JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

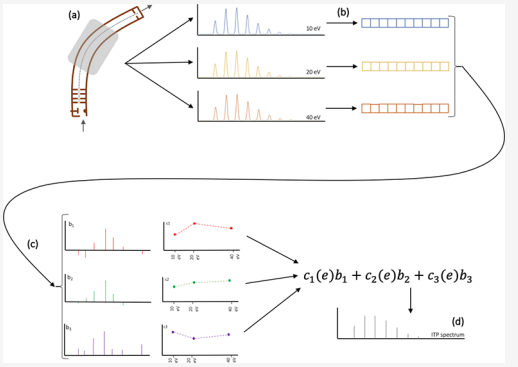
# Augmentation of MS/MS Libraries with Spectral Interpolation for Improved Identification

Ethan King, Richard Overstreet, Julia Nguyen, and Danielle Ciesielski*

Cite This: J. Chem. Inf. Model. 2022, 62, 3724−3733 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Tandem mass spectrometry (MS/MS) is a primary tool for the identification of small molecules and metabolites where resultant spectra are most commonly identified by matching them with spectra in MS/MS reference libraries. The high degree of variability in MS/MS spectrum acquisition techniques and parameters creates a significant challenge for building standardized reference libraries. Here we present a method to improve the usefulness of existing MS/MS libraries by augmenting available experimental spectra data sets with statistically interpolated spectra at unreported collision energies. We find that highly accurate spectral approximations can be interpolated from as few as three experimental spectra and that the interpolated spectra will be consistent with true spectra gathered from the same instrument as the experimental spectra. Supplementing existing spectral databases with interpolated spectra yields consistent improvements to identification accuracy on a range of instruments and precursor types. Applying this method yields significant improvements (∼10% more spectra correctly identified) on large data sets (2000−10 000 spectra), indicating this is a quick yet adept tool for improving spectral matching in situations where available reference libraries are not yet sufficient. We also find improvements of matching spectra across instrument types (between an Agilent Q-TOF and an Orbitrap Elite), at high collision energies (50−90 eV), and with smaller data sets available through MassBank.




## INTRODUCTION

Mass spectrometry (MS) is a gold-standard compound identification method used in many fields such as food safety, wastewater/environmental analysis, clinical and forensic toxicology, metabolic profiling, lipid and peptide analysis, and many more.[1−13] The technique came to prominence with the use of standardized hard ionization methods (fixing the electron ionization source at 70 eV) coupled with gas chromatography (GC-MS), facilitating the development of extensive GC-MS spectral libraries and associated look-up techniques that quickly match experimental spectra to known reference spectra. Because GC-MS is limited in its ability to analyze small molecules, metabolites, and nonvolatile substances, researchers developed alternate MS methods to analyze these compounds. A common alternative is to use liquid chromatography sample preparation, soften the ionization method, and connect a series of mass spectrometers in tandem to refine how compounds are ionized. By coupling two or more mass analyzers, analysts can first ionize the molecules to separate the ions by their mass-to-charge ($m/z$) ratio and then identify ions having a particular $m/z$ ratio to split into smaller fragment ions. This level of detail allows analysts to elucidate molecule structure from molecular weight discovery and fragmentation behavior, enhancing the ability to identify unknown unknowns.[14−16] Electrospray ionization-liquid chromatography-tandem mass spectrometry (ESI-LC-MS/MS) has come to the forefront as a specialized MS method that deserves a place next to GC-MS as highly rigorous, specific, and sensitive compound identification methods. As a soft ionization technique, ESI-LC reduces sample preparation demands, allowing analysts to study nonvolatile and larger substances and making it invaluable for the study of small molecules and metabolites that fragment beyond recognition under hard ionization conditions.

As MS/MS technology becomes ubiquitous, researchers are calling for standardization of techniques and data.[17−19] With so much variability in MS/MS acquisition techniques, building the spectral reference libraries and rigorously validated workflows that were essential to the mainstreaming of GC-MS is a massive challenge for the MS community. Most laboratories use the high-quality, value-enhanced commercial National Institute of Standards and Technology (NIST) and Wiley libraries,[20] which have greatly increased their collection of MS/MS spectra over the years. Value is added to these

Received: May 17, 2022
Published: July 29, 2022





https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

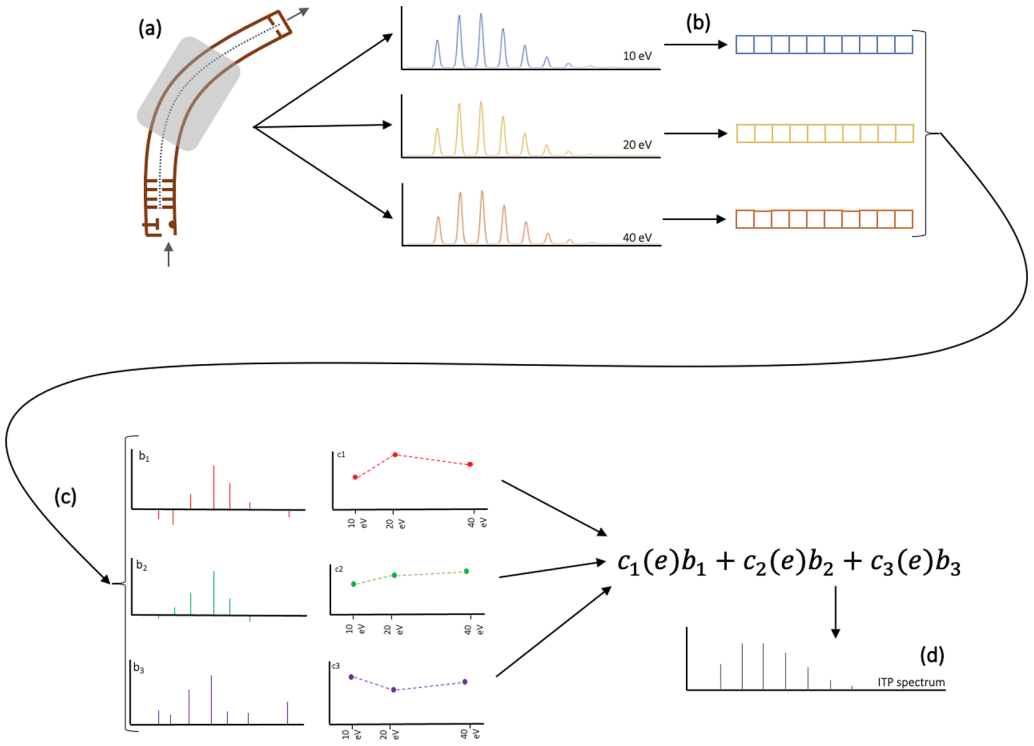



**Figure 1.** (a) Three experimental spectra are collected from a consistent mass spectrometry workflow at collision energies spanning the energies of interest. (b) The experimental spectra are binned to the desired level of detail, forming vector-formatted spectra that are suited for mathematical analysis. (c) The vector-formatted spectra are joined into a matrix, and the SVD of the matrix is computed. The resulting SVD provides a set of basis vectors $b_k$ and the weight coefficients $c_k$ at the known collision energies. The weight coefficients for the desired spectra are interpolated from the known weight coefficients. (d) Finally, for a desired collision energy $e$, the interpolated weights are applied as a linear combination with the basis vectors to determine the anticipated spectrum at the unknown collision energy.

libraries because they are curated by experienced mass spectrometrists who manually inspect and correct spectra, remove noise and artifacts, add structures and CAS numbers, build consensus spectra, add peak annotations, and perform interlibrary comparisons.[21,22] NIST's spectra, in particular, are generated on a number of different instruments and at many different collision energies, the latter being one of the most important factors in a resulting MS/MS spectrum. Realistically, however, NIST and Wiley will never be able to keep up with the modular nature of MS/MS workflows or the quantity and diversity of molecules of interest to researchers.

Researchers are using a number of approaches to generate more diverse spectral libraries. One approach is to gather all the available spectra into a communal database, like MassBank,[23] but the quality assurance (QA) task for such an endeavor is monumental, and appropriate QA standards are under debate.[18,19] A popular approach is to generate spectra in silico for both known and suspected molecules using quantum-chemical properties and/or machine-learning techniques.[24−36] This is a highly active area of research with varying degrees of success in both quality of the prediction and length of time required to generate a quality prediction, but it rarely accounts for the differences between spectra produced by different instruments. Also, for disciplines requiring rigorous validation for legal purposes, in silico libraries may not be considered valid for a confirmatory analysis. A limiting but pervasive approach is for laboratories to curate and maintain their own internal libraries that match their own validated workflows. While the community in general may be disappointed at the loss of communal resources, this final option is quite popular and allows laboratories to maintain libraries that perform best with the spectra they generate.

To improve the utility of available MS/MS spectral databases—whose contents are invaluable due to the diversity of instrumentation, collision energies, other instrument parameters, and molecules of interest—we developed a first of its kind tool that takes existing spectra, applies principle component analysis (PCA) to fill in gaps in the libraries, and allows researchers to do a spectral comparison and compound identification with high accuracy. Analyzing high-resolution mass spectrometry (HRMS) data available in NIST20,[37] we show that filling in unavailable collision energies with PCA-interpolated spectra results in ∼10% improvement in compound identification than comparing with just the existing database. Our method reduces the need for libraries to cover a broader range of collision energies with the costly and time-consuming collection of spectra. Additionally, unlike many in silico library tools, this interpolation method generates spectra quickly and accounts for the spectral differences caused by the particular instruments and settings used to generate the spectra. This tool can be broadly applied as a quick and simple way to improve accuracy when performing spectral matching for compound identification.

## METHODS

**Spectral Interpolation.** Because we start with fine-grain HRMS data, we bin the peaks of the spectra to increase our ability to identify key spectral features. This peak-binning process naturally creates a vector representation for spectra, so we can leverage linear algebra tools and construct a method to



https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

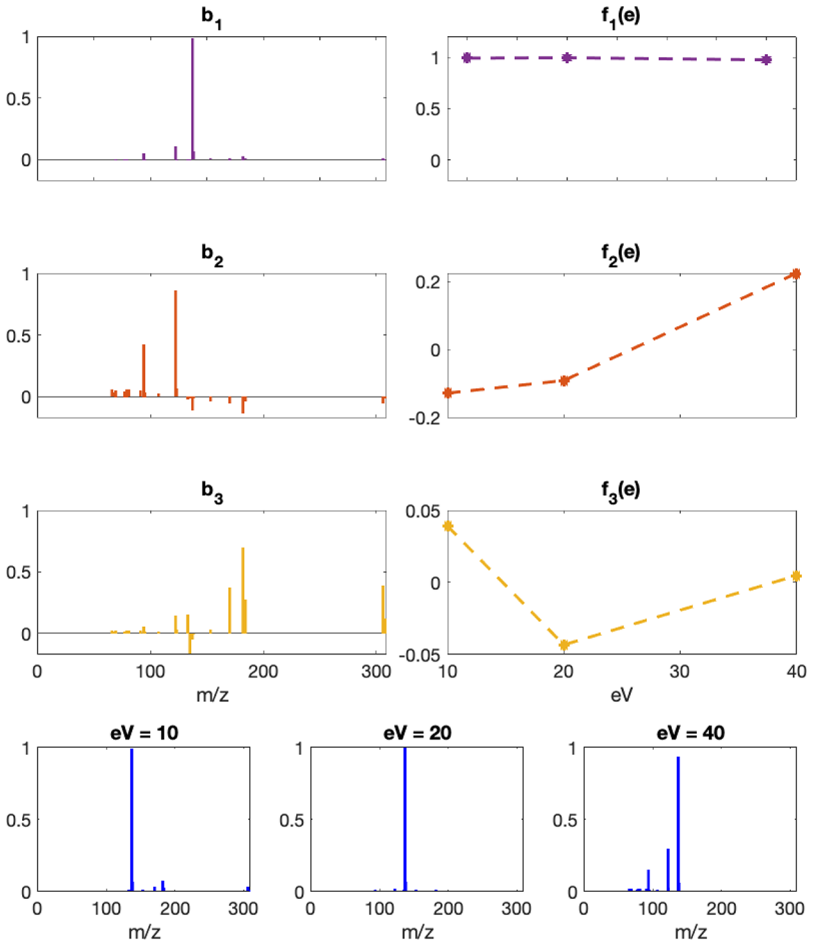



**Figure 2.** (top) The left-hand plots titled $b_k$ show the basis vectors generated from the capsaicin spectra shown in the bottom row of the figure. The right-hand plots titled $f_k(e)$ give the coefficients that reconstruct the spectrum at a given electronvolt value, $e$, with a linear interpolation (dashed line) plotted across collision energies. For each basis vector, these interpolations are the functions $f_k$. Across the range of collision energies, the contribution of $b_1$ remains relatively constant, as it contains the prominent base peak across all electronvolts. At 10 eV, there are small influences from basis vector $b_3$, and the negative value of $f_2$ decreases peak intensity at the positive value in basis vector $b_2$ while increasing the intensity of the peaks shown as negative. At 20 eV, the contributions of basis vector $b_3$ switch sign, and the coefficient for basis vector $b_2$ begins to increase. Finally at 40 eV, the influence of basis vector $b_2$ increases, corresponding with the appearance of strong peaks with lower $m/z$ values than the base peak observed at 10 and 20 eV. This represents the fragmentation that occurs between 20 and 40 eV. (bottom) Normalized MS/MS Agilent Q-TOF spectra for capsaicin from the NIST20 database at collision energies of 10, 20, and 40 eV are shown in the bottom row.

interpolate spectra across collision energies. We first lay out the notation for this process and then discuss details of the approach. A high-level overview is illustrated in Figure 1.

Let a spectrum $s$ with a set of peaks $P$ be given by the set

$$s = \{(m_p, i_p)\}_{p \in P} \tag{1}$$

where $i_p$ is the measured intensity of a peak at $m/z$ value $m_p$. To apply principal component analysis (PCA) to a set of spectra, they must be represented in the same vector space. To conform a set of spectra, we first choose a $Q_{max}$ value such that all relevant $m/z$ values are in the interval $[0,Q_{max}]$ and partition the interval into $N$ uniform bins $\{[q_{min^n},q_{min^n})\}_{n \in \{0,\cdots,N\}}$. For our purposes, $Q_{max}$ is determined for each trial independently by the highest nonzero $m/z$ value in the set of spectra being analyzed, and the value of $N$ is set to bin the peaks to the nearest integer. This coarse binning is chosen to allow many trials to be run quickly to validate the process, though the mathematical details still apply for the much finer detail required for a real-world analysis. We then represent a spectrum $s$ as a vector $v \in \mathbb{R}^N$ where, at each index $n$, the vector value $v^n$ becomes either the intensity of the highest peak in that section of the partition or zero if there are no peaks in that bin. This modified binning method is designed to ignore noise around prominent peaks. If a moderate height peak is surrounded by many very small peaks, the common method of binning by summing the peaks may allow the peak to appear more prominent. This binning method preserves peak prominence and reduces noise. Mathematically, this is expressed as follows.

$$v^n = \begin{cases} \max\{i_p | m_p \in [q_{min^n}, q_{max^n})\} & \text{if there exists } m_p \in [q_{min_n}, q_{max_n}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

To capture how the spectra for a given molecule progress as collision energy changes, we seek an optimal representation of the set of spectra using singular value decomposition (SVD). Commonly in PCA, SVD is used to identify a set of component vectors that is smaller than the full data set but still represents all of the data well enough to make strong



https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

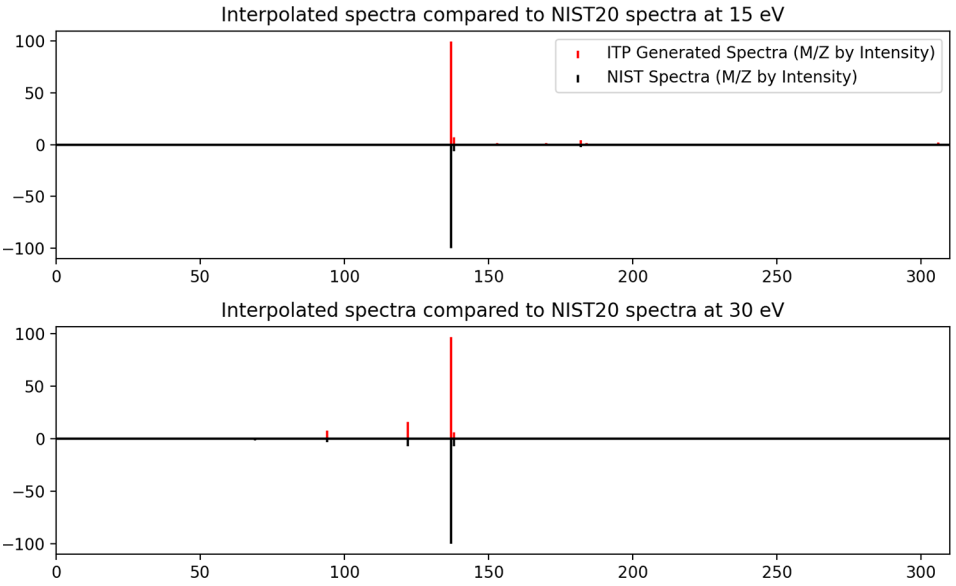



**Figure 3.** Sample interpolation predicted (ITP) Q-TOF capsaicin spectra compared to the known spectra available in NIST20 at collision energies of 15 and 30 eV. Note that the methods used to generate predictions preclude accurate predictions outside the range of provided collision energies. These spectra were generated with samples at 10, 20, and 40 eV, so ITP spectra can only be generated for collision energies between 10 and 40 eV.

predictions. However, because spectral prediction is incredibly nuanced, we generate a full set of basis vectors to retain as much spectral detail as possible. In this analysis, a basis is a minimal set of vectors required to be able to recreate any spectrum in the data set through a linear combination of the basis vectors (in linear algebra terms, the basis spans the original data set). The basis vectors also have the properties of being linearly independent and orthogonal. These properties make PCA a powerful tool but also make the basis vectors purely statistical artifacts, no longer representative of actual spectra. To represent a given set of $J$ known, vectorized spectra $V = \{v_j\}_{j \in \{1,\cdots,J\}}$ taken at collision energies $\{e_j\}_{j\in\{1,\cdots,J\}}$, we use SVD to construct an orthonormal basis $\{b_k\}_{k\in\{1,\cdots,K\}}$ of the span of $V$, where $K$ is the dimension of the span. Because of the complexity of the HRMS data, $K$ will be equal to $J$ in most cases for this method.

Now because $\{b_k\}_{k \in\{1,\cdots,K\}}$ is a basis, there exist a set of coefficients $\{c_{k,j}\}_{k \in\{1,\cdots,K\},j \in\{1,\cdots,J\}}$ such that each vectorized spectrum $v_j \in V$ can be written as a linear combination of basis vectors.

$$v_j = c_{1,j}b_1 + c_{2,j}b_2 + \ldots + c_{K,j}b_K \tag{3}$$

That is, each spectrum can be represented as a weighted sum of the basis vectors $b_k$, where the coefficient $c_{k,j}$ can be understood as the contribution of the vector $b_k$ to the spectrum $v_j$. In this view, the changes in spectra across collision energies can be described by the changes to the contributions (coefficients) of each basis vector. For example, a given basis vector may have a small contribution at low collision energy but a large contribution at higher collision energies. It is important here to note that the coefficients $c_{k,j}$ can be positive or negative because the basis vectors do not necessarily correspond to any physical phenomenon (e.g., fragment structure/stability); they are statistical in nature.

Finally, to generate the interpolations for all missing collision energies, we need to build functions that map how the contributions for each basis vector change as a function of the collision energy. These functions are represented by the dotted lines in Figure 2. Ideally we would use a function that takes in a scalar collision energy $e$ and outputs the corresponding continuous, HRMS spectrum $g(e)$ for a given molecule. While we cannot determine the true function $g$, we can construct an approximation $\hat{g}$ from $\mathbb{R}$ to $\mathbb{R}^N$ that outputs an $N$-dimensional vectorized spectrum in the span of the basis $\{b_k\}_{k\in\{1,\ldots,K\}}$ of the form

$$\hat{g}(e) = f_1(e)b_1 + f_2(e)b_2 + \ldots + f_K(e)b_K \tag{4}$$

where we initially define

$$f_k(e_j) = c_{k,j} \tag{5}$$

for all $j \in \{1,\ldots,J\}$ and $k \in \{1,\cdots,K\}$ such that the approximation exactly satisfies (5) with the values satisfying (3) for our $J$ known vectorized spectral representations. We then estimate the values of $f_k$ at all other $e \in [e_{\min},e_{\max}]$ by linear interpolation, where we have the following.

$$e_{\min} := \min_{j\in\{1,\ldots,J\}} \{e_j\} e_{\max} := \max_{j\in\{1,\ldots,J\}} \{e_j\} \tag{6}$$

By this definition, the vectorized spectrum approximations of the form 4 may include a negative intensity value. To make sensible spectrum estimates, all negative values in $\hat{g}(e)$ are set to zero.

Figure 2 (top) shows the three basis vectors $b_k$ for a set of three capsaicin spectra (Figure 2 (bottom)) along with the known contribution values $c_{k,j}$ (the dots on the right-hand plots) and how they change as a function of collision energy, $f_k$. The basis vector $b_1$ represents a peak prominent across all collision energies, and the associated coefficients, $c_{1,j}$ and $f_1$, remain close to constant. In contrast, $b_2$ represents peaks that are more prominent in the highest collision energy spectra, estimated as linear increases to the contribution of $b_2$ in the interpolation $f_2$. Head-to-tail comparisons of interpolated spectra against experimental spectra from NIST20 are shown in Figure 3. While this method shows strong results within the range $[e_{\min}, e_{\max}]$, it is important to note that, as an interpolation, extrapolating to spectrum estimates at collision


https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

**Table 1. This Table Gives the Number of Molecules with Spectra for Each Precursor and Instrument Pair, along with the Acronym That Each Pair Is Referred to in This Text**

| | Agilent Q-TOF | Elite Orbitrap | Velos Orbitrap | Orbitrap Fusion Lumos |
|---|---|---|---|---|
| $[M + H]^+$ | QHP: 2,603 | EHP: 15,081 | VHP: 558 | LHP: 6,191 |
| $[M-H]^-$ | QHN: 246 | EHN: 8,932 | VHN: 13 | LHN: 2,066 |
| $[M + Na]^+$ | QNa: 173 | ENa: 4,087 | VNa: 107 | LNa: 927 |

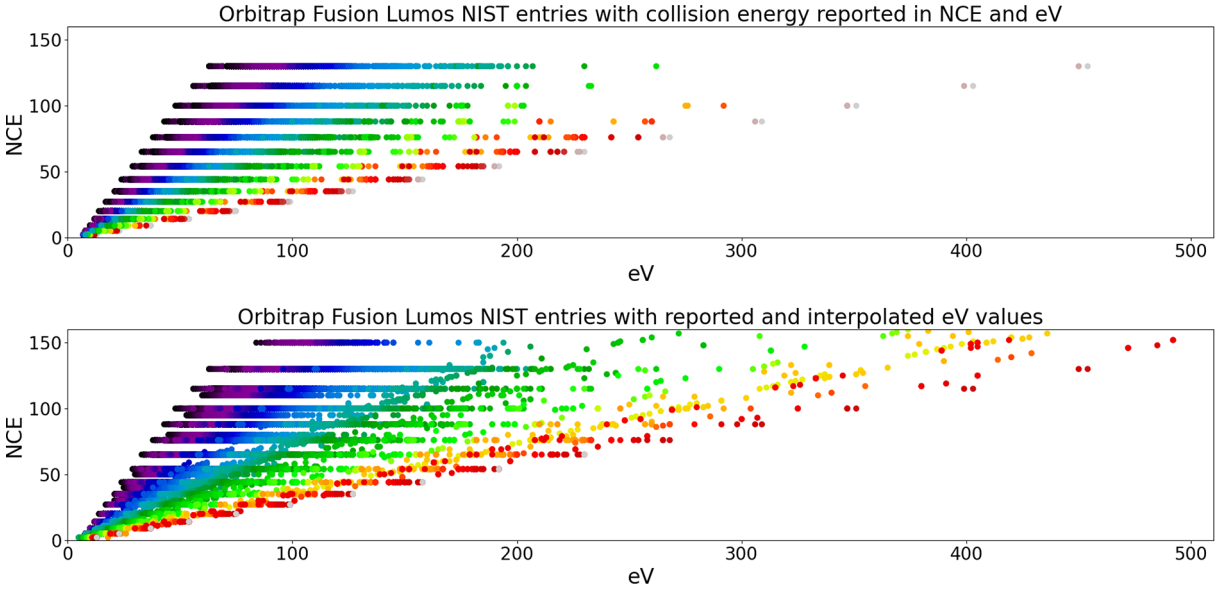



**Figure 4.** (top) Comparison of reported electronvolt readings against NCE for the Orbitrap Fusion Lumos, color graded by precursor $m/z$. A strong linear trend suggests the potential to apply linear regression to approximate missing electronvolt values. (bottom) All Orbitrap Fusion Lumos scores are shown with interpolated electronvolt values where none were reported. Variation in the reported data set leads to imperfect alignment, though a strong correlation is still obvious. The subset of data used for most of this analysis is restricted to spectra collected at or below 40 eV, including only the bottom left-hand corner of these images.

energies outside the range $[e_{min}, e_{max}]$ is not possible, as the approximations are not meaningfully defined for such values.

**Spectra Curation.** We assess the accuracy of our interpolation methodology for molecular identification using spectra from the HRMS and APCI libraries available in NIST20.[37] The instruments used to generate these spectra are the Agilent 6530 Q-TOF, the Thermo Finnigan Velos Orbitrap, Thermo Finnigan Elite Orbitrap, and the Orbitrap Fusion Lumos. Spectra at multiple collision energies for each molecule are only available for high-energy collision dissociation (HCD) and quadrupole time-of-flight (Q-TOF) measurements; therefore, an analysis of our method is limited to these. We further restrict our analysis to small molecules with molecular weight in the range of 100−500 Da and stratify by precursor type to focus on molecules that do not respond well to standard hard ionization methods. We selected the positive ion and negative ion mode precursors with the most available data, which are $[M + H]^+$ and $[M-H]^-$, respectively. In addition we test the $[M + Na]^+$ precursor across all instruments. We include this variety of precursor types to ensure our method works for different analytical workflows. Table 1 gives the number of molecules available in each subset. Throughout we will refer to the instrument precursor pairs by the acronyms given in the table.

For all Q-TOF spectra and HCD spectra from the Thermo Finnigan Velos and Elite Orbitraps, NIST20 reports the collision energy in electronvolts. For the Orbitrap Fusion Lumos, however, only approximately one-third of the spectra have an electronvolt value listed alongside their normalized collision energy (NCE). When an instrument implements NCE, the collision energy is dynamically adjusted based on the expected precursor weight so that small ions are not ejected from the trap and large ions can be sufficiently fragmented.

In Figure 4 (top), we inspect the relationship between reported electronvolt and NCE values from the Orbitrap Fusion Lumos. The reported electronvolts are listed on the $x$-axis, and the reported NCE values are on the $y$-axis. By adding a color gradient based on the precursor $m/z$ of each molecule, it becomes apparent that NCE has a generally linear relationship with estimated precursor $m/z$ and electronvolts that can be used to approximate the electronvolt value when it is not explicitly reported. The spectra represented in Figure 4 (top), when grouped by reported NCE and precursor $m/z$, display a preexisting variation in the reported electronvolt with a standard deviation of ∼0.2539 eV across all molecules. We treat this variation as negligible for our purposes and assign electronvolt values to spectra without explicit measurement based on the following algorithm. First, all Orbitrap Fusion Lumos spectra having a precursor $m/z$ within one unit of the listed precursor $m/z$ with electronvolt values provided were collected. Next, a linear regression was computed on this subset of spectra to predict the linear trend for that precursor $m/z$. Finally, the spectrum is assigned an electronvolt approximation based on its listed NCE value and this localized regression. The average difference in computed versus reported electronvolt means for a given NCE and precursor $m/z$ was ∼0.1210 eV, indicating a strong correlation for our estimates. Figure 4 (bottom) shows the same data from Figure 4 (top)


https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

with the interpolated values added to the plot. Here we can see that many of the spectra whose precise electronvolt values were unknown will be excluded from this work due to our focus on small molecules and low collision energies.

**Interpreting Added Value from Interpolation-Predicted Spectra.** To measure accuracy gained by using our interpolation-predicted (ITP) spectra method, we employ two experiments of similar format. In each case, we subset spectra from NIST20 to represent a limited number of available database spectra. The remaining spectra serve as unknown test spectra that can be searched against the subset database.

To evaluate the quality of our ITP spectra, we compare only the test spectrum, the ITP spectrum, and the database spectra used to generate the ITP spectrum. For a test spectrum at collision energy $t_{eV}$, we construct an ITP spectrum at collision energy $t_{eV}$ and compute the cosine similarity between the test spectrum and ITP spectrum, denoted by $ITP_{sim}$. We also compute the cosine similarity between the test spectrum and the closest matching database spectrum, denoted by $DB_{sim}$. That is, if there are three database spectra used to generate an ITP spectrum, the cosine similarity is computed between the ITP spectrum and each spectra used to generate the ITP spectrum. The highest cosine similarity score is retained as $DB_{sim}$. Finally, the gain in cosine similarity when the ITP spectra are included in the subset database can be measured as the difference between these scores, $\Delta_{sim} = ITP_{sim} - DB_{sim}$.

To determine the benefit added by including ITP spectra in a limited database search, we modify the previous experiment slightly. We first assume both the precursor molecular weight and collision energy are known for each test spectrum. To identify candidate molecules for our subset database, we screen the NIST20 database for all spectra matching the instrument, precursor type, and target collision energies used to generate the ITP spectra. We then identify all molecules in the candidate list within 10 Da of the test spectrum's precursor weight. For a test spectrum at collision energy $t_{eV}$, we use the spectra in our subset database to construct ITP spectra at the values

$$[t_{eV} - 10, t_{eV} - 9, ..., t_{eV}, ..., t_{eV} + 9, t_{eV} + 10] \quad (7)$$

for each candidate molecule. Finally, we identify the highest cosine similarity score between our test spectrum and (1) all known spectra identified as appropriate for our subset database ($DB_{sim}$) and also (2) the subset database spectra combined with the ITP spectra—but not replacing any known subset spectra—denoted $maxITP_{sim}$. To assess performance of the ITP enhanced database, we report the difference between these values.

$$\max \Delta_{sim} = maxITP_{sim} - DB_{sim} \quad (8)$$

For each trial, we also report the percentage of spectra for which the highest cosine similarity score matches the molecule of the test spectrum when matching to the ITP spectra enhanced database (IM) and the percentage of spectra likewise correctly identified when matching to just subset database spectra (DBM). We also record the difference in electronvolts between the closest matching interpolated spectrum and the test spectrum, $\Delta_{eV}$. That is, for the test spectrum with collision energy $t_{eV}$ and the closest matching ITP spectrum with collision energy $I_{eV}$, we compute

$$\Delta_{eV} = I_{eV} - t_{eV} \quad (9)$$

Finally, we record the percentage of spectra that were correctly identified with only the subset database spectra but misidentified when using the ITP spectra and denote this value as IF.

## RESULTS AND DISCUSSION

**Adds Benefit When Few Spectra Are Available.** We first determine how many known spectra are required to create an accurate ITP spectrum by testing ITP spectra generated by as few as two spectra and as many as eight. To generate a subset database containing $n$ known spectra per molecule, we select $n$ uniformly spaced electronvolt values between 10 and 45 as database collision energies. This range was chosen to represent reasonable energies for small molecules. Then we construct the database from the closest spectra within $(45-10)/2n$ eV of each of the uniformly spaced energies.

For this trial, we consider the Agilent Q-TOF and the Elite Orbitrap spectra with either a $[M + H]^+$ or $[M\text{-}H]^-$ precursor to ensure the behavior is similar on the Q-TOF and Orbitrap instruments. Data sets are referenced by the acronyms as given in Table 1. Within each instrument/precursor pair, we screen for molecules with at least 10 spectra in the 10−45 eV range. The number of unique molecules satisfying the criteria for each of QHP, QHN, EHP, and EHN is 991, 101, 1112, and 460, respectively. The data sets for the Velos Orbitrap and Orbitrap Fusion Lumos are too small to provide sufficient test data under these requirements, so they are omitted.

As shown in Figure 5, the average cosine similarity between the ITP and test spectra quickly approaches 1 as more spectra are added to the database. However, the ITP spectra achieve high accuracy with as few as three spectra in the database, reporting an average cosine similarity over 0.95. Furthermore, with just three known spectra at a range of electronvolts, generating an ITP spectrum to target the test electronvolt in a specific reference library gives a stronger match on average, across all molecules, than the closest known spectrum. These results are shown in Figure 6, where a positive difference ($\Delta_{sim}$) indicates a stronger match with the ITP spectrum and a

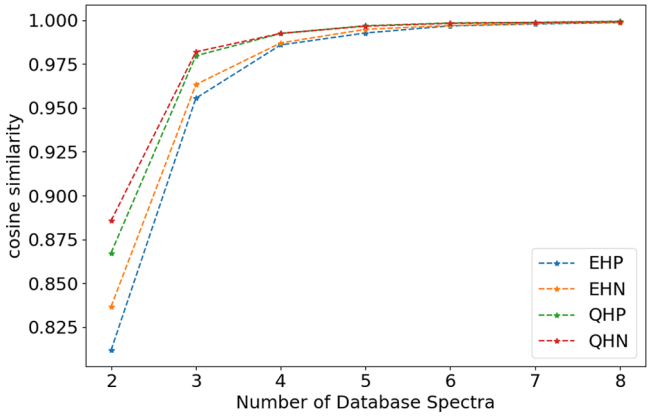



**Figure 5.** Average cosine similarity between ITP spectra and true spectra ($ITP_{sim}$) as more known, database spectra are used to generate the interpolation. Results are shown for NIST20 spectra on a Q-TOF and Elite Orbitrap with $[M + H]^+$ precursor (QHP and EHP, respectively) as well as with an $[M\text{-}H]^-$ precursor (QHM, EHM). As more database spectra are available ITP spectra quickly approach true spectra.



https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

negative score indicates a stronger match to a spectrum in the database.

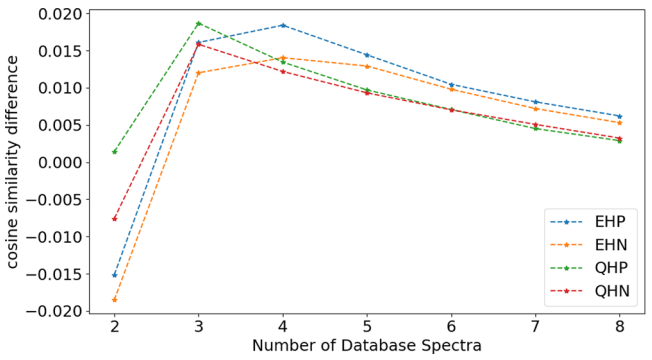



**Figure 6.** Average difference in cosine similarity of ITP spectra and database spectra to test spectra ($\Delta_{sim}$) as more spectra are available for interpolation. Results are shown for NIST20 spectra on a Q-TOF and Elite Orbitrap with $[M + H]^+$ precursor (QHP and EHP, respectively) as well as with an $[M\text{-}H]^-$ precursor (QHN, EHN). A positive $\Delta_{sim}$ indicates the interpolated spectra are closer matches than the known spectra used to generate the interpolations, while a negative value indicates the reverse.

With the exception of the QHP case, including only spectra at 10 and 45 eV is insufficient to produce ITP spectra that match test spectra better than the available database on average. By including just one intermediate spectrum, however, the ITP spectra offer a consistent advantage. Because the ITP spectra are mathematically determined, the interpolations can only include peaks that are present in the spectra provided to the model. This means the interpolations cannot contain spurious peaks but do require enough initial data—at least three spectra—to ensure all expected peaks are found. For molecule classes where known, anomalous fragmentation occurs, researchers must be deliberate when selecting the representative collision energies. Figure 6 plots the average $\Delta_{sim}$ value for the tested cases, showing that our method offers the most advantage in cases where limited data are available. The ITP spectra yield diminishing returns when more spectra are known for a molecule in the electronvolt range of interest. This likely happens because the increased availability of known spectra evenly dispersed through the range of electronvolts increases the chances of a known spectra being closer in electronvolts to the entries in the library. While the gains of this method are more modest when more data are available, including the ITP spectra still increases matching scores over the database spectra on average.

**Performs Consistently across Instruments and Precursor Types.** On the basis of the results of the previous section, we now test how much value is added to a subset database of three spectra by augmenting it with ITP spectra for all nine instrument and precursor types outlined in Table 1. Using the NIST20 libraries, we construct subset databases that contain only spectra at ∼10, 20, and 40 eV by identifying the closest spectrum for each available molecule within 5 eV of each target energy. All other spectra serve as unknown test spectra that can be searched for in the subset database. The number of molecules with spectra within the target electronvolt values for each instrument precursor pair are reported in Table 2.

For each instrument/precursor pair, we test identification of the held out test spectra against the associated database after augmenting it with interpolated spectra. For each instrument/precursor pair, Table 2 also reports the percentage of spectra for which the highest cosine similarity score matches the identity of the test spectrum when matching to the ITP spectra enhanced database (IM) and the percentage of spectra likewise correctly identified when matching to just subset database spectra (DBM).

For all cases, the percent of spectra correctly identified improved when using interpolation estimates. The $\Delta_{eV}$ and $ITP_{sim}$ results indicate that the interpolation is providing improvements in line with our expectation: test spectra at a collision energy $t_{eV}$ find a better match with an ITP spectrum within 1 or 2 eV of $t_{eV}$ on average in all cases. Note that interpolation does not strictly improve identification—a small proportion of spectra is correctly identified matching to just the subset database but is misidentified when ITP spectra are added. These instances are far outweighed by those in which

**Table 2. Summary of ITP Spectra Enhanced Database Performance across Instrument and Precursor Types: The Number of Molecules Available in NIST20 for the Given Instrument/Percursor Type Pairing (# mol.), the Average Highest Cosine Similarity between a Test Spectrum and and ITP Spectra Enhanced Database ($maxITP_{sim}$), the Average Highest Cosine Similarity between a Test Spectrum and the Subset Database ($max\Delta_{sim}$), the Percent of Test Spectra Accurately Identified by Their Best Match Either by Directly Matching against the Database (DBM) or Matching against ITP Spectra (IM), the Average Distance the Test Spectrum's Electronvolt and the Closest Match's Electronvolt ($\Delta_{eV}$), and the Percent of Spectra Correctly Identified without Interpolation That Were Incorrectly Identified with Interpolation (IF)**

| | # mol. | $mxITP_{sim}$ | $mx\Delta_{sim}$ | IM | DBM | $\Delta_{eV}$ | IF |
|---|---|---|---|---|---|---|---|
| QHP | 1995 | 0.97 | 0.029 | 95.3% | 88.1% | 0.84 | 0.6% |
| QHN | 128 | 0.97 | 0.021 | 96.5% | 94.3% | 1.4 | 0.1% |
| QNa | 30 | 0.97 | 0.023 | 99.5% | 95.7% | 1.3 | 0.5% |
| EHP | 10 059 | 0.96 | 0.026 | 87.3% | 76.0% | 0.65 | 1% |
| EHN | 4,105 | 0.98 | 0.035 | 87.8% | 78.3% | 1.1 | 1.5% |
| ENa | 589 | 0.93 | 0.020 | 100% | 100% | 1.3 | 0% |
| VHP | 394 | 0.95 | 0.028 | 93.1% | 85.6% | 0.92 | 0.4% |
| VHN | 6 | 0.97 | 0.021 | 100% | 98.2% | 1.4 | 0% |
| VNa | 22 | 0.87 | −0.019 | 100% | 100% | 1.3 | 0% |
| LHP | 3935 | 0.95 | 0.026 | 87.6% | 78.3% | 0.56 | 0.8% |
| LHN | 893 | 0.95 | 0.018 | 93.8% | 88.5% | 1.4 | 0.5% |
| LNa | 135 | 0.96 | 0.021 | 99.1% | 97.7% | 1.1 | 0% |



https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

identification is improved but should be kept in mind for researchers attempting to use interpolation with molecules that are known to have complicated spectra across collision energies.

*Is Robust to Similarity Metrics.* Our interpolation method is robust to alternative similarity metrics. Here we report our findings using a novel spectral entropy metric that outperforms 42 alternative similarity metrics including cosine similarity for MS/MS library matching.[38] We test the capacity for spectral interpolation to improve identification with entropy similarity using the four instrument precursor pairs with the most data (QHP, EHP, EHN, LHP). Results are reported in Table 3.

**Table 3. Summary of Entropy Similarity Metric Analysis: IM Indicates the Percent of Test Spectra That Were Accurately Identified by the Highest Cosine Similarity Match in ITP Augmented Databases, DMB Indicates the Percent of Spectra Accurately Identified Using Just the Subset Database, and IF Indicates the Percent of Spectra That Become Misidentified When ITP Spectra Are Added to a Subset Database**

| | IM | DBM | IF |
|---|---|---|---|
| QHP | 97.2% | 95.0% | 0.5% |
| EHP | 95.7% | 91.2% | 0.6% |
| EHN | 94.6% | 89.6% | 0.8% |
| LHP | 96.4% | 92.9% | 0.5% |

Entropy similarity increases identification accuracy across the board, and supplementing a subset database with interpolation leads to further improvement in all cases. For the EHP case, use of the entropy similarity metric and interpolation results in a 20% increase in identification accuracy over cosine similarity demonstrating the pronounced effect the choice of methodology can have on performance.

*Strengthens Spectra Matching across Instrument Types.* There is a set of molecules in NIST20 with $[M + H]^+$ precursors and spectra on both the Agilent Q-TOF and Elite Orbitrap. We test the use of ITP spectra for molecule identification across instruments using this overlapping data set of 892 molecules. Results using both cosine similarity and entropy similarity for matching are shown in Table 4.

For both QHP and EHP, the use of ITP spectra improves cross-database identification, though the improvement is less than that observed in self-comparisons. For instance, when the entropy similarity metric is used, only a 1% improvement is seen in accuracy. This suggests there may be changes in the relationships between peaks and collision energies across instruments that limit the accuracy of interpolation approximations from one instrument to another.

**Table 4. Summary of Cross-Instrument Analysis: Percent of Test Spectra on the Test Instrument (Test Inst.) Accurately Identified Using the Reference Spectra from Another Instrument (DB Inst.) with the Given Comparison Metric Either Directly Matching against the Database (DM) or Matching against Interpolated Spectra (IM). The Column $\Delta_{eV}$ Reports the Average Difference between the Electronvolt of the Test Spectrum and the Electronvolt of the Closest Matching Interpolated Database Spectrum**

| Test Inst. | DB Inst. | Metric | IM | DM | $\Delta_{eV}$ |
|---|---|---|---|---|---|
| QHP | EHP | cosine | 89% | 85% | 4.7 |
| EHP | QHP | cosine | 84% | 82% | −3.4 |
| QHP | EHP | entropy | 93% | 92% | 4.8 |
| EHP | QHP | entropy | 91% | 90% | −4.0 |

An interesting result is also seen in the average $\Delta_{eV}$ value for the cross-instrument comparisons. There is a consistent shift with both similarity metrics in the collision energies of the closest matching spectra that suggests the Agilent Q-TOF spectra are most similar to the Elite Orbitrap spectra at ∼4 eV higher collision energies. A similar result was found experimentally in a systematic comparison of Q-TOF and Orbitrap HCD MS/MS spectra at varying collision energies for peptide MS/MS spectra.[39] Our analysis suggests such a shift may hold more generally across a diverse set of chemical classes.

*Works at High Collision Energies.* While most of the NIST20 spectra data lie in the range of 10−40 eV, there are a number of spectra taken at higher electronvolts on the Orbitrap instruments. We test our method with the same procedure at higher collision energies by constructing databases with three spectra at ∼50, 70, and 90 eV on the instrument/precursor pairs EHP, EHN, and LHP. The data sets at these energies will be limited due to our focus on small molecules (100−500 Da) and will generally be more representative of the heavier end of that range. Recall that the LHP trials will be using approximated electronvolt values as described in the Methods section. Results for the identification of test spectra using cosine similarity are shown in Table 5.

Again, the use of interpolation improves identification accuracy in all cases, demonstrating that this method can be effective at a range of collision energies. In fact both the average interpolation accuracy and identification performance are increased with the higher collision energy databases.

*Works with Different Workflows.* To test the robustness of our method, we also sourced spectra from MassBank. In particular, Dr. Nikolaos Thomaidis at the University of Athens has submitted a large number of ESI-LC-QTOF spectra from a Bruker maXis Impact using a $[M + H]^+$ precursor.[40] For 549 molecules with molecular weight between 100 and 500 Da, spectra at collision energies (10, 20, 30, 40, 50) are available. We construct a reduced database with three spectra for each molecule at 10, 30, and 50 eV and test identification of the 20 and 40 eV spectra using interpolation. Results for spectra identification using cosine similarity are in Table 6.

Interpolation improves the identification accuracy on the MassBank data, though the improvement is less than what we see in the NIST20 Q-TOF data, and the IF percent is also higher. The reduced improvement may be due to the increased distance between available energies in the database, making the interpolations less accurate. This hypothesis is supported by the lower average $mxITP_{sim}$ value.

## CONCLUSION

Spectral interpolation provides a quick, robust method to improve small-molecule identification with MS/MS reference matching from limited data sets. We found interpolation to consistently improve the percent of spectra correctly identified across instrument and precursor types with only three database spectra per molecule. The method offers the most benefit for instances where only a few spectra are available with diminishing returns as more database spectra are added. Augmenting databases using spectral interpolation offers a



https://doi.org/10.1021/acs.jcim.2c00620
J. Chem. Inf. Model. 2022, 62, 3724−3733

Table 5. Summary of High Collision Energy Analysis on Orbitrap Instruments: The Number of Molecules Available in NIST20 for the Given Instrument/Percursor Type Pairing (# mol.), the Average Highest Cosine Similarity between a Test Spectrum and and ITP Spectra Enhanced Database ($maxITP_{sim}$), the Average Highest Cosine Similarity between a Test Spectrum and the Subset Database ($max\Delta_{sim}$), the Percent of Test Spectra Accurately Identified by Their Best Match Either by Directly Matching against the Database (DBM) or Matching against ITP Spectra (IM), the Average Distance the Test Spectrum's Electronvolt and the Closest Match's Electronvolt ($\Delta_{eV}$), and the Percent of Spectra Correctly Identified without Interpolation That Were Incorrectly Identified with Interpolation (IF)

| | # mol. | $maxITP_{sim}$ | $max\Delta_{sim}$ | $\Delta_{eV}$ | IM | DBM | IF |
|---|---|---|---|---|---|---|---|
| EHP | 3348 | 0.99 | 0.047 | 0.77 | 96.1% | 84.2% | 0.3% |
| EHN | 1452 | 0.99 | 0.033 | 0.76 | 98.3% | 95.2% | 0.4% |
| LHP | 3935 | 0.96 | 0.038 | 0.28 | 96.4% | 92.8% | 0.5% |

Table 6. Summary of Analysis Using Spectra from MassBank: The Number of Molecules Available in NIST20 for the Given Instrument/Percursor Type Pairing (# mol.), the Average Highest Cosine Similarity between a Test Spectrum and and ITP Spectra Enhanced Database ($maxITP_{sim}$), the Average Highest Cosine Similarity between a Test Spectrum and the Subset Database ($max\Delta_{sim}$), the Percent of Test Spectra Accurately Identified by Their Best Match Either by Directly Matching against the Database (DBM) or Matching against ITP Spectra (IM), the Average Distance the Test Spectrum's Electronvolt and the Closest Match's Electronvolt ($\Delta_{eV}$), and the Percent of Spectra Correctly Identified without Interpolation That Were Incorrectly Identified with Interpolation (IF)

| # mol. | $maxITP_{sim}$ | $max\Delta_{sim}$ | $\Delta_{eV}$ | IM | DBM | IF |
|---|---|---|---|---|---|---|
| 549 | 0.94 | 0.039 | −0.12 | 91.2% | 89.5% | 2.6% |

transparent method for improving identification where inspection of both spectra estimates and predicted relationships between peaks and collision energy is straightforward as shown in Figure 2. The methodology is agnostic to the choice of comparison metric and can be used in any workflow where spectra at multiple collision energies are available.

**Data and Software Availability.** The Supporting Information for this article contains a set of Python functions that can be applied to mass spectrometry data to generate the interpolations described. A sample set of three Q-TOF spectra of capsaicin is included. Our benchmarking work was performed as described using data from NIST20 and can be replicated by licensed users. The methods described here can be applied to any available mass spectrometry data set.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00620.

Python code file containing primary functions and examples, Sample data pickle file to run code (ZIP)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Danielle Ciesielski** − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0003-0190-4530; Email: danielle.ciesielski@pnnl.gov

### Authors

**Ethan King** − *Computing and Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**Richard Overstreet** − *Signature Science and Technology Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**Julia Nguyen** − *Computing and Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00620


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the Pacific Northwest National Laboratory (PNNL) Laboratory Directed Research and Development Program and is a contribution of the *m/q* Initiative. A portion of the research was performed using resources available through Research Computing at PNNL. PNNL is a multiprogram national laboratory operated by Battelle for the Department of Energy under Contract No. DE-AC05-76RLO 1830.

## ■ REFERENCES

(1) Reilly, C. A.; Crouch, D. J.; Yost, G. S.; Fatah, A. A. Determination of capsaicin, dihydrocapsaicin, and nonivamide in self-defense weapons by liquid chromatography-mass spectrometry and liquid chromatography-tandem mass spectrometry. *Journal of Chromatography A* **2001**, *912*, 259−267.

(2) Krauss, M.; Singer, H.; Hollender, J. LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* **2010**, *397*, 943−951.

(3) Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods* **2013**, *10*, 755−758.

(4) Lebedev, A. T.; Polyakova, O. V.; Mazur, D. M.; Artaev, V. B. The benefits of high resolution mass spectrometry in environmental analysis. *Analyst* **2013**, *138*, 6946−6953.

(5) Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripolles Vidal, C.; Hollender, J. Strategies to characterize polar organic contamination in wastewater: Exploring the capability of high resolution mass spectrometry. *Environ. Sci. Technol.* **2014**, *48*, 1811−1818.

(6) González-Mariño, I.; Gracia-Lor, E.; Bagnati, R.; Martins, C. P.; Zuccato, E.; Castiglioni, S. Screening new psychoactive substances in urban wastewater using high resolution mass spectrometry. *Anal. Bioanal. Chem.* **2016**, *408*, 4297−4309.

(7) Pasin, D.; Cawley, A.; Bidny, S.; Fu, S. Current applications of high-resolution mass spectrometry for the analysis of new psycho-

active substances: a critical review. *Anal. Bioanal. Chem.* **2017**, *409*, 5821−5836.
(8) Mogollón, N. G. S.; Quiroz-Moreno, C. D.; Prata, P. S.; de Almeida, J. R.; Cevallos, A. S.; Torres-Guiérrez, R.; Augusto, F. New advances in toxicological forensic analysis using mass spectrometry techniques. *J. Anal. Methods Chem.* **2018**, *2018*, 1.
(9) Merkley, E. D.; Wunschel, D. S.; Wahl, K. L.; Jarman, K. H. Applications and challenges of forensic proteomics. *Forensic Science International* **2019**, *297*, 350−363.
(10) Brown, H. M.; McDaniel, T. J.; Fedick, P. W.; Mulligan, C. C. The current role of mass spectrometry in forensics and future prospects. *Royal Society of Chemistry Analytical Methods* **2020**, *12*, 3967−4102.
(11) Fabregat-Safont, D.; Sancho, J. V.; Hernández, F.; Ibáñez, M. The key role of mass spectrometry in comprehensive research on new psychoactive substances. *Journal of Mass Spectrometry* **2020**, *56*, e4673.
(12) Gilbert, N.; Antonides, L. H.; Schofield, C. J.; Costello, A.; Kilkelly, B.; Cain, A. R.; Dalziel, P. R.; Horner, K.; Mewis, R. E.; Sutcliffe, O. B. Hitting the Jackpot - development of gas chromatography-mass spectrometry (GC-MS) and other rapid screening methods for the analysis of 18 fentanyl-derived synthetic opioids. *Drug Testing and Analysis* **2020**, *12*, 798−811.
(13) Liebal, U. W.; Phan, A. N. T; Sudhakar, M.; Raman, K.; Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10*, 243.
(14) Weissberg, A.; Dagan, S. Interpretation of ESI(+)-MS-MS spectra-Towards the identification of "unknowns. *Int. J. Mass Spectrom.* **2011**, *299*, 158−168.
(15) Janesko, B. G.; Li, L.; Mensing, R. Quantum Chemical Fragment Precursor Tests: Accelerating de novo annotation of tandem mass spectra. *Anal. Chim. Acta* **2017**, *995*, 52−64.
(16) Steckel, A.; Schlosser, G. An Organic Chemist's Guide to Electrospray Mass Spectrometric Structure Elucidation. *Molecules* **2019**, *24*, 611.
(17) Colby, J. M.; Thoren, K. L.; Lynch, K. L. Optimization and Validation of High-Resolution Mass Spectrometry Data Analysis Parameters. *Journal of Analytical Toxicology* **2017**, *41*, 01−05.
(18) Deutsch, E. W.; Perez-Riverol, Y.; Chalkley, R. J.; Wilhelm, M.; Tate, S.; Sachsenberg, T.; Walzer, M.; Käll, L.; Delanghe, B.; Böcker, S.; Schymanski, E. L.; Wilmes, P.; Dorfer, V.; Kuster, B.; Volders, P.-J.; Jehmlich, N.; Vissers, J. P.; Wolan, D. W.; Wang, A. Y.; Mendoza, L.; Shofstahl, J.; Dowsey, A. W.; Griss, J.; Salek, R. M.; Neumann, S.; Binz, P.-A.; Lam, H.; Vizcaíno, J. A.; Bandeira, N.; Röst, H. Expanding the use of spectal libraries in proteomics. *J. Proteome Res.* **2018**, *17*, 4051−4060.
(19) Oberacher, H.; Sasse, M.; Antignac, J.-P.; Guitton, Y.; Debrauwer, L.; Jamin, E. L.; Schulze, T.; Krauss, M.; Covaci, A.; Caballero-Casero, N.; Rousseau, K.; Damont, A.; Fenaille, F.; Lamoree, M.; Schymanski, E. L. A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environ. Sci. Eur.* **2020**, *32*. DOI: 10.1186/s12302-020-00314-9
(20) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **2018**, *37*, 513−532.
(21) Yang, X.; Neta, P.; Stein, S. E. Quality control for building libraries from electrospray ionization tandem mass spectra. *Anal. Chem.* **2014**, *86*, 6393−6400.
(22) Wallace, W. E.; Ji, W.; Tchekhovskoi, D. V.; Phinney, K. W.; Stein, S. E. Mass spectral library quality assurance by inter-library comparison. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 733−738.
(23) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **2010**, *45*, 703−714.
(24) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **2010**, *11*, 148−160.
(25) Kangas, L. J.; Metz, T. O.; Isaac, G.; Schrom, B. T.; Ginovska-Pangovska, B.; Wang, L.; Tan, L.; Lewis, R. R.; Miller, J. H. In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics* **2012**, *28*, 1705−1713.
(26) Shen, H.; Dührkop, K.; Böcker, S.; Rousu, J. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics* **2014**, *30*, i157−i164.
(27) Allen, F.; Greiner, R.; Wishart, D. Competitive Fragmentation Modeling of ESI-MS/MS spectra for metabolite identification. *Metabolomics* **2015**, *11*, 98.
(28) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS* **2015**, *112*, 12580−12585.
(29) Ásgeirsson, V.; Bauer, C. A.; Grimme, S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chemical Science* **2017**, *8*, 4879−4895.
(30) Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; et al. ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries. *Anal. Chem.* **2019**, *91*, 4346−4356.
(31) Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D. S. CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites* **2019**, *9*, 72.
(32) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16*, 299−302.
(33) Ruttkies, C.; Neumann, S.; Posch, S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics* **2019**, *20*. DOI: 10.1186/s12859-019-2954-7
(34) Wei, J. N.; Belanger, D.; Adams, R. P.; Sculley, D. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Central Science* **2019**, *5*, 700−708.
(35) Zohora, F. T.; Rahman, M. Z.; Tran, N. H.; Xin, L.; Shan, B.; Li, M. DeepIso: a deep learning model for peptide feature detection from LC-MS map. *Nature Scientific Reports* **2019**, *9*. DOI: 10.1073/pnas.1509788112
(36) Cao, L.; Guler, M.; Tagirdzhanov, A.; Lee, Y.; Gurevich, A.; Mohimani, H. MolDiscovery: Learning Mass Spectrometry Fragmentation of Small Molecules. *bioRxiv* 2020. DOI: 10.1101/2020.11.28.401943
(37) NIST/NIH/EPA *Mass spectral library, standard reference database;* Standard Reference Data Program, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020.
(38) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **2021**, *18*, 1524−1531.
(39) Szabó, D.; Schlosser, G.; Vékey, K.; Drahos, L.; Révész, Á. Collision energies on QToF and Orbitrap instruments: How to make proteomics measurements comparable? *Journal of Mass Spectrometry* **2021**, *56*, e4693.
(40) Alygizakis, N. A.; Gago-Ferrero, P.; Hollender, J.; Thomaidis, N. S. Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater. *Journal of Hazardous Materials* **2019**, *361*, 19−29.