Harnessing data science to improve molecular structure elucidation from tandem mass spectrometry

Structural Chemistry (2023) 34:1935–1950
https://doi.org/10.1007/s11224-023-02192-2

RESEARCH

Harnessing data science to improve molecular structure elucidation from tandem mass spectrometry

E. Harris[1] · Lilian Gasser[1] · M. Volpi[1] · F. Perez-Cruz[1] · Saša Bjelić[2] · G. Obozinski[1]

Received: 3 April 2023 / Accepted: 17 May 2023 / Published online: 10 June 2023

Abstract

Compound structural identification for non-targeted screening of organic molecules in complex mixtures is commonly carried out using liquid chromatography coupled to tandem mass spectrometry (UHPLC-HRMS/MS and related techniques). Instrumental developments in recent years have increased the quality and quantity of data available; however, using current data analysis methods, structures can be assigned to only a small fraction of compounds present in typical mixtures. We present a new data analysis pipeline, "MSEI", that harnesses data science methodologies to improve structural identification capabilities from tandem mass spectrometry data. In particular, feature vectors for fingerprint calculation are found directly from tandem mass spectra, strongly reducing computational costs, and fingerprint comparison uses an optimised methodology accounting for uncertainty to improve distinction between matching and non-matching compounds. MSEI builds on the identification of a small number of compounds through current state-of-the-art data analysis on UHPLC-HRMS/MS measurements and uses targeted training and tailored molecular fingerprints to focus identification to a particular molecular space of interest. Initial compound identifications are used as training data for a set of random forests which directly predict a custom 75-digit molecular fingerprint from a vectorised MS/MS spectrum. Kendrick mass defects (KMDs) for peaks as well as "lost" fragments removed during fragmentation were found to be useful information for fingerprint prediction. Fingerprints are then compared to potential matches from the PubChem structural database using Euclidean distance, with fingerprint digit weights determined using an SVM to maximise distance between matching and non-matching compounds. Potential matches are additionally filtered for hydrophobicity based on measured retention time, using a newly developed machine learning method for retention time prediction. MSEI was able to correctly assign $>$ 50% of structures in a test dataset and showed $>$ 10% better performance than current state-of-the-art methods, while using an order of magnitude less computational power and a fraction of the training data.

Keywords Structural identification · Tandem mass spectrometry · Machine learning · Liquid chromatography

Introduction

Identification of organic molecules in complex mixtures is a key question facing many scientific fields ranging from metabolomics to renewable energy, with samples from diverse sources including plants, blood, aerosols, soil, or biofuels. Ultra-high performance liquid chromatography coupled to high-resolution tandem mass spectrometry (UHPLC-HRMS/MS) is a widely used, sensitive method for analysis of these complex mixtures. This method first involves separation of the mixture by retention time using UHPLC, followed by high resolution mass spectrometric analysis of the compounds in each retention time window—the "MS1" mass spectrum. Furthermore, a "parent ion" peak from each MS1 is selected for fragmentation (usually the most intense peak), followed by mass spectrometric analysis of the fragmentation products—the "MS2" or "fragment" mass spectrum. Instrumental developments over the past decades leading to high mass resolving power mean that the exact sum formula for hundreds or thousands of compounds may be detected in a single sample [1, 2]; however, molecular structures can only be confidently assigned to a very small subset of these compounds [3, 4].

✉ E. Harris
eliza.harris@sdsc.ethz.ch

[1] Swiss Data Science Center, ETH Zürich and EPFL, Zürich, Switzerland

[2] Laboratory for Bioenergy and Catalysis, Paul Scherrer Institut, Villigen 5232, Switzerland

Retention time reflects molecular structure and functional groups; however, retention time is poorly generalizable, and standards are not often readily available [5, 6]. Retention time is determined by partitioning between the hydrophobic stationary phase and the hydrophilic eluent and thus reflects a compound's hydrophobicity [7], most commonly described using the n-octanol-water partition coefficients (XlogP). Currently, retention time and hydrophobicity are not widely used in molecular structure annotation. The development of machine learning algorithms to predict retention time and hydrophobicity based on molecular structure mean that retention time will be increasingly useful as an independent filter to complement mass spectral information and verify structural assignments [8–11].

Mass spectra and fragment mass spectra (MS1 and MS2) depend on molecular structure and functional groups and therefore contain information which can be used to assign compound identities. Fragment presence and intensity depend on ionization efficiency, collisional activation, energy transfer, kinetic shift, and dissociation, and can involve multi-step fragmentation and rearrangment; therefore, interpreting and predicting fragmentation spectra is challenging [12–14]. Similar fragment patterns generally indicate similar molecules or molecular subunits; however, to converse is not true: high molecular similarity does not necessarily lead to similar fragmentation patterns [15]. Some functional groups lead to characteristic neutral "lost fragments", while others do not cause characteristic losses; furthermore, some losses, such as dehydration, can suppress additional losses [13]. Lost fragments can reflect molecular similarity better than detected fragments through their direct relationship to functional groups [16]. Although tandem mass spectra contain a wealth of information reflecting molecular structure, complex computational methods are needed to relate spectra to structure.

Numerous computational tools are available to process tandem mass spectral data and assign compound identities (see [4] for a detailed review of methodologies). Many commonly used and commercial methods compare measured fragmentation spectra to spectral databases. However, spectral databases have limited numbers of compounds—orders of magnitude less than molecular structure databases. Spectral matching is therefore often complemented by prediction of *in silico* fragmentation spectra for potential matches; this method is however computationally expensive, challenging and uncertain [4, 17]. Leading methodologies therefore mainly use measured fragmentation spectra to predict molecular substructures and construct a "fragmentation tree" [16], which is used to generate a molecular fingerprint that can be compared to candidate structures or molecular families [18, 19]. Molecular substructures and families can additionally be investigated using the "Kendrick Mass Defect" [20, 21]. KMDs are used to group compounds based on the sum formula by selecting for repeating subunits known as bases, which can represent particular functional groups. Despite many advances in computational methodologies, state-of-the-art approaches are computationally intensive and can correctly assign $< 40\%$ of molecular structures [4, 18], showing the potential for further improvements.

In this study, we present a data-driven approach to structural identification ("MSEI": **M**olecular **S**tructure **E**lucidation using **I**ntegrated data science approaches), which uses integrated data science methodologies to predict molecular fingerprints directly from tandem mass spectra and to optimise the comparison of predicted fingerprints to potential structural matches. MSEI uses a small number of initial compound identifications to train an ensemble of random forests to predict a tailored molecular fingerprint based on a vector representation of key features from UHPLC-HRMS/MS measurements. Because MSEI is tailored to a subset of compounds of interest and directly predicts fingerprints from spectra, the approach requires much less computational power than equivalent methods, allowing both training and prediction to be conducted on a standard laptop. The predicted fingerprint is compared to potential matches using a weighted distance metric, with weights learnt using a support vector machine, to assign unknown structures with performance exceeding currently available methods. Retention time, predicted using a newly developed machine learning method for estimation of hydrophobicity, is further used to filter potential structural matches. This manuscript presents the approach as a "proof of concept", illustrating data science techniques that could be used individually or together to improve current metholodogies for structural identification or as the basis for a stand-alone data analysis pipeline.

Methods

UHPLC-HRMS measurements

Prior to analysis, all samples (from the datasets in Table 1) were filtered using PTFE filters (0.22 μm). Blank runs were performed between each analysis and used for background subtraction. ^{13}C Vanillin ($^{13}C_6C_2H_8O_3$, 10 μM) was added to each sample before analysis to check the internal instrument performance and measurement stability. The injection volume was 1 μL. The analysis was performed using a Thermo Scientific Dionex^{TM} Ultimate 3000 Series RS system (Thermo Scientific^{TM}, Switzerland) consisting of a pump, a column compartment, and an auto-sampler. The chromatographic separation of compounds was carried out with the use of Accucore^{TM} RP-MS column (150 mm × 2.1 mm, particle size 2.6 μm), from Thermo Scientific^{TM}, with a Uniguard pre-column (Accucore RP-MS Defender Guards; 2.1 × 10 mm, 2.6 μm). The following gradients were used

Table 1 Summary of datasets used in this study. "# of compounds" describes the total number of compounds assigned a sum formula by Compound Discover. "# of compounds identified" is the number of compounds assigned a molecular structure by CompoundDiscoverer. "Well-matched compounds" are defined as having an mz-cloud match quality of ≥ 0.85 determined by CompoundDiscoverer. "Total # of MS2 spectra" is the total number of MS2 spectra collected, while "# of valid MS2 spectra" is the number of MS2 spectra identified as relating to a compound and assigned a sum formula, thus filtering out spectra which appear to be invalid (e.g. instrumental noise). Datasets are described further in [25, 26]

Name	DiAcids	Fallopia	Ruthenium
Description	Mixture of dicarboxylic acid standards	Compound mixture from catalytic hydrothermal near-critical liquefaction of Fallopia Japonica [25]	Compound mixture from Ruthenium catalyst at different stages during catalytic gasification of organic wastes [26]
# of .mzML files	12	39	22
# of compounds	13	947	1013
# of compounds identified	13	293	76
# of well-matched compounds	13	124	17
Min. molecular weight	104.0	58.0	114.1
Max. molecular weight	285.2	686.5	1493.1
Total # of MS2 spectra	10486	41280	44688
# of valid MS2 spectra	200	19750	1033
# of averaged MS2 spectra	63	4607	465
Av. # peaks per MS2 spectrum	22	19	58
Av. # peaks per av. MS2 spectrum	14	18	40

for mobile phase A (1 vol.% MeOH, 1 vol.% acetonitrile and 0.02 vol.% formic acid in water) and mobile phase B (100 vol.% MeOH): 1 % B (0–1 min) 1 to 99% B (1–6 min), 99% B (6–8 min), followed by an equilibration step and 99 to 1% B (8–8.2 min), 1% B (8.2–10 min). The flow rate was set to 0.7 m$Lmin^{-1}$, and the temperature of the column was kept constant at 323 K. A heated electrospray ionization source (H-ESI, 3 kV spray voltage) was used for the ionization of the analytes in positive and negative modes. Data acquisition was performed with a Thermo Scientific^{TM} Q-Exactive^{TM} hybrid quadrupole-orbitrap mass spectrometer controlled by the Xcalibur 4.1 software. Mass spectra were acquired in full scan mode with an isolation window of 1 m/z from 50 to 750 m/z. The resolution was 70,000 at m/z = 200. Raw mass spectral data files were collected in triplicate, including the blanks between each run.

Data preprocessing

The UHPLC-HRMS data was imported into the CompoundDiscoverer^{TM} 3.2 software (Thermo Scientific^{TM}, Switzerland) and processed with standard settings except for mass tolerance (set to 2.5 ppm). Chromatographic peaks detected in one of the input files but missing in others were checked by the *Fill Gaps* option. The sum formula was predicted based on exact mass and isotopic patterns and evaluated against MS/MS spectra. Only features yielding formulas available in ChemSpider database were used [22]. The identity of the compounds was determined whenever possible with mzCloud library of fragmentation spectra (MS/MS) [23], to be used as training data for the MSEI method.

The initial compound assignment with CompoundDiscoverer^{TM} produces a data file containing a summary of all unique compounds present in the spectral dataset, including compounds with unknown structures. The purpose of this initial processing step is to (i) to identify "valid" spectra based on peak patterns, (ii) to have an initial assignment of sum formulas, and (iii) to provide initial identification for a subset of compounds, to be used as training and testing data. Key information taken from the compound identification input file is compound name, sum formula, molecular weight, and the match quality determined by Compound Discover (for unknown compounds, only the sum formula and molecular weight are given). The output of any similar initial spectral processing software providing this key information could be used.

Further data preprocessing is described in detail in the SI text (Section 1); only a brief overview of key points is given here. For all identified compounds, the `PubChemPy` package [24] was used to retrieve the isomeric and canonical SMILES, the IUPAC name, and the hydrophobicity (XlogP). UHPLC-HRMS/MS data (.mzML format) for the three datasets shown in Table 1 was imported into Python and matched to compound information from CompoundDiscoverer^{TM}. Sum formulas for all MS1 and MS2 peaks in all datasets were determined.

The next stage of the preprocessing procedure involved two parts: (i) establishing a method to determine whether two MS2 spectra "match", i.e. represent the same structure,

(ii) applying this method to identify and aggregate matching spectra. To determine the similarity of two spectra, a radial basis function kernel is used to compare the measured mass-to-charge ratios (mz) of all peaks in the two spectra, similar to the approached used by [27] (approach described in detail in the SI text (Sections 1 and 1.3):

$$S_{1,2} = \sum_{i=1}^{I} \sum_{j=1}^{J} \exp\left(-\frac{1}{8} \frac{\left(\frac{mz_i - mz_j}{\frac{mz_i + mz_j}{2}} \times 10^6 \right)^2}{\sigma_{mz}} \right) \quad (1)$$

where $S_{1,2}$ is the *similarity index* for Spectrum 1 (i peaks) and Spectrum 2 (j peaks) where mz is the mass of each peak in the spectra. σ_{mz} is the estimated standard deviation in peak mass for repeated measurements of the same mass, expressed in ppm, and $\frac{mz_i - mz_j}{(mz_i + mz_j)/2 \times 10^6}$ expresses the difference between peak masses in ppm. σ_{mz} was empirically determined to be 1.4 ppm or 0.0002 amu, based on the repeatability of peak mass for peaks corresponding to the same sum formula in the DiAcids dataset, described in detail in the SI text (Sections 1 and 1.3).

The similarity index will approach 0 for spectra with no peaks in common, while for identical spectra, the similarity index will equal the number of peaks. The normalised similarity index (S^{norm}) can be calculated to account for varying numbers of peaks in different spectra:

$$S^{\text{norm}}_{1,2} = \frac{S_{1,2}}{(S_{1,1})^{0.5}(S_{2,2})^{0.5}} \quad (2)$$

where $S_{1,1}$ is the similarity of spectrum 1 with itself, thus giving the number of peaks in spectrum 1, and analogously for $S_{2,2}$. As a further indicator of similarity between two spectra, the number of peaks in common between the two spectra is determined, based on a threshold of $5 \times \sigma_{mz}$ to determine if two peaks represent the same sum formula; this threshold is a robust measure of similar and different peaks as shown in Fig. S1. Unlike in [27], peak intensity is not used in the calculation of peak similarity, as intensity and relative intensity of peaks are much less reproducible than mass (see Fig. S5); peak masses are sufficient to distinguish between similar spectra in high resolution MS/MS.

An intercomparison of spectra with a high match quality, as labelled by CompoundDiscoverer, was carried out to characterise the spectral similarity index as a labelling tool for matching and non-matching spectra (see SI text, Sections 1 and 1.3). Significant overlap between normalised similarity index distributions for matching and non-matching spectra was found (Fig. S2). High spectral similarity was found to indicate at least moderately similar molecular structures (Fig. S4); however, high molecular similarity did not guarantee high spectral similarity. Ionization voltage and analyte concentration did not play a clear role in determining spectral variability. A threshold to classify matching and non-matching spectra based on both the normalised similarity index and the number of peaks in common between two spectra (Fig. S3) was defined as:

$$S_{\text{threshold}} = S^{\text{norm}}_{max} - a * n(\text{peaks in common}) \quad (3)$$

where S^{norm}_{max} is the intercept and corresponds approximately to the maximum normalised similarity possible for non-matching spectra (0.25), and a is the slope. An optimum value for a (0.01) was found empirically in order to minimise false positive matches while capturing most true positive matches (Fig. S3).

Spectra from the same molecular structure should be combined to: (i) identify and remove "noise" peaks which do not represent a true, reproducible fragment substructure, and (ii) aggregate true peaks to increase confidence in molecular weight and relative intensity, particularly for low intensity peaks. Equation 3 was used to identify groups of matching spectra, and clusters of peaks found across multiple spectra were averaged to produce a single spectrum from each group of matching spectra (SI text, Sections 1 and 1.3). Spectral averaging led to a reduction of 50–75% in the number of MS2 spectra and a reduction of one-third in the average number of peaks per MS2 spectrum (Table 1).

The final stage of preprocessing involved quality checking the data. All averaged spectra assigned to a particular compound were intercompared. Compounds with more than one non-matching group of averaged spectra (e.g. spectra 1–3 for compound Y match each other, as do spectra 4–5; however, 1–3 do not match 4–5) are flagged (Fig. S6), and only the largest group of matching spectra are used for further calculations. Retention time and hydrophobicity (XLogP) were also compared for all well-identified compounds (Fig. S7). We use hydrophobicities predicted using a graph neural network combined with multitask learning—whereby helper tasks are added to the model, such as related molecular properties, to improve generalization—as described in [11]. A sigmoid fit was used to estimate expected XLogP for all compounds based on measured retention time. All compound identifications where expected XLogP, predicted from retention time, did not match the compound XLogP were flagged as representing potential structural isomers.

Molecular fingerprints

Molecular fingerprints can be used to summarise molecular structure information in a machine readable format; however, there is no single "optimal" set of keys to generalise across all possible molecular structures [28]. We therefore created a molecular fingerprint tailored to the molecular structure space of interest in this study incorporating nine numeric identifiers (Table 2), in order to maximise discrimination between similar

Table 2 Description of the 75 digits used in the molecular fingerprint to create a machine-readable description of molecular structures. "Minimum" and "maximum" refer to the minimum and maximum values found across the 382 identified molecular structures in the three datasets

	Name	Description	Minimum	Maximum
i	**n_Al_COO**	Number of aliphatic carboxyl groups	0	3
ii	**n_Ar_COO**	Number of aromatic carboxyl groups	0	2
iii	**n_Al_OH**	Number of aliphatic hydroxyl groups	0	5
iv	**n_Ar_OH**	Number of aromatic hydroxyl groups	0	3
v	**n_OH**	Number of hydroxyl groups	0	5
vi	**n_CO**	Number of carbonyl O groups, excluding COOH	0	2
vii	**n_ArRing**	Number of aromatic rings	0	4
viii	**longestchain**	Number of carbons in the longest unbranched carbon chain	1	20
ix	**backbone**	Length of the carbon backbone	1	26
	MACCS	66 MACCS keys with 95:5% split between major and minor values	0	1

molecules while avoiding computation for models of fingerprint digits which contain no information. Unlike most molecular fingerprints [18, 29, 30], the nine digits are not binary but represent counts of certain functional units and therefore contain much more information than typical binary fingerprints. For each molecular structure assigned by CompoundDiscoverer, identifiers (i–vi) were found using the `Chem.Fragments` module from the `rdkit` package in Python [31]; identifier (vii) was found with the `Chem.rdMolDescriptors` module from `rdkit`; and (viii and ix) were directly found from the canonical SMILES of the molecule based on repeated C units.

In addition to the nine selected count digits, all Molecular ACCess System (MACCS) keys that had at least a 95:5% split between major and minor values for the 382 unique molecular structures identified in the three datasets were used as digits of the molecular fingerprint. A total of 66 of the available 167 MACCS keys fulfilled this condition.[1] MACCS keys were generated using the function `GetMACCSKeysFingerprint` from the `Chem.rdMolDescriptors` module in `rdkit`. A total of 75-digit molecular fingerprints were determined for the 382 molecular structures assigned by CompoundDiscoverer.

Conversion of MS1 and MS2 spectra to an input data vector

Key information contained in each MS1-MS2 spectra was converted to a vector, to simplify input to machine learning algorithms for molecular fingerprint prediction. Peak rankings were used to improve the comparability of input data vectors for different spectra, in order to compare fragment ions describing similar molecular structure information. Based on the priority ranking, key peak information (mz, intensity, relative intensity, number of C, H, O, N atoms in the peak sum formula, number of C, H, O, N atoms lost from the parent ion to form the fragment, KMD values for 9 key bases (see "Structural information represented by KMDs and lost fragments")) was concatenated for peaks with priorities 1 to 6 for each spectrum. Intensity was not directly used to rank peaks, as it is poorly reproducible between different spectra (Fig. S5); therefore, intensity bands were used. For each of the averaged MS1-MS2 spectra from the three datasets, peaks were ranked in priority for vectorisation (see example in Table S1):

- Peaks with between 10% and 100% of the maximum intensity in the spectrum were ordered according to the mass of the lost fragment and for relative intensity bands of 1–10%, 0.1–1%, 0.01–0.1%, etc.
- A combined priority ranking based on intensity group and lost fragment mass was found.
- The parent ion peak was given priority of 1.
- All MS2 peaks with no assigned sum formula were not given a priority.

The combined peak ranking procedure ensures that peaks within a similar intensity range and representing a similar fragment loss from the parent ion are compared, as these peaks likely contain similar structural information, rather than simply comparing peaks with the same mass.

In addition to the information from the highest priority peaks, general information regarding lost fragment presence and absence was added to the input data vector representing each spectrum. For the 10 key lost fragments (see "Structural information represented by KMDs and lost fragments"), the presence (true/false), as well as the mz, intensity, and the relative intensity of the remaining fragment,

[1] MACCS keys 50, 53, 54, 57, 66, 72, 74, 76, 82, 83, 89, 90, 91, 92, 93, 95, 96, 98, 99, 100, 101, 104, 105, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 121, 123, 125, 127, 128, 129, 131, 132, 136, 137, 139, 140, 143, 144, 146, 147, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 165.

were added to the input data vector for each spectrum. For all 574 lost fragments identified in the three datasets, the presence (true/false) was recorded in the input data vector for each spectrum.

The final input data vector for each spectrum therefore contained information about peaks with priorities 1 to 6 and about peaks representing the 10 key fragment losses, as well as presence and absence of all identified lost fragments, and additionally the retention time and MS1 intensity of the parent ion, totalling 772 input data features. All features of the input data matrix were rescaled to values between 0 and 1.

Selection of input data features to predict molecular fingerprint digits

The fingerprint labels ("Molecular fingerprints") and the input matrix ("Conversion of MS1 and MS2 spectra to an input data vector") were filtered so that no NaN values were present. Consistent averaged spectra with a match quality from CompoundDiscoverer of 0.85 and a match between measured retention time and compound hydrophobicity (SI text, Sections S1 and S1.4) were selected (1 051 spectra representing 190 unique compounds). The 190 unique compounds were randomly split into train/validate (75%) and test (25%) datasets, and the spectra were then assigned to train/validate and test groups. This ensured that molecular structures used for training and validation are not represented in the test dataset.

Each digit of the fingerprint, representing a different molecular subunit or functional group, was predicted separately from the input data matrix using random forest (RF) and extreme gradient boosting (XGB) algorithms. These two methods were chosen for comparison as they perform well on non-linear classification and regression problems, have similar architecture, and can both provide feature importances to indicate which fingerprint digits are used to distinguish molecular structures. Classifiers were used for digits with ≤ 3 unique labels (`sklearn.ensemble.RandomForestClassifier` or `xgboost.XGBClassifier`) and regressors for digits with > 3 labels (`sklearn.ensemble.RandomForestRegressor` or `xgboost.XGBRegressor`) (Table 2). The compounds in the train/validate dataset were split into training and validation datasets: for RF, out of bag sampling was used during training, and one quarter of compounds were used for validation, and for XGB, 4 cross-validation folds were applied. The procedure described below was repeated for each of the four folds, and the final results were averaged.

To select features relevant for prediction of each fingerprint digit, the procedure recommend by [32] was followed. Initially, all 772 input features were used, and the RF or XGB model was trained. The trained model was used to predict the validation data, and goodness of fit criteria were calculated (macro-averaged precision, recall, F1 score, micro-averaged F1 score). Feature importance was found using two methods: Gini importance, provided as an attribute of the trained classifier (`classifier.feature_importances_`), or permutation importance, calculated with `inspection.permutation_importance` from the `sklearn` package. The least important 25% of features were flagged with an importance score of 1, and the next iteration was carried out without these features. The least important 25% of features in the second iteration was flagged with a score of 2, and iterations continued until one feature remained. Standard recursive feature elimination was not used (one feature removed per iteration) because of the large number of models this would entail for every digit and feature and the high intercorrelation between some features. The group of features giving the best F1 score (precision and recall compared when F1 scores were equal) were found for each fingerprint digit (Figs. 3 and S10).

Prediction of molecular fingerprints

The RF-G and XGB-G (random forest and extreme gradient boosting with Gini importance) models using the input data features selected as described in the previous subsection were optimised with hyperparameter tuning (SI text, Section S2). The models were then trained on the full train/validate dataset and used to predict the test dataset for each digit of the fingerprint (Fig. S11). For regression models, output was rounded to the nearest integer, as these values represent counts of functional groups.

For all fingerprint digits predicted using classification, the *prediction probability* was estimated using the `predict_proba` attribute of the classifier model from `sklearn`, which reports the number of votes for each class divided by the number of trees in the forest. This method cannot be used for regression models, thus the *prediction quality* for each unique value of each fingerprint digit was additionally calculated: the frequency of the value was found and used to estimate the probability of a true prediction (the prediction quality) for each value of the digit, according to Bayes' theorem:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A)} \times \mathrm{P(B|A)}}{\mathrm{P(B)}} \tag{4}$$

where P(A) is the frequency of the value, P(B) is the frequency with which the value is predicted, P(B|A) is the probability that the value is predicted when it is true, i.e. the true positive rate, and thus P(A|B) is the prediction quality for the value of a digit, i.e. the probability that the digit has that value given a prediction of that value. The frequencies and prediction qualities for non-MACCS fingerprint digits are shown in Fig. S12.

Using fingerprint comparisons to rank and assign molecular structure matches

To assign structures based on the tailored molecular fingerprints estimated using MS/MS spectra data with the RF-G model, the `get_compounds` function from the `PubChemPy` package was used to retrieve all compounds matching a sum formula from the PubChem molecular structure database. Hydrophobicity (logP) was estimated for all potential matches using the GNN+multitask method described in the SI (Section S1.4.1, [11]). The fingerprints for all potential matches were calculated, and the potential match fingerprints were compared to the spectrum fingerprint using the Tanimoto distance, as recommended by [33]. All potential matches with Tanimoto distance scores equal to or lower than the 200th match were retained; if there were less than 200 potential matches, all were retained.

The Tanimoto distance comparison does not weight fingerprint digits according to their importance in distinguishing matching and non-matching fingerprints. The importance of digits could be based on factors such as the uniqueness of the digit in identifying a molecular substructure or family or dependent on the prediction quality of the value or the digit, thus not easily estimated a priori. Therefore, digit weights were derived using a linear support vector machine (SVM) to maximise the distance of matching and non-matching fingerprints from the theoretical hyperplane. For each spectrum in the train/validate dataset, the squared difference between each digit of the predicted fingerprint and each digit of the fingerprints of the potential matches was found. When the potential match structure was the same as the structure assigned to the spectrum by Compound-Discoverer, the vector of squared differences was labelled “1”; when the structures did not match, the label of “2” was assigned. The resulting matrix of squared differences had 290 103 fingerprint comparisons for the 868 train/validate spectra, with 859 comparisons labelled as matching. A linear SVM was trained on the fingerprint comparisons using the function `sklearn.svm.SVC(kernel = ‘linear’,C=1).fit()` with a four-fold cross-validation procedure. The digit weights in each fold were retrieved with the attribute `coef_,` and mean weights across the four folds were found. Negative values should be avoided for the representation of weights; therefore, negative weights were set to 0, and the SVM procedure was repeated for all digits not yielding a negative weight. Up to three iterations were needed until no negative weights were found.

The SVM approach estimates weights for fingerprint digits but does not consider the prediction probability for each predicted digit value. Therefore, a further optimization was carried out using the expectation values for the predicted fingerprints, instead of the integer predictions. For fingerprint digits predicted with classification, the prediction probability was used to find the expectation value, e.g. if the predicted value was 1 with a probability of 0.7, the expectation value was 0.7; if the predicted value was 0 with a probability of 0.9, the expectation value was 0.1. The expectation value for fingerprint digits calculated using regression was taken as the predicted integer value, as this is the best value estimate given the input data. The differences between digits were found as described in the previous paragraph, using the expectation values for predicted digits and comparing to potential matches using squared differences for each digit. The SVM procedure was repeated to determine optimal weights based on expectation value fingerprint comparisons.

Following optimisation of the digit weights, fingerprint comparisons for the train/validate, test, and unlabelled spectra datasets were carried out for potential matches using Tanimoto distance, Euclidean distance, Euclidean distance with SVM weights (hereafter “Weighted Euclidean distance”), and Euclidean distance using expectation values with SVM-expectation value weights (hereafter “Weighted Euclidean distance with expectation values”) to measure fingerprint similarities.

The “match quality” for each potential match was estimated based on (i) confidence in the fingerprint prediction and (ii) the confidence with which the fingerprint comparison can be attributed to the “matching” class. For classification digits, the prediction probability of each digit was used as an estimate of confidence in the digit; for regression digits, where prediction probability could not be obtained, the prediction quality for the digit value was used to estimate the confidence. The mean confidence across the 75 digits was found to estimate the overall confidence in the fingerprint ($P_{\text{fingerprint}}$), using optimised digit weights from the SVM. To estimate confidence in the fingerprint match, the distributions of fingerprint similarities for matching and non-matching classes for the test dataset (Fig. 4) were approximated with Gaussian distributions. The probability that a calculated fingerprint similarity belongs to the matching class was found with:

$$P_{\text{match-dist}} = \frac{1}{\sigma_{\text{match}}\sqrt{2\pi}} e^{-\frac{(D-\overline{D}_{\text{match}})^2}{\sigma_{\text{match}}}} \tag{5}$$

where $P_{\text{match-dist}}$ is the probability that a distance belongs to the “match” distribution, D is the distance between the predicted fingerprint and the potential match, $\overline{D}_{\text{match}}$ is the mean distance for matching fingerprints, and σ_{match} is the standard deviation of matching fingerprints distances (Fig. 4). $P_{\text{non-match-dist}}$ is calculated analogously. The probability that the distance is a match given $P_{\text{match-dist}}$ and $P_{\text{non-match-dist}}$ is found using a Bayesian estimate as $P_{\text{match}} = \frac{P_{\text{match-dist}}}{P_{\text{match-dist}}+P_{\text{non-match-dist}}}$. The overall “match quality” was estimated as $P_{\text{fingerprint}} \times P_{\text{match}}$.

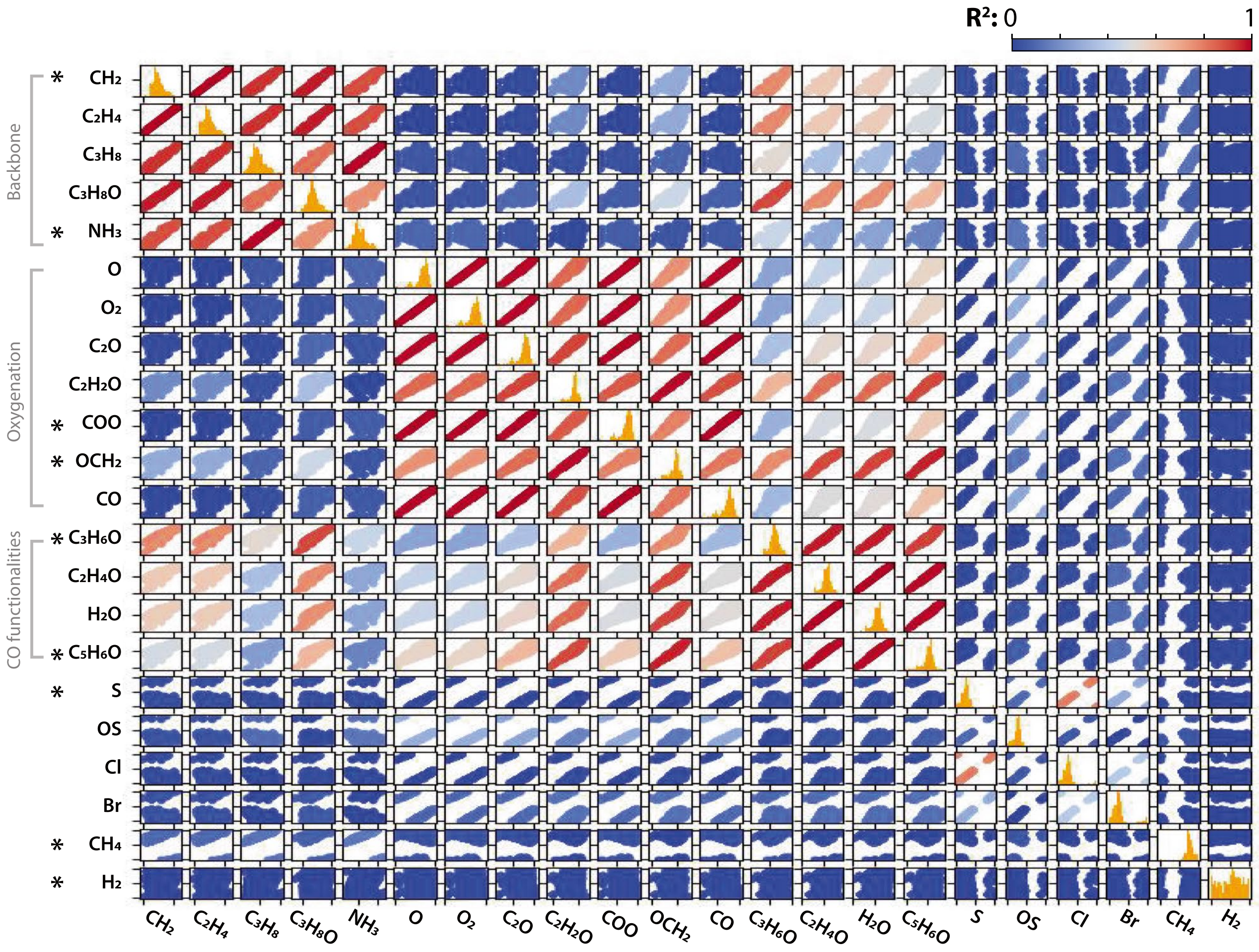

Fig. 1 Correlations between KMD values for different bases seen for all fragment and parent ion peaks in the three datasets used in this study. The colour scale shows R^2 to indicate the strength of the correlation. The diagonal shows a histogram of values for each KMD. Molecular patterns represented by three major KMD groupings are shown in grey at the left-hand side, and the nine KMD bases selected as input data are marked with asterixes

Results and discussion

Structural information represented by KMDs and lost fragments

Kendrick mass defects (KMDs) reflect molecular subunits and families and can therefore summarise key structural information from MS1/MS2 peak masses [20, 21]. We therefore investigated the potential of KMDs of MS1 and MS2 peak masses to act as "features"—input for the machine learning models used to predict molecular properties represented as fingerprint digits. However, KMDs for many bases show strong correlations as they describe similar molecular subunits (Fig. 1). For example, strong correlation is seen between KMDs for bases CH_2, C_2H_4, and C_3H_8, which all describe the aliphatic carbon backbone of a molecule. Similarly, KMDs for O, O_2, C_2O, and other oxygenated carbon bases all describe the oxygenation state of a molecule and the presence of C=O groups and thus show strong correlations. The importance of these subunits can directly be seen in the DiAcids dataset fragmentation patterns, which consistently show CO_2 losses. More complex C-O functionalities including ring structures are described with KMD bases C_3H_6O, C_5H_6O, and C_2H_4O. Based on KMD intercorrelations, nine KMD bases were selected which contain independent information about key molecular structures and substructures: CH_2, OCH_2, H_2, CO_2, S, NH_3, C_3H_6O, C_5H_6O, and CH_4.

"Lost fragments"—the mass difference between the parent ion and fragment ions—in MS2 spectra represent key molecular structure information and could also be useful features for molecular fingerprint prediction. A total of 574 unique lost fragments were identified in the three datasets, and 10 key lost fragments were selected: the 7 most frequent losses, as well as the smallest fragments H_2 and CO, and the non-oxygenated C_3H_8 fragment. The most common

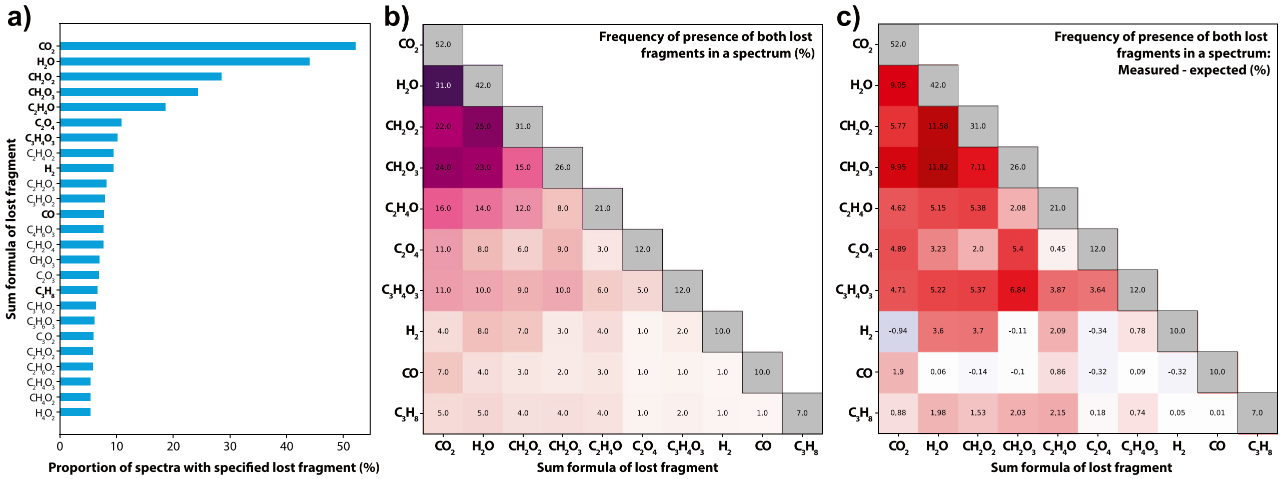

Fig. 2 Most common fragmentation losses across the three datasets used in this study. **a** Sum formula and frequency of the 25 most common lost fragments. **b** For 10 key lost fragments (in bold in *a*), the percentage of spectra containing *both* fragment losses is shown. **c** The difference between measured and predicted simultaneous fragment losses. For example, CO_2 is lost in 52% of spectra and H_2O in 42% of spectra. If losses were independent we would expect intersection in $0.42 \times 0.52 = 22\%$ of spectra. In (***b***), we see these losses occur in 31% of spectra, thus a difference of +9% is shown in (***c***). In (***b***) and (***c***), the diagonal shows the frequency of each lost fragment

lost fragments were CO_2, describing presence of a carboxyl or similar group, H_2O, showing potential for dehydration, e.g. an OH group, and often involving H-rearrangement, and other small carbohydrate fragments (Fig. 2). Simultaneous losses of both CO_2 and H_2O as well as other small carbohydrates within a single spectrum were more common than statistically predicted (Figs. 2*c* and S5), reflecting the occurrence of multiple-step fragmentation. Some double losses were less common than expected, for example H_2 losses occurred less often together with oxygenated fragments like CO, CO_2, and C_2O_4.

Input data features for prediction of fingerprint digits

For each of the algorithms (RF and XGB) and each of the two importance measures (Gini (RF-G, XGB-G) and permutation (RF-P, XGB-P)), the optimum features used to predict each fingerprint digit were found from the maximum macro-averaged F1 score for the digit, while also accounting for no major drop in either precision or recall, as shown in Fig. S10. The performance with the chosen features was compared among the algorithms (Fig. S10 and S11, Table S2). The best performing models were the RF and the XGB with Gini importance (RF-G and XGB-G); moreover, the feature selection based on Gini importance was around ten times faster. RF-G, XGB-G, and XGB-P models chose similar number of features and around 2/3 of features in common. The RF-P model was less consistent with other models and chose significantly more features, whereby only around half of features were chosen in common with other models.

A summary of most common chosen features for the RF-G and XGB-G models is shown in Fig. 3. KMD values describe the molecular structure and functional groupings of peaks and were chosen as useful input data for a large proportion of fingerprint digits. KMDs and other information were primarily chosen from peaks with priorities 1–4; lower priority peaks were rarely chosen as input, which may be due to the lower reproducibility of less intense peaks. The mz and relative intensity of the peak corresponding to a CO_2 loss was important for around a quarter of fingerprint digits in both RF-G and XGB-G models, showing the information quality of this loss for predicting carboxyl structures in particular. The H_2O loss was additionally chosen for many digits in the XGB-B model. Across most fingerprint digits in both RF-G and XGB-G models, retention time was a key input data feature; the strong relationship between retention time and hydrophobicity makes it an integrated descriptor of molecular structure (Fig. S7, [7, 8]).

Using the chosen features for each fingerprint digit, RF-G and XGB-G models were optimised (see SI Text Section S2), trained on the train/validation dataset, and used to predict the test dataset. Both models performed similarly (Fig. S10); models were able to make very good predictions for many MACCS digits, but prediction of non-binary digits and particularly of the carbon branching structure ("backbone" and "longestchain" digits) was challenging. The frequencies and prediction qualities for non-MACCS fingerprint digits using the RF-G model are shown in Fig. S12. The prediction of high OH and ring counts was challenging, likely due to the low frequency of these structures in training data and the difficulty of capturing high functional group counts in single fragments. Prediction quality of

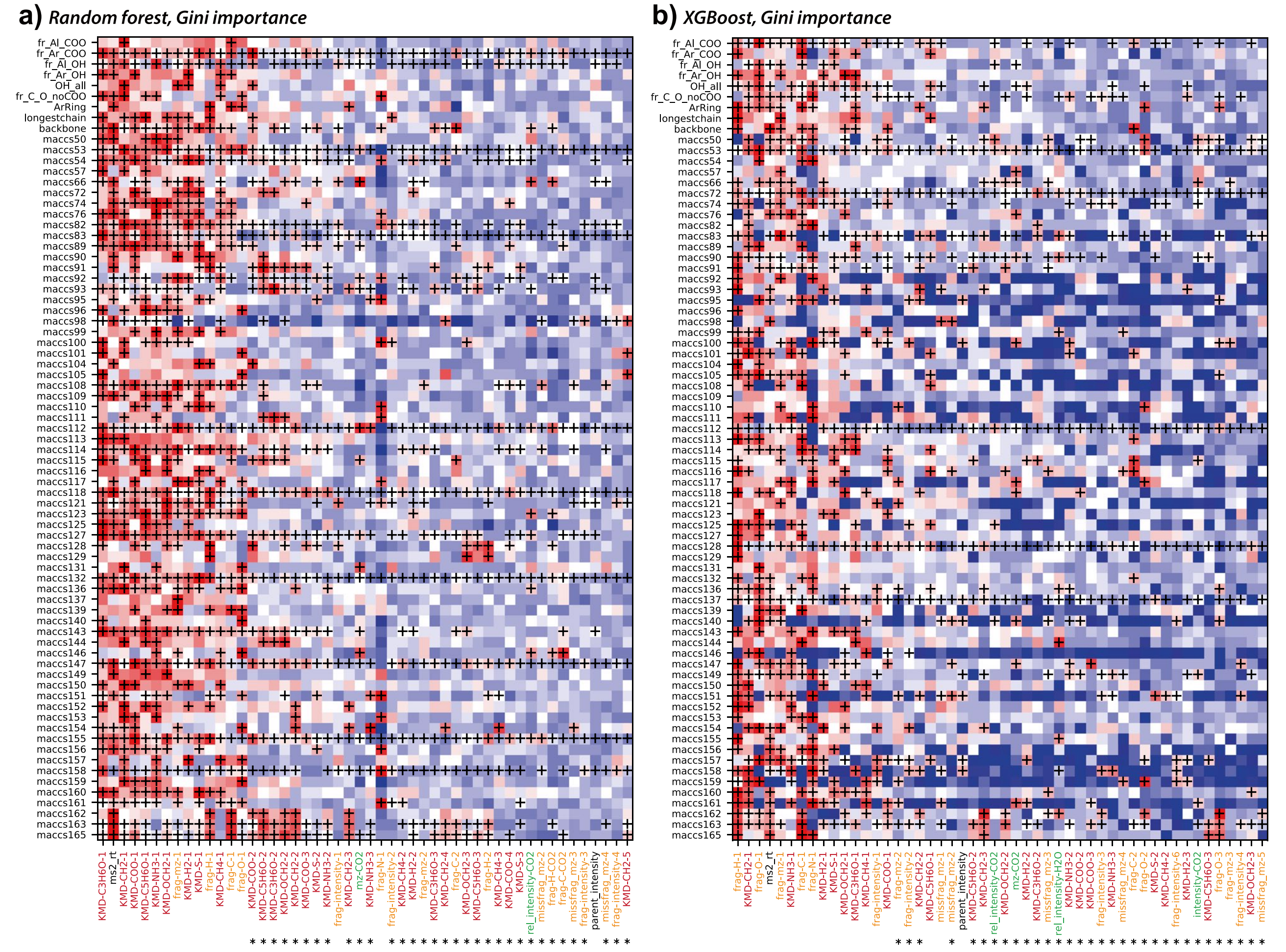

Fig. 3 The top 25 input data features for molecular property (fingerprint digit) predictions are shown on the x-axis, ordered by averaged importance, and the colour scale indicates their final importance ranking for each of the molecular subunits shown on the y-axis. Red indicates the most important features and dark blue the least important features. The black crosses mark features which were chosen for optimum model performance. The feature names are coloured: red text indicates KMD features (KMD-base-npeak where npeak is the peak priority ranking). Orange text indicates features relating to the priority ranked peaks (type-property-npeak, where type is "frag" for the measured fragment or "missfrag" for the lost fragment, and property is mz, atom (C, O, N, H) count, intensity, or relative intensity). Green text indicates features relating to particular lost fragments (property-lostfragment). Asterixes below the feature names indicate that the feature can only be obtained from the MS2; other features are from the parent ion spectrum and thus can obtained from the MS1

even-numbered chain and backbone lengths is much higher than odd-number chain and backbone lengths, despite similar frequencies, suggesting less ambiguous fragmentation patterns for even-C-numbered molecular structures. Averaged over the whole fingerprint, the RF-G model delivered 1% better F1 scores than the XGB-G model and was thus selected for final fingerprint prediction.

Assigning molecular structures based on predicted fingerprints

The sum formulas for the spectra in the three datasets yielded between 29 ($C_9H_4O_5$) and 10 912 ($C_{13}H_{16}O_3$) potential matches from the PubChem database. The top 200 potential matches were initially selected using the Tanimoto distance to compare the predicted fingerprint with fingerprints for potential matches, as recommended by [33] (see "Using fingerprint comparisons to rank and assign molecular structure matches"). Potential matches were additionally ranked using Euclidean, weighted Euclidean, and weighted Euclidean with expectation value methods. The distributions of similarities for fingerprints representing matching and non-matching molecular structures calculated with the different methods using the train/validate and the test datasets are shown in Fig. 4. The weighted Euclidean distances (with and without expectation values) perform best to distinguish fingerprints from matching and non-matching structures compared to the Tanimoto distance and the unweighted Euclidean distance: the Tanimoto distance has 31% overlap between matching and non-matching distributions for the test set, compared to 29% overlap for the

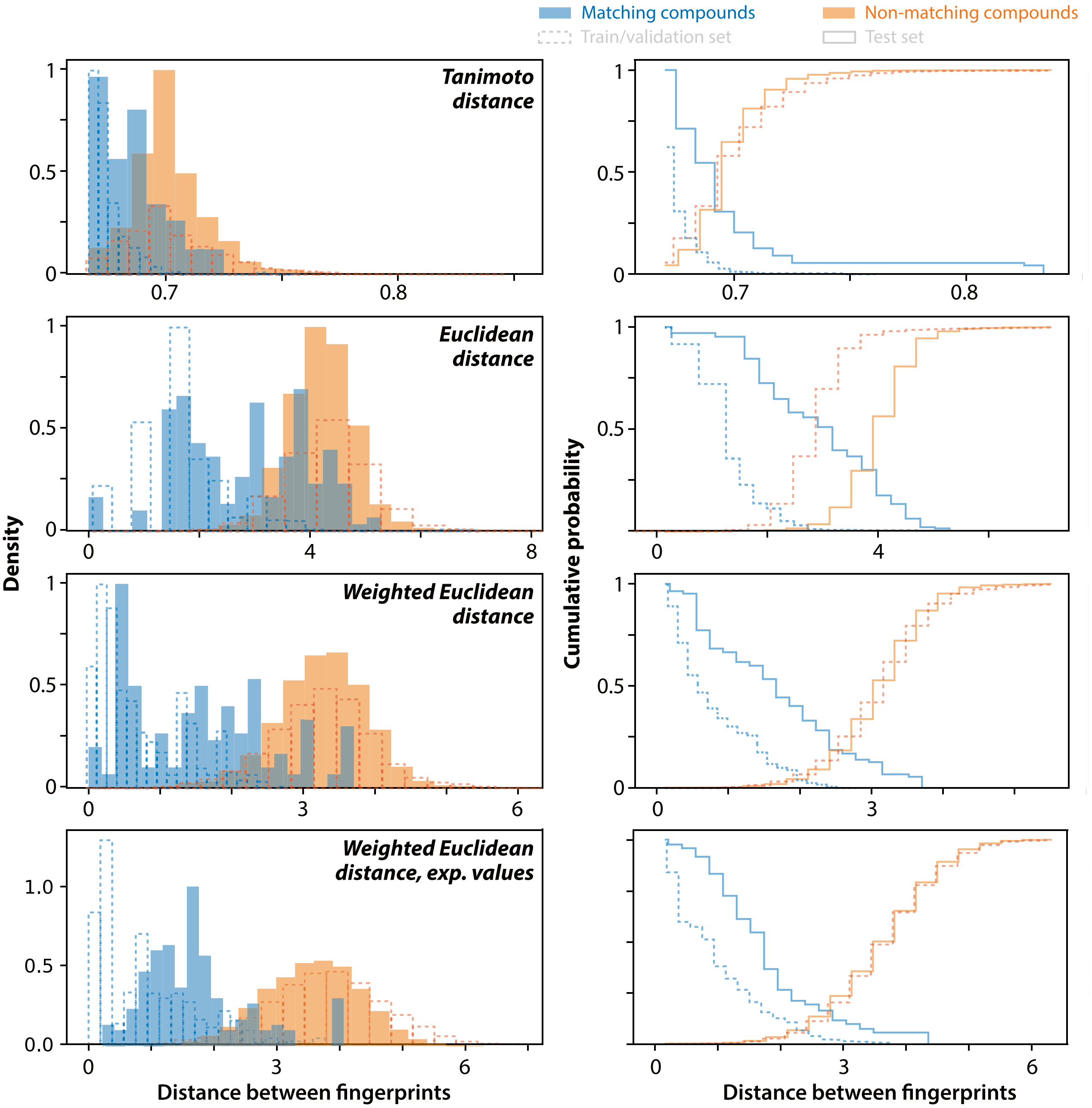

Fig. 4 Fingerprint similarity for matching (blue) and non-matching (orange) structures calculated using different distance measures: Tanimoto, Euclidean, weighted Euclidean, and weighted Euclidean using expectation values. Dashed lines and bars show results for the train/validate set; solid lines and filled bars show values for the test set. Left-hand panels show the density histogram and right-hand panels show the cumulative density histogram

Euclidean distance, 19% overlap for the weighted Euclidean distance, and 15% overlap for the weighted Euclidean distance with expectation values.

The estimated weights for different fingerprint digits depended on a combination of prediction quality and digit class frequency (Fig. S13). Digits with low F1 scores did not receive high weights, but not all well-predicted digits were weighted highly. Digits where the majority value had a frequency of less than 50% were not highly weighted, and in general digits with an imbalanced split between classes were weighted more highly, as they were presumably able to encode “unique” features of particular molecules and molecular families. The most highly weighted digits described small substructures within molecules related to branching and bonding patterns and heteroatom positions (Fig. S13). Some digits in the fingerprint carry similar or overlapping information, such as

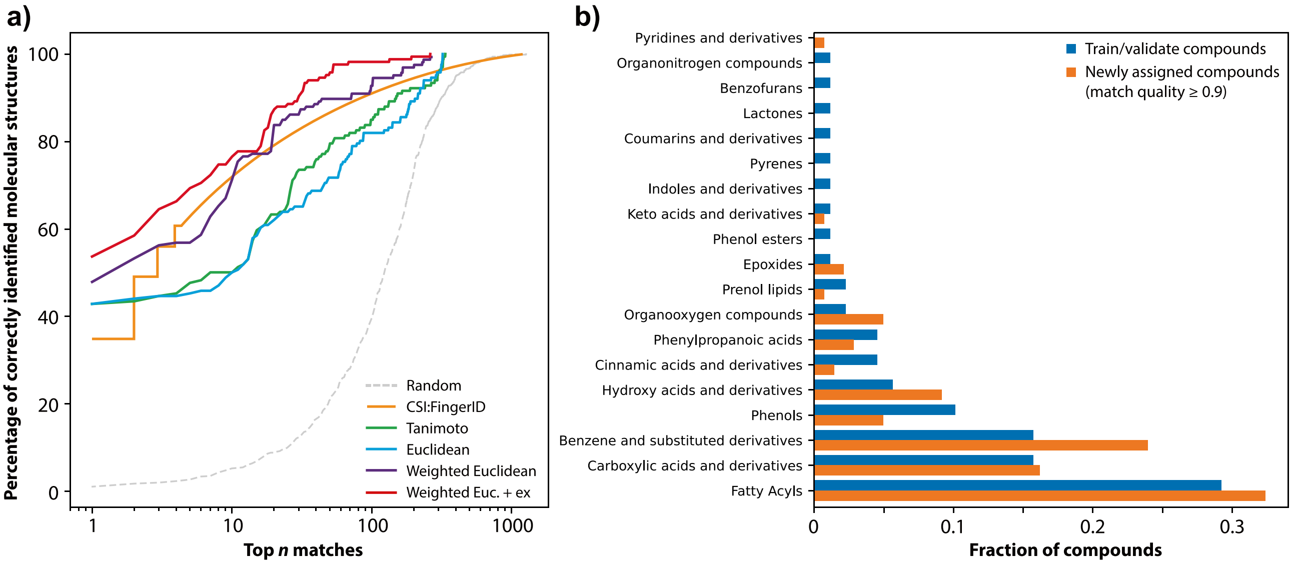

Fig. 5 Characterisation of the MSEI structure assignment method. **a** The percentage of instances in which the correct structure (y-axis) was identified within the top *n* matches (x-axis) for different fingerprint distance calculation methods for the test dataset. The results from CSI:FingerID taken from [18] are shown in orange for comparison, and the results for random matching are shown as a grey dashed line. **b** ClassyFire level 2 taxonomies [34] for compounds in the train/validate sets and for well-matched compounds assigned using the MSEI method

n_Al_OH and n_Ar_OH and n_OH, and some MACCS keys. In this case, we find low importances for all three OH-related digits. Despite this, we opt to retain digits describing similar or overlapping features: predictions are uncertain, particularly for non-binary digits; therefore, similar digits can still provide additional information.

The performance of the molecular structure assignment for each of the fingerprint distance calculation methods is summarised in Fig. 5a, and a detailed view of the best-performing method (weighted Euclidean with expectation values) is shown in Fig. 6. Despite the fingerprint only having 75 digits, equally well-matched compounds were rare, because the tailored fingerprint had 9 non-binary digits and very few redundant digits. These results may underestimate the performance, as the "correct" structures assigned by CompoundDiscoverer may be incorrect in an unknown number of cases. Using the weighted Euclidean distance with expectation values (red), the MSEI approach assigns around 10% more correctly identified molecular structures within the top *n* matches compared to previously published results for the state-of-the-art CSI:FingerID approach [18].

The estimated match quality for MSEI clearly relates to the goodness of prediction (Fig. 6): 50% of molecular structures are correctly assigned as the top match with match quality ≥ 0.85, and 70% of spectra are able to be assigned with match quality ≥ 0.85. For match quality ≥ 0.95, 80% of assignments are correct, and 25% of structures are able to be assigned at this level. The MSEI approach performs best for molecular families included in the training datasets: compounds assigned with high confidence by MSEI (MQ ≥ 0.9) belong to the same molecular taxonomies [34] as compounds in the train/validate dataset (Fig. 5b). The MSEI approach is able to learn MS1 and MS2 spectral features from a limited number of compounds and effectively predict other compounds within these molecular families; however, prediction outside of these molecular families is more challenging and uncertain and thus requires further training data.

Case study: the ruthenium dataset

The ruthenium dataset contains spectra measuring organic compounds present on a ruthenium catalyst at different stages during the catalytic gasification of organic wastes, described in detail by [26]. The catalyst stages are (i) initial active catalyst, (ii) inactive catalyst with organic molecule deposits on the surface, and (iii) catalyst reactivated by washing with organic solvents.

The dataset (Table 1) contained > 1000 valid MS2 spectra; however, only 17 well-matched compounds could be identified by CompoundDiscoverer. A total of 11 compounds corresponding to 24 averaged spectra from this dataset were used for training and validation of the MSEI approach. The dataset is illustrative of the MSEI approach, which is to assign compounds within a limited molecular structure space based on a small amount of training data from similar datasets. The full MSEI approach illustrated in this paper, including data preprocessing for all datasets in Table 1, fingerprint selection, feature selection, training and validation, and prediction of compound structures for

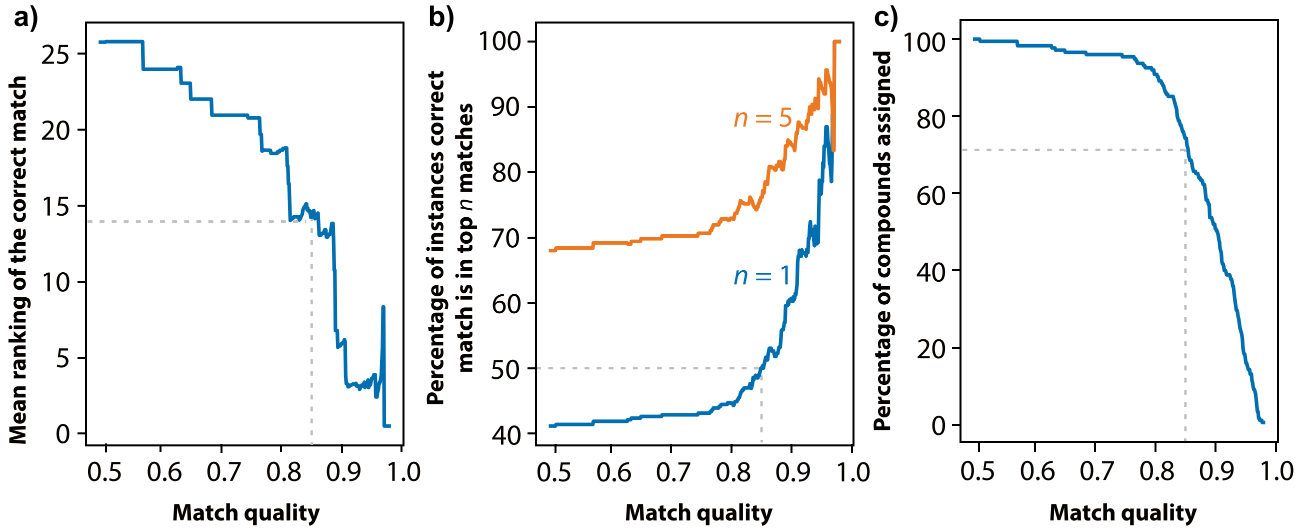

Fig. 6 Summary of match statistics based on match quality for all datasets in Table 1. The x axis shows the match quality; results are cumulative for all matches with match quality equal or above the indicated value. **a** The mean ranking of the correct match; **b** the percentage of instances where the correct match is within the top 5 matches (orange) or the correct match is the top match (blue); **c** the percentage of compounds that were able to be assigned at or above different match quality levels. A match quality of 0.85 corresponds to 50% of top matches correctly assigned in the test dataset, and this threshold is indicated with the grey dashed lines in each panel

the ruthenium dataset took around 48 h on a standard laptop (MacBook Pro, M1, 2020, 16 GB RAM), making the approach at least an order of magnitude faster than other currently available methodologies [18].

Using MSEI, 15% of spectra in the ruthenium dataset were assigned a molecular structure with a match quality ≥ 0.85; the test dataset has shown that at MQ ≥ 0.85, 50% of structures are correctly predicted (Fig. 6). MSEI was able to assign 81% of spectra were assigned with match quality ≥ 0.5 (42% correctly predicted). MSEI was able to identify 40 unique compounds with match quality ≥ 0.85, more than double the number identified by CompoundDiscoverer.

Well-matched compounds identified at the inactive (ii) and reactivated (iii) catalyst stages are shown in Tables S3 and S4. Three of the straight chain molecules on the inactivated catalyst contain C=C double bonds, which are present in none of the molecules on the activated catalyst. Structures with multiple linked rings such as tetracyclehexadecaoctaene and hydroxychromenone may also be implicated in catalyst deactivation. These results illustrate the strength of the MSEI approach to assign structures to unknown compounds in non-targeted analysis. Moreover, the fingerprints generated can provide significant direct information about the functional groups and composition of compounds, even when structures cannot be assigned.

Application and limitations

MSEI is a tailored approach to prediction of molecular structures from UHPLC-HRMS/MS measurements (see overview in Fig. 7) using a variety of data science techniques. With MSEI, we can achieve > 10% more correct structural identifications than similar methods, using a fraction of the computational power and training data volume [4, 17, 18]. We used 142 unique compound identifications generated for three datasets by CompoundDiscoverer as training data to generate a 75-digit molecular fingerprint, specific to the molecular space represented by the training data. The training data was then used to select relevant spectral features to predict each fingerprint digit using a random forest model, optimise hyperparameters, and train a final model for each digit. Intensity and mass-to-charge ratio of fragments corresponding to CO_2 and H_2O losses, as well as retention time and fragment elemental composition, were found to be key input data. KMDs for parent and fragment ions, describing molecular families based on molecular substructures, were also found to be useful features for prediction of all fingerprint digits. Existing structural identification methods could incorporate the MSEI direct fingerprint prediction approach to complement existing fingerprint estimation methods, for example, to speed up fingerprint calculation. The approach could also be used in applications required only fingerprints and not full structural information, such as in the prediction of toxicity of unknown compounds from non-targeted HRMS/MS analysis [35].

The optimised random forest models were used to predict fingerprints for spectra corresponding to both test data with known structures and spectra representing unknown molecular structures. Several methods were tested to compare predicted fingerprints to potential matches from the PubChem

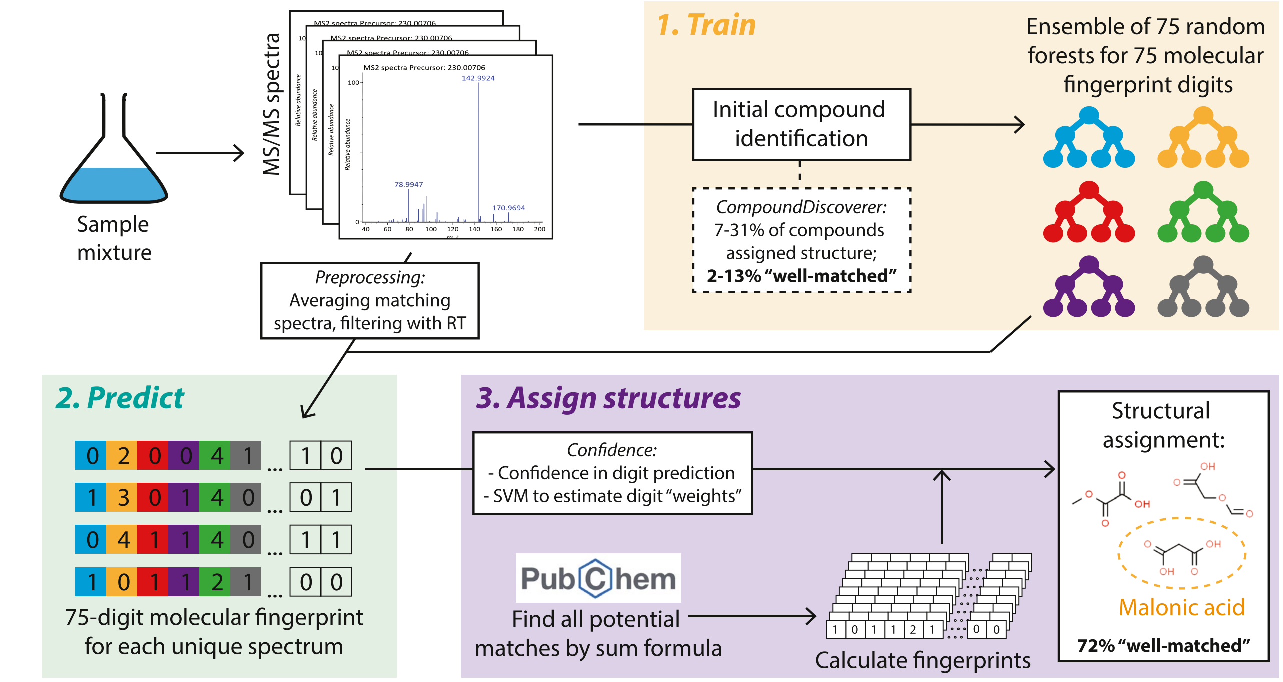
Sample mixture
MS/MS spectra
MS2 spectra Precursor: 230.00706
142.9924
78.9947
170.9694
Preprocessing: Averaging matching spectra, filtering with RT
1. Train
Initial compound identification
CompoundDiscoverer: 7-31% of compounds assigned structure; 2-13% "well-matched"
Ensemble of 75 random forests for 75 molecular fingerprint digits
2. Predict
0 2 0 0 4 1 ... 1 0
1 3 0 1 4 0 ... 0 1
0 4 1 1 4 0 ... 1 1
1 0 1 1 2 1 ... 0 0
75-digit molecular fingerprint for each unique spectrum
3. Assign structures
Confidence: - Confidence in digit prediction - SVM to estimate digit "weights"
PubChem
Find all potential matches by sum formula
1 0 1 1 2 1 ... 0 0
Calculate fingerprints
Structural assignment:
OH
Malonic acid
72% "well-matched"

Fig. 7 A schematic overview of the MSEI approach to structural identification based on UHPLC-HRMS/MS measurements. An initial set of structural identifications is used to **train** a set of random forests, each representing one digit of a tailored molecular fingerprint. The random forests are used to **predict** fingerprints for all valid spectra in a dataset. Comparison of predicted fingerprints to potential structural matches from the PubChem database is used to **assign structures** for unknown molecules. For the datasets used in this study (see Table 1), 2–13% of compounds had a "well-matched structure" following initial assignment, which increased to 72% with MSEI (Fig. 6)

database. Compared to unweighted Tanimoto and Euclidean distances, weighted distance metrics showed significantly improved performance. Optimal weights for each fingerprint digit were determined using an SVM; the method performed best when combining learnt weights with expectation values for each fingerprint digit. Fingerprint digits encoding less common functional groups were weighted more highly, showing the importance of tailored molecular fingerprints for structural identification. Furthermore, we developed a method to predict compound hydrophobicity based on molecular structure, to compare to measured retention times and act as an independent filter for molecular structure attribution [11]. MSEI was able to correctly assign > 50% of structures from the test dataset, with 80% of structures correctly identified in the top 10 matches. Incorporation of the SVM-weighted fingerprint comparison into current structural identification pipelines would improve performance while requiring minimal additional computational power, even for large fingerprints.

MSEI uses fingerprints specifically selected for the molecular space of interest and is trained on a subset of the data of interest; thus, it has limited ability to generalize to unknown molecular families. Moreover, the ability of MSEI to predict compounds for spectra measured on different instrumentation is currently unknown. The approach is envisaged as part of an analysis pipeline: a laboratory will be interested in compounds from particular sample types and will first analyse their data using their existing methodologies, for example, CompoundDiscoverer or CSI:FingerID. This will generate a dataset on which MSEI can be trained and then used to complement these existing approaches and predict structures for a much larger number of unknown compounds. Over time, MSEI can be retrained on further data and standards to gain greater predictive ability and widen the molecular space of interest. Furthermore, training data and/or trained models can be shared between laboratories working on similar sample types using equivalent instrumentation to improve compound identification. Database spectra can be used to widen the molecular space available for training, although—as with all approaches—caution should be taken when comparing spectra collected from different instrumental set-ups. MSEI is written in Python, and code is publicly available; thus, MSEI can be tailored to the needs of a specific pipeline or run from the command line, and internal processing steps can be investigated and visualised much more simply than with many existing spectral processing packages.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11224-023-02192-2.

Acknowledgements We would like to acknowledge the contribution of Prof. Tanja Käser to the initial stages of the project, as well as the contributions of ETHZ DAS-Capstone project student Patrik Friedlos to the investigation of algorithms for hydrophobicity prediction and EPFL visiting MSc student Malo Hillairet for the development of mathematical methods for spectral comparison based on peak intensity and mass-to-charge.

Author contribution S.B. collected the data. E.H., S.B., and L.G. carried out the data analysis. E.H. prepared the figures and wrote the manuscript. All authors provided expertise throughout the project and reviewed the manuscript.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich This study was funded by the Swiss Data Science Center (SDSC), EPFL, and ETH Zürich under project MSEI (grant number C18-02). The project has received funding from the Swiss Innovation Agency Innosuisse in the framework of the Swiss Competence Center for Energy Research (SCCER BIOSWEET). Part of this work was supported by and performed within the Energy System Integration Platform at the Paul Scherrer Institute.

Data availability The full analysis pipeline included data and results illustrated in this paper can be downloaded from: https://renkulab.io/gitlab/msei/msei-main-clean-v1_20221026. The data and code for hydrophobicity prediction [11] can be downloaded from: https://renkulab.io/gitlab/msei/capstone-project.

Declarations

Ethical approval Not applicable

Competing interests The authors declare no competing interests.

References

1. Cotton J, Leroux F, Broudin S, Marie M, Corman B, Tabet JC, Ducruix C, Junot C (2014) High-resolution mass spectrometry associated with data mining tools for the detection of pollutants and chemical characterization of honey samples. J Agric Food Chem62(46):11335–11345
2. Ludwig M, Nothias LF, Dührkop K, Koester I, Fleischauer M, Hoffmann MA, Petras D, Vargas F, Morsy M, Aluwihare L, Dorrestein PC, Böcker S (2020) Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. Nat Mach Intell 2(10):629–641
3. Böcker S (2017) Searching molecular structure databases using tandem MS data: are we there yet? Curr Opin Chem Biol 36:1–6
4. Hufsky F, Böcker S (2017) Mining molecular structure databases: identification of small molecules based on fragmentation mass spectrometry data. Mass Spectrom Rev 36:624–633
5. Klammer AA, Yi X, MacCoss MJ, Noble WS (2007) Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. Analytical Chemistry79(16):6111–6118
6. Naylor BC, Catrow JL, Maschek JA, Cox JE (2020) QSRR automator: a tool for automating retention time prediction in lipidomics and metabolomics. Metabolites10(6)
7. Witting M, Böcker S (2020) Current status of retention time prediction in metabolite identification. J Sep Sci 43(9–10):1746–1754
8. Bouwmeester R, Martens L, Degroeve S (2019) Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. Anal Chem 91(5):3694–3703
9. Yang Q, Ji H, Hongmei L, Zhang Z (2021) Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. Anal Chem 93(4):2200–2206
10. Qu C, Schneider BI, Kearsley AJ, Keyrouz W, Allison TC (2021) Predicting Kováts retention indices using graph neural networks. J Chromatogr A 1646
11. Friedlos P, Gasser L, Harris E (2022) Retention time prediction to facilitate molecular structure identification with tandem mass spectrometry. bioRxiv, pp 1–20
12. Sleno L, Volmer DA (2004) Ion activation methods for tandem mass spectrometry. J Mass Spectrom 39(10):1091–1112
13. Levsen K, Schiebel HM, Terlouw JK, Jobst KJ, Elend M, Preiß A, Thiele H, Ingendoh A (2007) Even-electron ions: a systematic study of the neutral species lost in the dissociation of quasi-molecular ions. J Mass Spectrom 42(8):1024–1044
14. De Vijlder T, Valkenborg D, Lemière F, Romijn EP, Laukens K, Cuyckens F (2018) A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. Mass Spectrom Rev 37(5):607–629
15. Rojas-Cherto M, Peironcely JE, Kasper PT, Van Der Hooft JJJ, De Vos RCH, Vreeken R, Hankemeier T, ReijmersT (2012) Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. Anal Chem 84(13):5524–5534
16. Rasche F, Svatos A, Maddula RK, Böttcher C, Böcker S (2010) Computing fragmentation trees from metabolite multiple mass spectrometry data.Anal Chem83:1243–1251
17. Ruttkies C, Neumann S, Posch S (2019) Improving MetFrag with statistical learning of fragment annotations. BMC Bioinformatics 20(1):1–14
18. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci USA 112(41):12580–12585
19. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, Petras D, Gerwick WH, Rousu J, Dorrestein PC, BöckerS (2020) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol 39(April)
20. Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K (2001) Kendrick mass defect spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal Chem 73(19):4676–4681
21. Roach PJ, Laskin J, Laskin A (2011) Higher-order mass defect analysis for mass spectra of complex organic mixtures. Anal Chem83:4924–4929
22. Royal Society of Chemistry (2018) ChemSpider. Royal Society of Chemistry, Thomas Graham House, Cambridge, UK

23. Ltd HC (2018) mzCloud - Advanced Mass Spectral Database. HighChem Ltd., Bratislava, Slovakia
24. SwainM (2014) PubChemPy package
25. Arturi KR, Kucheryavskiy S, Nielsen RP, Maschietti M, Vogel F, Bjelić S, Søgaard EG (2019) Molecular footprint of co-solvents in hydrothermal liquefaction (HTL) of Fallopia Japonica. J Supercrit Fluids 143:211–222
26. Gasser U, Bjelic S (2022) Deposits of non-volatile organics in methanation catalyst (Ru/C) studied with SANS. In preparation, pp. 1–7
27. Shen H, Duehrkop K, Boecker S, Rousu J (2014) Metabolite identification through multiple kernel learning on fragmentation trees. Bioinformatics 30(12)
28. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280
29. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
30. Yap CW (2010) PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. J Comput Chem 32:1466–1474
31. Landrum G (2022) The RDKit 2022.03.1: Open Source Cheminformatics and Machine Learning
32. Datta S, Depadilla LM (2006) Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. Stat Methodol 3(1):79–92
33. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminformatics 7(1):1–13
34. 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 (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminformatics 8(1):1–20
35. Katarzyna R (2022) Arturi and J Hollender. Mining Toxicity and High-Resolution Mass Spectrometry for Linking Exposures to Effects, EXPECTmine

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.