Title

LibGen: Generating High Quality Spectral Libraries of Natural Products for EAD–, UVPD–, and HCD-High Resolution Mass Spectrometers

Journal

Analytical Chemistry, 95(46)

ISSN

0003-2700

Authors

Kong, Fanzhou

Keshet, Uri

Shen, Tong

et al.

Publication Date

2023-11-21

DOI

10.1021/acs.analchem.3c02263

Peer reviewed

HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2024 November 21.

Published in final edited form as:

Anal Chem. 2023 November 21; 95(46): 16810–16818. doi:10.1021/acs.analchem.3c02263.

LibGen: Generating High Quality Spectral Libraries of Natural Products for EAD-, UVPD-, and HCD-High Resolution Mass Spectrometers

Fanzhou Kong,
Chemistry Department, One Shields Avenue and West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States

Uri Keshet,
West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States

Tong Shen,
West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States

Elys Rodriguez,
Chemistry Department, One Shields Avenue and West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States

Oliver Fiehn
West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States

Abstract

Compound annotation using spectral-matching algorithms is vital for (MS/MS)-based metabolomics research, but is hindered by the lack of high-quality reference MS/MS library spectra. Finding and removing errors from libraries, including noise ions, is mostly done manually. This process is both error-prone and time-consuming. To address these challenges, we have developed an automated library curation pipeline, LibGen, to universally build novel spectral libraries. This pipeline corrects mass errors, denoises spectra by subformula assignments, and performs quality control of the reference spectra by calculating explained intensity and spectral entropy. We employed LibGen to generate three high-quality libraries with chemical standards of 2241 natural products. To this end, we used an IQ-X orbital ion trap mass spectrometer to generate 1947 classic high-energy collision dissociation spectra (HCD) as well as 1093 ultraviolet-

Corresponding Author: **Oliver Fiehn –** West Coast Metabolomics Center, University of California–Davis, Davis, California 95616, United States; ofiehn@ucdavis.edu; Fax: +1-530754-9658.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c02263.

Additional methods and materials including program implantations, representative good quality spectra, library coverage, and statistics (PDF)

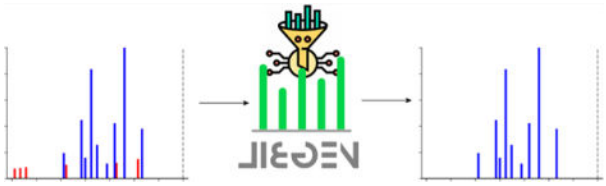
Complete annotation results (XLSX)

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c02263

The authors declare no competing financial interest.

photodissociation (UVPD) mass spectra. The third library was generated by an electron-activated collision dissociation (EAD) 7600 ZenoTOF mass spectrometer yielding 3244 MS/MS spectra. The natural compounds covered 140 chemical classes from prenol lipids to benzypyrans with >97% of the compounds showing <0.2 Tanimoto-similarity, demonstrating a very high structural variance. Mass spectra showed much higher information content for both UVPD- and EAD-mass spectra compared to classic HCD spectra when using spectral entropy calculations. We validated the denoising algorithm by acquiring MS/MS spectra at high concentration and at 13-fold diluted chemical standards. At low concentrations, a higher proportion of spectra showed apparent fragment ions that could not be explained by subformula losses of the parent molecule. When more than 10% of the total intensity of MS/MS fragments was regarded as noise ions, spectra were considered as low quality and were not included in the libraries. As the overall process is fully automated, LibGen can be utilized by all researchers who create or curate mass spectral libraries. The libraries we created here are publicly available at MassBank.us.

Graphical Abstract



INTRODUCTION

Liquid chromatography coupled with high-resolution tandem mass spectrometry (LC-MS/MS) is the leading tool in metabolomic studies for measuring small organic molecules in biological systems.[1,2] However, the metabolome is constituted of extremely diverse chemicals, unlike genes or proteins.[3,4] Mass spectral-based annotation of metabolites and exposome chemicals rely on collections of reference spectra acquired on a variety of LC-MS/MS systems and conditions.[5,6] The larger and the more diverse such spectral libraries are, the better the chances to identify unknowns in metabolomic screens. While current public and licensed libraries consist of millions of mass spectra, the total number of molecules is less than 2% of PubChem's 113 million compounds.[7–10] Moreover, many molecules give scarce mass spectra with few fragment ions, leading to high false discovery rates to disambiguate isomers, for example, for lipid double-bond positions. Here, we used new mass spectral fragmentation mechanisms, ultraviolet photodissociation (UVPD) and electron activated dissociation (EAD), to yield more informative and more exhaustive MS/MS spectra.[11,12]

Creating high-quality mass spectral libraries is not straightforward. Spectra from small molecules must account for a variety of in-addition to in-source fragmentation, isobaric interferences, satellite ions due to Fourier transformation processes,[13] electronic noise,[14] or unstable instrument conditions.[15] Typically, users create libraries by discarding low-intensity fragment ions and manually curating data to ensure high-quality mass spectra.[16] For example, the NIST MS interpreter tool is used for manual curation of NIST libraries,[17,18] but cannot be used in batch mode without human interaction. The batch-mode data curation

tool Curatr[16] does not perform any quality control. SIRIUS and MS-FINDER show decent performance on annotating fragment peaks with neutral losses, but have poor tolerance for recognizing a large number of radical ion losses in mass spectra.[19] While Rmassbank[14] employs spectra cleaning steps, it requires one compound per file, which disables the use of multicompound LC-MS/MS runs. We here present LibGen, an automatic tool to curate and quality-control mass spectra files from analyses of mixtures of chemical compounds, and apply this tool to generate libraries of new MS/MS fragmentation mechanisms using an improved denoising algorithm based on subformula assignments.

EXPERIMENTAL SECTION

Acquiring Tandem Mass Spectral Data for MS/MS Libraries.

For all libraries, authentic standards were prepared in mixtures of 25 nonisomeric compounds each at 1 mg/L in methanol and stored at −20 °C. (1) EAD spectra were acquired by using an Exion LC AD liquid chromatograph system (SCIEX, ON, Canada) coupled with a Sciex ZenoTOF 7600 quadrupole-time-of-flight (QTOF) mass spectrometer (Sciex, ON, Canada). Chromatographic separation was performed with a reversed-phase, 30 mm length × 2.1 mm inner diameter (i.d.) Kinetex pentafluorophenyl-based (PFP) column with 1.7 μm particle size (Phenomenex Inc., Torrance, CA), equipped with a 2.1 mm i.d. SecurityGuard ULTRA Cartridges precolumn (Phenomenex Inc., Torrance, CA). Flow rate was 0.8 mL/min on a water/acetonitrile gradient with 4 min data acquisitions from t_0 to t_{max}. Both water and acetonitrile were modified with 0.1% formic acid. MS/MS spectra were acquired by creating a precursor ion inclusion list comprising $[M + H]^+$, $[M + Na]^+$, and $[M+NH_4]^+$ adduct forms for each standard in each mixture using information-dependent analysis (IDA) mode with a 40 ms EAD reaction time. No dynamic exclusion window was used to ensure we captured multiple EAD-CID MS/MS spectra per reference standard, including isomers. Fragmentation was achieved with 12 eV kinetic energy for EAD and 30 eV collision energy for CID in series for each individual MS/MS spectrum. The electron current used for EAD was 8.0 uA. The Zeno trap was active for all MS/MS scans. All spectra were stored in centroid mode. (2) Both UVPD and HCD spectra were acquired on a ThermoFisher Orbitrap IQ-X tribrid mass spectrometer coupled with a Vanquish UHPLC system (ThermoFisher Scientific, San Jose, CA). Chromatographic separation was performed with a reversed-phase 130 Å, 1.7 μm particle size, 100 mm length × 2.1 mm i.d. Waters bridged ethylene hybrid C18 column (Waters, Milford, MA). Flow rate was 0.6 mL/min on a water/acetonitrile gradient with 10 min data acquisitions from t_0 to t_{max}. Electrospray was performed at 3.5 kV in positive mode and 2.5 kV in negative mode at a 275 °C capillary temperature. MS/MS spectra were acquired by creating a precursor ion inclusion list comprising $[M + H]^+$, $[M + NH_4]^+$, $[M - H]^-$, and $[M + CH_3COO]^-$ adduct forms for each standard in each mixture. In addition, the top 4 data-dependent MS/MS spectra were acquired by HCD at 30 NCE and HCD+UVPD fragmentation with HCD at 30 NCE concomitant with UVPD fragmentation using a 213 nm laser with 200 ms activation time. All spectra were acquired in profile mode and subsequently centroided using MSConvert.[20]

Acquiring Experimental Data Sets for Benchmarking Performance of Denoising Algorithm and Normalized Entropy.

Stock solutions at 10 mM concentrations for all target chemicals were prepared by dissolving a purchased aliquot of 0.1 mol and diluted in 1 mL of MeOH for a 10 mM solution. Six working standard solutions containing 40 compounds each were prepared by mixing 2.27 μL of each standard for a final concentration of 0.25 mM. Twelve dilutions were made from each working stock for final concentrations of 0.1, 0.05, 0.02, 0.005, 0.002, 0.001, 0.0005, 0.0002, 0.0001, 0.00005, 0.00002, and 0.00001 mM. All measurements were carried out on a Thermo IQ-X Exactive instrument (ThermoFisher Scientific, San Jose, CA). A total of 2 μL of sample was separated on a Waters Acquity UPLC BEH column (Waters, Milford, MA). The column was maintained at 30 °C with a flow rate of 0.4 mL/min. The mobile phases consisted of (A) water (100%) with formic acid (0.1%) and (B) acetonitrile (100%) with formic acid (0.1%). The Q Exactive MS instrument was operated by using positive mode electrospray ionization with the following parameters: Mass range, 60–1500 m/z; Sheath gas flow rate, 60; Aux gas flow rate, 25; Sweep gas flow rate, 2; Spray voltage (kV) 3.6; Capillary temp, 300 °C; S-lens RF level, 50; Aux gas heater temp, 370 °C. Full MS parameters: Microscans, 1; Resolution, 70000; AGC target, 1e6; Maximum IT, 100 ms; Number of scans, 1; Spectrum data type, Centroid. MS/MS spectra were acquired with the data-dependent mode using the following parameters: Microscans, 1; Resolution, 15000; AGC target, 1e4; Maximum IT, 100 ms; Loop count, 4; MSX count, 1; TopN, 4; Isolation window, 1.0 m/z; Isolation offset 0.0 m/z; (N)CE/stepped, N(CE): 25, 35, 65; Spectrum data type, Centroid. Target compounds were included in the inclusion to ensure MS2 acquisition.

Implementing LibGen.

Python v3.8 was used to create the LibGen framework. The first 14 characters of InChI keys provided by vendors were validated against PubChem[7] and GNPS[8] databases to retrieve SMILES codes of achiral (2D) structures. SMILES codes were used to calculate accurate masses using RDKit.[21] Chemical class information was acquired using ClassyFire and stored as metadata.[22] Mono-isotopic mass of precursors were calculated using an adduct calculator.[23] Spectral entropy and entropy similarity were computed with the flash-entropy module.[24] Raw spectra files were converted from instrument files to 64-bit mzML files using MSConvert in Proteowizard.[20] Distinct functionalities in Libgen are realized through discrete modules, promoting greater efficiency and facilitating code reusability. This modular architecture of Libgen provides advantages for both novice users seeking a simple all-in-one mode for high-throughput processing of mass spectrometry libraries, as well as experienced users who wish to tailor their workflows to meet specific requirements. Ions from MS/MS spectra were removed at >1.5 Da below the precursor ion, but not the precursor ion itself. As data was acquired with a 1 Da isolation window, most isotopic peaks of the precursor were not present in MS/MS spectra anyways. Then, fragment ions were binned to 0.020 Da windows to match the R = 15000 mass spectral resolving power of the instruments. Fragment ion intensities were sum-normalized to unity.

RESULTS AND DISCUSSION

LibGen Pipeline to Curate MS Libraries from Reference Standards.

Natural product libraries were acquired and curated: for electron activated dissociation (EAD), 1,614 reference standards were injected in nonisobaric mixtures into a 7600 ZenoTOF mass spectrometer in positive ESI mode. For both ultraviolet photodissociation (UVPD) and high collision energy (HCD), an Orbitrap IQ-X tribrid mass spectrometer was used, analyzing 2007 natural products in positive and negative ESI mode. Because UVPD and HCD spectra were obtained using an identical liquid chromatography system and reference standards, a similar number of MS1 features and statistics were derived from these two instruments. The curation process employed in LibGen is depicted in Figure 1. Following the preprocessing of raw spectral data and conversion into feature tables containing peak information, the curation process commenced by matching precursor masses against theoretical precursor mass values in a preestablished standard list, using a mass tolerance of 5 mDa. The curation statistics are summarized in Table 1. All three data sets yielded a match rate of approximately 85% to 88% for injected chemicals. Missed detection of chemical standards was likely due to failure to acquire MS/MS spectra due to poor ionization efficiency that led to low signal/noise in MS1 or poor quality MS/MS spectra. For ZenoTOF-EAD data acquisitions, chemicals may also have been missed if these were only ionizable in negative ESI mode and not in positive ESI.

LibGen was developed as an out-of-box solution for curating high-quality mass spectral libraries in a fully automatic manner. LibGen simplifies the input process by requiring only a list of InChIKeys[25] for injected chemical standards pointing to specific chromatogram file labels, and resulting raw mass spectra data files. LibGen uses SMILES codes retrieved from InChI keys to neutralize molecules if applicable, followed by recalculation of chemical information from updated SMILES. The prepared standard list can also be used as an inclusion list for DDA data collection for minimizing human input error. Currently, LibGen uses $[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$, and $[M - H_2O + H]^+$ adducts for calculating cation precursors, and $[M - H]^-$, $[M + acetate]^-$, $[M - H_2O - H]^-$, $[M + formate]^-$, $[M + Cl]^-$, and $[M + Na - 2H]^-$ for anions. Additional adducts can be incorporated by modifying the source code. For each injection mixture, the custom-built feature finding module, ff_droup, generates a single peak matrix to pair retention time and precursor m/z information with corresponding MS/MS spectra. The detailed description of the ff_droup module is given in Supporting Information, Method 1. As precursor ion abundance in MS/MS spectra depends on instrument conditions and dissociation energies, the residual abundance of precursor ions in MS/MS spectra may obfuscate the orthogonal information on fragmentation ions in MS/MS similarity scoring. We therefore bypassed precursor ions from the MS/MS spectra for entropy calculation and subsequent denoising process.

Relying solely on precursor masses for annotation generated many incorrect peak annotations. One frequently overlooked source of incorrect peak annotations is isomers produced alongside chemical reference standards, even when these were marketed and purchased as “pure standards”. Other potential reasons for false peak annotations by MS1 peak picking encompass in-source fragmentations, artifacts from chemical noise and

buffer chemicals, or cross-contamination between samples. To ensure the integrity and comprehensiveness of curated libraries, we first removed all MS1 candidates that did not yield information-rich MS/MS spectra, using spectral entropy $S < 0.5$ as the threshold. This step removed 0.03–0.5% of all candidates (Table 1). Next, we developed a deduplication algorithm to identify the authentic spectra associated with both target compounds and their isomers while excluding spectra originating from artifacts and other contaminant sources (Figure 2). As it is reasonable to assume that the most abundant peak ("major compound") corresponds to the pure reference chemicals, such peaks were assigned as the most probable source for the true MS/MS spectrum. Lower abundant chromatographic peaks within a 10 s range and within 5 mDa precursor mass error were labeled as isomeric impurities ("minor compounds"), if such peaks were not detectable in preceding or subsequent samples. Isomeric peaks with retention time differences >10 s were retained if peak intensities exceeded 0.33 base peak intensity, and if MS/MS spectra exhibited an entropy similarity of >0.75 to the major compound.

For the three data sets, the algorithm reduced the number of potential candidates MS1 features between 34–71% (Table 1). This deduplication step is essential for the subsequent recalibration process. By incorporation of major compounds together with their plausible isomeric analogs, the coverage of curated libraries is considerably expanded. Figure 2 also demonstrates the usefulness of using chromatographic separation instead of direct infusion in library generation, as commercial chemical standards should not be considered to merely consist of only one chemical structure. Peaks #1 and #3 had identical precursor masses but different MS/MS spectra than piceid_minor (peak #2) and piceid_major (Figure 2). In direct infusion library generation, these spectra would have been combined. We did not further investigate the difference in chemical structures of these peaks, but it is possible that cis/trans isomers and regioisomers of the hexose-unit exist in this commercial chemical.

All accurate mass instruments exhibit both systematic and stochastic mass errors, especially for low intensity compounds with low poor ion statistics.[26–28] To mitigate such technical errors during the curation of MS/MS libraries, LibGen recalibrates the fragment ion mass-to-charge (m/z) values for each data file from the complement of annotated MS^1 peaks in each mixture. Interestingly, a simple linear regression to model the relationship between MS1 error and MS1 intensity did not improve the overall accuracy. In the mass ranges used for our natural product libraries, no relationship between the m/z values and the m/z errors was observed. Instead, LibGen employs a random forest model to predict the mass errors depending on the ion intensities, which is essentially a nonlinear regression without specifying the kernel function. Incorporating retention time data did not enhance the model performance. This mass recalibration resulted in a substantial improvement in mass accuracy across all three libraries (Figure 3).

This MS/MS mass recalibration is essential for the next step in the LibGen algorithm, denoising MS/MS spectra. Spectral denoising selectively eliminates false positive fragment ions in MS/MS spectra and ensures the overall high quality of the library entries, especially for low abundant compounds. Here, LibGen employs a subformula-based spectral denoising algorithm.[14,29] All fragment ions are validated by chemical plausibility, determining whether the exact mass of each fragment can be logically associated with a subformula

loss from its parent molecular ion species. In addition, we expanded this logic to cover observations that in rare cases (<1%), the collision gas nitrogen may be added to formulas of fragments of aromatic compounds.[30] Similarly, two hydrogen atoms and an oxygen atom can be added to fragments of aromatic compounds during collision-induced dissociation.[30] When curating the natural product libraries presented here, we confirmed the rare occurrence of such fragments at <1%. Mass losses between molecular ion species and fragment ions were first searched against a database of 3.5 million chemical obtained from more than 26 chemical databases[31] that was increased to 10 million formulas by adjusting the number of hydrogens to cover radical ions (odd- and even-electron counterparts), in analogy to the BUDDY algorithm that was recently published to calculate elemental compositions for molecular ion species.[31] Only if fragments could not be rationalized against this formula database, the algorithm decomposes the molecular formula itself and comprehensively calculates all elemental subsets that could fit the mass loss while excluding implausible losses such as ones that only account for pure carbon or pure nitrogen losses (except N_2).[32] Such calculations are much more computationally expensive and therefore only used after direct matching. Hence, all fragments that cannot be rationalized by these two methods must logically come from another source. Such ions were removed from the experimental spectra. Consequently, for each library spectrum, we calculated the residual explained intensity (EI%) as a relative measure of spectral quality by examining the percentage of valid MS/MS fragment intensity over the raw (uncurated) MS/MS fragment intensity:

$$\text{explained intensity}(\%) = \frac{\sum_{\text{p,valid}} I_{\text{p,valid}}}{\sum_{\text{p}} I_{\text{p}}}$$

where I_p is the intensity of a given peak.

To remove low-quality spectra from MS/MS libraries, LibGen applies both two criteria: (1) $S > 0.5$, to ensure that spectra with sparse fragmentation do not get added to the library, e.g., if only one or two fragment ions are generated. (2) Explained intensity EI > 90%, to ensure that the vast majority of fragments in a spectrum can be assigned to the parent molecular structure.

Only few spectra were removed by the low entropy criterion (1), removing <0.5% of the total number of MS/MS spectra (Table 1). However, in comparison, a surprisingly high number of spectra showed a high number of fragment ions that could not be logically explained by elemental formula losses. For the EAD library, we found that 5.6% of the spectra were removed in the final denoising step (Table 1), giving an overall high quality of the spectra generated by this method. Surprisingly, the HCD library contained 40.8% spectra that had to be removed in the %EI denoising step (Table 1). Even more spectra were removed for the UVPD data set, with 67.8% of all spectra giving less than 90% of explained fragment ion intensities (Table 1).

EAD spectra may be interpreted as being of overall higher quality because the EAD mechanism combines two powerful fragmentation pathways, a reaction between small, singly charged ions and electrons followed by classic collision-induced dissociation.[33]

As a consequence, the population of precursor ions is more efficiently subjected to fragmentation than by CID alone, yielding relatively large contributions of noise and artifact ions in MS/MS spectra. Correspondingly, when the CID MS/MS spectra were analyzed, a high proportion of unfragmented precursor ions was observed. The less fragmentation of precursor ions occurs, the more notable are the relative contributions of (chemical and electronic) noise ions in the MS/MS spectra. This fact is specifically important if precursor ions are (virtually) removed from MS/MS similarity queries'. Similarly, UVPD ionizes ions through the absorption of high-energy photons with a predetermined wavelength, selectively breaking covalent bonds and producing unique fragmentation patterns at low fragmentation efficiency, often leaving much of the precursor ion population unfragmented.[12] Specifically, for compounds lacking conjugate systems, the energy absorption process is significantly reduced, yielding lower fragmentation efficiency for these compounds that lead to MS/MS spectra with a high ratio of noise ions, which are subsequently removed by the denoising algorithm.[12]

Validation of the Denoising Algorithm on High/Low Quality Data Sets.

Next, we tested the effectiveness of the LibGen denoising algorithm by comparing 240 compounds in six highly concentrated mixtures against 13-fold diluted mixtures using HCD fragmentation (Table 2, Supporting Information, Method 2). As expected, only about half of the compounds in the 13-fold diluted data set triggered MS/MS spectra acquisitions, yielding only about 1/3 of the MS/MS spectra compared to the positive control of all 214 compounds in the highly concentrated data set (Table 2). Here, we tested the hypothesis that spectra from diluted mixtures have a higher ratio of noise ions in MS/MS spectra, yielding a lower proportion of explained intensity. As comparison to LibGen's approach using chemical plausibility to differentiate between noisy and clean mass spectra, we also calculated normalized entropy as data-driven method.[28] In the highly concentrated MS/MS data set, LibGen labeled 77% of the spectra as clean spectra using an explained MS/MS abundance ratio of 90% as threshold (Table 2). For normalized entropy, we used a 0.8 cutoff as proposed previously, yielding an 89% rate of MS/MS spectra labeled as clean (Table 2).

Conversely, for the 13-fold diluted spectra, LibGen rejected 96% of the spectra, proving our hypothesis that noise ions in low-abundance MS/MS spectra have a significant contribution to overall MS/MS spectra and should therefore not be used in library building. In comparison, using normalized entropy, only 50% of the highly diluted MS/MS spectra were rejected. Using normalized entropy therefore might lead to the erroneous incorporation of poor-quality MS/MS spectra into libraries for compounds that inadvertently showed low ionization efficiency. As expected, MS/MS spectra with high normalized entropy values also generally showed lower explained intensity for both the highly concentrated and the 13-fold diluted data set (Figure 4A,B). However, a number of MS/MS spectra with very low explained intensity were still categorized as clean by normalized entropy. Such spectra contained many noise ions that could not be justified by their constituent elements, and the denoising algorithm accurately identified them as invalid (Figure 4C). Similarly, the LibGen denoising algorithm qualifies a range of spectra as good-quality, although they have high normalized entropy (Supporting Information, Data 1). This data exemplifies the utility of

the LibGen denoising algorithm as essential to exclude low-quality MS/MS spectra while avoiding to excluding information-rich true positive MS/MS spectra.

Overview and Comparison of Curated Libraries.

The three libraries generated here served slightly different purposes and were acquired separately. Therefore, the chemical diversity of the EAD library was different from the UVPD/HCD libraries, although a number of natural products were shared. Notably, the EAD library boasts the largest collection of reference standards. Moreover, UVPD and HCD libraries differ in ionization efficiency and spectra quality. Thus, only a subset of spectra that successfully passed LibGen quality control criteria was included in the curated library for each technique. Overall, 486 compounds were detected in all three libraries, while 368 were only tested and detected in the EAD library, while 50 compounds were exclusively found in the UVPD library and 95 compounds solely presented in the HCD library (Figure 5a). By utilizing the 2048-bit Morgan fingerprint and Tanimoto similarity index for chemometric analysis, all libraries displayed a high degree of structural diversity. More than 96% of all compounds in all three exhibited cumulative Tanimoto cosimilarity indices <0.2 (Supporting Information, Data 2). Around 22% of all compounds were detected as adducts other than $[M - H]^-$ or $[M + H]^+$, highlighting the necessity of considering all adduct types when building mass spectral libraries (Supporting Information, Data 3). Importantly, the curated EAD and UVPD libraries are the first metabolomic libraries of their kind and may serve as test data to study fragmentation mechanisms. Therefore, all three libraries are open-access and can be publicly downloaded from MassBank.us.

The use of UVPD and EAD has not been explored comprehensively for annotating natural compounds, specifically in comparison to classic collision induced dissociation. We found that EAD spectra showed far richer fragmentation spectra than UVPD or HCD spectra with an average spectral entropy of 4.42, with 90% of all spectra found between spectral entropy 3.3–5.5 (Figure 5b). Similar to other small molecule libraries, the natural products used here showed low entropy in HCD fragmentation, with an average of spectral entropy 2.1 and a 90% range from 0.8 to 3.6. UVPD spectra were more information-rich than HCD spectra but showed clearly less fragmentation than EAD spectra (Figure 5b), with a mean spectral entropy of 2.8 and a 90% quantile ranging from 1.1 to 4.1. Hence, both EAD and UVPD fragmentations are better suited than HCD to yield unique fragmentation patterns that can be used to identify natural products in untargeted metabolomics. Third, we found that MS/MS spectra of the same compounds showed overall little similarity across the three instrument types (Figure 5c). Here, we quantified the MS/MS similarity in a pairwise manner using entropy similarity. Few compounds showed highly similar MS/MS spectra at entropy similarity >0.75 between the different instrument types, while around half of the compounds showed at least moderate similarity between fragmentation types at entropy similarities between 0.6 and 0.75 (Figure 5c). Often, some fragment ions are shared between two fragmentation types, while even abundant ions may be absent from the MS/MS spectra (Figure 6). Hence, acquiring complementary and orthogonal fragmentation data through MS/MS analysis may significantly enhance confidence in compound identification in complex mixtures and add to our understanding of fragmentation.

ESI Positive and Negative Modes in UVPD.

While EAD is currently available only for positive mode electrospray ionization, we were also interested in evaluating the difference between positive and negative ESI mode spectra under UVPD fragmentation. For both, raw and curated MS/MS spectra, we found that UVPD fragmentation yielded about 4-fold more positive ESI spectra than negative ESI spectra (Supporting Information, Data 3). Also, UVPD produces more fragments in the positive mode (Figure 7). Collectively, spectra obtained in positive mode carry higher information content, with a mean spectral entropy of 2.39 compared to a mean spectral entropy of 1.48 in the negative mode.

Annotation Using Curated Libraries.

To validate the capability of libraries constructed by LibGen, compound annotation was performed on biological samples. Human GI tract samples[34] were analyzed by LC-EAD mass spectrometry using a ZenoTOF 7600 mass spectrometer, while the NIST bilberry SRM 3291 dietary supplement reference material was analyzed by LC-HCD/UVPD mass spectrometry using a Orbitrap IQ-X instrument. The latest version of MS-DIAL 4.9.2[35] was used for data preprocessing, and the MSP file was imported into LibGen for compound annotation via an identity search using entropy similarity. All MS/MS spectra were subject to cleaning prior to library searching, and a similarity score of 0.75 was employed for unambiguous compound annotation with a minimized false discovery rate.[28] Annotations using the curated libraries are listed in Supporting Information, Data 4, showing select examples in Figure 8.

CONCLUSION

Open-access spectral libraries play a crucial role in the advancement of metabolomics, particularly as the development of novel ionization techniques continues to gain momentum. In this study, we presented LibGen, a fully automated solution for curating MS/MS spectral libraries from mixes of reference standards with exceptional efficiency and reliability. Specifically, we demonstrated the superior performance of the subformula-based denoising algorithm compared to using normalized entropy for deciding which spectra to include in high quality libraries. Curated by LibGen, three specialized metabolite libraries comprising highly diverse chemical standards acquired by HCD-, EAD-, and UVPD-fragmentation were curated. The EAD and UVPD natural product libraries are the first of their kind as publicly available MS/MS libraries. Being freely available on MassBank.us, users can employ such libraries for nutritional or gut microbiome research and related fields. Open access MS/MS libraries such as shown here facilitate advancing informatics algorithms to better interpret fragmentation mechanisms and to identify unknown compounds, both of which require high-quality reference data and high-confidence identifications. The application of the curated EAD and UVPD libraries is demonstrated for compound annotation on biological samples. The distinct fragmentation manner of EAD and UVPD shows potential opportunities for future researchers to use these tools as complementary evidence to classic collision induced dissociation spectra for compound identifications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Funding was provided by NIH OD034497 and NIH ES030158. We thank Sciex for giving access to the EAD-mass spectrometer, specifically Dr. Mackenzie J. Pearson, and we thank ThermoFisher for giving access to the UVPD-HCD mass spectrometer, specifically Dr. Ameer Bashar.

Data Availability Statement

All curated libraries are uploaded onto MassBank of North America database (Massbank.us) and can be freely downloaded. The molecular formula database is uploaded to the project repository on GitHub (https://github.com/FanzhouKong/Libgen_stable/tree/main/db). Raw spectra files can be found upon request. The source code of LibGen for curating spectral libraries with all associated modules can be found at https://github.com/FanzhouKong/Libgen_stable, along with demo data, Jupyter notebooks, and documentation.

REFERENCES

(1). Fiehn O Plant Mol. Biol 2002, 48 (1–2), 155–71. [PubMed: 11860207]

(2). Vinaixa M; Schymanski EL; Neumann S; Navarro M; Salek RM; Yanes O TrAC, Trends Anal. Chem 2016, 78, 23–35.

(3). Wishart DS Bioanalysis 2011, 3 (15), 1769–1782. [PubMed: 21827274]

(4). Wishart DS Bioanalysis 2009, 1 (9), 1579–1596. [PubMed: 21083105]

(5). Johnson CH; Ivanisevic J; Siuzdak G Nat. Rev. Mol. Cell Biol 2016, 17 (7), 451–459. [PubMed: 26979502]

(6). Sumner LW; Amberg A; Barrett D; Beale MH; Beger R; Daykin CA; Fan TW; Fiehn O; Goodacre R; Griffin JL; Hankemeier T; Hardy N; Harnly J; Higashi R; Kopka J; Lane AN; Lindon JC; Marriott P; Nicholls AW; Reily MD; Thaden JJ; Viant MR Metabolomics 2007, 3 (3), 211–221. [PubMed: 24039616]

(7). Kim S; Chen J; Cheng T; Gindulyte A; He J; He S; Li Q; Shoemaker BA; Thiessen PA; Yu B; Zaslavsky L; Zhang J; Bolton EE Nucleic Acids Res. 2023, 51 (D1), D1373–d1380. [PubMed: 36305812]

(8). Wang M; Carver JJ; Phelan VV; Sanchez LM; Garg N; Peng Y; Nguyen DD; Watrous J; Kapono CA; Luzzatto-Knaan T; Porto C; Bouslimani A; Melnik AV; Meehan MJ; Liu WT; Crüsemann M; Boudreau PD; Esquenazi E; Sandoval-Calderón M; Kersten RD; Pace LA; Quinn RA; Duncan KR; Hsu CC; Floros DJ; Gavilan RG; Kleigrewe K; Northen T; Dutton RJ; Parrot D; Carlson EE; Aigle B; Michelsen CF; Jelsbak L; Sohlenkamp C; Pevzner P; Edlund A; McLean J; Piel J; Murphy BT; Gerwick L; Liaw CC; Yang YL; Humpf HU; Maansson M; Keyzers RA; Sims AC; Johnson AR; Sidebottom AM; Sedio BE; Klitgaard A; Larson CB; Boya P CA; Torres-Mendoza D; Gonzalez DJ; Silva DB; Marques LM; Demarque DP; Pociute E; O'Neill EC; Briand E; Helfrich EJN; Granatosky EA; Glukhov E; Ryffel F; Houson H; Mohimani H; Kharbush JJ; Zeng Y; Vorholt JA; Kurita KL; Charusanti P; McPhail KL; Nielsen KF; Vuong L; Elfeki M; Traxler MF; Engene N; Koyama N; Vining OB; Baric R; Silva RR; Mascuch SJ; Tomasi S; Jenkins S; Macherla V; Hoffman T; Agarwal V; Williams PG; Dai J; Neupane R; Gurr J; Rodríguez AMC; Lamsa A; Zhang C; Dorrestein K; Duggan BM; Almaliti J; Allard PM; Phapale P; Nothias LF; Alexandrov T; Litaudon M; Wolfender JL; Kyle JE; Metz TO; Peryea T; Nguyen DT; VanLeer D; Shinn P; Jadhav A; Müller R; Waters KM; Shi W; Liu X; Zhang L; Knight R; Jensen PR; Palsson BO; Pogliano K; Linington RG; Gutiérrez M; Lopes NP; Gerwick WH; Moore BS; Dorrestein PC; Bandeira N Nat. Biotechnol 2016, 34 (8), 828–837. [PubMed: 27504778]

(9). NIST/EPA/NIH NIST 20 Tandem Mass Spectral Libraries. https://chemdata.nist.gov/. Accessed on December 2022.

(10). Guijas C; Montenegro-Burke JR; Domingo-Almenara X; Palermo A; Warth B; Hermann G; Koellensperger G; Huan T; Uritboonthai W; Aisporna AE; Wolan DW; Spilker ME; Benton HP; Siuzdak G Anal. Chem 2018, 90 (5), 3156–3164. [PubMed: 29381867]

(11). Baba T; Ryumin P; Duchoslav E; Chen K; Chelur A; Loyd B; Chernushevich IJ Am. Soc. Mass Spectrom 2021, 32 (8), 1964–1975.

(12). Brodbelt JS; Morrison LJ; Santos I Chem. Rev 2020, 120 (7), 3328–3380. [PubMed: 31851501]

(13). Kaufmann A; Walker S Rapid Commun. Mass Spectrom 2012, 26 (9), 1081–1090. [PubMed: 22467458]

(14). Stravs MA; Schymanski EL; Singer HP; Hollender JJ Mass Spectrom. 2013, 48 (1), 89–99.

(15). Yang X; Neta P; Stein SE Anal. Chem 2014, 86 (13), 6393–400. [PubMed: 24896981]

(16). Kind T; Wohlgemuth G; Lee DY; Lu Y; Palazoglu M; Shahbaz S; Fiehn O Anal. Chem 2009, 81 (24), 10038–10048. [PubMed: 19928838]

(17). Burke MC; Zhang Z; Mirokhin YA; Tchekovskoi DV; Liang Y; Stein SE J. Proteome Res 2019, 18 (9), 3223–3234. [PubMed: 31364354]

(18). Wallace WE; Ji W; Tchekhovskoi DV; Phinney KW; Stein SE J. Am. Soc. Mass Spectrom 2017, 28 (4), 733–738. [PubMed: 28127680]

(19). Xing S; Huan T Anal. Chim. Acta 2022, 1200, No. 339613. [PubMed: 35256147]

(20). Kessner D; Chambers M; Burke R; Agus D; Mallick P Bioinformatics 2008, 24 (21), 2534–2536. [PubMed: 18606607]

(21). RDKit: Open-source cheminformatics. https://www.rdkit.org/. Accessed on December 2022.

(22). Djoumbou Feunang Y; Eisner R; Knox C; Chepelev L; Hastings J; Owen G; Fahy E; Steinbeck C; Subramanian S; Bolton E; Greiner R; Wishart DS J. Cheminform 2016, 8 (1), 61. [PubMed: 27867422]

(23). Mass Spectrometry Adduct Calculator. https://fiehnlab.ucdavis.edu/staff/kind/metabolomics/ms-adduct-calculator/. Accessed on December 2022.

(24). Li Y; Fiehn O Nat. Methods 2023, 20, 1475–1478. [PubMed: 37735567]

(25). Heller SR; McNaught A; Pletnev I; Stein S; Tchekhovskoi DJ Cheminform 2015, 7 (1), 23.

(26). Petyuk VA; Jaitly N; Moore RJ; Ding J; Metz TO; Tang K; Monroe ME; Tolmachev AV; Adkins JN; Belov ME; Dabney AR; Qian WJ; Camp DG 2nd; Smith RD Anal. Chem 2008, 80 (3), 693–706. [PubMed: 18163597]

(27). Lysiak A; Fertin G; Jean G; Tessier D BMC Bioinformatics 2021, 22 (2), 65. [PubMed: 33902435]

(28). Li Y; Kind T; Folz J; Vaniya A; Mehta SS; Fiehn O Nat. Methods 2021, 18 (12), 1524–1531. [PubMed: 34857935]

(29). Dührkop K; Fleischauer M; Ludwig M; Aksenov AA; Melnik AV; Meusel M; Dorrestein PC; Rousu J; Böcker S Nat. Methods 2019, 16 (4), 299–302. [PubMed: 30886413]

(30). Shaffer CJ; Schröder D; Alcaraz C; Žabka J; Zins E-L ChemPhysChem 2012, 13 (11), 2688–2698. [PubMed: 22693155]

(31). Xing S; Shen S; Xu B; Li X; Huan T Nat. Methods 2023, 20 (6), 881–890. [PubMed: 37055660]

(32). Böcker S; Dührkop K J. Cheminform 2016, 8 (1), 5. [PubMed: 26839597]

(33). Baba T; Rajabi K; Liu S; Ryumin P; Zhang Z; Pohl K; Causon J; Le Blanc JCY; Kurogochi MJ Am. Soc. Mass Spectrom 2022, 33 (9), 1723–1732.

(34). Folz J; Culver RN; Morales JM; Grembi J; Triadafilopoulos G; Relman DA; Huang KC; Shalon D; Fiehn O Nat. Metab 2023, 5 (5), 777–788. [PubMed: 37165176]

(35). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M Nat. Methods 2015, 12 (6), 523–526. [PubMed: 25938372]

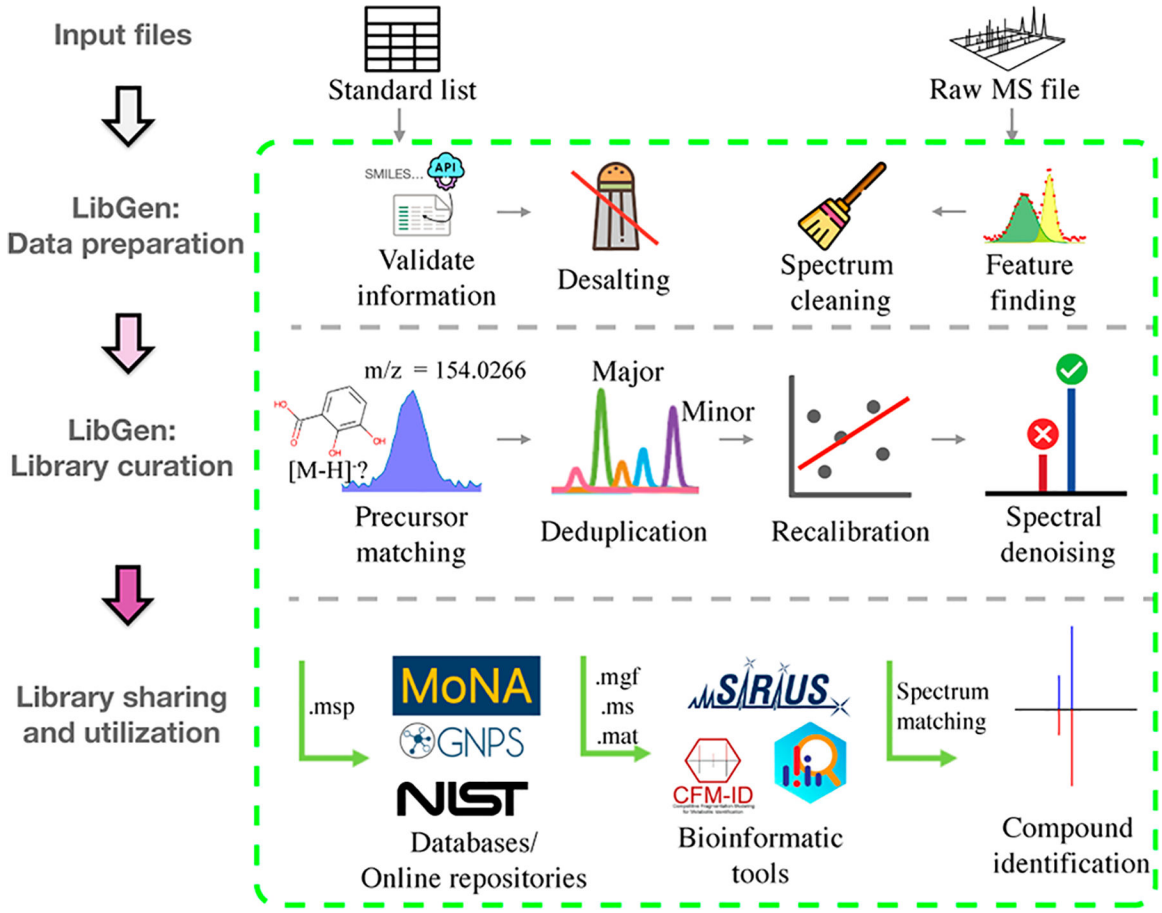


Figure 1.

Workflow for curating high-quality mass spectral libraries in LibGen. The software needs a list of target chemicals with the corresponding raw mass spectral data. LibGen performs all curation processes and exports a curated library into various formats for upload to MS/MS libraries, downstream bioinformatics tools, or advanced compound identification software.

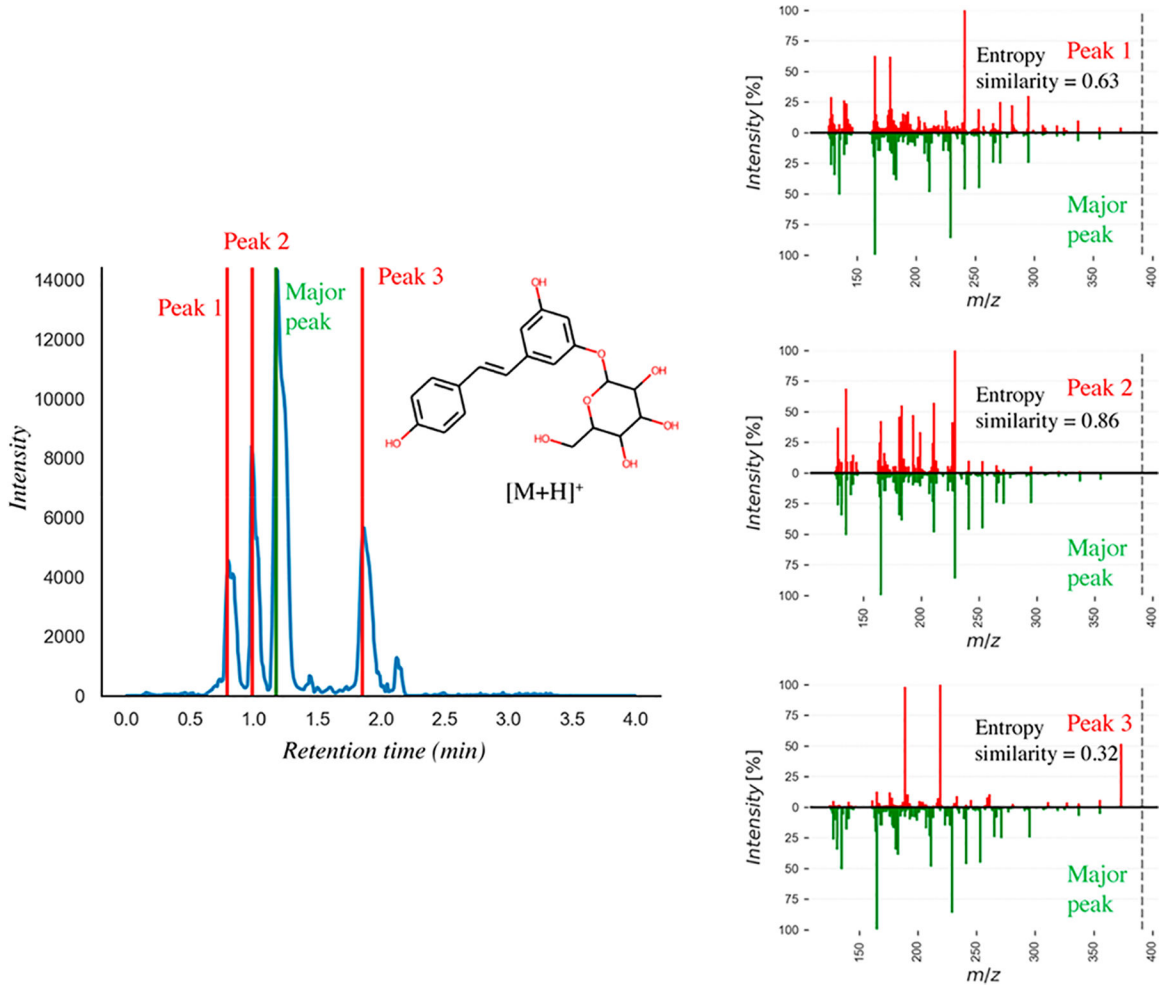


Figure 2.

Deduplication of isomeric peaks in library curation. Left: extracted ion chromatogram *m*/*z* 391.139 for piceid as the chemical reference standard, yielding four matching MS1 peaks. The major peak at 1.2 min is annotated as reference compound, while minor peaks at 0.7, 1.0, and 1.8 min represent isomers. Right: corresponding MS/MS spectra of minor peaks are matched by the entropy similarity to the major peak. Isomers at entropy similarity >0.75 are included into the final library.

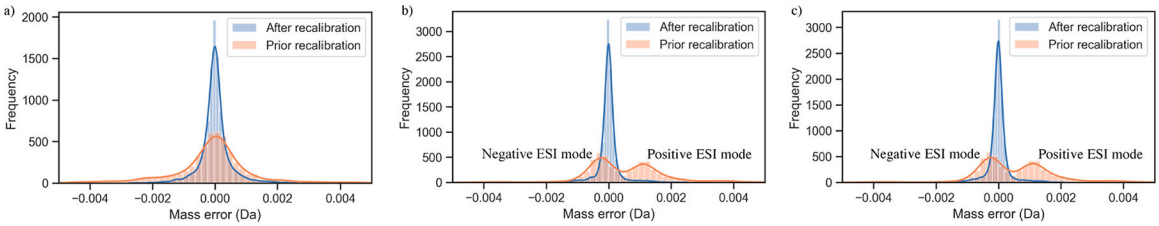


Figure 3.

Kernel-density plots of MS^1 mass errors before and after mass recalibrations using random forest models of MS^1 precursor intensities: (a) EAD data set, (b) UVPD data set, and (c) HCD data set.

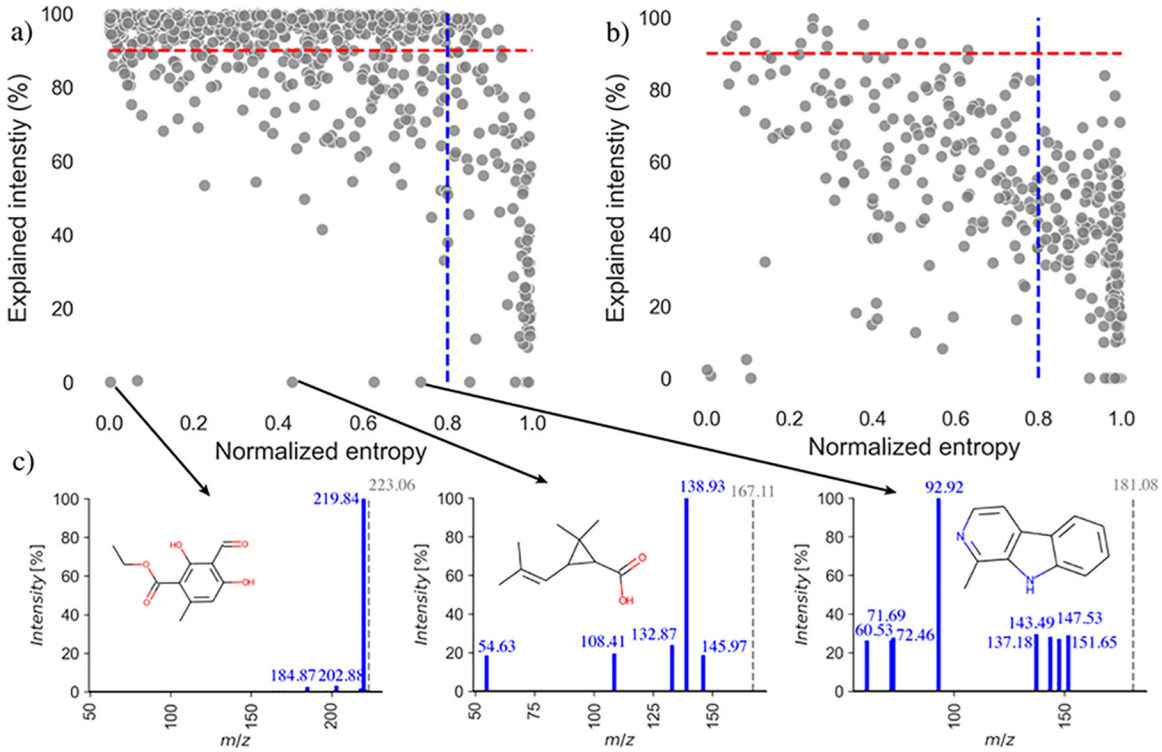


Figure 4.

Benchmarking results for MS/MS quality assessments using a data set of highly concentrated compound mixture (a) and a 13-fold diluted data set of the same compounds (b). MS/MS quality thresholds are used as 0.8 for normalized entropy (blue dashed lines) and 90% explained intensity (red dashed lines) for the LibGen subformula denoising algorithm. (c) Three example MS/MS spectra are given for the highly concentrated data set with low normalized entropy but no chemically plausible ions (from left to right: hematommic acid ethyl ester $[M - H]^-$, chrysanthemic acid $[M - H]^-$, and harmane $[M - H]^-$).

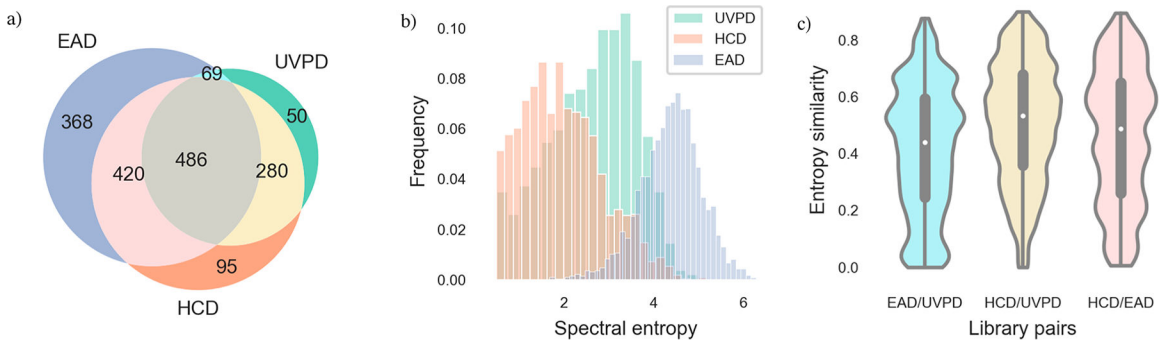


Figure 5.

Key statistics of three curated libraries: (a) Compound overlap between the EAD, HCD, and UVPD libraries. (b) Normalized distribution of the spectral entropies of the MS/MS libraries. (c) MS/MS similarity of compound spectra across libraries.

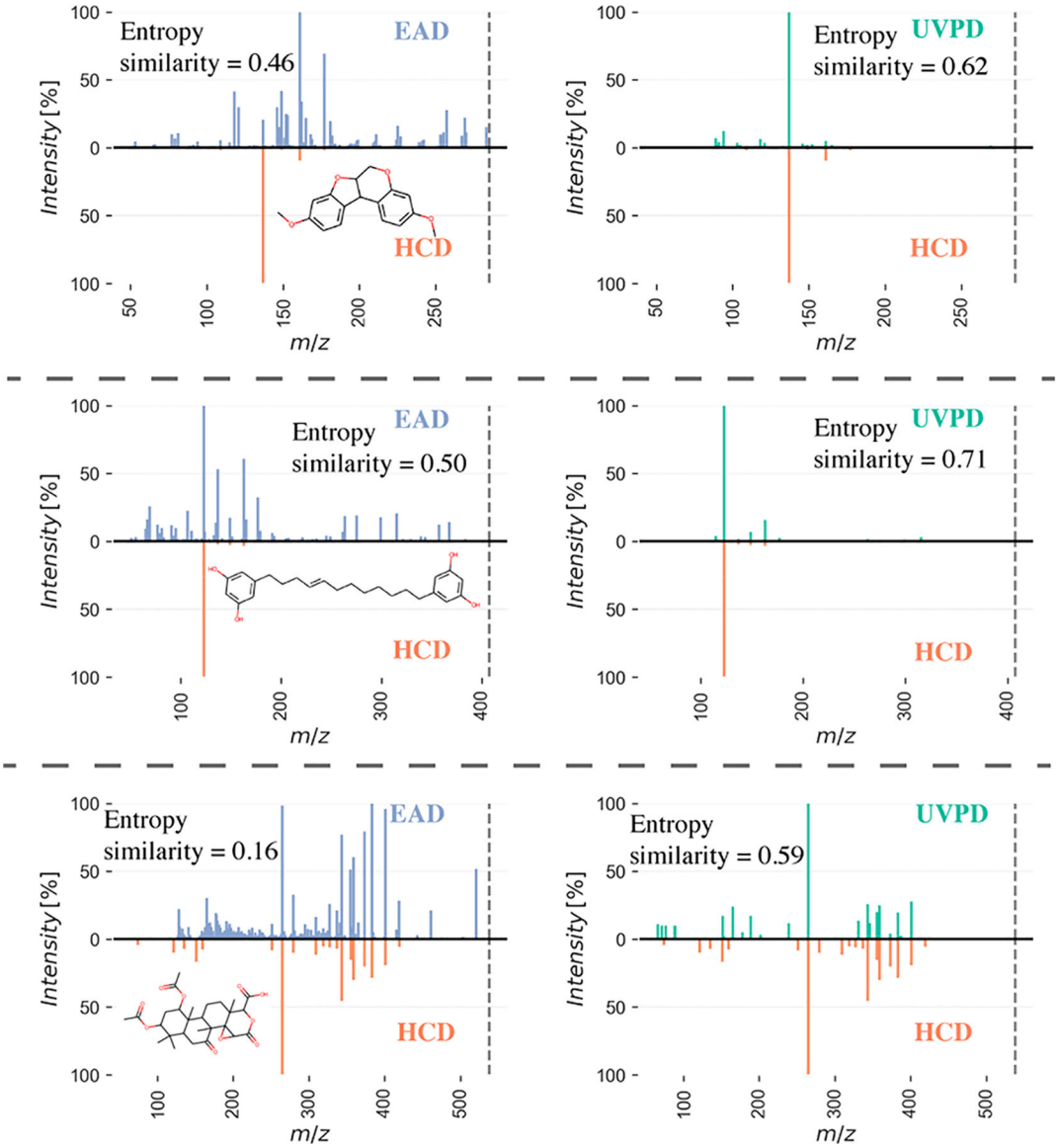


Figure 6.

Head-to-tail comparisons of EAD/HCD mass spectra (left panels) and UVPD/HCD mass spectra (right panels) for three example compounds. Top: Homopterocarpin, $[M + H]^+$. Middle: 5-[(*Z*)-12-(3,5-dihydroxyphenyl)dodec-8-enyl]benzene-1,3-diol, $[M + NH_4]^+$. Bottom: Deacetoxy(7)-7-oxokhivorinic acid, $[M + NH_4]^+$. Dashed lines indicate precursor *m*/*z*.

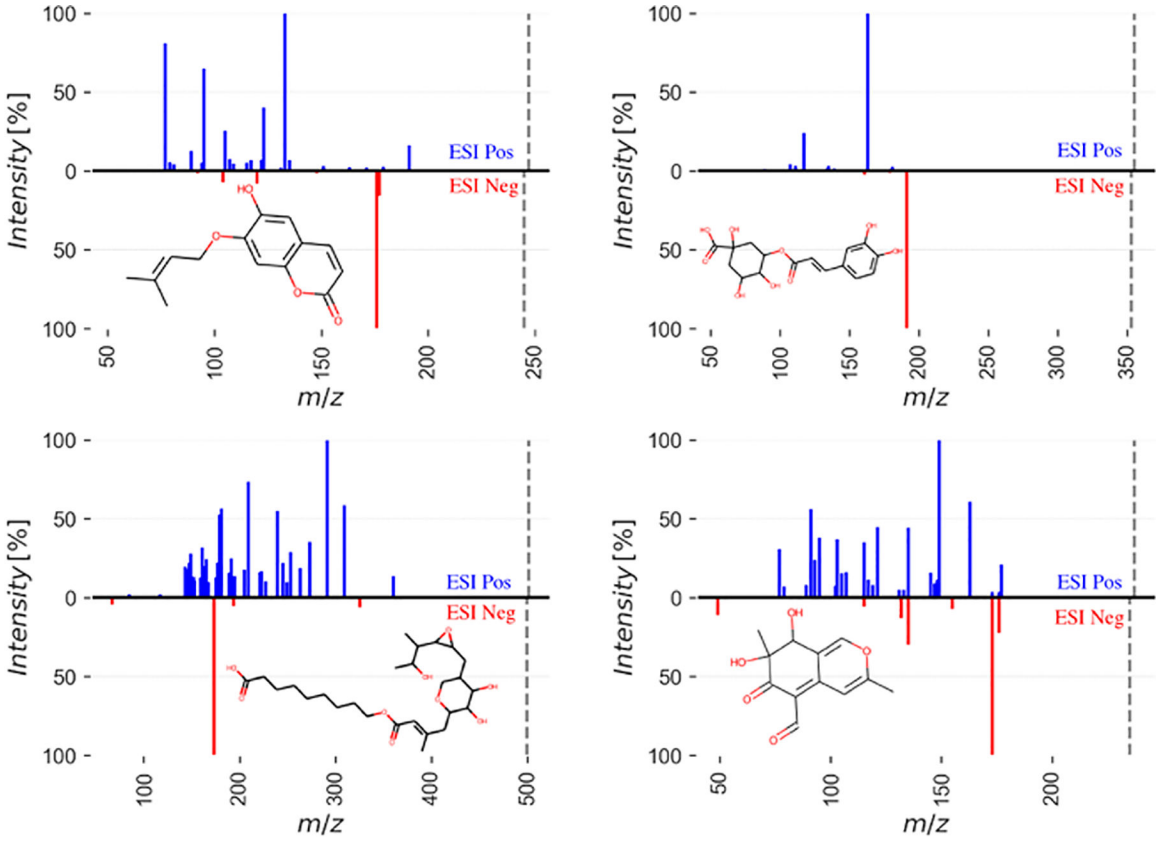


Figure 7.

Comparisons of UVPD MS/MS spectra across ESI positive ($[M + H]^+$, blue) and negative ($[M - H]^-$, red) modes with selected compounds: (a) prenyletin, (b) chlorogenic acid, (c) 2-amino-5-[2-[[2,3-dihydroxy-2-(1-hydroxyethyl)butanoyl]oxymethyl]-4-hydroxyanilino]-5-oxopentanoic acid, and (d) furanone. The dashed line indicates the precursor *m*/*z*.

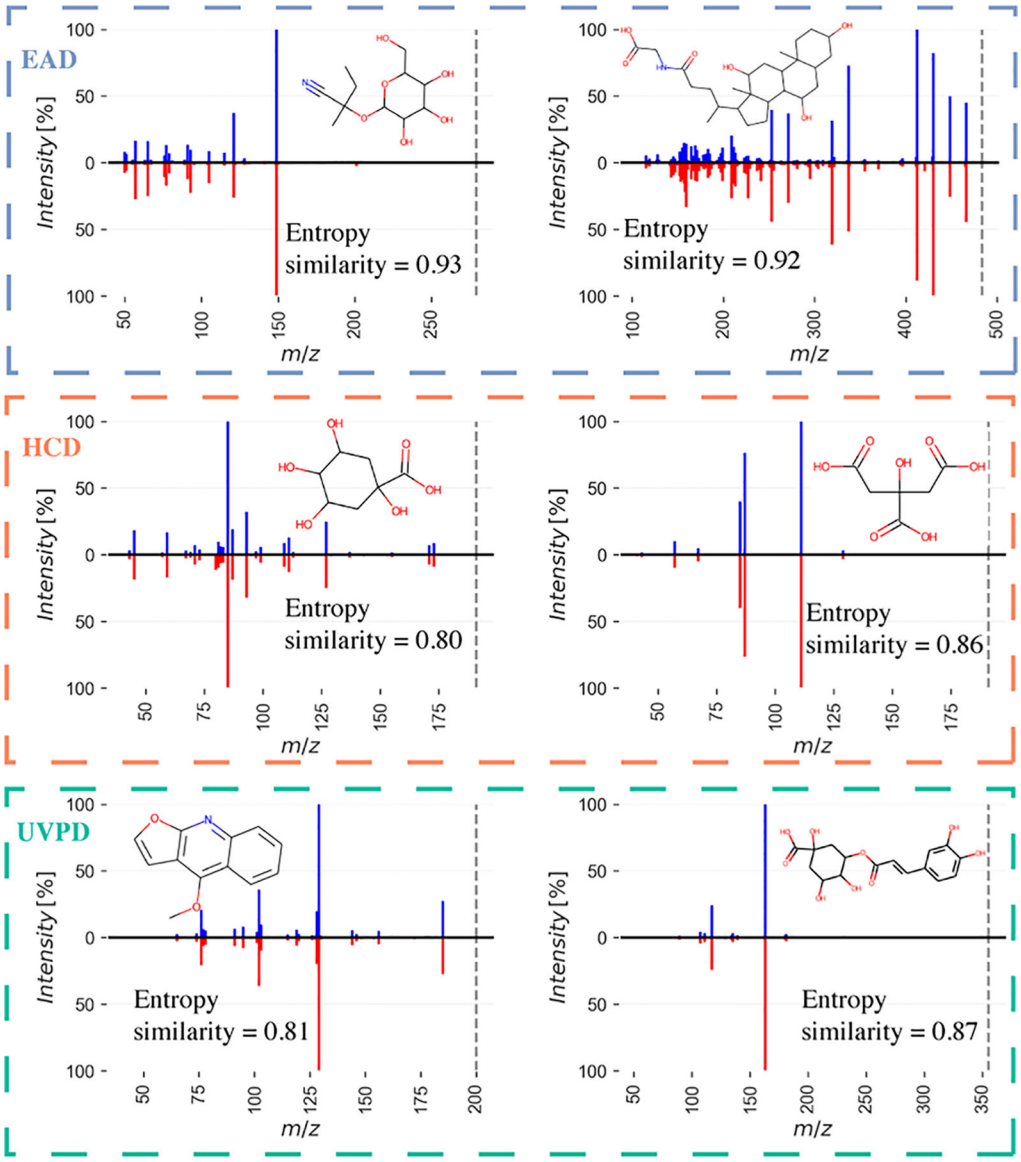


Figure 8.

EAD spectra on human GI tract samples and UVPD/HCD spectra on bilberries are annotated with curated libraries. The experimental spectrum is always on the top of the head-to-tail plot and the library spectrum is on the bottom. Top: EAD library, isomeric lotaustralin ($[M + NH_4]^+$, left) and glycocholic acid ($[M + NH_4]^+$, right). Middle: HCD library, quinic acid ($[M - H]^-$, left) and citric acid ($[M - H]^-$, right). Bottom: UVPD library, dictamnine ($[M + H]^+$, left), and chlorogenic acid ($[M + H]^+$, right). The minimum spectral similarity for an annotation is 0.75.

Table 1.

Overview of Retrieving and Curating Library Entries Using LibGen for EAD, HCD, and UVPD Libraries[a]

libraries	EAD library			UVPD			HCD		
No. of compounds injected	**1614**			**2007**			**2007**		
libraries overviews	**No. of unique compounds**	**No. of MS^1 candidates**	**match rate (%)**	**No. of unique compounds**	**No. of MS^1 candidates**	**match rate (%)**	**No. of unique compounds**	**No. of MS^1 candidates**	**match rate (%)**
precursor matched	1423	5737	88.2	1729	13009	86.1	1729	13009	86.1
low entropy curated	1423	5735	88.2	1727	12919	86.0	1717	12635	85.6
deduplicated	1423	3808	88.2	1727	3797	86.0	1717	3750	85.6
curated library	1343	3524	83.2	885	1223	44.1	1281	2220	63.8

[a]Match rate is defined as the number of reference compounds present/number of total number of reference compounds injected.

Table 2.

Performance of Quality Control by Subformula Denoising Algorithm and Normalized Entropy on the Concentrated and Diluted Datasets

	No. of compounds	No. of spectra	No. of clean spectra by normalized entropy	No. of clean spectra by subformula denoising
highly concentrated set	214	1049	935	807
13-fold diluted set	106	347	175	13