,167 (12.9 %) were present in the CSI training dataset. We now filter the 24,554 structure annotations for high confidence (score threshold 0.64), resulting in 911 structure annotations. Of these high-confidence annotations, 475 (52.1 %) were present in the CSI training dataset, leaving us with 436 (47.9 %) high-confidence novel structure annotations. Finally, we excluded all hits with structures in the HMDB structure database, resulting in 21,128 unique structure annotations, 436 high-confidence structure annotations, and 315 high-confidence structure annotations without reference MS/MS data (Fig. 7.2). Of the 315 novel structures, 48 were proteinogenic peptides (peptides made from proteinogenic amino acids), which are not considered novel metabolite structures. Text-based lists including all structures shown in Fig. 7.2 are available from [70].

We searched 14-character InChIKeys of all 267 novel metabolite structures in the current version of HMDB (Feb 2021) and found that at least 23 of these structures are present in the current HMDB version. The exact number may be slightly higher, as structures from the current HMDB version were not standardised using the PubChem standardisation procedure. Notably, the recent inclusion of structures in HMDB does not mean that reference MS/MS data are available for these structures. It does however indicate, that many of the novel structures are indeed present in human samples. Michael Witting manually verified the 315 structures by checking common neutral losses and fragments, and by comparison of spectra against reference spectra from similar compounds. Hits are available from `https://bio.informatik.uni-jena.de/cosmic/`; users can view, discuss and verify annotated structures there. Based on characteristic fragmentation patterns, different acyl-carnitines and N-acyl-amino acids not part of HMDB were annotated. N-acyl amino acids play an important biochemical role in mitochondria [97]. From 30 spectra annotated as acyl-carnitines with high confidence, 21 were presumably correct based on manual verification. The verification of this analysis was performed by Michael Witting, see Hoffmann et al. [70] for details.

7.4 Repository-scale Annotation

In the previous two applications, we used the COSMIC workflow to annotate structures in a sort of confined context, focusing on either Bile Acid conjugates or structures in human samples. To demonstrate that COSMIC can be applied at a repository scale, we searched the Orbitrap dataset with 17,414 LC-MS/MS runs against the biomolecule structure database; this resulted in 979,521 hits. Again, multiple hits can annotate the same structure; the above hits correspond to 77,932 unique annotated structures, of which 8,172 (10.5 %) were present in the CSI training dataset. We now filter the 77,932 structure annotations for high confidence (score threshold 0.64), resulting in 3,530 structure annotations. Of these high-confidence structure annotations, 1,815 (51.4 %) were present in the CSI training dataset, leaving 1,715 (48.6 %) high-confidence novel structure annotations (Fig. A.8). Again, all hits of the Orbitrap dataset can be accessed via a web interface available from `https://bio.informatik.uni-jena.de/cosmic`.

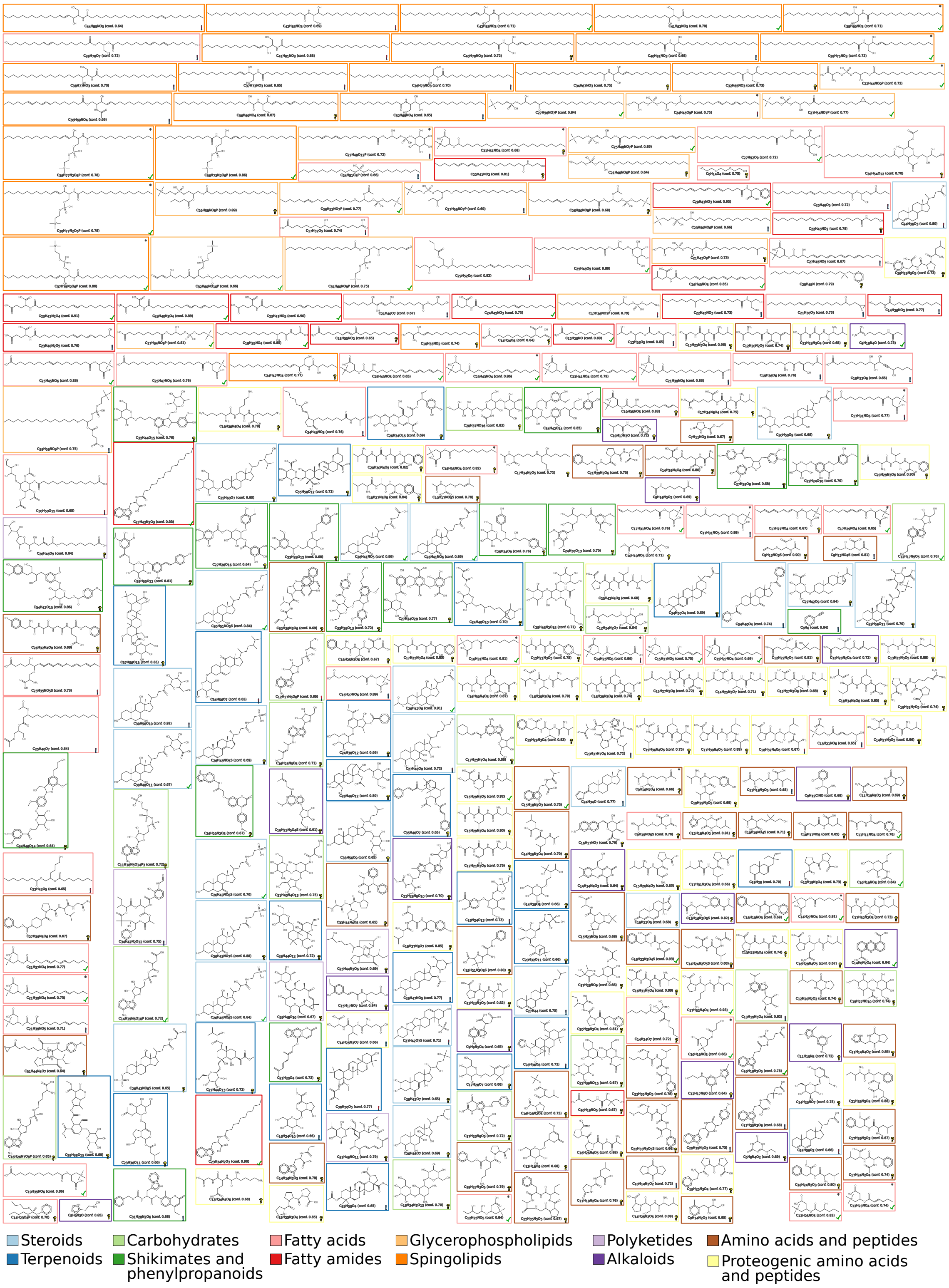


Figure 7.2: The 315 molecular structure not contained in HMDB annotated with high confidence in the human dataset. Confidence score threshold 0.64 was used. For none of these structures, reference MS/MS data are available. Structures are shown with identification number (ID), molecular formula and COSMIC confidence score. Structures present in the latest version of HMDB (Feb 2021) are marked by an asterisk. Colours indicate compound classes. Notably, 48 compounds were annotated as proteinogenic peptides; these structures were absent from HMDB but are clearly no novel metabolite structures. Lipid structures must be interpreted with some care: It is understood that neither COSMIC nor any other method can deduce, say, the position of the double bond in a carbon chain from MS/MS data alone; rather, this happens to be the candidate present in our biomolecule structure database.

8 Conclusion

In this thesis, we presented the COSMIC (Confidence Of Small Molecule IdentifiCations) confidence score, a machine learning-based approach that assigns confidence to structure annotations by CSI:FingerID. Integrated into SIRIUS, it is part of the COSMIC workflow. We established that hit scores returned by popular tools such as CSI:FingerID, MetFrag, MAGMa+ and CFM-ID are unable to separate correct from incorrect annotations. The COSMIC confidence score outperforms these hit scores by a wide margin for the task of score separation. On the CASMI 2016 contest data, our confidence score was used to correctly annotate over 46% of queries with a false discovery rate below 10% when searching the biomolecule structure database. In comparison, no other scoring method was able to correctly annotate even 5% of queries below a FDR of 10%. We extensively evaluated the COSMIC confidence score on training data as well as independent data. Since reference data used in training and evaluation datasets is usually of higher quality than real-world data from actual complex mixtures, we artificially noisified evaluation data. We showed that the separation power of the COSMIC confidence score is only minorly affected when presented with this noisified data. Reference spectra are often also measured using multiple collision energies, which can then be combined for structure elucidation. We evaluated the COSMIC confidence score for four different collision energy settings, and show that measuring multiple collision energies is consistently beneficial for annotation as well as separation performance.

Rather surprisingly, COSMIC outperformed spectral library search for dereplication, even though it was neither designed nor optimised for this task. On the independent dataset, COSMIC correctly annotated over 40% of queries below an FDR of 5%, while spectral library search correctly annotated below 2% of queries.

The COSMIC confidence score is based on linear support vector machines. This simple machine learning model was deliberately chosen to retain classifier interpretability and reduce overfitting. SVMs were trained for four collision energy settings individually. We developed an E-value estimation using proxy decoys from PubChem, which on its own showed improved separation power over the CSI:FingerID hit scoring. We integrated these estimated E-values into our SVM-based models as features. Other features are strongly intertwined with SIRIUS and CSI:FingerID, such as the relative percentage of peak intensities explained by the fragmentation tree of a molecular formula candidate and many structure candidate list features. The first support vector machine models we trained showed clear signs of overfitting. In particular, feature weight signs often defied common sense, which we corrected by enforcing feature directionality. This was only possible because of the simple, interpretable classifier. We showed that in those cases, where incorrect annotations received a high confidence score, the incorrect structure and the true structure are exceedingly similar. On the contrary, incorrect annotations receiving a very low confidence score proved to be very dissimilar from the true structure. We showed that the confidence score is not restricted to very small masses or a high number of peaks in the spectrum.

The COSMIC confidence score offers no statistical interpretability. We showed how to transform confidence values into FDR estimates, but unfortunately these estimates are of poor quality. This however does not impede COSMIC's ability to rapidly test biological hypotheses and even annotate previous unknown, truly novel structures. We demonstrated this by processing publicly available mice fecal data, and searching fingerprints in a structure database of hypothetical bile acid conjugate structures. Of the twelve novel structures that were assigned the highest confidence, two were experimentally verified by synthesis and another nine were manually validated by an expert. This automated discovery of novel structures was performed within mere weeks, and up until validation done completely *in-silico*. We then showed the power of the COSMIC workflow on a larger, fully automated scale. We processed ten datasets containing measurements of samples taken from humans, and were able to annotate 267 metabolite structures that were missing from the Human Metabolome Database (HMDB). Lastly, we processed 123 publicly available datasets consisting of 17,414 LC-MS/MS runs in a repository-scale annotation study. We did not restrict ourselves to specific species or compound classes, and were able to annotate 1,715 novel structures with high confidence. These annotations are accessible for the community via a dedicated website, where one can rate and comment on annotation results. The COSMIC workflow is integrated in SIRIUS and available to the public.

Ongoing Research and Future Work

While the COSMIC confidence score already is a large performance increase over existing scorings for hit separation, multiple improvement opportunities exist. Parallel to the work shown in this thesis, Dührkop et al. presented CANOPUS [48], a method that predicts compound classes for an LC-MS/MS spectrum without the need of structure databases. CANOPUS results may also be used as features in the COSMIC SVMs in the future, to assess if the compound class of the hit structure and the compound class predicted by CANOPUS match. In a similar spirit, Kai Dührkop also implemented "Epimetheus" into SIRIUS, which uses combinatorial fragmentation to assess how well a CSI:FingerID structure candidate fits to the input spectrum. For that, *in-silico* generated substructures are matched to the spectrum peaks and a scoring is used to evaluate the match. This scoring could also be integrated into COSMIC as a feature, to have a more direct relation to the input data.

When we applied COSMIC to real-world biological data, the correct molecular formula is not known to us, like it is in evaluation. In this thesis we used the predicted fingerprint and fragmentation tree features of the molecular formula candidate that produced the highest scoring structure candidate. A different approach would be to merge structure candidate lists of all molecular formulas that lie above some threshold. This would make sense, because the confidence score can be somewhat influenced by the size and completeness of the structure candidate lists. We have already started implementation of this approach and integrated it into SIRIUS, large scale evaluations of it are however still to be done. Similarly, it might make sense to train additional SVM models for different structure candidate list sizes. Currently, we differentiate between list sizes of one and "two or more" solely based on some features being incomputable for size one lists. Since the completeness of a large list is often times much higher than the completeness of a small list, we conjecture that the feature weight of for example the "Score Diff" feature could highly vary between different list sizes.

COSMIC was designed to be structure-precise, meaning that if multiple structure candidates exist, that are extremely similar to the top scoring candidate, it will usually receive a very low confidence. This makes sense, as mass spectrum and molecular fingerprint are also near indistinguishable in those cases. In application however, the approximate structure of a query spectrum can be very valuable information. In the future, we want to implement such an “approximate structure mode” into SIRIUS, which would merge structure candidates that are very similar to each other into a cluster and treat them as a singular candidate.

Fleming Kretschmer is currently working on predicting the retention time of small molecule structures, which could lead to important orthogonal information on structure candidates. Some structures that are very similar in two-dimensional space exhibit distinguishable retention times. This information could then also be used in COSMIC as a feature. Just like how the bile acid conjugate study was conducted, we are looking for other metabolites that consist of larger building blocks and can be combinatorially constructed to generate novel compound structure databases. Searching publicly available datasets in these novel structure databases might reveal more truly novels and aid in the understanding of biological processes. In addition to creating hypothetical structure databases combinatorially, we are also trying to create hypothetical novel structures using machine learning, specifically autoencoder. On a bigger scale, our next project is what we internally call “Project Harvester”. Using the COSMIC workflow, we want to process all available public LC-MS/MS datasets and use hits with high confidence as additional training data for CSI:FingerID and COSMIC. Even if some annotations are incorrect, we conjecture that the molecular fingerprint is sufficiently similar to what would be the true structure’s fingerprint. In that fashion, available training data for CSI:FingerID and other machine learning approaches in the field could be greatly increased. Going back to the topic of my masters thesis, chimeric spectra are still highly present in every day LC-MS/MS data with currently no good way to resolve them. The approach we used in my masters thesis showed promise, but we were lacking real-world evaluation data to further develop and evaluate the approach. Michael Witting in Munich has declared his interest in the topic, and might be able to measure some data in the near future.

9 Data and Code Availability

9.1 Data Availability

Input mzML/mzXML files are available at MassIVE (`https://massive.ucsd.edu/`) with accession nos. MSV000082973 (mice fecal dataset); MSV000084630 (mass spectrometry analysis of the synthetic standards for Phe-CDCA and Trp-CDCA); MSV000083559, MSV000079651, MSV000080167, MSV000080469, MSV000080533, MSV000080627, MSV000081351, MSV000082261, MSV000082629, MSV000082630 (human dataset). See Appendix Table A.1 for accession numbers of the Orbitrap dataset. The bile acid conjugate structure database is available at `https://github.com/lfnothias/Combinatorial_BileAcids_DB_COSMIC`. Spectral libraries generated from the high-confidence COSMIC annotations of the mice fecal, human and Orbitrap datasets are available from `https://bio.informatik.uni-jena.de/cosmic/`. For further data availability of spectra involved in the manual verification of the bile acid conjugates, we refer to [70].

9.2 Code Availability

COSMIC is written in Java, and is integrated into the current release version of SIRIUS 4; it is open source under the GNU General Public License (version 3). It is available for Windows, macOS X, and Linux operating systems. We also provide source code, executable binaries, living documentation, training videos, sample data as well as the public part of the training data on the SIRIUS website (`https://bio.informatik.uni-jena.de/sirius/`); a source copy is hosted on GitHub (`https://github.com/boecker-lab/sirius/`). Scripts for generating the bile acid conjugate structure database are available from `https://github.com/lfnothias/Combinatorial_BileAcids_DB_COSMIC`.

Bibliography

[1] F. Allen, R. Greiner, and D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015. doi: 10.1007/s11306-014-0676-4.

[2] F. Allen, A. Pon, R. Greiner, and D. Wishart. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Anal Chem*, 88(15):7689–7697, 2016. doi: 10.1021/acs.analchem.6b01622.

[3] T. Alon and A. Amirav. Isotope abundance analysis methods and software for improved sample identification with supersonic gas chromatography/mass spectrometry. *Rapid Commun Mass Spectrom*, 20(17):2579–2588, 2006. doi: 10.1002/rcm.2637.

[4] R. Altenburger, S. Ait-Aissa, P. Antczak, T. Backhaus, D. Barceló, T.-B. Seiler, F. Brion, W. Busch, K. Chipman, M. L. de Alda, et al. Future water quality monitoring—adapting tools to deal with mixtures of pollutants in water resource management. *Science of the total environment*, 512:540–551, 2015.

[5] D. C. Anderson, W. Li, D. G. Payan, and W. S. Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res*, 2 (2):137–146, 2003.

[6] M. Askenazi and M. Linial. ARISTO: ontological classification of small molecules by electron ionization-mass spectrometry. *Nucleic Acids Res*, 39(Web Server issue): W505–W510, 2011. doi: 10.1093/nar/gkr403.

[7] L. Bahiense, G. Manić, B. Piva, and C. C. De Souza. The maximum common edge subgraph problem: A polyhedral investigation. *Discrete Applied Mathematics*, 160 (18):2523–2541, 2012.

[8] D. Bajusz, A. Rácz, and K. Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminformatics*, 7(1):20, 2015.

[9] M. Baker. Metabolomics: from small molecules to big ideas. *Nature Methods*, 8(2): 117–121, 2011.

[10] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, and M. Dunkel. Super Natural II - a database of natural products. *Nucleic acids research*, 43:D935–D939, 2015.

[11] D. A. Barkauskas and D. M. Rocke. A general-purpose baseline estimation algorithm for spectroscopic data. *Anal Chim Acta*, 657(2):191–197, 2010. doi: 10.1016/j.aca.2009.10.043.

[12] S. Becker. *Inorganic mass spectrometry: principles and applications.* John Wiley & Sons, 2008.

[13] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How similar are similarity searching methods? a principal component analysis of molecular descriptor space. *J Chem Inf Model*, 49(1):108–119, 2009. doi: 10.1021/ci800249s.

[14] S. Böcker. Searching molecular structure databases using tandem MS data: are we there yet? *Curr Opin Chem Biol*, 36:1–6, 2017. doi: 10.1016/j.cbpa.2016.12.010.

[15] S. Böcker. Algorithmic mass spectrometry: From molecules to masses and back again. `https://bio.informatik.uni-jena.de/textbook-algoms/`, Friedrich-Schiller-Universität Jena, Jena, Germany, 2019. Version 0.8.2.

[16] S. Böcker and K. Dührkop. Fragmentation trees reloaded. *J Cheminformatics*, 8:5, 2016. doi: 10.1186/s13321-016-0116-8.

[17] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. doi: 10.1093/bioinformatics/btn270. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[18] S. Böcker, M. Letzel, Zs. Lipták, and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect Notes Comput Sci*, pages 12–23. Springer, Berlin, 2006. doi: 10.1007/11851561_2.

[19] S. Böcker, Zs. Lipták, M. Martin, A. Pervukhin, and H. Sudek. DECOMP–from interpreting mass spectrometry peaks to solving the Money Changing Problem. *Bioinformatics*, 24(4):591–593, 2008. doi: 10.1093/bioinformatics/btm631.

[20] S. Böcker, M. Letzel, Zs. Lipták, and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009. doi: 10.1093/bioinformatics/btn603.

[21] R. Bonner and G. Hopfgartner. Swath data independent acquisition mass spectrometry for metabolomics. *TrAC Trends in Analytical Chemistry*, 120:115278, 2019.

[22] J. Boström, A. Hogner, and S. Schmitt. Do structurally similar ligands bind in a similar fashion? *J Med Chem*, 49(23):6716–6725, 2006. doi: 10.1021/jm060167o.

[23] A. Bouslimani, R. da Silva, T. Kosciolek, S. Janssen, C. Callewaert, A. Amir, K. Dorrestein, A. V. Melnik, L. S. Zaramela, J.-N. Kim, G. Humphrey, T. Schwartz, K. Sanders, C. Brennan, T. Luzzatto-Knaan, G. Ackermann, D. McDonald, K. Zengler, R. Knight, and P. C. Dorrestein. The impact of skin care products on skin chemistry and microbiome dynamics. *BMC Biol*, 17(1):47, 2019. doi: 10.1186/s12915-019-0660-6.

[24] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, and J. Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12): i28–i36, 2016. doi: 10.1093/bioinformatics/btw246. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2016).

[25] C. Brouard, E. Bach, S. Böcker, and J. Rousu. Magnitude-preserving ranking for structured outputs. In *Proc. of Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 407–422. PMLR, 2017.

[26] C. Brouard, A. Bassé, F. d'Alché Buc, and J. Rousu. Improved small molecule identification through learning combinations of kernel regression models. *Metabolites*, 9(8), 2019. doi: 10.3390/metabo9080160.

[27] N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *J Chem Inf Model*, 59(3):1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

[28] M. L. Brownawell and J. S. Filippo Jr. A program for the synthesis of mass spectral isotopic abundances. *Journal of Chemical Education*, 59(8):663–65, 1982.

[29] T. Cakir, K. R. Patil, Z. I. Önsan, K. Ö. Ülgen, B. Kirdar, and J. Nielsen. Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular systems biology*, 2(1):50, 2006.

[30] L. Cao, M. Guler, A. Tagirdzhanov, Y.-Y. Lee, A. Gurevich, and H. Mohimani. MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nat Commun*, 12(1):3718, 2021. doi: 10.1038/s41467-021-23986-0.

[31] C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. *ACM Trans Intell Syst Technol*, 2(3):27:1–27:27, 2011. doi: 10.1145/1961189.1961199.

[32] J. Colinge, A. Masselot, M. Giron, T. Dessingy, and J. Magnin. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3:1454–1463, 2003. doi: 10.1002/pmic.200300485.

[33] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

[34] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *J Mach Learn Res*, 13(1):795–828, 2012.

[35] R. R. da Silva, P. C. Dorrestein, and R. A. Quinn. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A*, 112(41):12549–12550, 2015. doi: 10.1073/pnas.1516878112.

[36] R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes, and P. C. Dorrestein. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput Biol*, 14(4):e1006089, 2018.

[37] R. Daly, S. Rogers, J. Wandy, A. Jankevics, K. E. V. Burgess, and R. Breitling. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics*, 30(19):2764–2771, 2014. doi: 10.1093/bioinformatics/btu370.

[38] T. De Vijlder, D. Valkenborg, F. Lemière, E. P. Romijn, K. Laukens, and F. Cuyckens. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass spectrometry reviews*, 37(29120505):607–629, 2018.

[39] Y. Djoumbou-Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, and D. S. Wishart. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminformatics*, 8(1):61, 2016. doi: 10.1186/s13321-016-0174-y.

[40] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la Fuente, R. Greiner, C. Manach, and D. S. Wishart. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminformatics*, 11(1):2, 2019. doi: 10.1186/s13321-018-0324-5.

[41] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, and D. S. Wishart. CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites*, 9(4):72, 2019. doi: 10.3390/metabo9040072.

[42] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman. Analysis and comparison of 2d fingerprints: Insights into database screening performance using eight fingerprint methods. *J Mol Graphics Modell*, 29(2):157–170, 2010. doi: 10.1016/j.jmgm.2010.05.008.

[43] A. Z. Dudek, T. Arodz, and J. Gálvez. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High T Scr*, 9(3):213–228, 2006.

[44] K. Dührkop and S. Böcker. Fragmentation trees reloaded. Technical report, 2014.

[45] K. Dührkop, M. Ludwig, M. Meusel, and S. Böcker. Faster mass decomposition. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2013)*, volume 8126 of *Lect Notes Comput Sci*, pages 45–58. Springer, Berlin, 2013. doi: 10.1007/978-3-642-40453-5_5.

[46] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A*, 112(41):12580–12585, 2015. doi: 10.1073/pnas.1509788112.

[47] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, and S. Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 16(4):299–302, 2019. doi: 10.1038/s41592-019-0344-8.

[48] K. Dührkop, L. F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein, and S. Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*, 39(4):462–471, 2021. doi: 10.1038/s41587-020-0740-8.

[49] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*, 42(6):1273–1280, 2002.

[50] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86, 2002.

[51] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, 2007. doi: 10.1038/nmeth1019.

[52] M. Ernst, L.-F. Nothias, J. J. J. van der Hooft, R. R. Silva, C. H. Saslis-Lagoudakis, O. M. Grace, K. Martinez-Swatson, G. Hassemer, L. A. Funez, H. T. Simonsen, M. H. Medema, D. Staerk, N. Nilsson, P. Lovato, P. C. Dorrestein, and N. Rønsted. Assessing specialized metabolite diversity in the cosmopolitan plant genus *Euphorbia* l. *Front Plant Sci*, 10:846, 2019. doi: 10.3389/fpls.2019.00846.

[53] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *J Mach Learn Res*, 9:1871–1874, 2008.

[54] A. R. Fernie, R. N. Trethewey, A. J. Krotzky, and L. Willmitzer. Metabolite profiling: From diagnostics to systems biology. *Nat Rev Mol Cell Biol*, 5(9):763–769, 2004. doi: 10.1038/nrm1451.

[55] O. Fiehn, J. Kopka, R. N. Trethewey, and L. Willmitzer. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal Chem*, 72(15):3573–3580, 2000. doi: 10.1021/ac991142i.

[56] O. Fiehn, D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L. Sumner, R. Goodacre, N. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J. Lindon, and S.-A. Sansone. The metabolomics standards initiative (MSI). *Metabolomics*, 3(3):175–178, 2007. doi: 10.1007/s11306-007-0070-6.

[57] E. F. Fornasiero, S. Mandad, and H. e. a. Wildhagen. Precisely measured protein lifetimes in the mouse brain reveal differences across tissues and subcellular fractions. *Nat Commun*, 9(4230), 2018. doi: https://doi.org/10.1038/s41467-018-06519-0.

[58] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[59] M. Gerlich and S. Neumann. MetFusion: integration of compound identification strategies. *J Mass Spectrom*, 48(3):291–298, 2013. doi: 10.1002/jms.3123.

[60] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proc. of the International conference on Artificial Intelligence and Statistics (AISTATS 2011)*, pages 315–323, 2011.

[61] J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu, and X. Xu. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One*, 8(4): 1–10, 2013.

[62] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, and A. Wassermann. MOLGEN 5.0, a molecular structure generator. In *Advances in mathematical chemistry and applications*, pages 113–138. Elsevier, 2015.

[63] V. D. Hähnke, S. Kim, and E. E. Bolton. PubChem chemical structure standardization. *J Cheminformatics*, 10(1):36, 2018. doi: 10.1186/s13321-018-0293-8.

[64] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res*, 41(Database issue):D456–D463, 2013. doi: 10.1093/nar/gks1146.

[65] K. Haug, K. Cochrane, V. C. Nainala, M. Williams, J. Chang, K. V. Jayaseelan, and C. O'Donovan. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res*, 48(D1):D440–D444, 2019. doi: 10.1093/nar/gkz1019.

[66] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola, and J. Rousu. FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, 2008. doi: 10.1002/rcm.3701.

[67] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics*, 28(18):2333–2341, 2012. doi: 10.1093/bioinformatics/bts437.

[68] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. Inchi, the iupac international chemical identifier. *J Cheminformatics*, 7:23, 2015. doi: 10.1186/s13321-015-0068-4.

[69] A. W. Hill and R. J. Mortishire-Smith. Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun Mass Spectrom*, 19(21):3111–3118, 2005. doi: 10.1002/rcm.2177.

[70] M. A. Hoffmann, L.-F. Nothias, M. Ludwig, M. Fleischauer, E. C. Gentry, M. Witting, P. C. Dorrestein, K. Dührkop, and S. Böcker. High-confidence structural annotation of metabolites absent from spectral libraries. *Nature biotechnology*, 40(3): 411–421, 2022. doi: 10.1038/s41587-021-01045-9. https://doi.org/10.1038/s41587-021-01045-9.

[71] N. Hoffmann and J. Stoye. ChromA: Signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics*, 25(16):2080–2081, 2009. doi: 10.1093/bioinformatics/btp343.

[72] A. F. Hofmann and L. R. Hagey. Key discoveries in bile acid chemistry and biology and their clinical applications: history of the last eight decades. *J Lipid Res*, 55(8): 1553–95, 2014. doi: 10.1194/jlr.R049437.

[73] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka.

MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, 2010. doi: 10.1002/jms.1777.

[74] S. M. Houten. Metabolomics: Unraveling the chemical individuality of common human diseases. *Ann Med*, 41(6):402–407, 2009. doi: 10.1080/07853890902729794.

[75] M. P. Jedrychowski, E. L. Huttlin, W. Haas, M. E. Sowa, R. Rad, and S. P. Gygi. Evaluation of hcd-and cid-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Molecular & Cellular Proteomics*, 10(12), 2011.

[76] J. G. Jeffryes, R. L. Colastani, M. Elbadawi-Sidhu, T. Kind, T. D. Niehaus, L. J. Broadbelt, A. D. Hanson, O. Fiehn, K. E. J. Tyo, and C. S. Henry. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminformatics*, 7:44, 2015. doi: 10.1186/s13321-015-0087-1.

[77] M. A. Johnson and G. M. Maggiora. *Concepts and applications of molecular similarity.* Wiley, New York, 1990.

[78] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*, 4(11):923–925, 2007. doi: 10.1038/nmeth1113.

[79] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*, 7(1):29–34, 2008. doi: 10.1021/pr700600n.

[80] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Posterior error probabilities and false discovery rates: Two sides of the same coin. *J Proteome Res*, 7(1):40–44, 2008. doi: 10.1021/pr700739d.

[81] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30:42–46, 2002.

[82] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1): D457–D462, 2016.

[83] L. J. Kangas, T. O. Metz, G. Isaac, B. T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R. R. Lewis, and J. H. Miller. In silico identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*, 28(13):1705–1713, 2012. doi: 10.1093/bioinformatics/bts194.

[84] M. Kaplan. DNA has a 521-year half-life. *Nature*, 2012. doi: https://doi.org/10.1038/nature.2012.11555.

[85] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*, 25 (2):197–206, 2007. doi: 10.1038/nbt1284.

[86] E. Kenar, H. Franken, S. Forcisi, K. Wörmann, H.-U. Häring, R. Lehmann, P. Schmitt-Kopplin, A. Zell, and O. Kohlbacher. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol Cell Proteomics*, 13(1):348–359, 2014. doi: 10.1074/mcp.M113.031278.

[87] R. A. Khan. Natural products chemistry: The emerging trends and prospective goals. *Saudi Pharm J*, 26(5):739–753, 2018.

[88] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant. PubChem substance and compound databases. *Nucleic Acids Res*, 44:D1202–D1213, 2016. doi: 10.1093/nar/gkv951.

[89] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8:105, 2007. doi: 10.1186/1471-2105-8-105.

[90] H. Kubinyi. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal Chim Acta*, 247:107–119, 1991.

[91] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, and S. Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*, 84(1):283–289, 2012. doi: 10.1021/ac202450g.

[92] E. Lange, C. Gröpl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, and K. Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, 2007. doi: 10.1093/bioinformatics/btm209. Proc. of *International Conference on Intelligent Systems for Molecular Biology* and *European Conference on Computational Biology* (ISMB/ECCB 2007).

[93] I. Laponogov, N. Sadawi, D. Galea, R. Mirnezami, K. A. Veselkov, and J. Wren. Chemdistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics*, 1:7, 2018.

[94] S. Li, Y. Park, S. Duraisingham, F. H. Strobel, N. Khan, Q. A. Soltow, D. P. Jones, and B. Pulendran. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol*, 9(7):e1003123, 2013. doi: 10.1371/journal.pcbi.1003123.

[95] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for Support Vector Machines. *Mach Learn*, 68(3):267–276, 2007.

[96] J. L. Little, A. J. Williams, A. Pshenichnov, and V. Tkachenko. Identification of "known unknowns" utilizing accurate mass data and ChemSpider. *J Am Soc Mass Spectrom*, 23(1):179–185, 2012. doi: 10.1007/s13361-011-0265-y.

[97] J. Z. Long, K. J. Svensson, L. A. Bateman, H. Lin, T. Kamenecka, I. A. Lokurkar, J. Lou, R. R. Rao, M. R. Chang, M. P. Jedrychowski, J. A. Paulo, S. P. Gygi, P. R. Griffin, D. K. Nomura, and B. M. Spiegelman. The secreted enzyme PM20D1 regulates lipidated amino acid uncouplers of mitochondria. *Cell*, 166(2):424–435, 2016. doi: 10.1016/j.cell.2016.05.071.

[98] M. Loos, C. Gerber, F. Corona, J. Hollender, and H. Singer. Accelerated isotope fine structure calculation using pruned transition trees. *Anal Chem*, 87(11):5738–5744, 2015. doi: 10.1021/acs.analchem.5b00941.

[99] S. Lowry, T. Isenhour, J. Justice, F. McLafferty, H. Dayringer, and R. Venkataraghavan. Comparison of various k-nearest neighbor voting schemes with the self-training interpretive and retrieval system for identifying molecular substructures from mass spectral data. *Analytical Chemistry*, 49(12):1720–1722, 1977.

[100] M. Ludwig, K. Dührkop, and S. Böcker. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics*, 34(13):i333–i340, 2018. doi: 10.1093/bioinformatics/bty245. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2018).

[101] M. Ludwig, L.-F. Nothias, K. Dührkop, I. Koester, M. Fleischauer, M. A. Hoffmann, D. Petras, F. Vargas, M. Morsy, L. Aluwihare, P. C. Dorrestein, and S. Böcker. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nat Mach Intell*, 2(10):629–641, 2020. doi: 10.1038/s42256-020-00234-6.

[102] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *B Math Biophys*, 5(4):115–133, 1943.

[103] F. W. McLafferty. Mass spectrometric analysis. molecular rearrangements. *Analytical chemistry*, 31(1):82–87, 1959.

[104] M. Meringer, S. Reinker, J. Zhang, and A. Muller. MS/MS data improves automated determination of molecular formulas by mass spectrometry. *MATCH Commun Math Comput Chem*, 65:259–290, 2011.

[105] H. Mohimani, A. Gurevich, A. Mikheenko, N. Garg, L.-F. Nothias, A. Ninomiya, K. Takada, P. C. Dorrestein, and P. A. Pevzner. Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol*, 13(1):30, 2017.

[106] H. Mohimani, A. Gurevich, A. Shlemov, A. Mikheenko, A. Korobeynikov, L. Cao, E. Shcherbin, L.-F. Nothias, P. C. Dorrestein, and P. A. Pevzner. Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications*, 9(1):4035, 2018. doi: 10.1038/s41467-018-06082-8.

[107] M. E. Monge, J. N. Dodds, E. S. Baker, A. S. Edison, and F. M. Fernández. Challenges in identifying the dark molecules of life. *Annu Rev Anal Chem (Palo Alto Calif)*, 12(1):177–199, 2019. doi: 10.1146/annurev-anchem-061318-114959.

[108] R. E. Moore, M. K. Young, and T. D. Lee. Qscore: an algorithm for evaluating sequest database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[109] K. Morreel, Y. Saeys, O. Dima, F. Lu, Y. Van de Peer, R. Vanholme, J. Ralph, B. Vanholme, and W. Boerjan. Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. *Plant Cell*, 26(3):929–945, 2014. doi: 10.1105/tpc.113.122242.

[110] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[111] G. Neglur, R. L. Grossman, and B. Liu. Assigning unique keys to chemical compounds for data integration: Some interesting counter examples. In B. Ludäscher and L. Raschid, editors, *Data Integration in the Life Sciences*, pages 145–157, Berlin, Heidelberg, 2005. Springer.

[112] A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, S. Baginsky, and R. Aebersold. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*, 5(4):652–670, 2006. doi: 10.1074/mcp.M500319-MCP200.

[113] D. J. Newman and G. M. Cragg. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod*, 2020. doi: 10.1021/acs.jnatprod.9b01285.

[114] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka. SIMPLE: Sparse interaction model over peaks of molecules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics*, 34(13):i323–i332, 2018. doi: 10.1093/bioinformatics/bty252. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2018).

[115] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka. ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*, 35(14):i164–i172, 2019. doi: 10.1093/bioinformatics/btz319.

[116] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.*, 22(9-10):1006–1026, 2003. doi: 10.1002/qsar.200330831.

[117] L.-F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, A. M. Caraballo-Rodriguez, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, C. Martin H, L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. Tripathi, J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang, and P. C. Dorrestein. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods*, 17(9):905–908, 2020. doi: 10.1038/s41592-020-0933-6.

[118] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *J Mass Spectrom*, 44(4):485–493, 2009. doi: 10.1002/jms.1545.

[119] N. M. O'Boyle. Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi. *J Cheminformatics*, 4(1):22, 2012. doi: 10.1186/1758-2946-4-22.

[120] N. M. O'Boyle and R. A. Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *J Cheminformatics*, 8:36–36, 2016.

[121] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminformatics*, 3:33, 2011. doi: 10.1186/1758-2946-3-33.

[122] D. B. Parker. *Learning logic: casting the cortex of the human brain in silicon.* PhD thesis, Massachusetts Institute of Technology, 1985.

[123] G. J. Patti, O. Yanes, and G. Siuzdak. Metabolomics: The apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012. doi: 10.1038/nrm3314.

[124] H. E. Pence and A. Williams. ChemSpider: An online chemical information resource. *J Chem Educ*, 87(11):1123–1124, 2010. doi: 10.1021/ed100697w.

[125] D. Petras, I. Koester, R. Da Silva, B. M. Stephens, A. F. Haas, C. E. Nelson, L. W. Kelly, L. I. Aluwihare, and P. C. Dorrestein. High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter. *Front Mar Sci*, 4:405, 2017. doi: 10.3389/fmars.2017.00405.

[126] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, chapter 5. MIT Press, Cambridge, Massachusetts, 2000.

[127] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11:395, 2010. doi: 10.1186/1471-2105-11-395.

[128] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, and A. Zhavoronkov. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. Technical report, 2018.

[129] R. A. Quinn, A. V. Melnik, A. Vrbanac, T. Fu, K. A. Patras, M. P. Christy, Z. Bodai, P. Belda-Ferre, A. Tripathi, L. K. Chung, M. Downes, R. D. Welch, M. Quinn, G. Humphrey, M. Panitchpakdi, K. C. Weldon, A. Aksenov, R. da Silva, J. Avila-Pacheco, C. Clish, S. Bae, H. Mallick, E. A. Franzosa, J. Lloyd-Price, R. Bussell, T. Thron, A. T. Nelson, M. Wang, E. Leszczynski, F. Vargas, J. M. Gauglitz, M. J. Meehan, E. Gentry, T. D. Arthur, A. C. Komor, O. Poulsen, B. S. Boland, J. T. Chang, W. J. Sandborn, M. Lim, N. Garg, J. C. Lumeng, R. J. Xavier, B. I. Kazmierczak, R. Jain, M. Egan, K. E. Rhee, D. Ferguson, M. Raffatellu, H. Vlamakis, G. G. Haddad, D. Siegel, C. Huttenhower, S. K. Mazmanian, R. M. Evans, V. Nizet, R. Knight, and P. C. Dorrestein. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*, 579:123–129, 2020. doi: 10.1038/s41586-020-2047-9.

[130] L. M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, et al. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature biotechnology*, 19(1):45–50, 2001.

[131] I. Rauf, F. Rasche, F. Nicolas, and S. Böcker. Finding maximum colorful subtrees in practice. *J Comput Biol*, 20(4):1–11, 2013. doi: 10.1089/cmb.2012.0083.

[132] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, and J. Vervoort. Substructure-based annotation of high-resolution multistage MS^n spectral trees. *Rapid Commun Mass Spectrom*, 26(20):2461–2471, 2012. doi: 10.1002/rcm.6364.

[133] A. L. Rockwood, S. L. Van Orden, and R. D. Smith. Rapid calculation of isotope distributions. *Anal Chem*, 67:2699–2704, 1995. doi: 10.1021/ac00111a031.

[134] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *J Chem Inf Model*, 50 (5):742–754, 2010. doi: 10.1021/ci100050t.

[135] F. B. Rogers. Communications to the editor. *Bull Med Libr Assoc*, 51(1):114–116, 1963.

[136] S. Rogers, R. A. Scheltema, M. Girolami, and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009. doi: 10.1093/bioinformatics/btn642.

[137] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, 65(6):386, 1958.

[138] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods*, 13(9):741–748, 2016. doi: 10.1038/nmeth.3959.

[139] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, 1985.

[140] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminformatics*, 8:3, 2016. doi: 10.1186/s13321-016-0115-9.

[141] K. Scheubert, F. Hufsky, D. Petras, M. Wang, L.-F. Nothias, K. Dührkop, N. Bandeira, P. C. Dorrestein, and S. Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun*, 8:1494, 2017. doi: 10.1038/s41467-017-01318-5.

[142] C. Schiffman, L. Petrick, K. Perttula, Y. Yano, H. Carlsson, T. Whitehead, C. Metayer, J. Hayes, S. Rappaport, and S. Dudoit. Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics*, 20(1):334, 2019. doi: 10.1186/s12859-019-2871-9.

[143] J. Schmidhuber. Deep learning in neural networks: an overview. *Neural networks*, 61:85–117, 2015.

[144] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2001.

[145] A. Schüller, G. Schneider, and E. Byvatov. SMILIB: Rapid assembly of combinatorial libraries in SMILES notation. *QSAR Comb Sci*, 22(7):719–721, 2003. doi: 10.1002/qsar.200310008.

[146] A. Schüller, V. Hähnke, and G. Schneider. SmiLib v2.0: A java-based tool for rapid combinatorial library enumeration. *QSAR Comb Sci*, 26(3):407–410, 2007. doi: 10.1002/qsar.200630101.

[147] E. L. Schymanski and S. Neumann. The critical assessment of small molecule identification (CASMI): Challenges and solutions. *Metabolites*, 3(3):517–538, 2013.

[148] E. L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. R. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, and S. Neumann. Critical Assessment of Small Molecule Identification 2016: Automated methods. *J Cheminformatics*, 9:22, 2017. doi: 10.1186/s13321-017-0207-1.

[149] W. Shao and H. Lam. Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass Spectrom Rev*, 36(5):634–648, 2017. doi: 10.1002/mas.21512.

[150] H. Shen, K. Dührkop, S. Böcker, and J. Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(12):i157–i164, 2014. doi: 10.1093/bioinformatics/btu275. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).

[151] H. Shen, S. Szedmak, C. Brouard, and J. Rousu. *Soft Kernel Target Alignment for Two-Stage Multiple Kernel Learning*, pages 427–441. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-46307-0_27.

[152] X. Shen, R. Wang, X. Xiong, Y. Yin, Y. Cai, Z. Ma, N. Liu, and Z.-J. Zhu. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun*, 10(1):1516, 2019. doi: 10.1038/s41467-019-09550-x.

[153] B. W. Silverman. *Density estimation for statistics and data analysis.* Routledge, 2018.

[154] P. R. Spackman, B. Bohman, A. Karton, and D. Jayatilaka. Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: assessment against experimental reference data and comparison to competitive fragmentation modeling. *Int J Quantum Chem*, 118(2), 2018.

[155] M. Spivak, J. Weston, L. Bottou, L. Käll, and W. S. Noble. Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *J Proteome Res*, 8(7):3737–3745, 2009. doi: 10.1021/pr801109k.

[156] S. E. Stein and D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom*, 5(9):859–866, 1994. doi: 10.1016/1044-0305(94)87009-8.

[157] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J Chem Inf Comput Sci*, 43:493–500, 2003.

[158] M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni. Msnovelist: De novo structure generation from mass spectra. *Nature Methods*, pages 1–6, 2022.

[159] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274: 141–145, 2016.

[160] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, S. Sumner, and S. Subramaniam. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res*, 44(D1):D463–D470, 2015. doi: 10.1093/nar/gkv1042.

[161] D. Szabó, G. Schlosser, K. Vékey, L. Drahos, and Á. Révész. Collision energies on qtof and orbitrap instruments: How to make proteomics measurements comparable? *Journal of Mass Spectrometry*, 56(1):e4693, 2021.

[162] S. Tan, K. C. Sim, and M. Gales. Improving the interpretability of deep neural networks with stimulated learning. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 617–623. IEEE, 2015.

[163] R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti, and G. Siuzdak. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol*, 30(9):826–828, 2012. doi: 10.1038/nbt.2348.

[164] R. Thakare, J. A. Alamoudi, N. Gautam, A. D. Rodrigues, and Y. Alnouti. Species differences in bile acids i. plasma and urine bile acid composition. *J Appl Toxicol*, 38:1323–1335, 2018. doi: 10.1002/jat.3644.

[165] A. Tripathi, A. V. Melnik, J. Xue, O. Poulsen, M. J. Meehan, G. Humphrey, L. Jiang, G. Ackermann, D. McDonald, D. Zhou, R. Knight, P. C. Dorrestein, and G. G. Haddad. Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome. *mSystems*, 3(3):e00020–18, 2018. doi: 10.1128/mSystems.00020-18.

[166] H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, T. Cajka, K. Saito, O. Fiehn, and M. Arita. Hydrogen rearrangement rules: Computational ms/ms fragmentation and structure elucidation using MS-FINDER software. *Anal Chem*, 88(16):7946–7958, 2016. doi: 10.1021/acs.analchem.6b00770.

[167] H. Tsugawa, R. Nakabayashi, T. Mori, Y. Yamada, M. Takahashi, A. Rai, R. Sugiyama, H. Yamamoto, T. Nakaya, M. Yamazaki, R. Kooke, J. A. Bac-Molenaar, N. Oztolan-Erol, J. J. B. Keurentjes, M. Arita, and K. Saito. A

cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat Methods*, 16(4):295–298, 2019. doi: 10.1038/s41592-019-0358-2.

[168] D. Tziotis, N. Hertkorn, and P. Schmitt-Kopplin. Kendrick-analogous network visualisation of ion cyclotron resonance fourier transform mass spectra: Improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur J Mass Spectrom*, 17(4):415–421, 2011. doi: 10.1255/ejms.1135.

[169] J. J. van der Hooft, J. Wandy, F. Young, S. Padmanabhan, K. Gerasimidis, K. E. Burgess, M. P. Barrett, and S. Rogers. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Analytical chemistry*, 89(14):7569–7577, 2017.

[170] J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci USA*, 113(48):13738–13743, 2016. doi: 10.1073/pnas.1608041113.

[171] V. Vapnik. Pattern recognition using generalized portrait method. *Autom Remote Control*, 24:774–780, 1963.

[172] V. Vapnik. A note one class of perceptrons. *Autom Remote Control*, 1964.

[173] D. Verdegem, D. Lambrechts, P. Carmeliet, and B. Ghesquiére. Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. *Metabolomics*, 12(6):1–16, 2016. doi: 10.1007/s11306-016-1036-3.

[174] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, and O. Yanes. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *Trends Anal Chem*, 78:23–35, 2016. doi: 10.1016/j.trac.2015.09.005.

[175] R. C. H. D. Vos, S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino, and R. D. Hall. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Protocols*, 2(4):778–791, 2007. doi: 10.1038/nprot.2007.95.

[176] M. Wang, G. Audi, A. Wapstra, F. Kondev, M. MacCormick, X. Xu, and B. Pfeiffer. The AME2012 atomic mass evaluation (II). tables, graphs and references. *Chinese Physics C*, 36(12):1603–2014, 2012.

[177] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute,

E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, and N. Bandeira. Sharing and community curation of mass spectrometry data with Global Natural Products Social molecular networking. *Nat Biotechnol*, 34(8):828–837, 2016. doi: 10.1038/nbt.3597.

[178] M. Wang, A. K. Jarmusch, F. Vargas, A. A. Aksenov, J. M. Gauglitz, K. Weldon, D. Petras, R. da Silva, R. Quinn, A. V. Melnik, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, L. F. Nothias, C. M. Aceves, M. Panitchpakdi, E. Brown, F. Di Ottavio, N. Sikora, E. O. Elijah, L. Labarta-Bajo, E. C. Gentry, S. Shalapour, K. E. Kyle, S. P. Puckett, J. D. Watrous, C. S. Carpenter, A. Bouslimani, M. Ernst, A. D. Swafford, E. I. Zúñiga, M. J. Balunas, J. L. Klassen, R. Loomba, R. Knight, N. Bandeira, and P. C. Dorrestein. Mass spectrometry searches using MASST. *Nat Biotechnol*, 38(1):23–26, 2020. doi: 10.1038/s41587-019-0375-9.

[179] S. Wang, T. Kind, D. J. Tantillo, and O. Fiehn. Predicting in silico electron ionization mass spectra using quantum chemistry. *Journal of cheminformatics*, 12(1):1–11, 2020.

[180] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37(Web Server issue):W623–W633, 2009. doi: 10.1093/nar/gkp456.

[181] Y. Wang, G. Kora, B. P. Bowen, and C. Pan. MIDAS: A database-searching algorithm for metabolite identification in metabolomics. *Anal Chem*, 86(19):9496–9503, 2014. doi: 10.1021/ac5014783.

[182] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A*, 109(26):E1743–E1752, 2012. doi: 10.1073/pnas.1203689109.

[183] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

[184] P. J. Werbos. *Beyond regression: new tools for prediction and analysis in the behavioral sciences.* PhD thesis, Harvard University, 1974.

[185] C. L. Wilkins and M. Randić. A graph theoretical approach to structure-property and structure-activity correlations. *Theor Chim Acta*, 58(1):45–68, 1980. doi: 10.1007/BF00635723.

[186] P. Willett. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*, 11(23-24):1046–1053, 2006. doi: 10.1016/j.drudis.2006.10.005.

[187] P. Willett. Similarity searching using 2D structural fingerprints. *Methods Mol Biol*, 672:133–158, 2011. doi: 10.1007/978-1-60761-839-3_5.

[188] P. Willett and V. Winterman. A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. *Quant Struct-Act Relat*, 5(1):18–25, 1986. doi: 10.1002/qsar.19860050105.

[189] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminformatics*, 9 (1):33, 2017. doi: 10.1186/s13321-017-0220-4.

[190] W. Windig, J. M. Phalp, and A. W. Payne. A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Anal Chem*, 68(20):3602–3606, 1996. doi: 10.1021/ac960435y.

[191] D. S. Wishart. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature reviews. Drug discovery*, 15(7):473–484, 2016. doi: 10.1038/nrd.2016.32.

[192] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*, 46(D1):D608–D617, 2018. doi: 10.1093/nar/gkx1089.

[193] M. Witting and S. Böcker. Current status of retention time prediction in metabolite identification. *J Sep Sci*, 43(9–10):1746–1754, 2020. doi: 10.1002/jssc.202000060.

[194] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010. doi: 10.1186/1471-2105-11-148.

[195] C. Wu, M. J. Gales, A. Ragni, P. Karanasou, and K. C. Sim. Improving interpretability and regularization in deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):256–265, 2017.

[196] X. Yang, P. Neta, and S. E. Stein. Quality control for building libraries from electrospray ionization tandem mass spectra. *Anal Chem*, 86(13):6393–6400, 2014. doi: 10.1021/ac500711m.

A Appendix

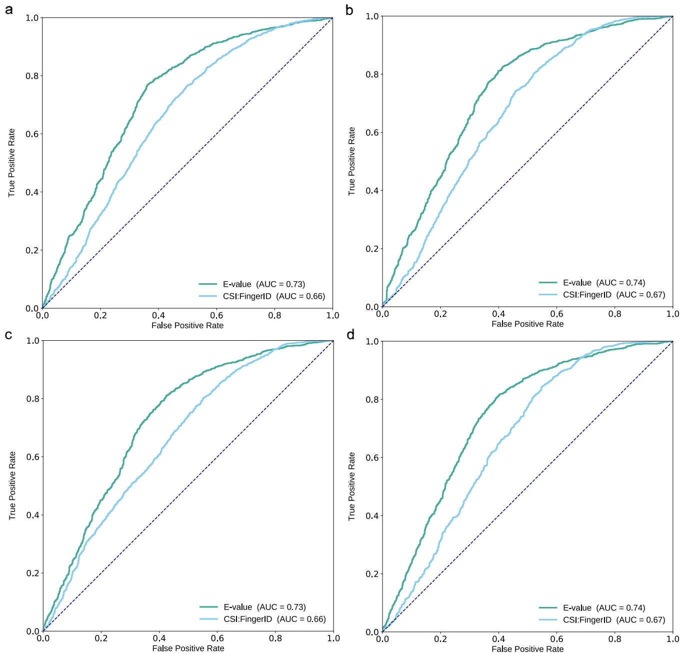


Figure A.1: (a–d) Comparison of CSI:FingerID score and E-value score. ROC curves, structure-disjoint evaluation, independent data and no artificial noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra (“all collision energies”)

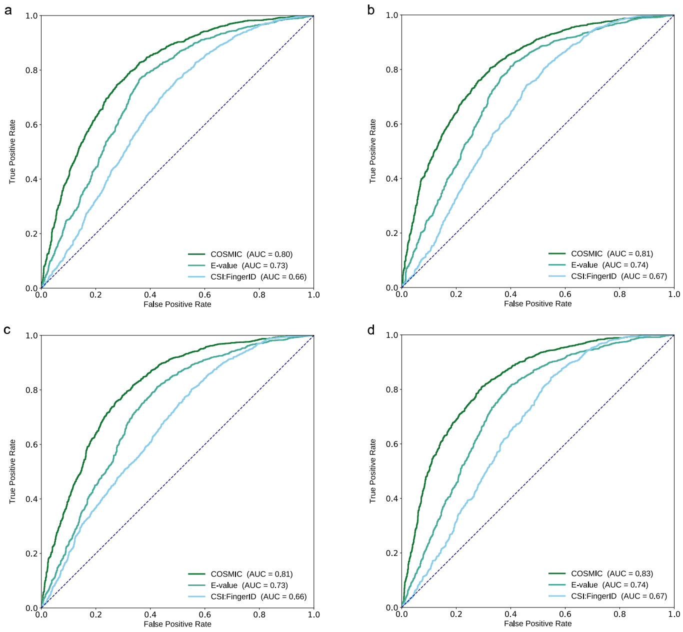


Figure A.3: (a–d) Comparison of CSI:FingerID score, E-value score and SVM-based COSMIC confidence score. ROC curves, structure-disjoint evaluation, independent data and no artificial noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra ("all collision energies")

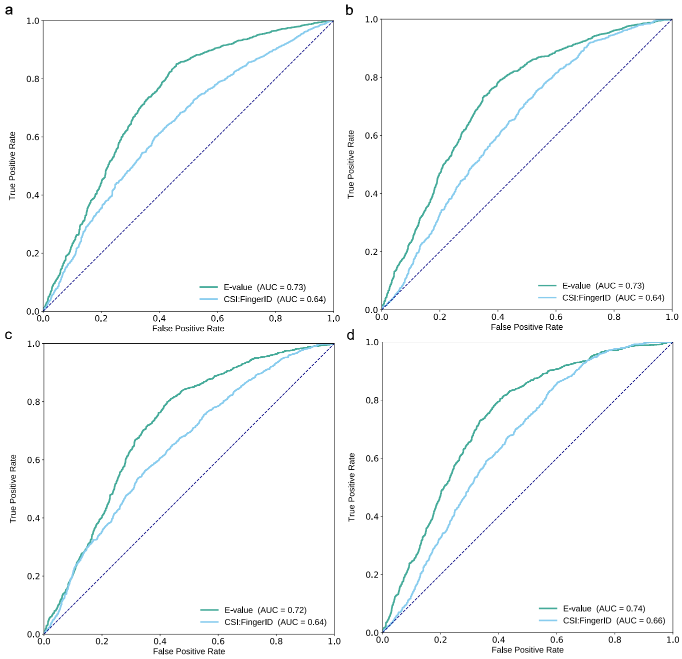


Figure A.2: (a–d) Comparison of CSI:FingerID score and E-value score. ROC curves, structure-disjoint evaluation, independent data and high noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra ("all collision energies")

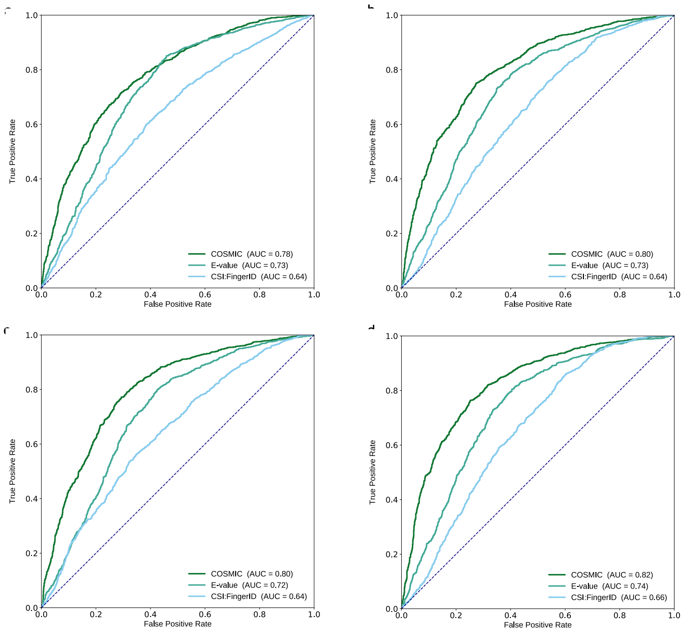


Figure A.4: (a–d) Comparison of CSI:FingerID score, E-value score and SVM-based COSMIC confidence score. ROC curves, structure-disjoint evaluation, independent data and high noise, biomolecule structure database, $N = 3\,013$. (a) 10 eV, (b) 20 eV, (c) 40 eV, (d) merged spectra ("all collision energies")

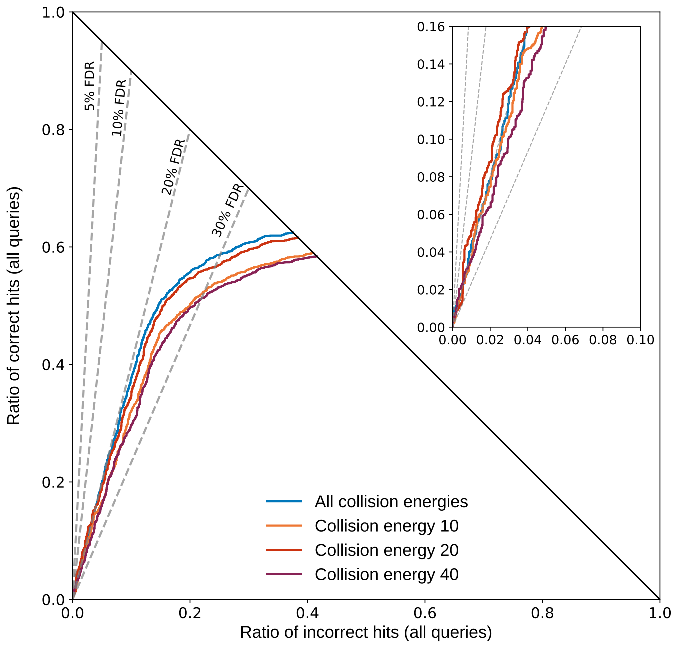


Figure A.5: Hop plots of E-value scores for different collision energies. Independent data with structure-disjoint evaluation, no artificial noise, queries are QTOF MS/MS data, $N = 3,013$. FDR levels shown as dashed lines; FDR levels are exact, not estimated

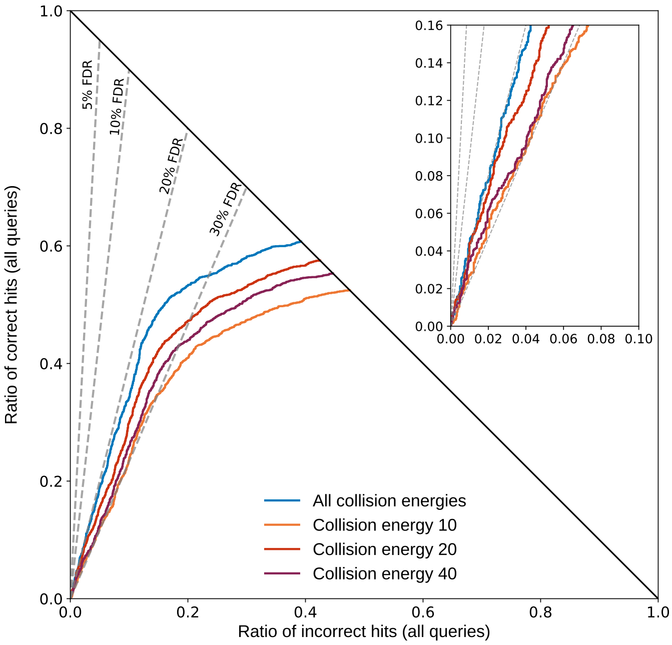


Figure A.6: Hop plots of E-value scores for different collision energies. Independent data with structure-disjoint evaluation, medium noise, queries are QTOF MS/MS data, $N = 3,013$. FDR levels shown as dashed lines; FDR levels are exact, not estimated

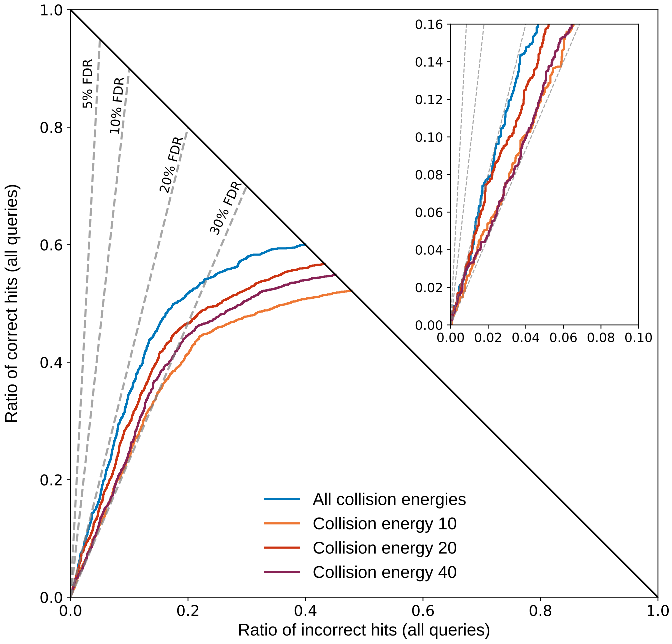


Figure A.7: Hop plots of E-value scores for different collision energies. Independent data with structure-disjoint evaluation, high noise, queries are QTOF MS/MS data, $N = 3,013$. FDR levels shown as dashed lines; FDR levels are exact, not estimated

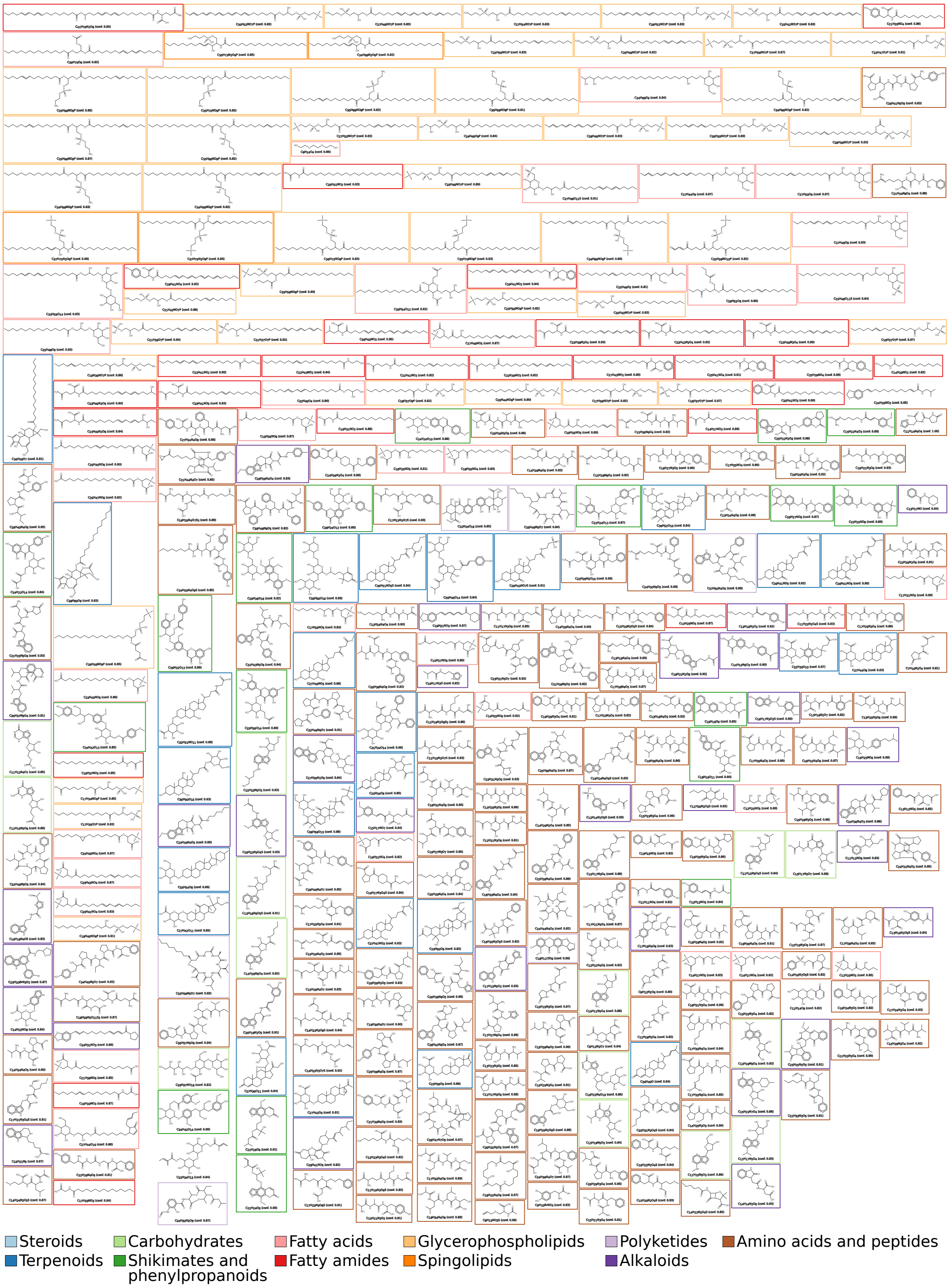


Figure A.8: The 1,715 novel molecular structures annotated with high confidence in the Orbitrap dataset. Confidence score threshold 0.64 was used. Structures are shown with identification number (ID), molecular formula and COSMIC confidence score. Colours indicate compound classes. Lipid structures must again be interpreted with some care.

Figure A.9: The 1,715 novel molecular structures annotated with high confidence in the Orbitrap dataset (cont.).

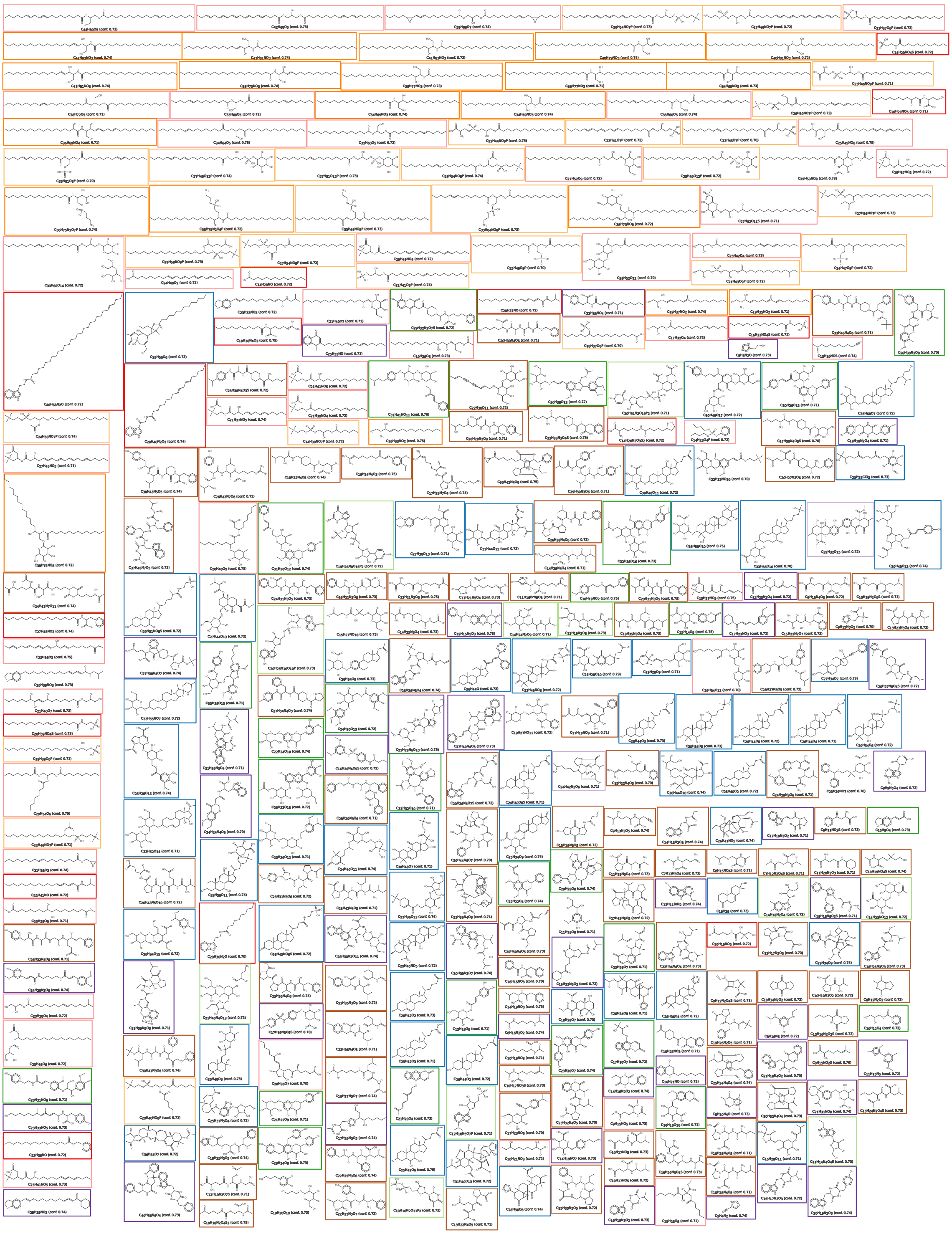


Figure A.10: The 1,715 novel molecular structures annotated with high confidence in the Orbitrap dataset (cont.).

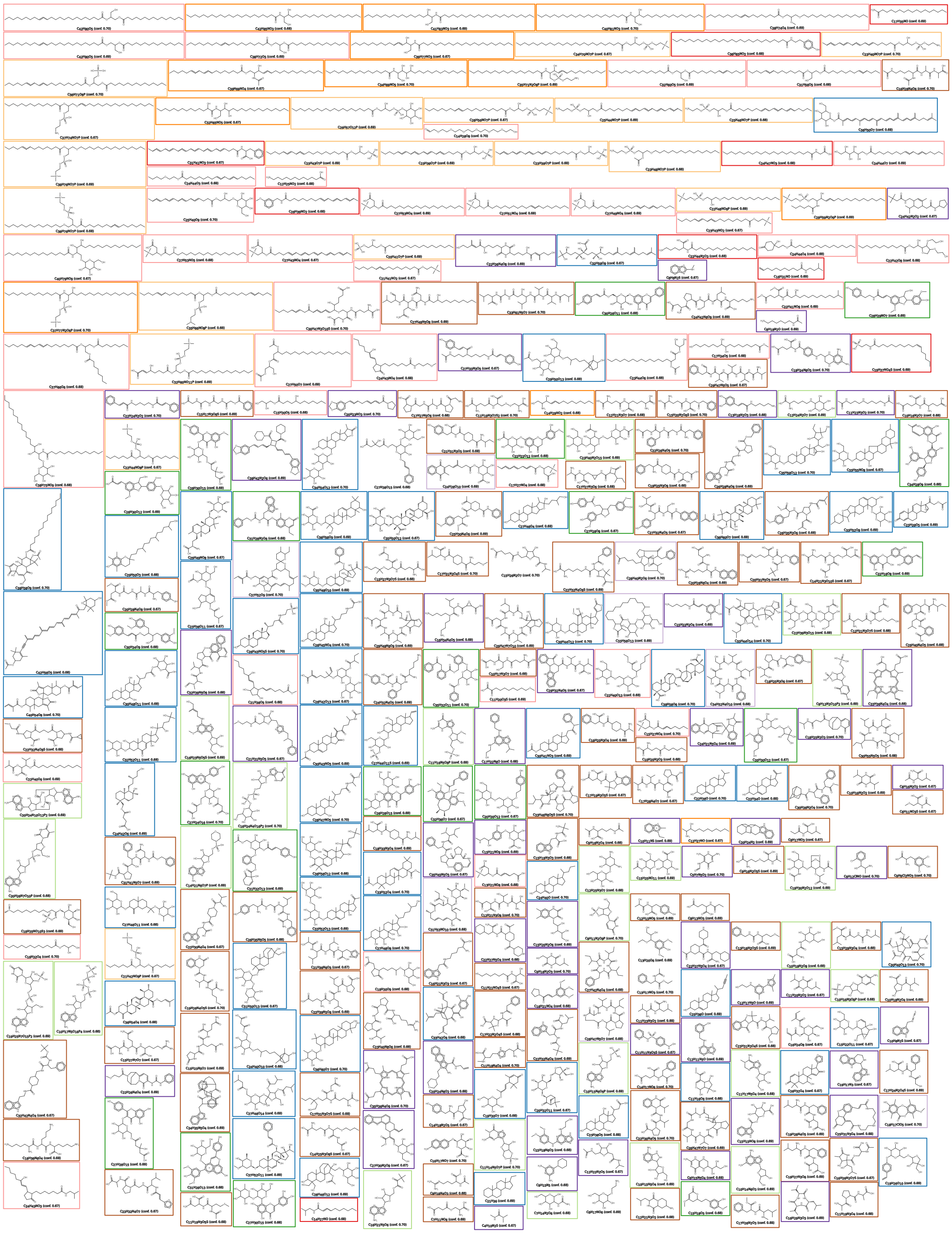


Figure A.11: The 1,715 novel molecular structures annotated with high confidence in the Orbitrap dataset (cont.).

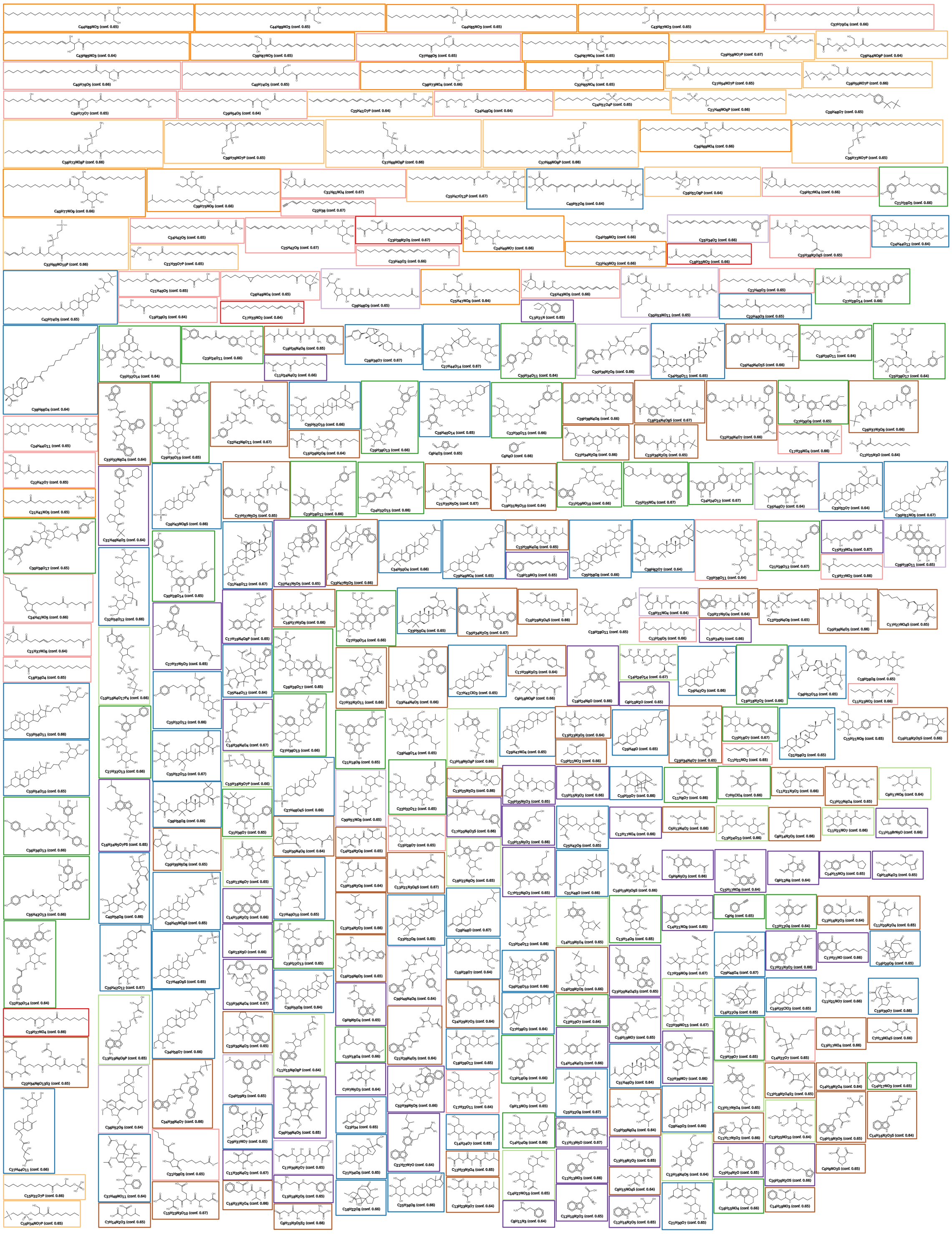


Figure A.12: The 1,715 novel molecular structures annotated with high confidence in the Orbitrap dataset (cont.).

Table A.1: MassIVE accession numbers for the Orbitrap dataset. Corresponding mzML/mzXML files are available from MassIVE (`https://massive.ucsd.edu/`).

MSV000084873, MSV000084753, MSV000084744, MSV000084741, MSV000084738, MSV000084674, MSV000084630, MSV000084628, MSV000084585, MSV000084576, MSV000084556, MSV000084312, MSV000084289, MSV000084278, MSV000084237, MSV000084143, MSV000084132, MSV000084119, MSV000084118, MSV000084117, MSV000084112, MSV000084107, MSV000084102, MSV000084072, MSV000084062, MSV000084045, MSV000084030, MSV000084020, MSV000084016, MSV000083889, MSV000083888, MSV000083791, MSV000083773, MSV000083749, MSV000083705, MSV000083660, MSV000083651, MSV000083647, MSV000083632, MSV000083631, MSV000083612, MSV000083541, MSV000083523, MSV000083522, MSV000083521, MSV000083483, MSV000083481, MSV000083482, MSV000083475, MSV000083472, MSV000083471, MSV000083470, MSV000083469, MSV000083411, MSV000083396, MSV000083395, MSV000083387, MSV000083383, MSV000083372, MSV000083365, MSV000083306, MSV000083300, MSV000083275, MSV000083274, MSV000083272, MSV000083134, MSV000083110, MSV000083106, MSV000083098, MSV000083094, MSV000083083, MSV000083077, MSV000083073, MSV000082999, MSV000082952, MSV000082869, MSV000082650, MSV000082649, MSV000082647, MSV000082633, MSV000082618, MSV000082616, MSV000082614, MSV000082612, MSV000082608, MSV000082602, MSV000082582, MSV000082480, MSV000082463, MSV000082433, MSV000082402, MSV000082385, MSV000082384, MSV000082383, MSV000082382, MSV000082380, MSV000082379, MSV000082378, MSV000082377, MSV000082331, MSV000082157, MSV000082086, MSV000082085, MSV000082084, MSV000082083, MSV000082082, MSV000082081, MSV000080249, MSV000082048, MSV000081957, MSV000081952, MSV000081949, MSV000081808, MSV000081804, MSV000081492, MSV000081482, MSV000081456, MSV000081097, MSV000080905, MSV000080630, MSV000080628, MSV000079900, MSV000081160

Ehrenwörtliche Erklärung

Hiermit erkläre ich

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angfertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt:
Sebastian Böcker, Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Fleming Kretschmer, Louis-Félix Nothias und Michael Witting

Ich habe weder die gleiche, noch eine in wesentlichen Teilen ähnliche bzw. eine andere Abhandlung bereits bei einer anderen Hochschule als Dissertation eingereicht.

Jena, den 14.06.2022

Martin André Hoffmann