OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

Finding Potential Synthetic Cannabinoids in Forensic Drug Analysis Data

Master Thesis

Ostbayerische Technische Hochschule Regensburg
Fakultät Informatik & Mathematik
Master of Science Informatik

Submitted by: Sebastian Strasser
Matriculation number: 3301047
Semester: MIN4

Supervisor: Prof. Dr. Johannes Schildgen
Date of submission: March 31, 2023

Abstract

Synthetic cannabinoids represent the largest class of designer drugs. They pose a challenge on forensic laboratories as new substances emerge quickly. Classical approaches in forensic drug analysis rely on using mass spectrometry to gather information on the substance at hand and query databases of known illicit drugs with this information. For novel substances, this is not possible because there are no entries linked to them. Thus, forensic laboratories are in need of a way to check if mass spectrometry data potentially contains measurements of synthetic cannabinoids.

This thesis tackles the challenge by developing a method to trace potential synthetic cannabinoids in mass spectrometry data. The approach bases on matching mass spectra with a candidate database consisting of generated virtual molecules which fit the description of synthetic cannabinoids and their fragments which were predicted based on a fragmentation model built on known synthetic cannabinoids. To test and make the method accessible for users, an analysis tool in the form of a web application was implemented.

The implemented method shows good results. In an evaluation where a database consisting only of generated compounds and fragments was used, 86.8% of synthetic cannabinoids were identified correctly in the matching process. Only 8 out of 53,629 negative spectra were wrongfully classified as positive, resulting in a specificity of 99.985%.

Zusammenfassung

Synthetische Cannabinoide stellen die größte Gruppe von Designerdrogen dar. Sie stellen forensische Labore vor Herausforderungen, da laufend neue Substanzen entstehen. Forensische Drogenanalysen verwenden Massenspektrometrie, um Informationen zu den vorliegenden Substanzen zu ermitteln. Mithilfe dieser Informationen werden Datenbanken abgefragt, die bekannte illegale Drogen enthalten. Dieser Ansatz liefert bei neu auftretenden Substanzen keine Treffer, da keine zugehörigen Einträge in den Datenbanken vorhanden sind. Deshalb brauchen forensische Labore Lösungen, um herauszufinden, ob Massenspektrometriedaten möglicherweise synthetische Cannabinoide enthalten.

Diese Arbeit entwickelt eine Methode, die mögliche synthetische Cannabinoide in Massenspektrometriedaten feststellt. Dafür wird zunächst eine Kandidaten-Datenbank erstellt, die aus generierten virtuellen Molekülen besteht. Diese Moleküle entsprechen der Beschreibung von synthetischen Cannabinoiden. Außerdem werden Fragmente zu den Molekülen generiert. Diese beruhen auf einem Modell, das anhand von bekannten synthetischen Cannabinoiden erstellt wurde. Um die Analyse für Benutzer zugänglich zu machen, wurde ein Web-Tool entwickelt.

Die Methode zeigt gute Ergebnisse. Für die Evaluierung wurden nur generierte Moleküle und Fragmente verwendet. 86,6% der synthetischen Cannabinoide wurden als solche erkannt. Nur 8 der 53.629 Negativ-Spektren wurden fälschlicherweise als synthetische Cannabinoide klassifiziert, was einer Spezifizität von 99,985% entspricht.

Contents

List of Figures

List of Abbreviations

CSV Comma-separated values

Da dalton

EMCDDA European Monitoring Centre for Drugs and Drug Addiction

eV electronvolt

HTML Hypertext Markup Language

m/z mass-to-charge ratio

MGF Mascot Generic Format

MI monoisotopic mass

ppm parts per million

SMILES simplified molecular-input line-entry system

THC Tetrahydrocannabinol

u unified atomic mass unit

List of Listings

1 Introduction

Synthetic cannabinoids are substances which are supposed to mimic the effects of Tetrahydrocannabinol (THC), the psychoactive component of cannabis [Auw+21, p.2]. Initially, the development of synthetic cannabinoids focused on its usage as a medicine. However, in the mid-2000s, synthetic cannabinoids emerged as an ingredient of the product "Spice". This was thought of as a legal alternative to natural cannabis. Since then, a number of substances have been developed and brought to market by clandestine laboratories [Auw+21, p.6].

There is a number of problems in association with synthetic cannabinoids. One is the unpredictability of the effects on the user as there are no scientific tests on short- or long-term side effects [Auw+21, p.21]. Another problem is the high potency of synthetic cannabinoids due to them being designed to fully bind to cannabinoid-receptors in the body [Auw+21, p.21]. This means that synthetic cannabinoids are often much stronger than natural cannabis. This presents itself in stronger and more frequent unwanted effects after consumption compared to cannabis. These effects include hallucinations, panic attacks, or psychosis [Auw+21, p.31]. There are not only mental effects, also physical problems can arise from the consumption of synthetic cannabinoids. Heart problems and kidney damages were linked to the consumption of certain synthetic cannabinoids [Auw+21, p.25–26]. Even death cases are reported where the consumption of synthetic cannabinoids were assumed to be the main cause [Auw+21, p.26–27].

As an answer to the increasing emergence of synthetic cannabinoids, countries have put efforts into controlling them [Auw+21, p.9]. This led to the black market producers creating new substances which are structurally slightly different to existing ones. Obviously, countries and other institutions have reacted to this development. For example, the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) implemented the so-called *EU Early Warning System* which not only monitors synthetic cannabinoids,

but all so-called *new psychoactive substances* [DA21]. Its goal is to rapidly detect and respond to substances newly discovered in the European Union.

However, forensic laboratories still face the challenge of detecting synthetic cannabinoids in drug-analysis data, as there are potentially novel substances in the datasets. These go unnoticed in the analysis due to them not being existent in substance databases. Hence, they are in need of solutions which can indicate whether analysis data potentially contains synthetic cannabinoids. The thesis at hand addresses this problem by implementing a method which searches for potential synthetic cannabinoids in drug-analysis data, to be more precise, mass-spectrometry data. The approach is based on the generation of virtual molecules which fit to the description of synthetic cannabinoids and their corresponding fragments arising in the mass-spectrometry process. This generation process uses knowledge from existing synthetic cannabinoids and builds rule-sets on these observations. An analysis tool in the form of a web application then faciliates the matching of mass-spectrometry data with the generated molecules and fragments.

This thesis enables forensic laboratories to spot potential synthetic cannabinoids, including novel ones, in large mass-spectrometry datasets in a fast way. Thus, new synthetic cannabinoids do no not go unnoticed in drug screenings. The developed approach can be seen as an addition to measures currently in place to spot novel substances.

In the following, the structure of this thesis is elucidated. Chapter 2 explains the underlying problem more detailed. Afterwards, important terms and concepts needed for the understanding of the method are introduced in Chapter 3. This chapter also looks at literature which discussed approaches similar to this thesis. Then, the implemented method is presented in Chapter 4 and evaluated in Chapter 5. Chapter 6 gives a conclusion and looks at future applications.

2 Problem description

Forensic drug analysis works, roughly speaking, in the following way: A sample of blood, urine, or hair sample from a patient is provided. This sample is then analyzed in a labratory. A common technique for the analysis is liquid chromatography coupled with mass spectrometry [KA12]. In this process, containing molecules are separated according to their chemical properties. For each molecule, a mass spectrometer generates a so-called mass spectrum. The remarkable thing about a mass spectrum is that it gives sufficient information to conclude about the chemical structure of the substance at hand. This process is also reproducible which means that the mass spectrum is a strong indicator for a compound. Hence, there exist databases containing mass spectra of many known molecules. Examples are the NIST mass spectral library and the Wiley Registry [SHB13, p.3] An important stage in the drug screening is then finding the analyte molecule in the spectral database by searching it with the acquired mass spectrum.

A limitation of this approach is the requirement that the substances to be found needs to have an corresponding entry in the spectral database. This is especially a challenge when dealing with designer drugs like the substance class this thesis focuses on, synthetic cannabinoids. Clandestine laboratories develop and bring them to the market rapidly and forensic laboratories always have to be up to date with latest developments. This is a very challenging task. Substances new to the market are not in chemical databases and thus not traceable. As there is also indication of a tendency to new and thus not legally controlled products consumed more often than established ones [Fra+18, p.66], this challenge becomes even harder. Hence, forensic labratories are in need of means to filter mass spectrometry data for spectra potentially containing synthetic cannabinoids.

This thesis tackles this challenge and aims to enable forensic laboratories to spot novel synthetic cannabinoids in mass-spectrometry data.

3 Foundations

This chapter introduces various terms and concepts that are used throughout the thesis. First, a short introduction of the representation of molecules in graphical and in machine-readable form is given. The next section describes the class of molecules this thesis focuses on: synthetic cannabinoids. Then two important chemical techniques for finding information on a substance are introduced, namely chromatography and mass spectrometry. After a general explanation of the functionality, there is an emphasis on drug testing with the means of mass spectrometry and finding novel substances with a mass spectrometer. The last section presents literature that aims to examine undiscovered substances with data extracted from mass spectrometry, a task similar to the one this thesis emphasizes on.

3.1 Molecules and Their Representation in Cheminformatics

Molecules consist of *atoms* and chemical *bonds*. They arise from atoms building bonds between one another by sharing electron pairs. When a molecule has atoms of more than one element, it is also referred to as a *compound*. All organic molecules consist of only a small number of elements. These are carbon (C), hydrogen (H), oxygen (O), nitrogen (N), sulfur (S), phosphorus (P), silicon (Si), fluorine (F), chlorine (Cl), bromine (Br), and iodine (I). Carbon plays an important role here, as carbon is capable of building carbon-carbon bonds which makes huge carbon skeletons possible [SGI16, Section 2.1]. The first property of a molecule that comes into mind is the *elemental composition* which names the elements it contains. One step further comes the *molecular formula*, indicating how many atoms of each element are present in the molecule [SGI16, Section 2.1]. A simple

example for this is water. A water molecule consists of two hydrogen atoms and one oxygen atom. Thus, its elemental composition is described by the elements hydrogen and oxygen and its molecular formula is H_2O.

The molecular formula gives the number of atoms of the different elements that are present in the molecule, but holds no information on how these atoms are connected with each other. Depending on the counts of the elements, a lot of different connectivities are possible for the same molecular formula. Compounds with the same molecular formula, but different arrangements of atoms are called *isomers*. To differentiate between these isomers, you have to examine their *structure*, i.e. the description which atoms are connected by chemical bonds. Knowing the structure is important, as isomers may have very different chemical properties [SGI16, Section 2.1]. For the sake of completeness, there are also molecules having the same structure, but different spatial arrangements. These are called *stereoisomers* [SGI16, Section 2.1].

Several conventions in the symbolizing of molecule structures exist [SGI16, Section 2.1]. A very common one is displayed in Figure 1. It shows the molecular structure of aspirin. As carbon and hydrogen are the most abundant elements in organic compounds, only the skeletons of carbon bonds are drawn. Thus, at the end of each bond where no element is displayed, a carbon atom is assumed. Carbon-hydrogen connections are also omitted, as it can be assumed that each carbon atom has full bonding (i.e. four bonds) and there are as many hydrogen atoms as required for that. Atoms of elements other than carbon and hydrogen are always displayed.

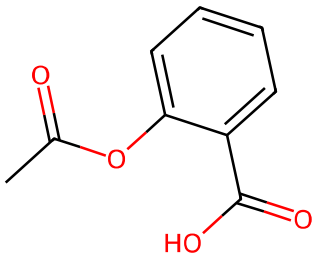

Figure 1: Molecular structure of aspirin (based on [Bio23b])

Obviously, the representation shown in Figure 1 is not interpretable for computers in this form. Thus, machine-readable representations of molecular structures are required.

Various models representing molecules are mentioned in literature. Wigh, Goodman, and Lapkin [WGL22, pp.2–3] differentiate between feature-based, computer learned, chemical table, and string respresentations. The first two do not represent the complete molecule structure, but rather exhibit its features. Chemical tables arrange the contained atoms in x-,y-, and z-coordinates and depict the chemical bonds in a connection table. String-based representations include registry and structure-based systems. Registry systems store relevant information of a molecule and give a unique ID to each compound in their databases which eases communication. Examples for registry systems are the CAS registry [Ser23] and PubChem [Bio23a].

A very prominent structure-based string respresentation is *simplified molecular-input line-entry system (SMILES)* [Sys19b]. Although it is a relatively old standard, SMILES is still widely used due to its simplicity and human readability [WGL22, p.4–5]. For instance, the SMILES string for aspirin is CC(=O)OC1=CC=CC=C1C(=O)O (the structure is depicted in Figure 1). Atoms in SMILES strings are simply represented by their atomic symbols. As in the notation in Figure 1, hydrogen is usually suppressed. Single bonds between atoms are not symbolized in most cases, symbols being adjacent imply that they have a single bond. However, double bonds are explicitly shown with the symbol "=" and triple bonds are signaled with "#". Rings are expressed by breaking one bond in the ring and appending an identifiying integer to the opening and closing atoms. In the example, the substring C1=CC=CC=C1 expresses the ring that is shown in the right upper corner in Figure 1. Parentheses in the string indicate a branch point. Looking at the example, after the first carbon-carbon bond, there are two branches: one is the double bond to an oxygen atom and the other one is a single bond to another oxygen atom. This is expressed by putting the =O into parentheses.

A challenge when working with SMILES is its noncanonical nature which means that there are many ways to build a SMILES string for a given molecule [LG07, p.6]. This is especially a problem when searching for exact structures or substructures. A solution is constructing a canonical SMILES string that is unique and thus comparable. This can be achieved in the following way: the atoms of molecule graph are labelled canonically. When two molecules are identical, this numbering is the same for both. In the next step, the graph is traversed with a traversal algorithm like depth-first-search [OBo12, p.3]. This results in the canonical SMILES strings. An example for a canonicalization is proposed by Schneider, Sayle, and Landrum [SSL15]. This implementation is used in the open-source cheminformatics application RDKit [RDK22c]. It is based on a stable sort-

ing algorithm in conjunction with invariants which consider stereochemistry [SSL15].

3.2 Synthetic Cannabinoids

Synthetic cannabinoids mimic the effects of the major psychoactive component of cannabis, THC [Auw+21, p.2]. They represent the largest class of *new psychoactive substances* which are substances having effects similar to illicit drugs, but are partially not controlled by international drug conventions. An often used synonym for these substances is "legal highs" [DA22, p.38]. The first occurrences of synthetic cannabinoids as legal highs are documented in 2004. They were sold as a product called "Spice" and were first identified in 2008 [Auw+21, p.6].

The rising popularity of synthetic cannabinoids due to their legal image and the undectability in the first years led to legal control in many countries [Fra+18, p.62]. Producers of synthetic cannabinoids reacted by creating new substances that are structurally slightly different compared to already known ones. A lot of novel substances emerged that way. The European Drug Report of 2022 [DA22, p.38] records 224 new synthetic cannabinoids since 2008. The rapid development of new synthetic cannabinoids poses a challenge for forensic laboratories, as they always have to be up-to-date which substances are currently in circulation and new to the market [Fra+18, p.62].

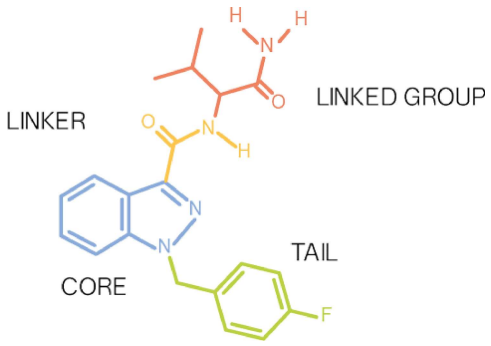

Figure 2: The four building blocks of synthetic cannabinoids shown on the compound AB-FUBINACA [DA17]

While the name "synthetic cannabinoids" suggests that the molecules are closely related to cannabinoids occurrring in natural cannabis, many compounds of different substance

classes mimic the effects of THC and thus classify as synthetic cannabinoids [Pul+22, p.1]. Nonetheless, many of the synthetic cannabinoids recorded by the EMCDDA (91% according to [Pul+22, p.3]) follow a certain pattern. Figure 2 illustrates the basic structure of these molecules; consisting of four building blocks: core, linker, linked group, and tail. This observation leads to the assumption that many synthetic cannabinoids were created with methods from combinatorial chemistry [Pul+22, p.1]. This is an approach commonly used in drug discovery. The idea is to create chemical libraries of molecules by systematically simulating the bonding between the building blocks. After an assessment of biological properties, this can lead to a library of potential synthetic cannabinoids [Pul+22, p.1].

3.3 Chromatography

Chromatography enables the separation of molecules in a mixture [Coş16, p.156]. It is based on two major components: a stationary phase which is solid or a liquid smeared on a solid, and a mobile phase which is liquid or gaseous. The mixture is applied to the stationary phase and moved with the aid of the mobile phase which is flowing over the stationary phase. Due to the molecules in the mixture having different chemical properties (e.g. molecular weight, absorption, and affinity), the time they stay in the stationary phase varies [Coş16, p.156]. The results of this process are the molecules in separated form. There are two types of chromatography, the classification depends on the mobile phase. Liquid chromatography indicates that the mobile phase is liquid while gas chromatography means that it is gaseous. The choice which one to use depends mainly on the mixture. Gas chromatography is mainly used for volatile liquids or solid materials while liquid chromatography is utilized when trying to separate non-volatile and thermal-unstable samples [Coş16, p.156]. A result additional to the molecule is the retention time thereof which shows how long it stayed in the chromatograph. It can be an indicator for a specific compound. However, there can be significant variation in retention time for compounds due to a number of reasons, e.g. minimally varying configurations of the chromatograph [Dol14].

3.4 Mass Spectrometry

Mass spectrometry is a technique with a wide range of applications in many scientific fields, like material sciene, astronomy, biology, and medical research [Was15]. It is used to analyze different characteristics of molecule, like their weight or formula. Another possible application is the determination of unknown compounds by using their properties extracted from the mass spectrometry analysis and comparing them with compound databases [Aha+22]. In this thesis, mass spectrometry is used to determine the mass of analyte compounds and matching it with compounds in a generated database which contains potential synthetic cannabinoids. While mass spectrometry gives qualified information for pure substances, data of mixtures is more difficult or even impossible to interpret. Thus, when dealing with mixtures, chromatographs can be used as an inlet system. Similar compounds are isolated and analyzed individually [WS07, p.43].

The following section introduces basic concepts in the context of mass spectrometry, the components of a mass spectrometer, and the data mass spectrometers produce. To illustrate the concepts, following example is used: a patient who has consumed cannabis has to provide a hair sample. This sample is given into a solvent in order to extract the substance. The resulting substance is then analyzed with a chromatograph coupled with a mass spectrometer to potentially find traces of illicit drugs. As mentioned in Section 3.2, the major psychoactive component in cannabis is THC (the molecule structure is depicted in Figure 3).

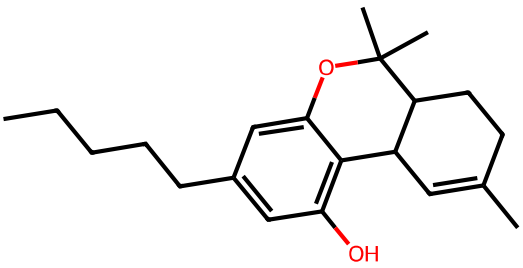

Figure 3: Structure of Tetrahydrocannabinol (based on [Bio23c])

3.4.1 Basic Concepts

As the name suggests, mass spectrometry concerns itself with masses, more precisely with the masses of *isotopes*. Isotopes are atoms of the same element (i.e. have the same number of protons and electrons) with different numbers of neutrons. They have identical chemical properties (i.e. the ability of a substance to react to form new substances), but differ in mass, as neutrons add only mass and no charge [Dow04, p.4].

When measuring the exact masses of elements, you look at the *monoisotopic mass (MI)*. That is a term for the mass of the most abundant stable isotope of an element. Consequently, the MI of a molecule is the sum of the MIs of all its elements [WS07, p.273]. The unit of measurement when determining exact masses of atomic particles is the unified atomic mass unit (u). One u is defined as one twelfth of the mass of the most abundant stable carbon isotope ^{12}C [LXB96, p.996]. A commonly used synonym is dalton (Da) [WS07, p.273].

A mass spectrometer is unable to measure the masses of molecules directly. They can, however, dectect *ions*. While neutral molecules are unresponsive, charged analytes can be controlled with electronic and magnetic fields [Dow04, p.10]. That means the analyte molecule has to be charged (positively or negatively) in order to be recognized by the mass spectrometer. You then have two factors: the mass and the charge. Thus, the mass-to-charge ratio (m/z) of the ion is measured. The z stands for an absolute multiple of the charge of an electron, so if an ion possesses two charges, z would be equal to 2 [Dow04, p.10].

3.4.2 The Mass Spectrometer

Figure 4 shows the basic components of a mass spectrometer: The *ion source*, the *m/z analyzer* (also commonly called *mass analyzer*), and the *ion detector*.

The ion source introduces the analyte molecule into the mass spectrometer and converts it into a ionized form [Dow04, p.22]. There are many ionization techniques available. What all of them have in common is the output, which are gas-phase ions of the analyte [NF15, p.6]. They differ in the physical state of the analyte, the internal energy that

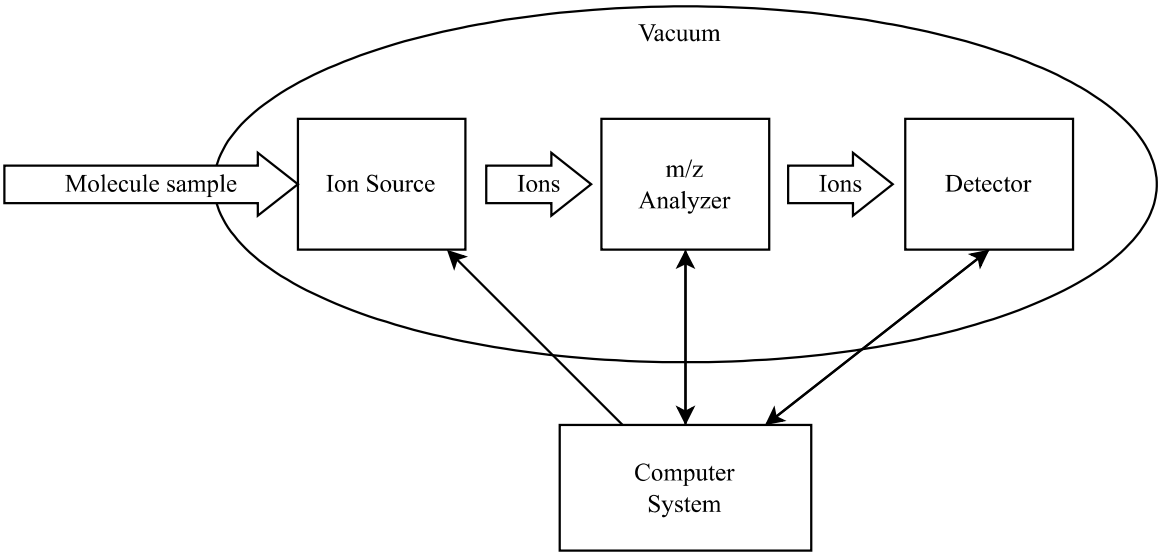

Figure 4: Schematic illustration of the components of a mass spectrometer (based on [WS07, p.2])

is transferred into the molecule, and the type of ions generated in the process [NF15, p.7]. A common technique which can be used for substances like THC, is *electron ionization*. This approach forces a collision of the sample molecule with an electron. A requirement is that the analyte is already in gaseous phase. During the process, an electron is added or removed. The result is a molecular ion, which is an ion with an odd number of electrons and positive charge (in contrast to a molecule which has an even number of electrons). If sufficient energy is put into the ionization, chemical bonds in the molecular ions break. This reaction is called fragmentation. An illustration of a fragmentation process is shown in Figure 5. The ions formed in this process are called fragment ions. They mostly have lesser mass and the same charge as its precursor ion, which is the ion it arises from. Secondary fragmentation is also possible, that means fragment ions can be formed through the fragmentation of fragment ions [WS07, p.22–23]. This is also apparent in Figure 5, where one path leads through four fragmentations. An important fact to note is that the fragmentation pathway is a strong indicator for a compound and also reproducible. Thus, it is often referred to as the "chemical fingerprint" of a compound [WS07, p.317]. The fragmentation pathway shown in Figure 5 depicts what happens in the ion source when THC, the major psychoactive component

of the cannabis the patient of our example consumed, is analyzed.

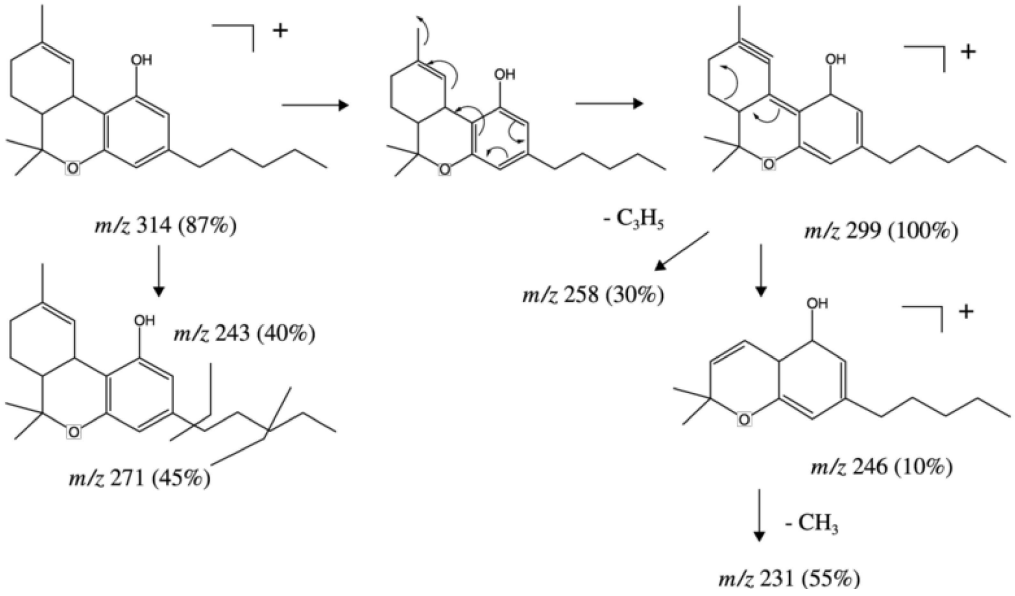

Figure 5: Fragmentation pathway of Tetrahydrocannabinol in mass spectrometry with electron ionization [Cho+04]

After the gas-phase ions have been generated, they are forwarded to the mass analyzer. The ions are present in the prior presented form. However, mass spectrometers only work with masses and charges and give no structural information. The mass analyzer thus separates the incoming ions according to their mass and charge, i.e. according to their m/z value. The separation process is based on the fact that ions show characteristic behavior in electric and magnetic fields [WS07, p.61]. Analogous to the ion sources, there is a number of mass analyzers you can use, Niessen and Falck [NF15, pp.13–18] differ between six types. The choice of a suitable mass analyzer is depending on a lot of parameters, like m/z range, mass accuracy, and resolving power [WS07, pp.25–26].

The last component of a mass spectrometer is the ion detector. As input, the instrument gets the separated ions from the mass analyzer. Koppenaal et al. [Kop+05, p.419] call detectors the “eyes” of mass spectrometry. The reason for this is their task which is to register the incoming ions producing an electrical current that indicates the intensity of the ion beam [Dow04, p.22]. There are different detectors, the type you use is heavily

dependent on the spectrometer design, mainly on the mass analyzer [Kop+05, p.420]. The output of the detector is a mass spectrum which is explained further in the following subsection.

3.4.3 The Mass Spectrum

As noted earlier, the result of the mass spectrometry process is a *mass spectrum*. An often used depiction of a mass spectrum is a histogram. An example for this representation is shown in Figure 6. This is the mass spectrum that is generated when the analyte molecule is THC, as it is the case in the example that was introduced in the beginning of this section.

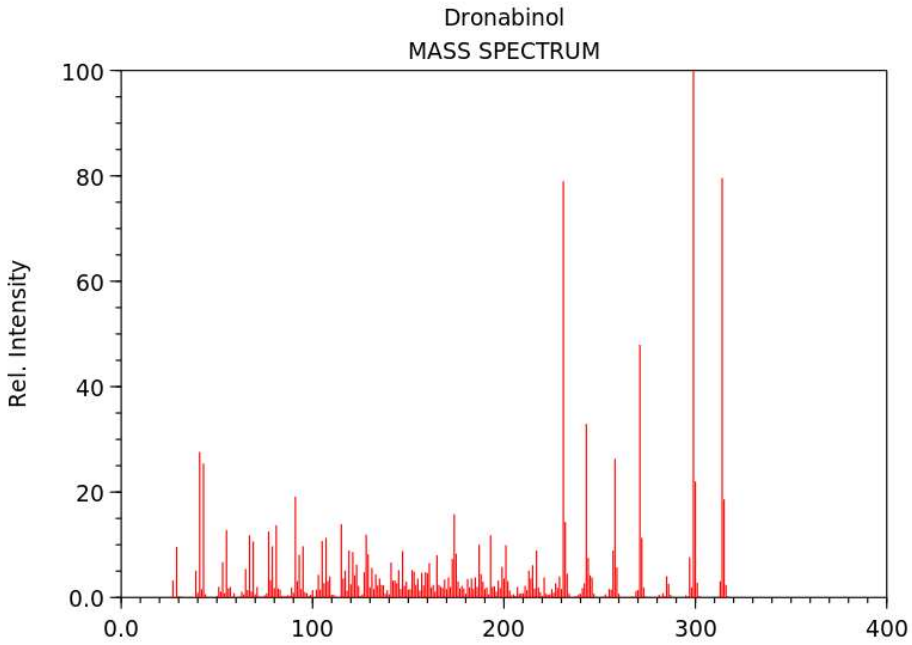

Figure 6: Mass spectrum of Tetrahydrocannabinol [Cen]

The mass spectrum shows peaks which "represent the ions formed in the mass spectrometer" [WS07, p.23]. The x-axis shows the m/z-values of the ions that were separated by the m/z analyzer while the y-axis indicates the intensity of the peaks. The highest peak (in Figure 6: peak with m/z-value of 299) in a mass spectrum is called base peak. The intensity can either be an absolute or a relative value. The absolute intensity shows the signal strength of the current for the corresponding ion. For the relative intensity you

normalize the absolute intensity in such a way that the base peaks equals 100% [WS07, p.23]. The base peak in the mass spectrum shown in Figure 6 corresponds with the fragment in the upper right corner in Figure 3.

When looking at mass spectra, one has to consider a few metrics describing the mass spectrometer to assess the data. A very important metric is the *mass accuracy* which is calculated from the *exact mass* and the *accurate mass*. The exact mass is the theoretical m/z of an ion and the accurate mass describes the "experimentally determined m/z" which was measured in the mass spectrometer. The absolute mass accuracy is given as the error function $accurate - exact$ (result is given in mu). You can also calculate a relative error with the formula $\frac{accurate-exact}{exact} \times 10^6$ (result is given in parts per million (ppm)) [NF15, p.4]. Thus, the mass accuracy describes how exact a mass spectrometer can measure the m/z of an ion. Another metric is the *mass resolution*. It shows the ability of a mass spectrometer how well it can differentiate between ions of almost identical mass [WB20, p.1747]. *High-resolution mass spectrometry* enables the differentation of peaks that have identical m/z up to several decimal places.

3.5 Analysis of Synthetic Cannabinoids

Section 3.4 introduced basic concepts and components of mass spectrometry with the aid of the example THC. In this example, gas chromatography in conjunction with electron ionization is used. This approach is not suitable for synthetic cannabinoids. In this context, liquid chromatography in conjunction with electrospray ionization proved to be a suitable tool [KA12]. Next to the fact that liquid chromatography instead of gas chromatography is used, other factors are slightly different in the process. The molecules introduced by the chromatograph are not bombarded with electrons, but a strong electric field is created [BM12, p.3]. This causes the sample solution to disperse into charged aerosol droplets. These droplets are then evaporated which result in gas-phase ions. In contrast to electron ionization, these ions remain intact and do not undergo fragmentation [BM12, p.3]. This has the advantage that the molecular weight can be determined very precisely. However, structural information can not be concluded. To get this additional information, collision-induced dissociation is performed [BM12, p.18]. The molecular ion is selected from the electrospray ionization results. This ion is also called *precursor ion* [BM12, p.18]. The internal energy of the precursor ion is

then increased which results in fragmentation. Results of electrospray ionization mass spectrometry thus include the molecular weight and the mass spectrum. Two sequential mass analyzers are needed in this process, one after the electrospray ionization and one after the collision-induced dissociation. Thus, such methods are also referred to as *tandem mass spectrometry.*

3.6 Identification of Drugs with Mass Spectrometry

Mass spectrometry plays an important role in drug testing and identification. Harper, Powell, and Pijl [HPP17, p.2] describe mass spectrometry as the drug testing technique that identifies a substance most accurately and call it the "gold standard in forensic drug analysis". Drug testing with mass spectrometry can be seen as an analytical application where you identify "known unknowns" [NF15, p.32]. This is a term for the analysis of a unknown substance where the goal is to find the substance at hand in literature or in databases. The approach of how to find the substance in a database depends on the ionization technique. When using a technique like electron ionization or electrospray ionization in conjunction with tandem mass spectrometry, the analyte molecule fragments. As the fragmentation pattern of a molecule is known, you can search for the molecule species by using the generated mass spectrum to match structures of known molecules in a spectral database. This approach has proven to work well for small molecules. Most illicit drugs, notably synthetic cannabinoids, classify as small molecules [HPP17, p.2]. As in many other application fields, chromatography techniques are performed on the analyte before the mass spectrometry process to separate the mixture into isolated molecules [HPP17, p.2].

An important component that makes the identification of known compounds possible are reference databases containing the mass spectra and other metadata of the compounds retrieved from previous experiments [SHB13, p.3]. There are several of these databases available. Two important commercial ones are the NIST mass spectral library and the Wiley Registry, mostly containing electron ionization mass spectra and divided by compound type [SHB13, p.3]. The most comprehensive library for synthetic cannabinoids is the database "Mass Spectra of Designer Drugs" from Wiley Science Solutions. In 2022, this database contains 32,855 compounds in total and 1,779 mass spectra of cannabinoids [Rös22]. While these commercially available libraries contain more compounds

than open databases, they also have major drawbacks. The access often requires special software and can be expensive. Furthermore, there are mostly no automatic continuous updates [Sto+12, p.2]. This is especially a problem with designer drugs such as synthetic cannabinoids because new substances are developed and brought to market rapidly.

Searching such libraries for the unknown substance can be done by building a similarity function to spectra in the database. A common scoring for the distance is the peak count of matching peaks (i.e. peaks with the same m/z). Another approach is building a dot product, taking the intensities into consideration [SHB13, p.4]. This is also what would happen in the example drug screening process introduced before. The generated mass spectrum depicted in Figure 6 or the peaks of it are matched against a database containing known illicit drugs. This would result in a high similarity between the spectrum at hand and substance in the database and thus in a positive result.

3.7 Identification of Novel Substances with Mass Spectrometry

Until now, it was assumed that you search for substances that are acquainted in literature and relevant databases. The example looked at the well-known compound THC which is a component of cannabis. When looking at designer drugs like synthetic cannabinoids, the need for the identification of undiscovered substances emerges because of the rapid development of new compounds in this field. Obviously, relevant databases do not contain these novel or undiscovered substances. While the prior section describes drug testing as a search for "known unknowns", the identification of novel substances can be referred to as searching for "unknown unknowns" [NF15, p.32]. The main task here is the structure elucidation of the molecule [HB16, p.625]. Different methods for automatically finding structural information to novel substances are mentioned in literature: Searching for similar compounds, mass spectral classifiers, and *in-silico fragmentation* [SHB13, pp.9-14]. The first approach bases on the assumption that spectral similarity corresponds to structural similarity of compounds. Thus, this method searches for similar spectra in a spectral library of known compounds. Mass spectral classifiers build classes from different mass spectrum properties and use these for a search in a spectral library. A drawback of approaches based on similarity and classifiers is the necessity of mass

spectral libraries and their limitations explained in Section 3.6.

Therefore, an approach often used is in-silico fragmentation. In-silico fragmentation refers to a technique where fragmentation processes are simulated in the computer. Here, instead of querying spectral libraries, you use a database containing molecule structures [SHB13, p.11]. In the case of undiscovered molecules, you can not use a publicy available or commercial structural database, as they do not contain structures yet to be found. A possible idea is to create a molecule structure generator which produces a database of candidate structures [SHB13, p.11]. Here, a lot of structures are created. Thus, when querying the candidate database, a simple search with the precursor ion mass of a spectrum produces a lot of matches. Here in-silico fragmentation comes into play. The idea is to simulate the fragmentation process and thus to predict the resulting fragments. These fragments are then stored together with the compound [SHB13, p.11]. Thus, querying not only by the precursor mass, but also by the peaks of a mass spectrum is possible. Peaks are matched with the predicted fragments which corresponds to the idea of the similarity search in finding "known unknowns". In the context of this thesis, this simulation of the fragmentation can be simplified. As it is assumed that novel synthetic cannabinoids are structurally similar to known ones, a model based on the fragmentation patterns of known synthetic cannabinoids is used and applied on novel ones.

3.8 Related Work

Several papers concern themselves with the identification of novel molecules, particularly small molecules, based on mass spectrometry in conjunction with in-silico fragmentation. Hufsky and Böcker [HB16] reviewed methods used to approach this challenge. Most approaches mentioned in the review follow the fundamental process of building molecule candidate sets and predicting the fragmentation in a mass spectrometer. The process consists of three stages: molecule-structure generation, in-silico fragmentation, and matching.

There are various ways to produce novel molecule structures. Stravs et al. [Str+22] introduced the tool "MSNovelist" which generates novel compound structures as SMILES based on the mass spectrum of an analyte. They do this by creating a fingerprint based

on the spectrum and generating the structure by making use of a recurrent neural network which is trained on a dataset of over a million structures. This approach is designed for a large class of compounds, namely metabolites, and delivered reasonable molecule structures in more than half of the test spectra [Str+22, p.869]. Another generative approach based on deep learning was presented by Skinnider et al. [Ski+21]. They use a specialized form of a recurrent neural network, namely a *Long Short-Term Memory model*. The model was trained for predicting structures of novel psychoactive substances. It showed results similar to MSNovelist, with a perfect accuracy of 51% and ranking the right structure among the top three in 71% test cases [Ski+21, p.979]. While these accuracies about 50% do not look convincing at first, one must consider that structure elucidation is a highly challenging task as there are so many theoretically possible structures, even if you know the molecular formula. For instance, when you only consider the formula $C_8H_6N_2O$, there exist more than 100 million possible structures [HB16, p.625]. This number scales up tremendously for synthetic cannabinoids, as these mostly contain more atoms, which means more possible combinations.

Therefore, it is not only important to generate reasonable compounds, but also to model the fragmentation process accurately which enables the filtering of reasonable candidate molecules. An often used method is the rule-based fragmentation spectrum prediction. Here, the fragmentation process for the generated molecules is simulated by applying known fragmentation rules which mostly are provided by mass spectrometry experts. While providing general fragmentation rules for a large class of molecules is difficult, finding rules for the fragmentation of structurally similar molecules becomes easier [HB16, p.626]. Kind et al. [Kin+13] used rule-based in-silico fragmentation to predict the fragmentation processes of prior generated lipids by defining experimentally aquired rules for lipid subclasses. This approach shows a sensitivity of 89%, a specificity of 96%, and a false positive rate of 4% when validating it with a then-current official spectral library. This shows that a rule-based approach can definitely be a viable choice for predicting fragmentation processes of a known compound class.

4 Implementation

Current research shows that a fully automated approach for accurately identifying novel substances with mass spectrometry remains a challenge. However, it also indicates that reasonable results can be achieved by taking fragmentation processes into account and restricting the set of possible targets to a small compound class. These observations fit to the goal of this thesis: filtering mass spectrometry data of forensic laboratories for "potential synthetic cannabinoids", i.e. novel molecules that would fit the building block description elucidated in 3.2.

However, simply stating that a molecule analyzed in the mass spectrometer could potentially be a synthetic cannabinoid is not sufficient for forensic laboratories. Thus, the implementation should support further manual analysis and therefore provide necessary information regarding the mass spectra, the generated molecule, and its fragments. This should especially eliminate the number of false positives, as an expert can filter these out based on the given information rather easily. On the other hand, specificity should also be reasonably high to minimize the manual effort. However, a high sensitivity is of higher priority, as overlooking positives would be worse in drug screenings for obvious reasons.

Another important requirement is that the implementation should not only support the analysis of a single spectrum, but rather enable the filtering of multiple mass spectrometry files for potential synthetic cannabinoids, each file containing approximately 2000–3000 spectra. This necessity also leads to a non-functional requirement: the analyis of single spectrum should be accomplished in a few seconds to enable a quick analysis of multiple files. Also, efforts for setting up the analysis tools should optimally be low.

This implementation is mainly meant for forensic laboratories which use mass spectrometry as a means for drug analysis. It enables the tracing and further analysis of not only

known, but also potentially novel synthetic cannabinoids in mass spectrometry data. An important property of the implementation is the low runtime for the analysis which enables the inspection of large mass spectrometry datasets that are common in forensic laboratories.

The following chapter describes the implementation in detail. After an overview of the solution outline, the project architecture is presented. Subsequently, the components of the architecure are explained in detail. This explanation includes utilized software packages and technologies, but also own developments.

4.1 Solution outline

Before the actual implementation, some general design decisions on the overall approach had to be made. The general process for tracing potential synthetic cannabinoids in mass spectrometry data consists of three steps: (i) generating structures of potential synthetic cannabinoids, (ii) generate their fragments according to the rules of the fragmentation model and (iii) matching mass spectrometry data with the generated molecules and their fragments by their mass or molecular formula.

The computational generation of potential synthetic cannabinoids and their fragments can take a long time as many possible structures adhere to the building block description. Because of this, it was decided to separate the first two steps from the analysis application. A script takes care of these two stages and provides the results of the process in the form of a database to the application where the end user does the analysis.

This analysis part can be seen as the third and last stage, the mass spectra reading and matching with the database. This is done by extracting the precursor ion and of the peaks of a mass spectrum and then querying the database for molecules (or rather their ionized form) that match the precursor ion and has fragments that match with peaks of the mass spectrum. There are two possibilities that were considered for the matching part: matching ions based on their monoisotopic mass, or calculating possible molecular formulas for the ions contained in the mass spectrum and match compounds and their fragments based on these calculated molecular formulas. The latter option has the potential to filter certain formulas based on the count of elements the formula should

have. A comparable implementation is the "MF Finder (Molecular Formula Finder)" from ChemCalc [PB13] where you can define ranges for the elements. However, the filtering through element count ranges is mostly based on empirical values which can not be assumed to be universally valid for all kinds of synthetic cannabinoids. Filtering out potentially relevant compounds by giving wrong ranges is not unlikely. On top of that, finding molecular formulas to a specific monoisotopic mass is an additional computing task which means additional runtime in the analysis stage. Due of these reasons, the first option was preferred.

4.2 Architecture

After the conception phase, an architecture was defined to chart components necessary for the implementation. A graphical depiction is shown in Figure 7.

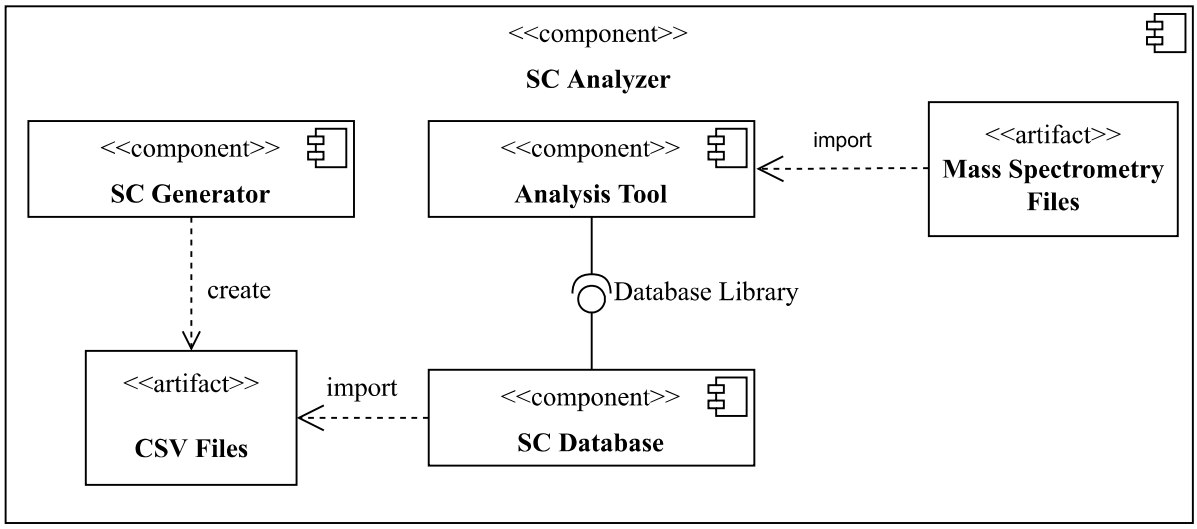

Figure 7: Illustration of the project architecture

The first main component is the *SC Generator*. As the name suggests, it generates structures of molecules that can be classified as synthetic cannabinoids. Additionally, it generates corresponding fragments based on a rule-based fragmentation model. The SC Generator writes the results of these two steps into Comma-separated values (CSV) files

which are then imported into the *SC Database*. There are multipe reasons for writing CSV files and importing them into the database instead of inserting the results directly. Firstly, this approach does not assume the underlying database which is used and thus decreases dependence from the underlying technology. A second reason is that it is not necessary for users to execute the generation process themselves. Also, if new rules for the generation and fragmentation are implemented into the SC Generator, which leads to updates of existing and inserts of new compounds, the user can simply download the new datasets and import them instead of manually initiating the generation process after an update.

The last component of the architecture is the *Analysis Tool*. This component is responsible for the third stage mentioned in Section 4.1. Here, the user can import multiple mass spectrometry files and search for potential synthetic cannabinoids within them. If there are matches with the database, the user can perform further analysis by inspecting the structure of the suggested compound, the fragments, the spectrum, or important numbers like mass measurement deviations.

4.3 Components

The following section explains the implementations of the components shown in Figure 7 in detail. The component first presented is the database because other components must adhere to the data model defined there. Thus, defining a sensible data model in the beginning is essential. Another constituent is a database library which provides functions to query and alter the contents of the database. This is especially important for the analysis tool. Next, the basic functionality of the SC Generator is explained. Lastly, the analysis tool, which can be seen as the main component of this thesis, is elucidated.

4.3.1 Database

As mentioned in Section 4.2, the choice of a database technology does not affect the generator, as there is no direct interaction between the generator and the database.

However, the analysis tool interacts with the database with the help of a database library. Thus, the choice of a database technology has to be considered. As the last requirement mentioned in the introduction of this chapter mentions, the effort for setting up the database should be low. The architecture shown in Figure 7 also does not require the usage of a client/server-database. Due to these reasons, it was decided to use SQLite. To make sure SQLite fits the needs of a project, the SQLite project provides a "When to Use" [SQL22] which also includes a checklist when not to use SQLite. There, it is recommended to stick to a client/server database if the accessed data is separated by a network, if there are many concurrent writers, or if the data size is huge. The first conditions definitely do not hold here, as the database is not supposed to be a central component and writing only happens when importing the current datasets. The last condition of the data not being too big should also not be a problem. While it is true that many compounds are created in the generation process, the database size will not come close to the terrabyte range where it is recommended to use a client/server database. Also, when the need to use a client/server-database arises, the effort for adjustments is low. Thus, SQLite can be considered a suitable choice as a database system here.

The following subsection describes the data model of the database and the library which is provided for querying and altering the data. The focus is not only on the structure, but also on the information each field is supposed to give.

Data Model

An illustration of the database schema is shown in Figure 8. The two most important entities are *Compound* and *Fragment*.

A compound is identified by its canonical SMILES string. Additionally, the atomic mass, the molecular formula and the m/z of its positively ionized form (named ion mass) are stored together with a compound. Another field is the name of the compound. As with many other chemical compounds, the nomenclature of synthetic cannabinoids is not always consistent. For instance, one of the first detected type of synthetic cannabinoids is named JWH-018, which are the initials of the researcher John W. Huffman who first synthesized this particular compound. Other nomenclatures reference the marketed names [Pul+22, p.2]. An idea to standardize the naming in order to enable clear and

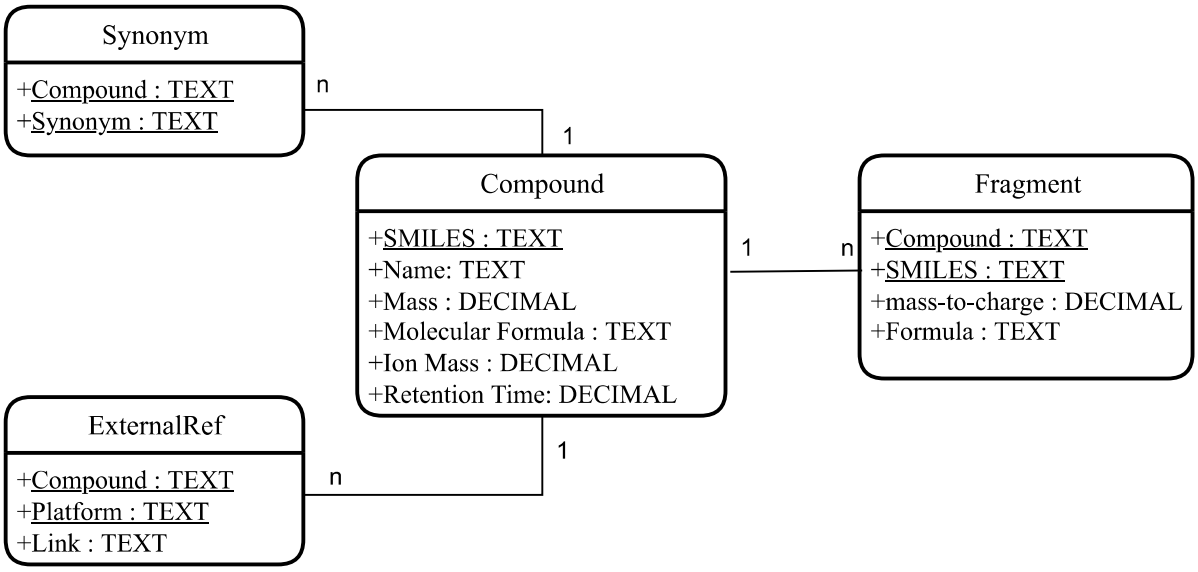

Figure 8: Illustration of the database schema

unambiguous differentation is to use a semi-systematic naming framework proposed by Pulver et al. [Pul+22, p.16–19]. This method focuses on the building blocks of synthetic cannabinoids and uses existing names for these blocks or generates new ones if they do not exist yet. The drawback of this method is that it is mostly limited to discovered compunds and thus can not guarantee a consistent naming of novel molecules [Pul+22, p.20]. The last field is the retention time of the compound in the chromatograph. This serves as a placeholder for eventual implementations of a retention time prediction. The reason for predicting the retention time is that using it as a parameter in the matching process can improve accuracy of the matching [Jew+20, p.8]. As the matching bases on the ion mass of a compound, queries are likely to involve this field.

There are two more entities which add information to a compound: *Synonym* und *ExternalRef*. Synonyms can be thought of as an addition to the semi-systematic naming as means to tackle the challenge of inconsistent naming. There are often different terms for the same substance in literature. For instance, PubChem lists over 300 synonyms for Dronabinol that were used by different data contributors [Bio23c]. The entity Synonym is supposed to cover at least some of these different terms for a compound to make molecules more recognizable to the users. What is important to note, however, is the fact that only compounds known in literature and compound databases can be recorded here.

The same is true for the entity ExternalRef. Here, references to compound databases are stored. This can be useful for finding additional information. Also, if the simple fact that there is an external reference is a good indication that the compound at hand is known in literature. The table references different database, e.g. PubChem [Bio23a] or the CAS registry [Ser23]. An example for an entry in this table is depicted in Listing 1 which shows the insert of the PubChem ID for Dronabinol.

```
INSERT INTO ExternalRef (compound, platform, link)
    VALUES ('CCCCCC1=CC(=C2C3C=C(CCC3C(OC2=C1)(C)C)C)O',
            'PUBCHEM', '16078');
```

Listing 1: Example for an insert into the table *ExternalRef*

Analogous to compounds, fragments contain their SMILES representation as an attribute. To identify a single fragment, however, the SMILES string of the compound it arises from is also necessary as a foreign key and as a second part of a composite primary key. Otherwise, a mapping between compounds and fragments would not be possible. Analogous to compounds, the formula and the m/z of a fragment is stored.

Database Library

To create, alter, or query the prior introduced tables, a simple SQL wrapper library in the form of a Python package was developed. As one of the main requirements mentioned in the introduction of this chapter is the performance in the matching process, little overhead when performing database reads is an important factor. Thus, data from the database is read (and also written) with the lightweight standard module sqlite3 [Fou23b]. To make a potential switch to other databases easier in the future, the functions in this wrapper library are simplistic. Thus, most of the business logic was implemented in other components. This library merely contains simple functions for creating compounds, fragments, synonyms, and external references. Furthermore, functions for querying compounds or fragments by ion mass or SMILES string are provided by the library.

4.3.2 Synthetic Cannabinoid Generator

The next component is responsible for the generation of potential synthetic cannabinoids and the simulation of the fragmentation process. It comes in the form of a Python script which follows the process depicted in Figure 9.

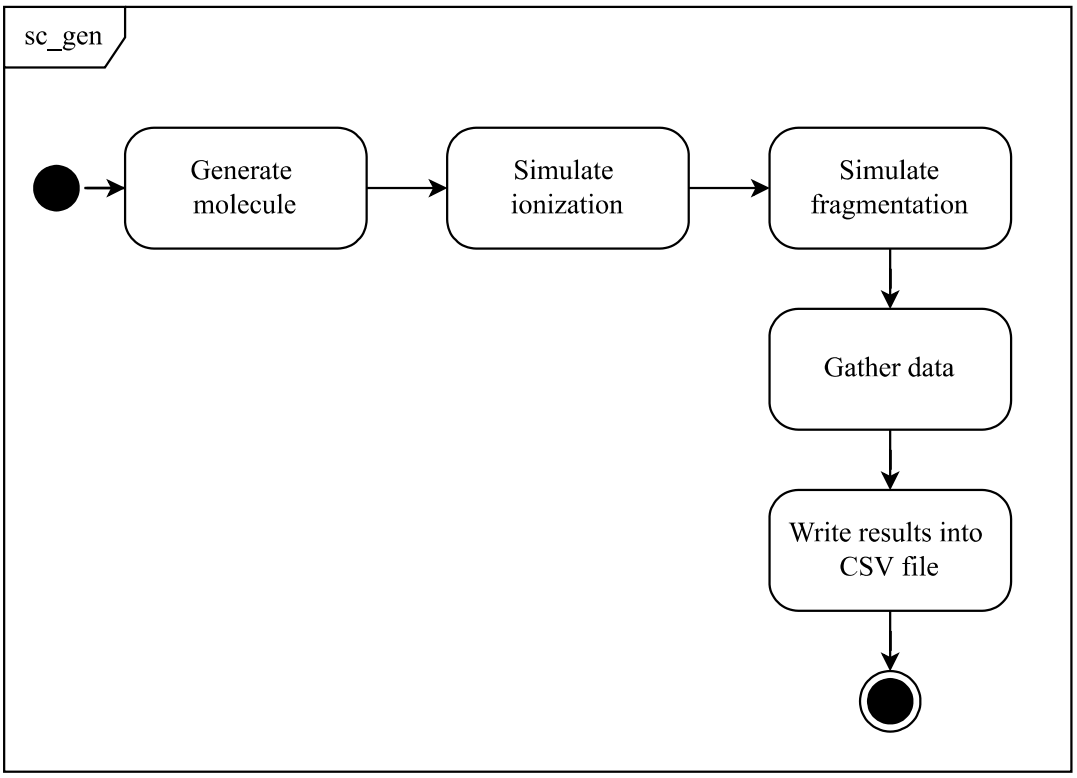

Figure 9: Activity diagram for the generation process of a single compound

Molecule Generation

As shown in the data model in Figure 8, the structures of molecules (and also fragments of them) are represented as SMILES strings. Consequently, the molecule generation works mostly with strings. A major part in the generation process is based on the building-block definition of synthetic cannabinoids. It is assumed that all synthetic cannabinoids follow this definition. The basic thought of generating the compounds is shown in Listing 2. First, lists of possible structures for each building block are defined. Here, sound knowledge of the chemical composition of synthetic cannabinoids

is essential to build reasonable compounds. The last step is the concatenation of the generated building blocks into one compound. This step corresponds to building the cartesian product of the building block sets. Because a lot of compounds are created here, a generator expression is used to allow lazy evaluation and to avoid out-of-memory errors in the further steps. The challenge in the generation process is to find all possible candidates for the building blocks.

```
1  def generate_syn_cans():
2      heads = [...]
3      cores = [...]
4      linkers = [...]
5      tails = [...]
6      yield from (f'{head}.{core}.{linker}.{tail}'
7                  for head in heads
8                  for core in cores
9                  for linker in linkers
10                 for tail in tails)
```

Listing 2: Demonstration of the basic idea for generating synthetic cannabinoids

Ionization and Fragmentation Simulation

After a compound is generated, the mass spectrometry process as described in Subsection 3.4.2 is simulated. All chemical reactions are executed with the rdChemReactions module from the RDKit project [RDK22a]. Here, a reaction is defined as a single SMARTS string. SMARTS is a language for specifying substructures of molecules and is compatible with SMILES. It also supports reaction queries [Sys19a]. Reactions are simulated in three steps: formulating a SMARTS query for matching the substructure which is supposed to be changed, defining what reactions are supposed to happen (e.g. adding atoms or breaking bonds), and running the reaction. RDKit works with an own class for molecules named Mol.

The first step of the in-silico mass spectrometry process is ionization. This is simulated by simply adding a positively charged H-atom to the molecule (protonation). The result is a molecular ion which is also positively charged. It is assumed here that sufficient energy is put into the molecule, resulting in fragmentation. As the compound generation is based on building blocks, it is known which substructures arise from it. Thus, one can

define fragmentation rules depending on the appearance of specific substructures. These rules are built from empirical findings on the fragmentation of synthetic cannabinoids.

Data Gathering and Export

The result of the prior described mass spectrometry process includes the original compound, the molecular ion, and the ions which arose from the fragmentation. All of them are present as RDKit Mol objects. To extract the information needed which corresponds to the fields of the two tables Compound and Fragment in Figure 8, RDKit provides methods to get chemical information to a molecule. This includes the fields SMILES, molecular formula, and mass. The ion mass (or more precise, the m/z) can be extracted from the molecular ion. The same goes for the fragment ions. For the fields name and retention time, default values are given for now. In the future, these fields can be used for a systematic naming scheme and for the result of a retention time prediction.

After the data gathering process is finished, the resulting compound and fragment datasets are exported to CSV files. These two datasets will contain a lot of rows. It is not assessable how many, as the rules for the generation and fragmentation can always change, but the number will be a multiple of millions or even billions. Thus, a efficient writing mechanism is needed to keep the run time at an acceptable level. PyArrow, which is the Python implementation of Apache Arrow [Fou23a], proved to be suitable for this task as it implements multi-threaded writing innately.

4.3.3 Analysis Tool

The last component of the architecture is the analysis tool. Its subcomponents are illustrated in Figure 10. It generally provides three core functionalities: reading mass spectrometry files, matching spectra of these files with the database, and presenting the results in a clear and purposeful manner. Correspondingly, these tasks are mainly handled by three subcomponents: a mass spectrometry data parser, a matcher, and an user interface.

The mass spectrometer data parser and the matcher provide functions to read mass spec-

trometry files and match a spectrum with the database elucidated in Subsection 4.3.1, respectively. These functions are called by the user interface when a user initiates a new analysis. The database is included in the component diagram in Figure 10 as it is a necessity for the matching process. An important requirement of the analysis tool is that it is supposed to handle multiple files. As showing all analysis results at once would be confusing for the user, only selective results are shown. To implement this requirement, state handling is needed. Thus, a model which holds the current state is added as a fourth subcomponent.

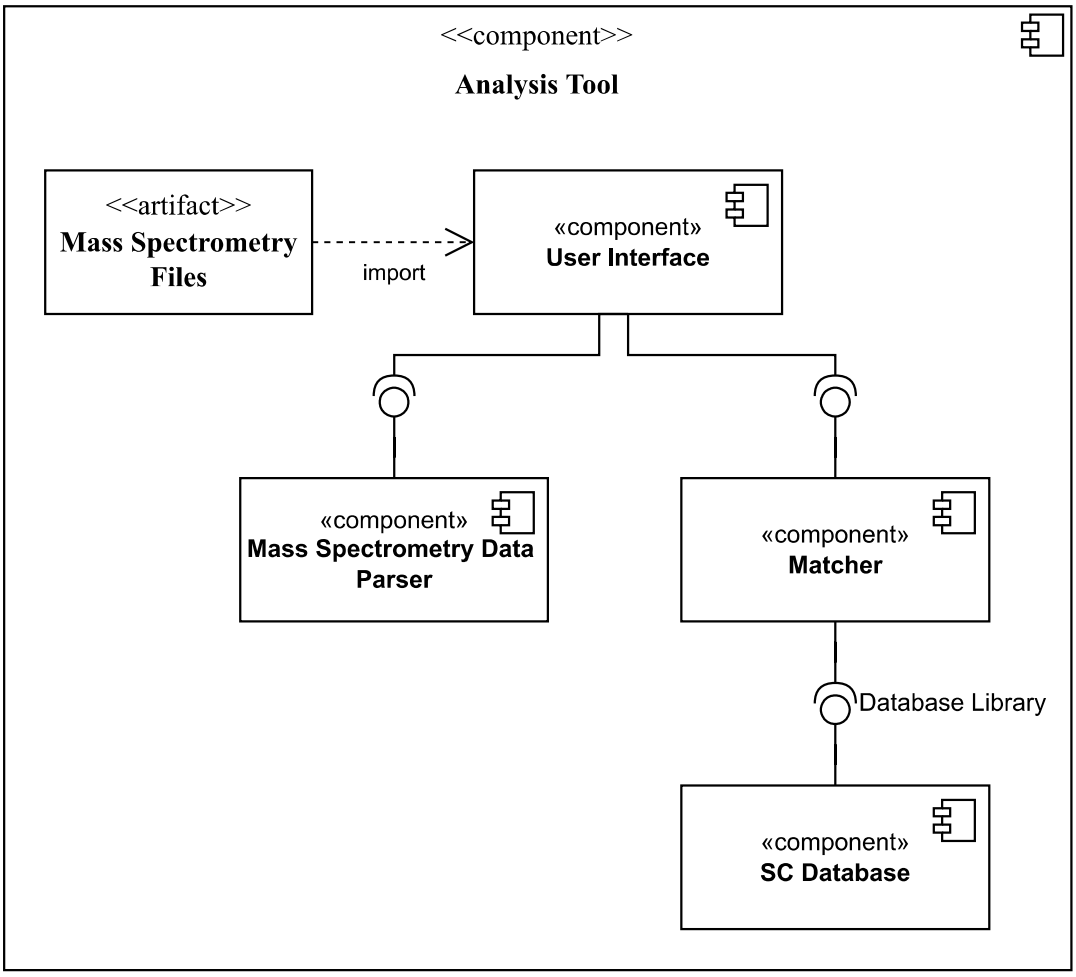

Figure 10: Component diagram of the analysis tool

In terms of technology, the analysis tool is designed as a web application that is meant to be run locally on the user's computer. If needed, it can also be deployed on a central webserver. To account for the low effort needed to set up the application, the lightweight

web development framework *Flask* [Pro23a] is used. For local usage, Flask provides a webserver. On the frontend side, plain Hypertext Markup Language (HTML) with vanilla JavaScript is used. To show dynamic content, the template engine Jinja [Pro23b] is used.

Mass Spectrometry Data Parser

Data produced in mass spectrometry can come in various formats. Mass spectrometer vendors have their own proprietary formats. Alongside, open data formats exist [Deucs, p.1612] which are definitely to be preferred in this implementation as the tool is not supposed to be dependent on a specific vendor. Another important factor is the use case where one can differentiate between files for preprocessing, mass spectrometer output files which contain mainly the mass spectra, and files which represent results of further analysis [Deucs, p.1613]. In this implementation, files containing the mass spectra are needed. To keep things simple, a plain-text format is used. One of the most common formats is the Mascot Generic Format (MGF) [Sci21].

Listing 3 shows an exemplary excerpt of a MGF file. Each MGF file can contain information to multiple mass spectra. A single spectrum is delimited by the pair of statements *BEGIN IONS* and *END IONS*. The field *PEPMASS* is an abbreviation for peptide mass, as MGF was originally designed for proteomics [Sci21]. More generally, it describes the precursor ion mass. *TITLE* is a field that applies to a single spectrum and thus is supposed to be a unique description thereof. The field *RTINSECONDS* indicates the retention time. *CHARGE* corresponds to the charge that was added to the precursor ion during the mass spectrometry process. After these parameters are defined, a peak list describing the fragment ions follows. Several defintions of a peak in the list is possible. However, the parser assumes that each line of the peak list has two entries: the first corresponds to the m/z of the fragment ion, the second to the intensity.

```
BEGIN IONS
PEPMASS=379.18132782
TITLE=Spectrum 01, MS/MS at 2.12571666667 mins
RTINSECONDS=127.543
CHARGE=1+
55.05476  99.99997
56.18750  1.02602
91.03173  63.41049
91.08336  1.01154
116.05015 17.49912
144.04569 76.55125
144.10567 1.40017
145.04730 2.06304
158.05801 1.17788
198.09361 11.76957
234.06959 63.87904
235.07039 1.14591
END IONS

BEGIN IONS
...
END IONS
```

Listing 3: Example for the structure of Mascot Generic Format file

The parser first splits the file into the single spectra. For each of them, it then searches for the precursor ion mass, the retention time, and the peak list. This is done with regular expressions. Depending on the output format of the mass spectrometer, intensities of the peaks are often given as absolute values in the MGF file. In literature and in practice, however, mostly relative intensities are used. As both formats could be relevant, for each peak list the relative intensities are calculated additionally to the absolute intensities by determining the intensity of the base peak and normalizing the intensities with the formula $relative\_intensity = \frac{absolute\_intensity}{intensity\_basepeak} \times 100$.

Different mass spectra do not have equal numbers of peaks. Depending on the mass spectrometer, the resulting mass spectra can contain much noise. To account for this, the parser has two additional parameters to filter the mass spectrum for peaks that are likely to be the most relevant: a threshold for the relative intensity and a maximum number of peaks to investigate.

Matching

The next subcomponent of the analysis tool is responsible for the matching of the prior acquired spectra with the database. If a spectrum has a match, it is called a *suspicious spectrum* in the analysis tool, as it does not guarantee the validity, further analysis by an expert is necessary. An important parameter for the matching process is the mass accuracy of the mass spectrometer the spectra were acquired from and has to be considered in all matching queries involving masses. The mass accuracy is given as an absolute value in Da. Another parameter is the minimal number of peaks that are supposed to match with the fragments that correspond to the compound in the database.

```
1  def match_spectrum(spectrum, mass_accuracy, min_matching_fragments):
2      candidates = []
3      compounds = search_compounds(spectrum.precursor_ion_mass,
4                                   mass_accuracy)
5      for compound in compounds:
6          matching_fragments = []
7          fragments = search_fragments_to_compound(compound)
8          for fragment in fragments:
9              for peak in spectrum.peak_list:
10                 if absolute(peak.mz - fragment.mz) <= accuracy:
11                     matching_fragments.append(fragment)
12         if len(matching_fragments) >= min_matching_fragments
13         or spectrum.base_peak in matching_fragments:
14             candidates.append(compound + matching_fragments)
15
16     if len(candidates) > 1:
17         return filter_candidates(candidates)
18     return candidates
```

Listing 4: Outline of the matching algorithm

Listing 4 shows an outline of the matching algorithm. It takes the prior discussed mass accuracy and minimal number of matching fragments as parameters. First, the database is searched for compounds that match with the precursor ion by querying the database with the ion mass and taking the mass accuracy into account. This step already restricts the number of potential matches to a small size. For each of these compounds, corresponding fragments are searched in the database. Fragments that match with peaks

in the spectrum are appended to a list of matching fragments. This approach results in a nested loop in the implementation. The alternative would be a single loop over the peak list and searching for fragments for each of the peaks. However, this would result in a lot of additional database reads which are more expensive. Hence, the former approach was preferred.

There are two possible conditions for a compound to be considered as a match to the spectrum at hand. The first one is if enough corresponding fragments match with peaks in the spectrum. The second possibility is that one of the matching fragments corresponds to the base peak of the spectrum. If one of these conditions hold, the compound is considered a potential match and thus added to the so-called list `candidates`. However, when both conditions do not hold, the compound is ignored in the further analysis of the spectrum.

The last step of the matching process is the ranking of candidates. This is done when multiple compounds are in line with the spectrum, i.e. when the list `candidates` in Listing 4 has multiple entries. The determination of the suggested compound is carried out by comparing following criteria (in the order they are enumerated):

1. Number of matching fragments

2. m/z of the peaks corresponding to the matching fragments (peaks with higher m/z have higher priority)

3. Intensity of the peaks corresponding to the matching fragments (peaks with higher intensity have higher priority)

To compare two compounds in step 2 and 3, the matching fragments are sorted by the m/z in descending order for step 2, or respectively sorted by the intensity for step 3.

Another possibility to further filter for sensible compounds is by using the retention time as an additional parameter in the matching process. This could even be a part of the first step, the function querying compounds then would have an additional parameter `retention_time`. As this would require a retention time prediction which is yet to be implemented, the filter is not in effect yet.

The matching and ranking process is further explained with the help of a fictional ex-

ample. For this, the process is demonstrated on a simple mass spectrum depicted in Figure 11. Additionally to the peak list, the precursor ion mass and the retention time is given, as it is read by the parser described before. It is assumed that the matching function is called with the parameter `min_matching_fragments` set to 2.

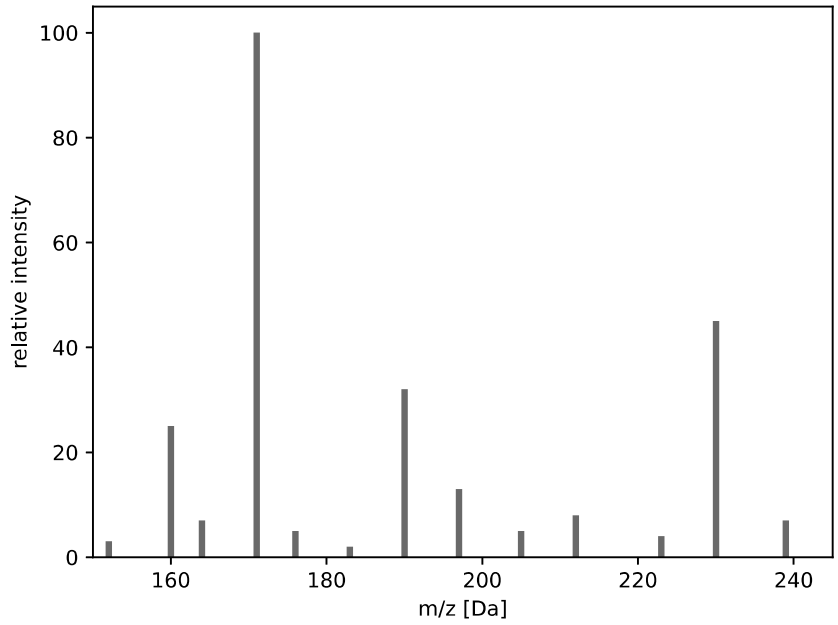

Figure 11: Example spectrum for demonstrating the matching and ranking process

It is assumed now that the database search initiated by the call of the function `search_compounds` in line 3 of Listing 4 has four compounds as a result for the example. For each of these compounds, the database is searched for fragments corresponding to the compound. Then the actual matching of the spectrum is happening. This is done by iterating through the fragments found to the compound and through the peak list. If they match by their m/z within the mass accuracy, the peak is considered to be "matching" with the fragment. The results of this process for the example is depicted in Figure 12.

Then the prior discussed conditions for a compound to be considered a potential match are checked. The first compound has only one matching fragment which does not correspond to the base peak. Thus, it is not considered as a candidate for the spectrum. All other compounds are still considered as possible matches. The second compound

has only one matching fragment, but as this fragment corresponds to the base peak, the compound is also considered as a candidate.

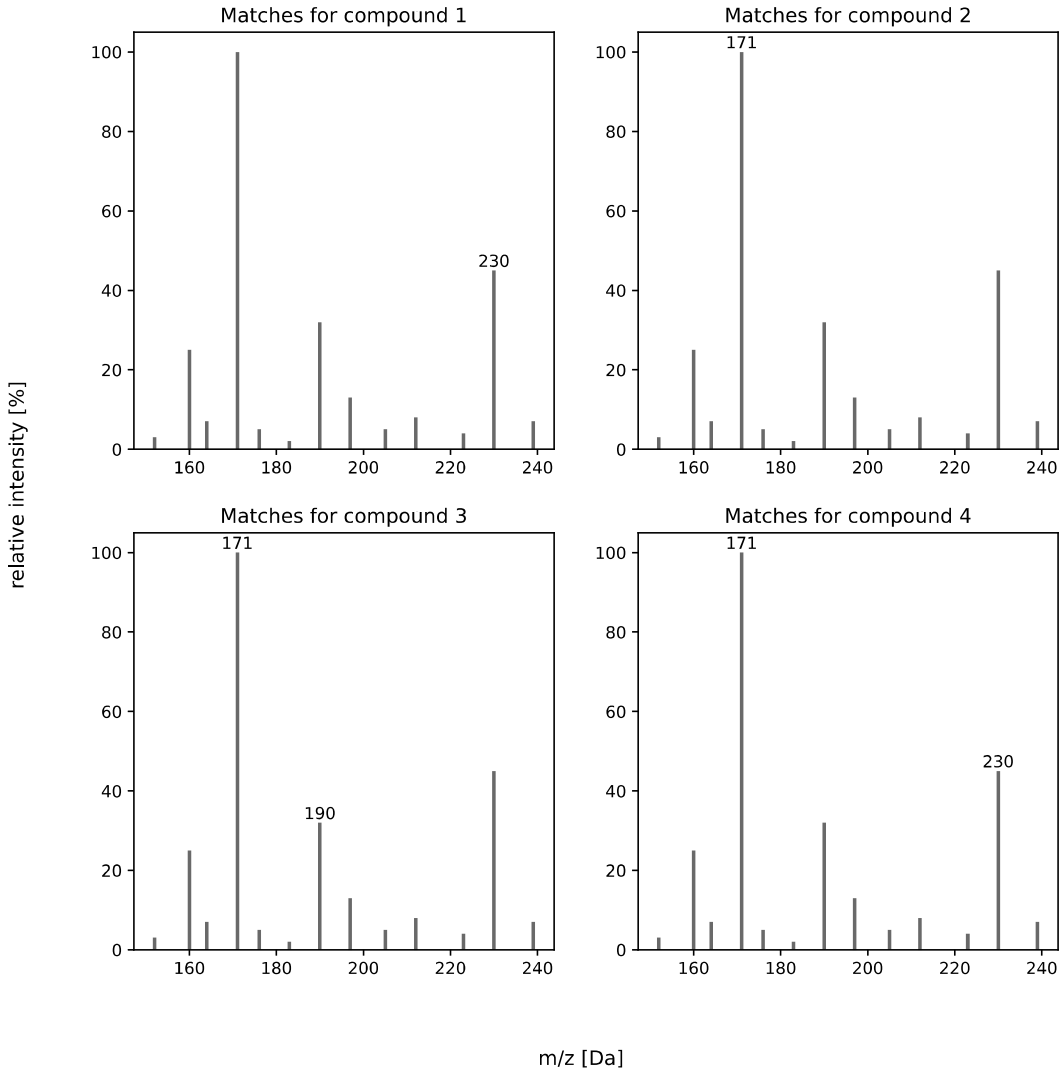

Figure 12: Illustration of the matching fragments in the example mass spectrum (matches are labelled with the m/z value)

In the last step, the candidates are filtered to only return compounds that are most likely to be eligible. For this, the candidates are prioritized. Only the candidates with

the highest priority are returned as matches. In the example, the second compound is filtered out in the first step, as it has only one matching fragment in contrast to the other two compounds which each have two matching fragments. The second filtering criteria is the m/z of the fragments where fragments with a higher m/z value have higher priority. In the example, sorting m/z values of the third compound results in the list `[190, 171]` while doing the same for compound 4 results in `[230, 171]`. As the first fragment of compound 4 has a higher m/z than the first fragment of compound 3, it has a higher priority. Thus, the function call of `filter_candidates` in line 17 in Listing 4 yields one compound, namely compound 4. Based on the matching rules, it is assumed that this compound fits the input spectrum the best. Therefore, this is the compound suggested to the user.

User Interface

As the sheer amount of compounds generated in the process described in 4.3.2 brings forth a risk of false positives by coincidentally matching non-relevant spectra, providing a good presentation of the results should support the user in performing further analysis. When starting the tool and accessing it with a web browser, a screen for inputting the MGF–files to analyze and parameters for the analysis is shown (see Figure 13). The first three parameters, namely mass accuracy, minimal number of matching fragments, and retention time accuracy are relevant for the matching process. The other two parameters, minimal relative abundance and maximal number of peaks, are passed to the MGF–file parser.

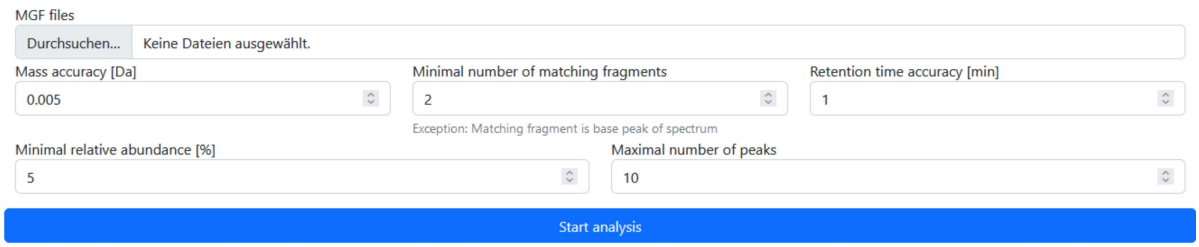

Figure 13: Start page of the analysis tool

The first step of the analyis is parsing the MGF files and optionally applying the two aforementioned filters. Afterwards, the matching process is conducted for each spectrum

in the files. The analysis results are then shown in an overview grouped by file. Figure 14 shows an example of this overview where five files, each containing > 2300 spectra, were analyzed. For each of these files, the number of spectra in total and the number of spectra that have potential matches are displayed. If there are suspicious spectra, a link is depicted. This link leads to a detailed analysis of a single file.

Figure 14: Overview of analysis results for all input files

The detailed file view consists of different subcomponents. The first one is a table displaying information to all suspicious spectra found in the file (see Figure 15). Here, information to the suggested compound, measurement deviations, the number of matching fragments is display for each spectrum. Each of these spectrum line also has a button indicated by a magnifying glass which selects the spectrum and shows further information for it.

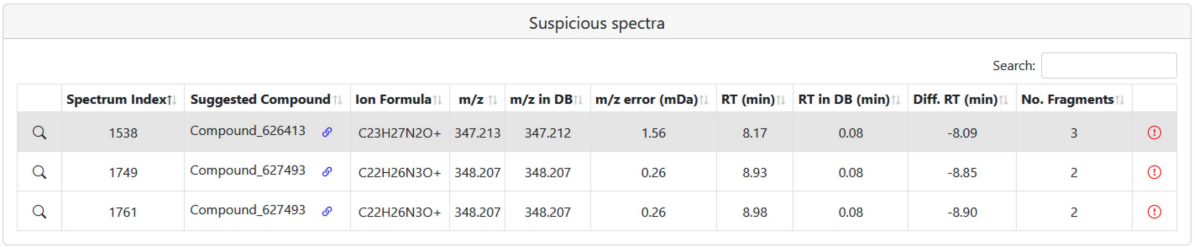

Figure 15: Overview of all suspicious spectra in a file

On the same page, three other subcomponents are shown which are specific for a spectrum, their content is loaded when clicking the magnifying glass corresponding to a

spectrum. The displayed information include the molecule structure of the suggested compound, a depiction of the spectrum, and a slideshow showing the molecular structure and other properties of all fragments matching to the spectrum. The structure of the compound and the fragments is created using the Draw module from RDKit [RDK22b]. The spectrum graphic is created with the means of matplotlib [tea23]. Additionally, the possibility to download the spectrum as a image or as a CSV–file is provided.

5 Evaluation

This chapter evaluates the method for tracing synthetic cannabinoids in mass spectrometry data presented in Chapter 4. Firstly, the test setup is elucidated. Then the results of the evaluation are presented and discussed.

5.1 Test Setup

For the evaluation, the generator presented in Subsection 4.3.2 is used to generate virtual synthetic cannabinoids and simulate their fragmentation. The resulting compounds and fragments are then input into a database which follows the data model and constraints elucidated in Subsection 4.3.1. What is important for the evaluation is that the database contains no other compounds and fragments than those produced by the generator. At the time of the evaluation, the database contains 685,880 compounds and 1,944,504 fragments.

For the evaluation, two separate datasets are used. The first dataset contains tandem mass spectra of compounds which are known to be synthetic cannabinoids. This dataset contains 76 compounds. For each compound, three MGF–files are provided. These three files differ in one important parameter of the mass spectrometer: the amount of energy that is put into the molecule in the ionization phase which is measured in electronvolt (eV). What is important to note is that this parameter can influence how much fragmentation is induced. The files include spectra with 10eV, 20eV, and 40eV. Here, sensitivity is measured which can be defined as the ability to predict true positives [SHT19, p.2]. In the case of this evaluation, it means to measure how many of the prior mentioned spectra are correctly identified as a synthetic cannabinoid.

The second dataset is meant to test the specificity of the method which can be defined as the ability to avoid predicting false positives [SHT19, p.2]. For this, 20 MGF–files, containing 2,681 spectra on average, which are known to not contain any synthetic cannabinoids or even compounds that would be suspicious, are used. Here, the eV as input parameter is not as important, as even strong fragmentation is not supposed to lead to matches in these files.

The implementation is tested with following parameters:

- Minimal number of matching fragments: 2
- Minimal relative abundance: 5%
- Maximal number of peaks per spectrum: 10

5.2 Results and Discussion

Firstly, the results of the test with the known synthetic cannabinoids are observed. Figure 16 shows the number of spectra that were correctly identified as synthetic cannabinoids grouped by the eV of the input spectra.

The method has a sensitivity of 60.5% for 10eV, 71% for 20eV, and 86.8% for 40eV. A notable observation is that the number of synthetic cannabinoids identified as such is increasing with the eV. Higher eV generally leads to stronger fragmentation. This is helpful here as more fragments arise which can potentially be matched. However, it is important to note that there can not be made a general recommendation what eV is best, based on these observations.

Next, the evaluation of spectra, which are known to not having any synthetic cannabinoids, is examined. In total, 53,629 spectra were analyzed in this evaluation. Only 8 of these showed matches with compounds in the database and thus can be considered as false positives. This results in a specificity of 99.985%.

The sensitivity is reasonable, especially for spectra with strong fragmentation where a large portion of synthetic cannabinoids were identified as such. However, further improvement needs to be done in order to come closer to the goal of spotting all synthetic

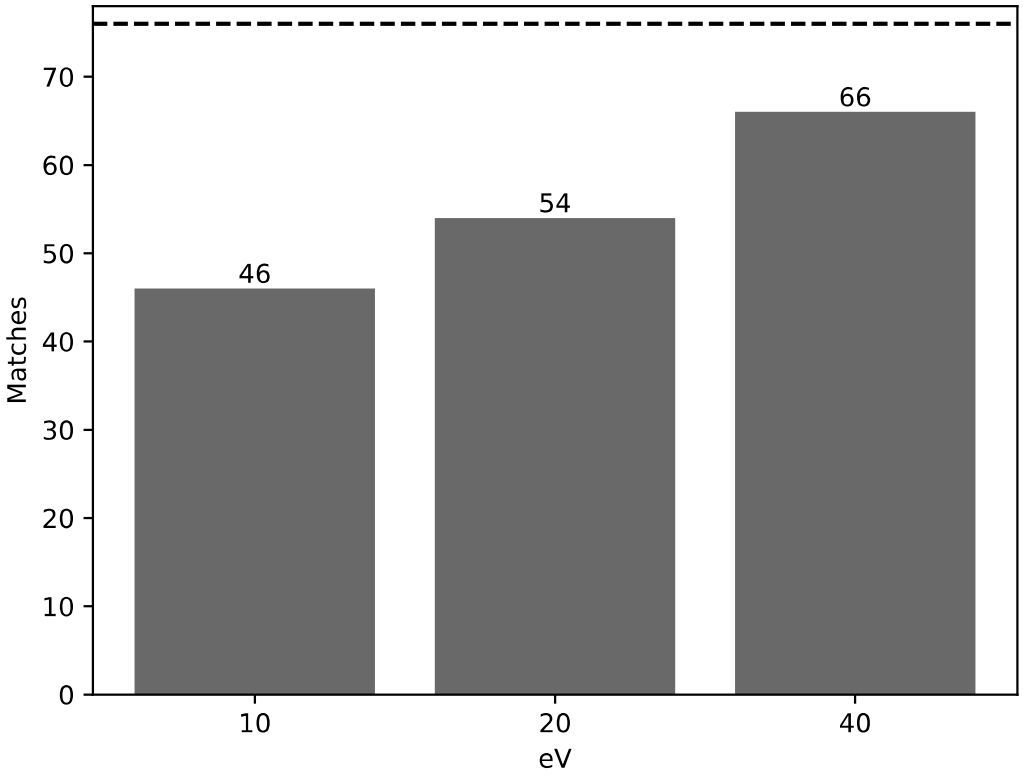

Figure 16: Matches in the evaluation of the method with known synthetic cannabinoids (dashed line indicates the total number of evaluated spectra which is 76)

cannabinoids or compounds that could fall into this category in mass spectrometry data. Reasons for not successfully spotting synthetic cannabinoids in the datasets can be manifold. Maybe the generator does not have rules to produce such a compound or does not produce suitable fragments. One possible reason could also be that the chosen parameters are too strict. More loose parameters can lead to a higher sensitivity, e.g. by setting the minimal number of matching fragments to 1. On the other hand, this would have implications for the specificity which is very good in the evalution. While it seems that lowering the specificity by a few percent is not bad, one has to consider that this would have a significant impact on the usability of the analysis tool. More false-positives means more manual effort to filter out coincidental matches. However, if a higher false-positive rate is accepted by the user, loosening the parameters is definitely a possibility. Thus an universal recommandation for choosing parameters can not be given, there

always is a trade-off. Based on the data, users have to proceed in a exploratively way. The analysis tool supports the user in this process by providing appropriate analysis parameters.

6 Conclusion

The implemented method tackles the challenge of tracing synthetic cannabinoids in mass spectrometry data which arises from the rapid development of new substances. It combines approaches from different fields known in literature and practice, namely structure generation based on combinatorial chemistry, rule-based in-silico fragmentation, and matching those compounds in conjunction with the resulting fragments with input spectra. The approach already shows promisingly good results in spotting synthetic cannabinoids in mass spectrometry data. The implemented user interface supports the user in further investigating potentially matching spectra. Besides general information on the suggested compound, structural information to the compound and its fragments is given. References to external compound databases are also given to further investigate the compound if it is already known in literature.

Although the analysis already delivers reasonably good results, the sensitivity is supposed to be higher. One way to achieve this could be to experimentally search for parameters more suitable to find all synthetic cannabinoids. What must be also considered in this approach is to limit the loss of specificity. Too many false positives would delimit the usability of the analyis tool. However, the specificity could then be kept high by introducing retention time prediction which would filter out non-relevant compounds. Another possible reason for the sensitivity not be closer to 100% is that the corrsponding synthetic cannabinoids are simply not in the generated database. This would mean that the generator needs further optimization. More rules to cover all possible synthetic cannabinoids are needed. Also, more fragmentation rules could also lead to more matches. These rules depend on basic research on synthetic cannabinoids and also on research on fragmentation processes thereof which means that this research as to be done in order to improve the detectability. Another possible improvement regarding the matching process could be to include the existence of external references. If a suggested compound corresponds to a molecule in a public compound database, the

likelihood that it is the correct one increases drastically. However, in order to enable this additional parameter, external references need to be in the database for all exisiting known synthetic cannabinoids. An automated process is not possible for this, a list of external references for known molecules would be a requirement. Rapidly emerging novel substances make this challenge even harder. Another important thing to point out is that there are still new structures of synthetic cannabinoids emerging, therefore rules for the compound generation have to be kept up to date.

There are other possible directions where future work can build on the results of this thesis. For instance, the implementation could be used for other classes of novel psychactive substances. The only component that would require adjustments is the generator. Obviously, the feasibility depends on the predictability of molecular structures. A relatively simple model for the structure of synthetic cannabinoids is used in this thesis. It is questionable if the same is possible for other classes of designer drugs.

In conclusion, the implemented method can be used as a good indicator for synthetic cannabinoids in mass spectrometry data. With relatively short run times, it is especially well suited for datasets with a large amount of spectra. Further improvements have to be done in order to increase the sensitivity of the method. Besides, including new findings in research of synthetic cannabinoid structures is essential in order to cover new developments.

List of References

[Aha+22] Javed Ahamad et al. “Basic Principles and Fundamental Aspects of Mass Spectrometry”. In: Mar. 2022, pp. 3–17. ISBN: 9781003091226. DOI: 10.1201/9781003091226-2.

[Auw+21] Volker Auwärter et al. *Synthetic cannabinoids in Europe – a review*. Tech. rep. European Monitoring Centre for Drugs and Drug Addiction, 2021. DOI: 10.2810/911833.

[Bio23a] National Center for Biotechnology Information. *PubChem*. 2023. URL: https://pubchem.ncbi.nlm.nih.gov/ (visited on Jan. 12, 2023).

[Bio23b] National Center for Biotechnology Information. *PubChem - Aspirin*. 2023. URL: https://pubchem.ncbi.nlm.nih.gov/compound/2244 (visited on Jan. 11, 2023).

[Bio23c] National Center for Biotechnology Information. *PubChem - Dronabinol*. 2023. URL: https://pubchem.ncbi.nlm.nih.gov/compound/16078 (visited on Jan. 25, 2023).

[BM12] Shibdas Banerjee and Shyamalava Mazumdar. “Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte”. In: *International Journal of Analytical Chemistry* (2012). DOI: 10.1155/2012/282574.

[Cen] NIST Mass Spectrometry Data Center. “Mass Spectra”. In: *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Ed. by P.J. Linstrom and W.G. Mallard. National Institute of Standards and Technology. DOI: 10.18434/T4D303. (Visited on Dec. 22, 2022).

[Cho+04] Young Choi et al. "NMR Assignments of the Major Cannabinoids and Cannabiflavonoids Isolated from Flowers of Cannabis Sativa". In: *Phytochemical analysis : PCA* 15.6 (Nov. 2004), pp. 345–54. DOI: 10.1002/pca.787.

[Coş16] Özlem Coşkun. "Separation techniques: Chromatography". In: *North Clin Istanbul* 3.2 (2016), pp. 156–160. DOI: 10.14744/nci.2016.32757.

[DA17] European Monitoring Centre for Drugs and Drug Addiction. *Synthetic cannabinoids in Europe (Perspectives on drugs)*. 2017. URL: https://www.emcdda.europa.eu/topics/pods/synthetic-cannabinoids_en (visited on Jan. 16, 2023).

[DA21] European Monitoring Centre for Drugs and Drug Addiction. *Early Warning System on NPS*. 2021. URL: https://www.emcdda.europa.eu/publications/topic-overviews/eu-early-warning-system_en (visited on Feb. 27, 2023).

[DA22] European Monitoring Centre for Drugs and Drug Addiction. *European Drug Report 2022: Trends and Developments*. Tech. rep. Publications Office of the European Union, 2022. DOI: 10.2810/75644.

[Deucs] Eric W. Deutsch. "2012". In: *Molecular & Cellular Proteomics* 11.12 (File Formats Commonly Used in Mass Spectrometry Proteomics). DOI: 10.1074/mcp.R112.019695.

[Dol14] John W. Dolan. "How Much Retention Time Variation Is Normal?" In: *LCGC North America* 32.8 (2014). URL: https://www.chromatographyonline.com/view/how-much-retention-time-variation-normal-0.

[Dow04] Kevin Downard. *Mass Spectrometry. A Foundation Course*. The Royal Society of Chemistry, 2004. ISBN: 978-0-85404-609-6. DOI: 10.1039/9781847551306.

[Fou23a] Apache Software Foundation. *SMARTS - A Language for Describing Molecular Patterns*. 2023. URL: https://arrow.apache.org/ (visited on Feb. 9, 2023).

[Fou23b] Python Software Foundation. *sqlite3 — DB-API 2.0 interface for SQLite databases (Python Standard Library Docs)*. 2023. URL: https://docs.python.org/3/library/sqlite3.html (visited on Feb. 7, 2023).

[Fra+18] Florian Franz et al. "Synthetic cannabinoids in hair – Pragmatic approach for method updates, compound prevalences and concentration ranges in authentic hair samples". In: *Analytica Chimica Acta* 1006 (May 2018), pp. 61–73. DOI: 10.1016/j.aca.2017.12.029.

[HB16] F. Hufsky and S. Böcker. "Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data". In: *Mass Spectrometry Reviews* 36.5 (2016). DOI: 10.1002/mas.21489.

[HPP17] L. Harper, J. Powell, and E.M. Pijl. "An overview on forensic drug testing methods and their suitability for harm reduction point-of-care services". In: *Harm Reduction Journal* 14.52 (2017). DOI: 10.1186/s12954-017-0179-5.

[Jew+20] Kevin S. Jewell et al. "Comparing mass, retention time and tandem mass spectra as criteria for the automated screening of small molecules in aqueous environmental samples analyzed by liquid chromatography/quadrupole time-of-flight tandem mass spectrometry". In: *Rapid Communications in Mass Spectrometry* 34.1 (2020). DOI: 10.1002/rcm.8541.

[KA12] Stefan Kneisel and Volker Auwärter. "Analysis of 30 synthetic cannabinoids in serum by liquid chromatography-electrospray ionization tandem mass spectrometry after liquid-liquid extraction". In: *Journal of Mass Spectrometry* 47.7 (2012). DOI: 10.1002/jms.3020.

[Kin+13] Tobias Kind et al. "LipidBlast - In-Silico Tandem Mass Spectrometry Database for Lipid Identification". In: *Nature Methods* 10.8 (2013). DOI: 10.1038/nmeth.2551.

[Kop+05] David W. Koppenaal et al. "MS Detectors". In: *Analytical Chemistry* 77.21 (2005), 418 A–427 A. DOI: 10.1021/ac053495p.

[LG07] Andrew R. Leach and Valerie J. Gillet. *An Introduction to Cheminformatics*. Springer Dordrecht, 2007. ISBN: 978-1-4020-6291-9. DOI: 10.1007/978-1-4020-6291-9.

[LXB96] H.P. Lehmann, Fuentes-Arderiu X., and L.F. Bertello. "Glossary of terms in quantities and units in Clinical Chemistry (IUPAC-IFCC Recommendations 1996)". In: *Pure and Applied Chemistry* 68.4 (1996), pp. 957–1000. DOI: 10.1351/pac199668040957. URL: https://doi.org/10.1351/pac199668040957.

[NF15] Wilfried M.A. Niessen and David Falck. "Introduction to Mass Spectrometry, a Tutorial". In: *Analyzing Biomolecular Interactions by Mass Spectrometry*. John Wiley & Sons, Ltd, 2015. Chap. 1, pp. 1–54. ISBN: 9783527673391. DOI: 10.1002/9783527673391.ch1.

[OBo12] Noel M. O'Boyle. "Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI". In: *Journal of Cheminformatics* 4.22 (2012). DOI: 10.1186/1758-2946-4-22.

[PB13] Luc Patiny and Alain Borel. "ChemCalc: A Building Block for Tomorrow's Chemical Infrastructure". In: *Journal of Chemical Information and Modeling* 5.53 (2013). DOI: 10.1021/ci300563h.

[Pro23a] The Pallets Projects. *Flask*. 2023. URL: https://palletsprojects.com/p/flask/ (visited on Feb. 18, 2023).

[Pro23b] The Pallets Projects. *Jinja*. 2023. URL: https://palletsprojects.com/p/jinja/ (visited on Feb. 18, 2023).

[Pul+22] Benedikt Pulver et al. "EMCDDA framework and practical guidance for naming synthetic cannabinoids". In: *Drug Testing and Analysis* (2022). DOI: 10.1002/dta.3403.

[RDK22a] RDKit. *RDKit Docs - rdChemReactions module*. 2022. URL: https://www.rdkit.org/docs/source/rdkit.Chem.rdChemReactions.html (visited on Feb. 9, 2023).

[RDK22b] RDKit. *RDKit Documentation - rdkit.Chem.Draw*. 2022. URL: https://www.rdkit.org/docs/source/rdkit.Chem.Draw.html (visited on Feb. 22, 2023).

[RDK22c] RDKit. *RDKit: Open Source Cheminformatics Software*. 2022. URL: https://www.rdkit.org/ (visited on Jan. 12, 2023).

[Rös22] P. Rösner. *Mass Spectra of Designer Drugs 2022*. Web Page - Description of the Database. 2022. URL: https://sciencesolutions.wiley.com/solutions/technique/gc-ms/mass-spectra-of-designer-drugs/ (visited on Jan. 3, 2023).

[Sci21] Matrix Science. *Data file format*. 2021. URL: https://www.matrixscience.com/help/data_file_help.html (visited on Feb. 14, 2023).

[Ser23] Chemical Abstracts Services. *CAS Registry*. 2023. URL: https://commonchemistry.cas.org/ (visited on Jan. 12, 2023).

[SGI16] R. P Schwarzenbach, P.M. Gschwend, and D.M. Imboden. *Environmental Organic Chemistry*. 3rd ed. John Wiley & Sons, Inc., 2016. ISBN: 978-1-118-76723-8.

[SHB13] K. Scheubert, F. Hufsky, and S. Böcker. "Computational mass spectrometry for small molecules". In: *Journal of Cheminformatics* 5.12 (2013). DOI: 10.1186/1758-2946-5-12.

[SHT19] Amelia Swift, Robert Heale, and Alison Twycrow. "What are sensitivity and specificity?" In: *Evidence Based Nursing* 23.1 (2019). DOI: 10.1136/ebnurs-2019-103225.

[Ski+21] Michael A. Skinnider et al. "A Deep Generative Model Enables Automated Structure Elucidation of Novel Psychoactive Substances". In: *Nature Machine Intelligence* 3 (2021), pp. 973–984. DOI: 10.1038/s42256-021-00407-x.

[SQL22] SQLite. *Appropriate Uses for SQLite*. 2022. URL: https://www.sqlite.org/whentouse.html (visited on Feb. 6, 2023).

[SSL15] Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. "Get Your Atoms in Order—An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm". In: *Journal of Chemical Information and Modeling* 55.10 (2015), pp. 2111–2120. DOI: 10.1021/acs.jcim.5b00543.

[Sto+12] P. Stout et al. *Expansion of a Cheminformatic Database of Spectral Data for Forensic Chemists and Toxicologists*. Tech. rep. 241444. National Criminal Justice Reference Service, 2012. URL: https://www.ojp.gov/library/publications/expansion-cheminformatic-database-spectral-data-forensic-chemists-and (visited on Jan. 3, 2023).

[Str+22] Michael A. Stravs et al. "MSNovelist: de novo structure generation from mass spectra". In: *Nature Methods* 19 (2022), pp. 865–870. DOI: 10.1038/s41592-022-01486-3.

[Sys19a] Daylight Chemical Information Systems. *SMARTS - A Language for Describing Molecular Patterns*. 2019. URL: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (visited on Feb. 9, 2023).

[Sys19b] Daylight Chemical Information Systems. *SMILES - A Simplified Chemical Language*. 2019. URL: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (visited on Jan. 12, 2023).

[tea23] The Matplotlib development team. *Matplotlib: Visualization with Python*. 2023. URL: https://matplotlib.org/ (visited on Feb. 22, 2023).

[Was15] William Wassmer. *The Operating Principle and Key Applications of Mass Spectrometry*. Apr. 2015. URL: https://www.azonano.com/article.aspx?ArticleID=3999 (visited on Dec. 10, 2022).

[WB20] Michael Witting and Sebastian Böcker. "Current status of retention time prediction in metabolite identification". In: *Journal of Separation Science* 43 (2020), pp. 1746–1754. DOI: 10.1002/jssc.202000060.

[WGL22] Daniel S. Wigh, Jonathan M. Goodman, and Alexei A. Lapkin. "A review of molecular representation in the age of machine learning". In: *WIREs Computational Molecular Science* 12.5 (2022). DOI: 10.1002/wcms.1603.

[WS07] J. Throck Watson and O. David Sparkman. *Introduction to Mass Spectrometry. Instrumentation, Applications and Strategies for Data Interpretation*. 4th ed. John Wiley & Sons, Ltd, Oct. 2007. ISBN: 9780470516898. DOI: 10.1002/9780470516898.