Database supported candidate search for Metabolite identification

Christian Hildebrandt[1,2], Sebastian Wolf[2], Steffen Neumann[2*]

[1]Anhalt University of Applied Sciences, Department of Computer Science, Lohmannstr. 23, 06366 Köthen (Anhalt), Germany, http://www.hs-anhalt.com

[2]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany, http://www.ipb-halle.de

Summary

Mass spectrometry is an important analytical technology for the identification of metabolites and small compounds by their exact mass. But dozens or hundreds of different compounds may have a similar mass or even the same molecule formula. Further elucidation requires tandem mass spectrometry, which provides the masses of compound fragments, but *in silico* fragmentation programs require substantial computational resources if applied to large numbers of candidate structures.

We present and evaluate an approach to obtain candidates from a relational database which contains 28 million compounds from PubChem.

A training phase associates tandem-MS peaks with corresponding fragment structures. For the candidate search, the peaks in a query spectrum are translated to fragment structures, and the candidates are retrieved and sorted by the number of matching fragment structures. In the cross validation the evaluation of the *relative ranking positions* (RRP) using different sizes of training sets confirms that a larger coverage of training data improves the average RRP from 0.65 to 0.72. Our approach allows downstream algorithms to process candidates in order of importance.

1 Introduction

Mass spectrometry is an important analytical technology in systems biology, and allows the detection of a large number of metabolites in biological samples. For a biological interpretation, their structures and/or accession numbers are required. Individual metabolites can be identified by their accurate mass, but dozens or hundreds of different compounds may have a similar mass or even the same molecular formula (and hence identical mass).

In tandem mass spectrometers, such as hybrid instruments like a triple quadrupole (QqQ), or quadrupole coupled to a time-of-flight analyser (QqTOF), the molecules of interest are isolated in the first quadrupole. This filter allows only the molecules within a narrow *precursor mass* window to pass through, and other molecules are discarded. These filtered molecules undergo collision induced dissociation (CID) in the second quadrupole (the so called collision cell), where they literally break apart. The masses (more correctly, the mass-over charge ratio m/z) of the resulting fragments are measured in the final mass analyser, either another quadrupole, or

*To whom correspondence should be addressed. Email: Steffen.Neumann@ipb-halle.de

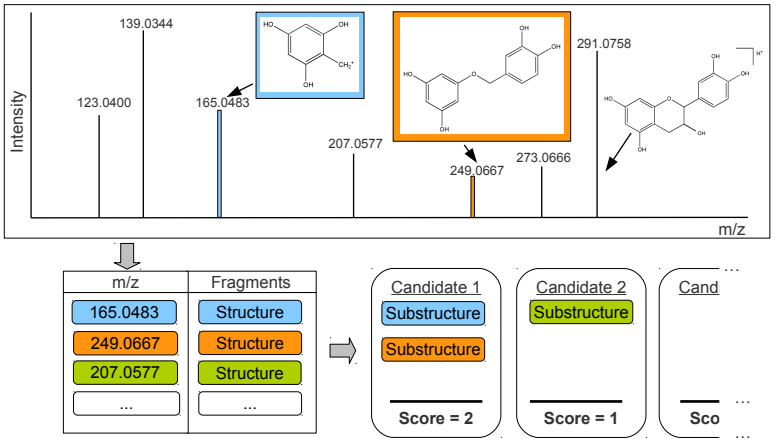

Figure 1: Top: tandem MS spectrum of Epicatechin, with some manually annotated fragment peaks. Below: database table with m/z → fragment association. Candidate compounds are scored, based on the number of substructures they contain.

a high-resolution time-of-flight (TOF) analyser. Other instruments such as Iontraps or Orbitrap perform these steps sequentially in time, rather than in different instrument compartments. A typical result of the fragmentation is shown in the tandem mass spectrum of Epicatechin in Figure 1.

For metabolite identification, a query spectrum can be compared with reference spectra from databases like MassBank [6] or commercial libraries provided by several vendors [7]. However, their chemical coverage is far from complete, especially in areas such as plant metabolomics, where most of the estimated 200 000 compounds are still uncharacterised [1].

If reference spectra are not available, the spectra can be interpreted using computational mass spectrometry methods, such as FiD [3], or the commercial ACD Fragmenter and HighChem's MassFrontier – see [7] for a review. These programs can also be used to search general purpose compound libraries, such as KEGG with about 14 215 metabolite structures or the much larger PubChem database with 28 million compounds [5, 10].

The MetFrag approach is designed to search online accessible compound databases with the accurate mass of the unfragmented metabolite. MetFrag obtains *candidates* from the compound databases, fragments these *candidates in-silico*, and scores the match between the query spectrum and the *in silico* fragments.

However, analysing thousands of candidate structures is a time-consuming process, especially for non-trivial compounds, and may take hours on a single machine. For example it takes ≈3 minutes to process about 1 672 candidates of strychnine N-oxide [10]. But for some spectra there are even more candidates. All hypotheses are processed in the (arbitrary) order determined by the candidate search. The correct compound might appear first, or towards the end of the list. If the candidate search would already pre-sort the corresponding candidates, it would be possible to process and display the correct one earlier.

The MetFrag web application will implement a dynamically updated user interface, and process all candidates in smaller batches. That way it is possible to present informative (but still preliminary) results almost from the beginning. The final result in MetFrag after completion of all candidates remains the same. Alternatively, the set of candidates can be filtered based on

the preliminary scores, and the subsequent MetFrag runtime would be reduced.

In this paper, we present the MassStruct approach to learn the association between the measured mass peaks and fragment structures, which allows to integrate the accurate molecule mass search with the score-based ordering. The next section describes the system architecture, the training phase and the candidate retrieval with dynamically generated SQL queries during operation. In section 3 we evaluate our approach on a dataset of 240 spectra from 218 unique compounds, and assess the runtime of the dynamically generated query.

2 Implementation

The MassStruct approach requires an offline preprocessing step to associate measured peak masses to the corresponding fragment structures in a set of training spectra. Afterwards these fragments are grouped by their mass. During a candidate search, the molecule mass and the peaks of a query spectrum are both combined into a single dynamically generated SQL query. If one or more fragment structures of a given mass exist within one candidate, one match is counted and added to the score of this candidate.

2.1 Learning the association between mass and fragment structure

The training spectra are processed with the MetFrag algorithm, to obtain a set of m/z $\rightarrow$ structure associations as shown in Figure 2. The training set of tandem MS spectra is synthetically fragmented with MetFrag and all annotated fragments are stored with their corresponding mass into a relational database.

MetFrag usually is not able to annotate every measured peak with a structure, and it is possible that one observed m/z value can be explained by different structures in different compounds, therefore all alternatives are stored.

We developed a batch import to store the fragments and their masses into a PostgreSQL 9.0 RDBMS[1] with the chemistry extension pgchem[2] 1.3-GiST [8]. All of the chemical algorithms and datatypes are handled by functions in the chemistry library OpenBabel[3] 2.3.0 [2]. The RDBMS integration allows chemical calculations, comparisons and predicates as a part of SQL statements. The ER diagram of the developed database is shown in Appendix B.

2.2 Multiple substructure database queries

The candidate retrieval query (see abbreviated query in Figure 3) selects all compounds within an error margin around the precursor mass. Then, any fragments matching the measured peak masses (within an error window) are joined with the condition `fragment.structure <= compound.structure`. This is provided by the OpenBabel chemistry algorithms and tests whether the fragment is a substructure of the candidate. The sum of the matched substructures is used as `score` for the `accession`.

[1] `http://www.postgresql.org`
[2] `http://pgfoundry.org/projects/pgchem`
[3] `http://www.openbabel.org/`

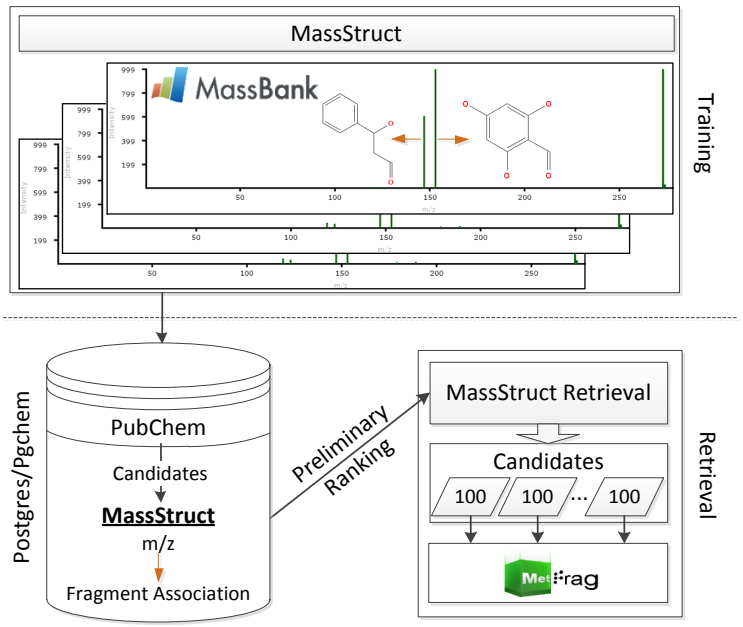

Figure 2: The training step (top) shows the MassBank spectra and the fragments predicted with MetFrag. The peaks annotated with a fragment structure are stored in the MassStruct database (lower left). In the operation phase the stored m/z → structure associations are used to retrieve the ordered candidates in batches of e.g. 100 structures.

To determine whether a fragment is a substructure of a molecule, their chemical fingerprints are compared. These 1536 bit fingerprints store characteristic chemical properties (such as bond- and atom counts or functional groups). For substructure searches, pgchem compares these fingerprints between the query molecule and the database content (primary filtering), using a Generalized Search Tree index (GiST) [4]. Afterwards, a time consuming substructure matching (secondary filtering) of the molecular structures on the previously selected records is done. All chemical operations benefit from the PostgreSQL query planning optimization.

The unabbreviated query in Appendix A also takes into account that 1) the database contains compounds from multiple compound libraries, and candidates can be restricted to a certain library 2) the compounds in the strcuture libraries might occur in multiple stereo conformations. Since mass spectrometry can hardly distinguish stereo isomers, MetFrag ignores the stereochemistry. Redundant candidates are removed by the query, such that only the first compound is considered. Because the fragments are measured with a certain error, the fragment masses are grouped into m/z cluster by hierarchical clustering analysis (HCA). The actual score counts at most one matching substructure per m/z cluster.

3 Results and Discussion

In the following we are going to present an example, and assess two separate performance aspects of the system. We evaluate the ability of the scoring to obtain the correct compound with a good rank, simulating various training set sizes. Second, we report the runtime on a snapshot (Q4 2010) of the PubChem compound database.

```
SELECT accession, count(fragment.id) AS score
FROM compound, fragment
WHERE compound.mass BETWEEN 290.2 AND 290.3
AND ( fragment.mass BETWEEN 123.0 AND 123.1
   OR fragment.mass BETWEEN 139.0 AND 139.1
   OR fragment.mass BETWEEN 165.0 AND 165.1
   OR fragment.mass BETWEEN 207.0 AND 207.1
   OR fragment.mass BETWEEN 249.0 AND 249.1
   OR fragment.mass BETWEEN 273.0 AND 273.1)
AND fragment.structure <= compound.structure
GROUP BY accession
ORDER BY score;
```

Figure 3: A SQL statement performing a combined search for the molecule's `mass`, and `fragments` which are a substructure (the `<=` predicate) ranked by `score`, where `<=` is the `chemical_substructure` operator. The peak data corresponds to the example spectrum in Figure 1.

3.1 Metabolite identification results

We used 240 metabolite spectra (see Appendix C or supplementary files as xls or csv hosted on `http://msbi.ipb-halle.de/msbi/massstruct`) with known PubChem accessions obtained from MassBank. These spectra contain data of several compounds, some of them were measured repeatedly with different instrument settings, so they covered 218 different compounds. Together, all spectra contained 2 083 peaks, and MetFrag was able to annotate 1 280 fragments with the parameters reported earlier [10]. The PubChem compound snapshot (Q4 2010) contained 28 838 421 structures. Including the indices, the database occupied ≈150 GB storage space.

To evaluate our approach, we annotated a randomly drawn sample of the 240 spectra, and used the remaining spectra as query spectra. For each query spectrum, we count the total number of candidates (TC), those with a better and those with a score worse than the correct compound (BC and WC, respectively). This allows to calculate a *relative ranking position* $RRP = 0.5\left(1 - \frac{BC-WC}{TC-1}\right)$, where the first position results in $RRP = 1$, and $RRP = 0$ in the worst case. A similar RRP was introduced in [9], where the authors used $RRP = 0$ for the best case. We modified the scoring to keep it consistent with the MetFrag scoring.

If all candidates have the same score, then $BC = WC = 0$, and hence $RRP = 0.5$. Similarly, a random score would also lead to an average $RRP = 0.5$ on a larger test set.

For evaluation we partitioned the set of spectra, again storing one subset of m/z → structure associations in the database, and used the remaining ones to evaluate the rank of the correct solution in the ordered result set. We used different ratios (1:1, 2:1, 3:1, 4:1 and 9:1) for partitioning, to simulate an increasing coverage of the training spectra in the dataset. The results are shown in Table 2. The average RRP increases from 0.65 to 0.72, and even more apparent the median RRP raises to 0.84 if the large training sets are used. An extract of an example for one evaluation run is summarized in Table 1.

Table 1: The best and worst examples from one of the evaluation runs.

CID	formula	mass	TC	RRP	runtime in s
13804	$C_{15}H_{16}O_9$	340.07	50 910	0.999	476
834	$C_7H_{14}N_2O_4S$	222.06	34 491	0.999	282
5319853	$C_{21}H_{22}O_{11}$	450.11	72 127	0.999	867
165627	$C_6H_{11}NO_4$	161.06	8 743	0.999	99
439155	$C_{14}H_{20}N_6O_5S$	384.12	105 691	0.998	949
442456	$C_{28}H_{34}O_{14}$	594.19	9 745	0.997	200
160556	$C_{11}H_{20}N_2O_6$	276.13	63 768	0.997	497
5318759	$C_{21}H_{18}O_{12}$	462.07	54 545	0.996	686
2901	$C_6H_{11}NO_2$	129.07	4 974	0.995	25
101781	$C_{21}H_{22}O_{11}$	450.11	72 127	0.995	1 147
5281673	$C_{21}H_{20}O_{12}$	464.09	59 828	0.995	469
5316673	$C_{21}H_{20}O_{10}$	432.10	78 553	0.995	638
...	...	...	...	...	...
637540	$C_9H_8O_3$	164.04	13 098	0.296	191
649	$C_4H_6N_2O_2$	114.04	4 175	0.261	41
70346	$C_7H_8N_4O_3$	196.05	22 378	0.250	208

Table 2: RRP of different partition sizes.

	RRP		
Partition	median	ø	Std. Err.
1:1	0.50	0.65	$\pm$ 0.018
2:1	0.70	0.71	$\pm$ 0.019
3:1	0.69	0.70	$\pm$ 0.019
4:1	0.75	0.71	$\pm$ 0.019
9:1	0.84	0.72	$\pm$ 0.019

3.2 Runtime and PostgreSQL database tuning

The query spectra result in 31 700 candidates on average (green circle in Figure 4), 16 630 in the median and in a few cases up to 100 000. The mean runtime of a query is 330s, or roughly 10ms per candidate.

The (virtual) database server had 2 CPUs, 2 GB RAM, and was hosted on a VMWare ESX cluster with 2.6 GHz Intel Xeon CPUs. The data partition was kept on a FC-SAN storage system.

The runtime clearly depends on the number of candidates. Therefore, any increase in e.g. instrument accuracy will decrease both the number of candidates and the runtime. The performance of an RDBMS often depends on the speed of the storage subsystem, but not in this case: the majority of time is spent in the actual sub-structure search, and the CPU speed is the limiting factor. Latencies for multiple concurrent queries can best be reduced using a server with a sufficient number of CPU cores.

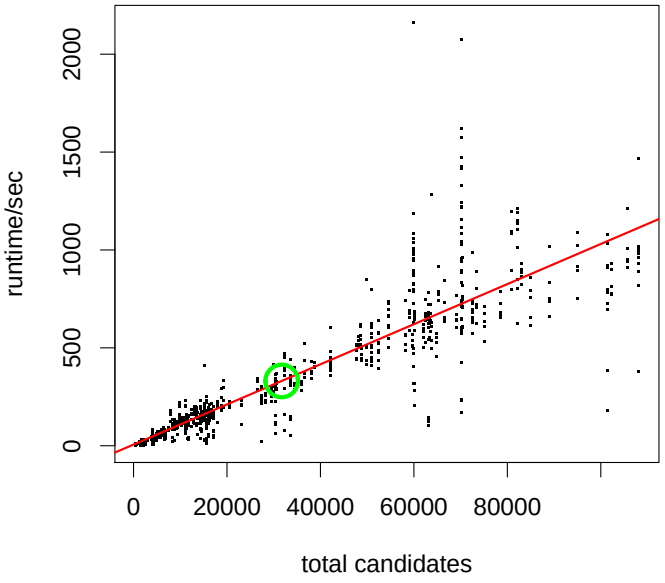

Figure 4: Runtimes of all candidate queries in the data set. The slope of the regression line is 10ms/candidate, the average (32 000 candidates in ≈ 5min) is encircled.

4 Conclusion

The process of structure elucidation with mass spectrometry data has been – and still is – the major bottleneck in metabolomics experiments. Starting from the mass of the molecule, dozens to thousands of candidates can be retrieved from compound databases like KEGG or PubChem, and subsequently analysed with computer aided structure elucidation (CASE) systems.

We introduced the MassStruct approach, improving the initial candidate query step to provide an *ordered* list of candidates. We evaluated the method with a medium sized test dataset. The benefit is that the interactive MetFrag web application can then process the candidates in batches of 100 or 1000 structures, and present intermediate results. Since the candidates are pre-sorted, the user might be satisfied after the first few iterations. The source code (including training procedure and dynamic queries) is available under the GNU General Public License from `https://github.com/childebr/MassStruct/`.

Future developments will be reducing the number of candidates to consider, e.g. by filtering the common ranges of ratios between elements for biological compounds. The growing number of spectral data in reference libraries such as MassBank will further improve the performance of the system by adding more m/z → structure associations.

References

[1] Oliver Fiehn. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2(3):155–168, 2001.

[2] Rajarshi Guha, Michael T Howard, Geoffrey R Hutchison, Peter Murray-Rust, Henry Rzepa, Christoph Steinbeck, Jörg Wegner, and Egon L Willighagen. The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model*, 46(3):991–998, 2006.

[3] Markus Heinonen, Ari Rantanen, Taneli Mielikäinen, Juha Kokkonen, Jari Kiuru, Raimo A. Ketola, and Juho Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, September 2008.

[4] Joseph M. Hellerstein, Jeffrey F. Naughton, and Avi Pfeffer. Generalized Search Trees for Database Systems. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, pages 562–573. Morgan Kaufmann, 1995.

[5] Dennis W Hill, Tzipporah M Kertesz, Dan Fontaine, Robert Friedman, and David F Grant. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem*, 80(14):5574–5582, Jul 2008.

[6] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, Jul 2010.

[7] Steffen Neumann and Sebastian Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem*, 398(7-8):2779–2788, Dec 2010.

[8] Ernst-Georg Schmid. *Database-driven procurement of substances in the researching chemical industry - An algorithmic optimization approach*. PhD thesis, Mercator School of Management - Fakultät für Betriebswirtschaftslehre - Technology and Operations Management - Wirtschaftsinformatik und Operations Research, June 2010.

[9] Emma L Schymanski, Markus Meringer, and Werner Brack. Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal Chem*, 81(9):3608–3617, May 2009.

[10] Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010.

A Unabbreviated SQL query

```
SELECT substance.accession, Score
FROM substance, compound, library,

    (SELECT inchi_key_1,
     COUNT(DISTINCT PeakFragments.mz_cluster_id) AS Score
     FROM substance, library, compound AS Candidates

        (SELECT MIN(compound_id)
         FROM compound
         WHERE exact_mass BETWEEN 290.24 AND 290.28
         GROUP BY inchi_key_1) AS FirstCandidates
         LEFT OUTER JOIN
        (SELECT fragments.structure, mz_cluster_id
          FROM fragments, mz_cluster
          WHERE fragments.mz_cluster_id = mz_cluster.id
          AND   ((mz_cluster.mass between 123.035 AND 123.045)
              OR (mz_cluster.mass between 139.030 AND 139.040)
              OR (mz_cluster.mass between 165.040 AND 165.050)
              OR (mz_cluster.mass between 207.050 AND 207.060)
              OR (mz_cluster.mass between 249.060 AND 249.070)
              OR (mz_cluster.mass between 273.060 AND 273.070)))
          AS PeakFragments
         ON (PeakFragments.structure <= Candidates.mol_structure)

     WHERE substance.compound_id = FirstCandidates.compound_id
     AND  substance.library_id = library.library_id
     AND  library_name = 'pubchem'
     AND  Candidates.compound_id = FirstCandidates.compound_id
     GROUP BY accession, inchi_key_1
     ORDER BY Score DESC) AS Results

WHERE exact_mass BETWEEN 290.24 AND 290.28
AND  substance.compound_id = compound.compound_id
AND  substance.library_id = library.library_id
AND  library_name = 'pubchem'
AND  compound.inchi_key_1 = results.inchi_key_1;
```

The full SQL statement performing a combined search for the intact `exact_mass`, and `fragments` which are a substructure (the <= predicate) ranked by `score`. where <= is the `chemical_substructure` predicate

B ER diagram of developed database

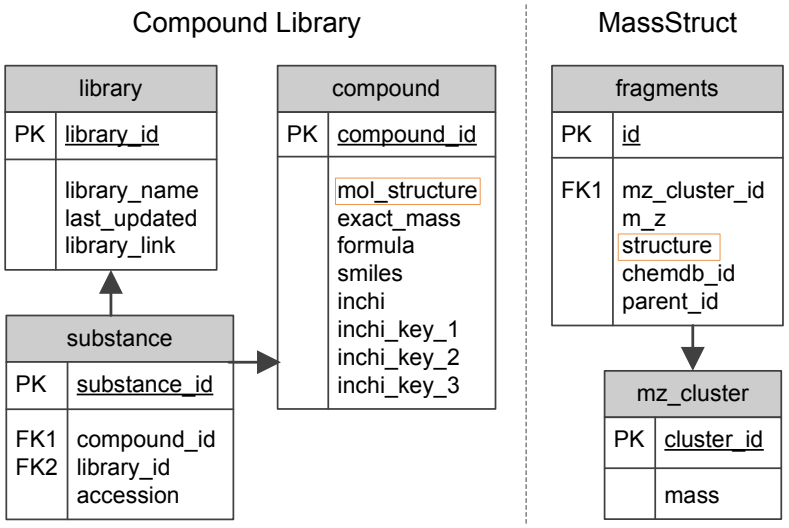

Figure 5: The tables `substance` and `library` contain the information from PubChem and other libraries, whereas `compound` contains the actual molecular structures. The `fragments` and `mz_cluster` tables contain the structure → mass associations which were created during the training step of the algorithm.

C Complete results for 1:1 evaluation

The following table shows the (hyperlinked) MassBank and PubChem accession numbers for all spectra in the training set. Concatenated IDs such as PR100001PR100002 denote merged spectra, as described in [10]. The set was sampled into two equally sized subsets, The MassStruct training was perfomed on one subset, and evaluated against the second. Then, training was repeated with the second subset, and evaluated on the first. The runtime column covers only the time of the query in the evaluation.

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100121PR100122	13804	1,000	0	50902	50910	477
PR100113PR100114	834	1,000	0	34475	34491	283
PR100317	13804	0,999	0	50852	50910	448
PR100239	5319853	0,999	0	72031	72127	867
PR100390	165627	0,999	5	8735	8743	100
PR100329	717531	0,999	0	29768	29828	252
PR100198	439155	0,999	48	105480	105691	950
PR101031	5274585	0,999	16	46540	46649	526
PR100359	6441269	0,999	35	71951	72127	793
PR100296	92136	0,998	7	8717	8743	84
PR100076	34755	0,998	12	107657	108138	932
PR100363	442456	0,998	12	9711	9745	200
PR100277	160556	0,997	98	63536	63768	497
PR100249	92794	0,997	35	80451	80855	898
PR100447	5320863	0,997	106	70812	71109	854
PR100395	65065	0,997	17	11783	11839	113
PR100256	5280459	0,997	118	69911	70272	587
PR101033	5280459	0,997	118	69911	70272	893
PR101047	5280459	0,997	118	69911	70272	423
PR100386	1029	0,996	19	11782	11855	140
PR101022	5318759	0,996	153	54264	54545	687
PR100001PR100002	2901	0,996	3	4936	4974	26
PR100351	101781	0,996	209	71749	72127	1147
PR100248	5281673	0,996	81	59419	59828	879
PR101027	5281673	0,995	196	59464	59828	469
PR101007	5316673	0,995	285	78101	78553	639
PR101024	5316673	0,995	313	78082	78553	1020
PR100243	5282102	0,995	256	69786	70272	1327
PR100253	5281643	0,994	208	59338	59828	405
PR101012	5281643	0,994	232	59356	59828	2163
PR101021	5481882	0,993	178	89873	90914	1483
PR100240	5318645	0,993	232	55942	56492	738
PR101025	5282102	0,993	351	69641	70272	832
PR100335	65127	0,993	161	36266	36630	350
PR100254	5280804	0,992	327	59166	59828	318

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100469	25674	0,991	39	8969	9090	97
PR100475	5281417	0,990	220	80697	82100	803
PR100175	6288	0,990	11	2027	2059	25
PR100315	99289	0,990	11	2027	2059	16
PR100366	5321576	0,990	120	26692	27134	425
PR100334	439574	0,988	5	2970	3036	28
PR100449	441031	0,988	5	2970	3036	25
PR100326	123938	0,988	427	41563	42144	400
PR100367	5321577	0,988	265	26740	27134	232
PR100436PR100437	439227	0,987	23	4863	4974	47
PR101030	5481224	0,982	351	68928	71109	436
PR100314	440018	0,982	1203	92875	95106	751
PR100221	137	0,981	10	4160	4316	6
PR100137	14982	0,981	0	1245	1296	17
PR100258	14982	0,981	0	1245	1296	22
PR100166	439579	0,980	512	47480	48887	443
PR100035PR100036	6057	0,980	0	13183	13733	55
PR100157PR100158	6322	0,979	165	14779	15256	184
PR100303	6322	0,979	165	14779	15256	160
PR100368	5320686	0,979	0	7828	8174	47
PR100163PR100164	5961	0,978	134	8505	8758	86
PR101023	5320686	0,978	14	7822	8174	119
PR100252	5484066	0,977	61	7750	8057	131
PR100322	5962	0,976	143	8158	8417	85
PR100153PR100154	439277	0,975	0	309	326	6
PR100420	70914	0,975	329	15535	15995	45
PR100093	637775	0,975	468	31226	32403	294
PR100349	5883291	0,975	84	8449	8815	183
PR100241	5481663	0,975	10	5710	6007	90
PR100244	5318767	0,974	21	8383	8815	173
PR100354	10621	0,972	12	7509	7950	125
PR100304PR100305	439232	0,971	280	14931	15546	86
PR100267PR100268	6950385	0,971	265	14002	14585	135
PR100162	33032	0,971	161	5843	6035	58
PR101009	5323562	0,970	304	10838	11203	89
PR100199	193653	0,970	35	10482	11125	109
PR101034	5323562	0,967	326	10792	11203	183
PR101046	5323562	0,967	326	10792	11203	57
PR100325	2724705	0,960	0	11343	12319	144
PR100448	9750	0,958	305	11068	11761	41
PR100456	5320835	0,956	187	4896	5160	86
PR100280	2761525	0,952	177	5843	6269	59
PR100306	439389	0,951	307	8199	8743	27
PR100320	10917	0,944	234	7721	8426	40
PR100299	439406	0,942	1486	27520	29428	225

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100290	2761558	0,941	128	3785	4147	35
PR100260	107982	0,938	119	34592	39387	333
PR100259	1548943	0,936	71	41713	47808	444
PR100263	182232	0,917	1655	54277	63106	544
PR100338	16211048	0,915	350	5552	6269	57
PR100161	88513	0,915	2501	52074	59736	568
PR101055	11953815	0,909	4901	72002	82100	1212
PR100211	23724461	0,906	215	3515	4068	65
PR100013	5280567	0,899	555	13063	15656	220
PR100220	119	0,897	77	1210	1427	13
PR100067PR100068	439217	0,892	2371	32269	38088	377
PR101041	8655	0,891	453	14331	17731	245
PR100272	99478	0,890	474	5176	6035	58
PR100242	5281693	0,888	83	1419	1724	24
PR100291	5706676	0,887	313	3132	3643	30
PR100286	1502076	0,883	71	1648	2059	20
PR100212	73323	0,871	6532	55016	65319	776
PR100336	152306	0,871	3525	34738	42078	360
PR100279	5780	0,856	2239	15931	19254	197
PR100169PR100170	21236	0,841	531	3355	4147	43
PR100282	5706673	0,830	313	6314	9090	96
PR100006	24405	0,768	14812	52467	70324	761
PR100324	24405	0,768	14812	52467	70324	608
PR100380	1662	0,758	1124	7471	12319	96
PR100365	5320844	0,756	1585	32262	59828	642
PR100441	5281576	0,741	15049	48426	69191	604
PR100222	564	0,710	847	2592	4147	21
PR100215	439656	0,682	619	1367	2059	21
PR100121PR100122	138	0,677	447	1357	2573	21
PR100332PR100333	7971	0,500	0	0	682	6
PR100048	8871	0,500	0	0	682	8
PR100094PR100095	439335	0,500	0	0	2155	24
PR100143	439225	0,500	0	0	4187	44
PR100134PR100135	65359	0,500	0	0	6324	74
PR100357	5490298	0,500	0	0	7169	34
PR100356	5317025	0,500	0	0	9292	187
PR100371	5282151	0,500	0	0	11203	94
PR100069PR100070	5280951	0,500	0	0	28152	220
PR100229	5281666	0,500	0	0	62470	526
PR100139PR100140	6802	0,500	0	0	64733	528
PR100246	114776	0,500	0	0	70272	652
PR100370	5280441	0,500	0	0	78553	682
PR100251	5281807	0,500	0	0	101329	780
PR100378	34755	0,500	0	0	108138	379
PR100399	493570	0,500	3	0	102257	772

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100127	7427	0,500	4	0	89036	735
PR100056PR100057	1052	0,500	1	0	16077	145
PR100100PR100101	6076	0,500	4	0	61985	555
PR100440	5282054	0,500	3	0	30008	268
PR100014PR100015	5202	0,500	2	0	15601	12
PR100362	442813	0,500	13	0	84869	616
PR100353	5281621	0,500	9	0	49954	847
PR100360	6450184	0,500	13	0	70272	331
PR100247	5280637	0,499	71	0	70272	171
PR100403	4644	0,499	12	0	11018	128
PR100227	5281654	0,499	93	0	60225	658
PR100023	227	0,499	7	0	4402	41
PR100312	5280378	0,499	86	0	52402	436
PR100072PR100073	439224	0,499	57	0	33551	339
PR100096PR100097	89	0,499	64	0	33591	148
PR100192	980	0,499	10	0	5213	58
PR100107PR100108	6115	0,499	1	0	520	6
PR100274	1051	0,498	84	0	27306	220
PR100323	449093	0,498	99	0	30440	106
PR100029	6106	0,498	16	0	4147	19
PR100275	1050	0,498	42	0	10541	114
PR100418PR100419	6723	0,498	52	0	12789	121
PR100043PR100044	24154	0,498	58	0	14257	81
PR100125PR100126	439213	0,498	58	0	14257	36
PR100091PR100092	6047	0,498	74	0	17006	151
PR101049	5748601	0,497	409	0	70272	251
PR100264	72276	0,497	431	0	63106	101
PR100262	9064	0,497	437	0	63106	602
PR100210	5610	0,496	32	0	4276	48
PR100219	14180	0,496	537	0	63501	675
PR100004PR100005	445858	0,496	197	0	23009	106
PR100089PR100090	91531	0,496	546	0	63336	653
PR100377	439498	0,495	167	0	16705	185
PR100385	64969	0,495	126	0	11476	35
PR100273	65059	0,494	808	0	73300	622
PR100165	165271	0,494	75	0	6621	63
PR100309	65110	0,494	848	0	72521	575
PR100054	10256	0,493	93	0	6578	85
PR100358	5319116	0,492	1151	0	70272	724
PR100423PR100424	6228	0,491	6	0	345	3
PR100177	4032	0,491	129	0	7376	86
PR100225	5281708	0,491	774	0	42047	413
PR100417	135	0,491	127	0	6810	55
PR100406	3845	0,491	275	0	14626	139
PR100010PR100011	1318	0,489	415	0	18859	198

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100330	5324677	0,489	671	0	29828	27
PR100042	637760	0,487	792	0	30570	367
PR101045	69867	0,486	243	0	8783	98
PR100339	2761537	0,485	415	0	14102	143
PR100289	2761554	0,485	557	0	18585	173
PR100201PR100202	5430	0,484	515	0	16548	156
PR100319	6804	0,484	2388	0	72634	740
PR100186	967	0,483	353	0	10303	132
PR100115PR100116	6175	0,483	1244	0	35937	282
PR100266	11250133	0,478	472	0	10608	115
PR100193	378	0,478	607	0	13515	144
PR100297PR100298	72924	0,477	235	0	5213	44
PR100217	67701	0,477	200	0	4311	45
PR100292	1051	0,477	1271	0	27306	217
PR100039	5280567	0,476	740	0	15656	37
PR100003	5280536	0,476	870	0	17982	214
PR100261	637542	0,474	677	0	13098	169
PR100059	637511	0,474	298	0	5711	45
PR100110	689043	0,473	930	0	17204	76
PR100271	736715	0,472	424	0	7478	80
PR100209	6440982	0,471	4788	0	83130	885
PR100147	800	0,471	541	0	9312	111
PR100307	199	0,470	363	0	5960	50
PR100318	6131	0,469	3661	0	58125	565
PR100288	2761550	0,468	2084	0	32750	309
PR100383	3469	0,467	645	0	9895	118
PR100077	89594	0,466	767	0	11426	129
PR100364	5280781	0,464	5447	0	75207	535
PR100049PR100050	351795	0,463	987	0	13296	192
PR100473	5280569	0,461	1165	0	15122	129
PR101044	398554	0,461	941	0	12036	159
PR101042	637775	0,458	2723	0	32403	459
PR101043	10256	0,454	603	0	6578	30
PR100472	5280460	0,453	1936	0	20491	196
PR100384	1145	0,453	36	0	380	4
PR100474	5281416	0,452	1451	0	15122	141
PR100233	5280343	0,452	5051	0	52403	530
PR100392	112072	0,450	3867	0	38762	406
PR100213	6119	0,450	143	0	1427	15
PR100197	40539	0,449	1502	0	14607	149
PR100234	5281691	0,448	6319	0	60225	318
PR100337	7618	0,444	620	0	5586	71
PR100409PR100410	5570	0,443	502	0	4402	40
PR100016PR100017	65040	0,432	519	0	3805	33
PR100188	6604563	0,432	3620	0	26509	285

MassBank ID	PubChem ID	RRP	BC	WC	TC	Runtime in s
PR100295	6613	0,431	4228	0	30449	224
PR100400	6613	0,430	4248	0	30449	251
PR100152	8582	0,426	9918	0	66605	636
PR100257	107971	0,414	17442	0	101329	178
PR100470	444795	0,414	8302	0	48047	535
PR100228	5280863	0,405	10376	0	54468	500
PR100398	6441567	0,404	1862	0	9738	112
PR100027PR100028	6274	0,399	1910	0	9452	102
PR100226	440735	0,397	12217	0	59060	587
PR100391	439224	0,381	7995	0	33551	337
PR100379	4396761	0,372	2937	0	11476	149
PR100321	6274	0,368	2502	0	9452	114
PR100223PR100224	5280443	0,353	14943	0	50991	572
PR100373	3611	0,350	4016	0	13391	181
PR100308	916	0,349	4893	0	16208	128
PR100415	70639	0,344	4491	0	14432	162
PR100425	6433206	0,336	3199	0	9765	39
PR100230PR100231	5280445	0,313	20331	0	54468	615
PR100408	763	0,305	1027	0	2641	23
PR100119PR100120	6132	0,302	19228	0	48529	561
PR100414	4687	0,301	7494	0	18873	186
PR100185	637540	0,297	5319	0	13098	191
PR101040	6433206	0,286	4180	0	9765	57
PR100394	649	0,261	1992	0	4175	42
PR100413	70346	0,250	11168	0	22378	209