

Supplementary information for:

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

E. Harris¹, Lilian Gasser¹, M. Volpi¹, F. Perez-Cruz¹, S. Bjelić², G. Obozinski¹,

¹Swiss Data Science Center, ETH Zürich, Switzerland

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

1 SI text: Data preprocessing

1.1 Data import and organization

Each dataset shown in Table 1 contains multiple .mzML files, corresponding to different measurement runs in the UHPLC-MS/MS set up. Each .mzML file is read into Python using the `run.Reader` function from the `pymzml` package with precision specified as 2.5×10^{-6} and 1×10^{-5} for MS1 and MS2 respectively. Key information regarding each spectra in the .mzML files is collated to a data frame, including scan type (positive or negative), retention time, parent peak intensity and mass-to-charge ratio from the MS1 spectrum, and intensity and mass-to-charge ratio for all peaks in the MS2 spectrum. MS2 spectra are cleaned: All peaks with an intensity of 0 are removed, all peaks closer than 0.001 mass units are combined, and peaks heavier than the parent ion are removed (there was no evidence for doubly-charged parent ions for any spectra considered). Compound descriptors for all compounds found in the dataset by CompoundDiscoverer (see Section 2.2 of the main article) are matched to each MS2 spectra based on retention time and molecular weight in order to filter for valid MS2 spectra, and where possible give an initial structural identification for the spectrum. Additionally, Kendrick mass defect [1] information is calculated for all peaks in MS1 and MS2 spectra for the bases: CH₂, OCH₂, C₂H₄O, CO, H₂, CO₂, H₂O, O, O₂, S, OS, Cl, Br, NH₃, C₂H₄, C₃H₈, C₃H₈O, C₅H₆O, CH₄, C₂O, C₂H₂O, C₃H₆O.

Three key dataframes are produced and saved as .csv files (in `/output/DatasetName/pt1/`) for each dataset during the first preprocessing step:

- **df_c**: A summary of all compounds (known and unknown) identified from the MS/MS spectra
 - *Columns*: `compound_id` (unique identifier of the compound), `compound_name` (*name), `compound_name_iupac` (*IUPAC name), `compound_sf` (sum formula), `compound_mw` (molecular weight), `compound_rt` (retention time), `compound_mq` (*match quality), `compound_iso_smiles` (*isomeric SMILES), `compound_can_smiles` (*canonical SMILES), `compound_xlogp` (*hydrophobicity as XlogP), KMD-related information
 - Columns indicated with * are only filled for compounds with an assigned structure
- **df_s**: A summary of valid MS/MS spectra
 - *Columns*: `spectrum_id` (unique identifier of the spectrum), `label` (1 if spectrum has an assigned compound, otherwise 0), `compound_id` (unique identifier of the compound as listed in `df_c`), `parent_full_mw` (molecular weight of the unionised parent ion), `ms1_rt` (retention time recorded in MS1 spectrum), `ms2_rt` (retention time recorded in the MS2 spectrum), `ms2_max_intensity` (maximum peak intensity in the MS2 spectrum), `parent_mz`

(mass-to-charge of the parent ion for the MS2), `parent_intensity` (intensity of parent ion), `parent_fromms1_mz` (mass-to-charge of the parent ion recorded in the MS1 spectrum), `parent_fromms1_intensity` (intensity of the parent ion recorded in the MS1 spectrum), `scan_type` (negative or positive), `mzml_file` (name of the mzML file), `mzml_file_short` (condensed name of the mzML file), `spectrum_folder` (dataset name)

- **df_p**: A description of every peak in all MS2 spectra in the dataset, created by appending all entries in the dictionary produced by the `pymzml.run.Reader` function:
 - *Columns*: `spectrum_id` (unique identifier of the spectrum as listed in `df_s`), `mz` (peak mass), `intensity` (peak intensity)

1.2 Assigning sum formulas to MS2 spectral peaks

The sum formula is found for each MS1 peak (in `df_s`) and each MS2 peak (in `df_p`) based on the following procedure:

- The sum formula of the parent ion for the MS2 spectrum is used as a basis to calculate the sum formulas of all peaks in the spectrum
- The exact masses and preferred valencies of all elements in the parent sum formula are found (preferred valencies: C: 4, H: 1, O: 2, N: 3, P: 3, S: 2, Cl: 1, K: 1)
- All possible sum formula combinations of the elements in the parent sum formula are found; for the sum formula $C_xH_yO_z$ this corresponds to $(x+1)(y+1)(z+1)$ combinations with formulas ranging from H to $C_xH_yO_z$
- Possible sum formulas are filtered to valid sum formulas containing at most 2 hydrogens per other element and a valid valency, as well as at least one carbon for organic molecules - around half of the possible sum formulas are removed
- Exact masses for each valid sum formula are found and compared to the peak mass
- If a matching sum formula is found within 40 ppm of the peak mass ($\pm 2 \times$ the specified MS2 precision of 10 ppm), the sum formula is accepted, and the mass discrepancy in mass units and in ppm is recorded

Columns with the sum formula of the peak (`formula`), mass difference between sum formula and peak in atomic mass units (`formula_massdiff`) and mass difference in parts per million (`formula_massdiff_ppm`) are added to `df_p`, and the resulting dataframes are saved as .csv files in `/output/DatasetName/pt2/`.

1.3 Identify and combine spectra representing the same molecular structure

The third stage of the preprocessing procedure involves two parts: i) establishing a method to determine whether two MS2 spectra match, ie. represent the same structure, ii) applying this method to identify and average matching spectra. Because of the novel nature of this procedure, a description is included in the main article and repeated in more detail here.

1.3.1 Quantifying spectral similarity

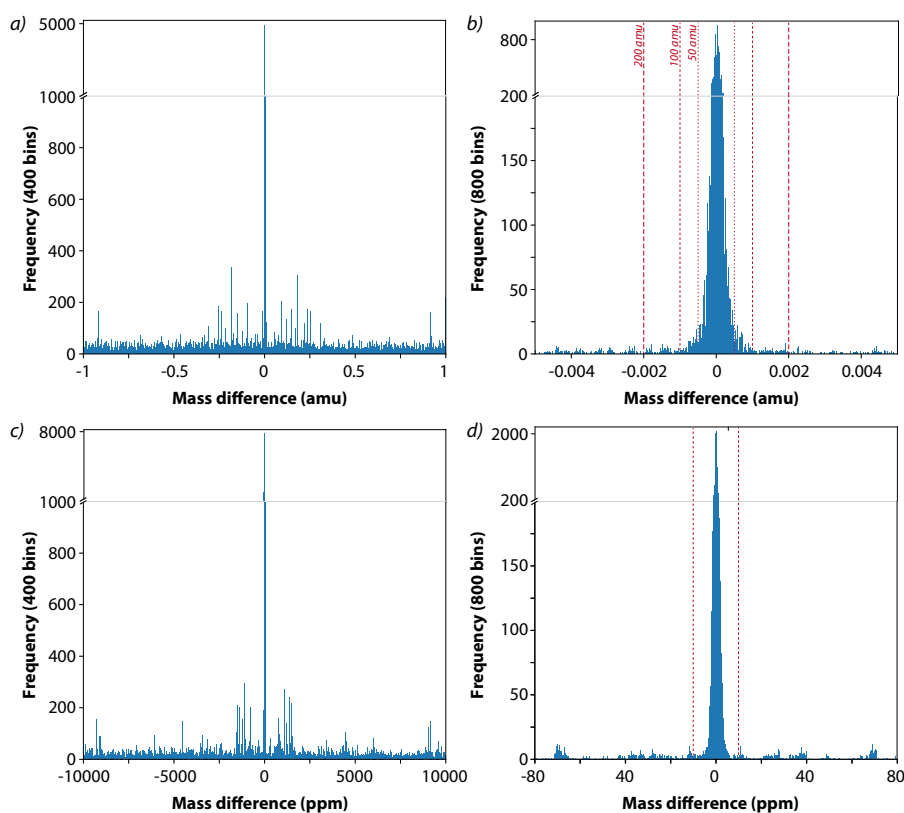
To determine the similarity of two spectra, a radial basis function kernel is used to compare the measured mass-to-charge ratios (`mz`) of all peaks in the two spectra, similar to the approach used by [2]:

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

Where $S_{1,2}$ is the similarity index for Spectrum 1 (i peaks) and Spectrum 2 (j peaks), mz is the mass of each peak in the spectra, and $\frac{mz_i - mz_j}{(mz_i + mz_j)/2 \times 10^6}$ expresses the difference between peak masses in ppm. σ_{mz} is the estimated standard deviation in peak mass for repeated measurements of the same mass, expressed in ppm. 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 can be calculated to account for varying numbers of peaks in different spectra:

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

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 $5\sigma_{mz}$ to determine if two peaks represent the same sum formula.



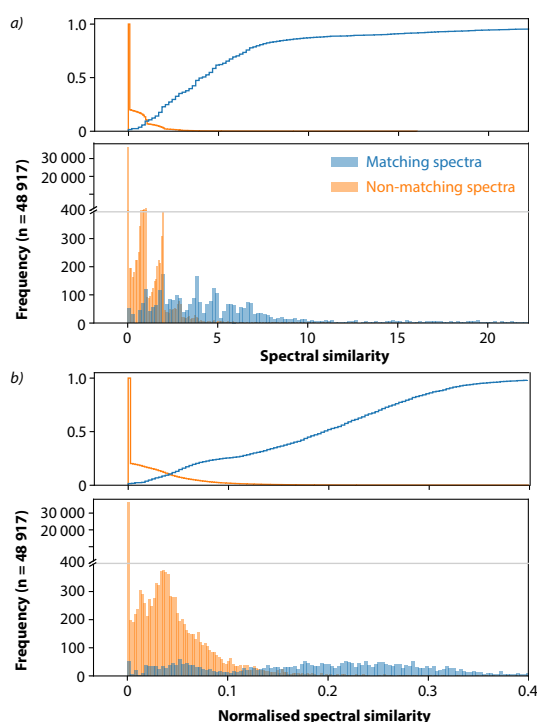
Supplementary Figure 1: Measured differences in peak mass for paired MS2 peaks in the DiAcids dataset, shown in amu (a/b) and ppm (c/d). c and d show a close-up version of the distributions in a and b. Note the split scale indicated on the Y-axes. The nominal MS2 precision of 10 ppm is shown with red dashed lines (with precision for different amu values shown in the upper panel).

In order to compare spectra using Eq. 1, the average mass difference between two peaks representing ions with the same sum formula measured in different MS2 spectra (hereafter, standard deviation of peak mass: σ_{mz}) must be determined. The DiAcids dataset was used to estimate the standard deviation of peak mass, as this dataset originates from a synthesised mixture of known compounds measured under a range of conditions (see Table 1). All spectra corresponding to each of the 13 compounds in this dataset were found and all peaks were intercompared. This resulted in 872 444 unique pairwise peak comparisons. The spectrum_id, mz, intensity, and relative intensity of the two peaks being compared were collated, as well as the compound_id represented by the spectra. The mass difference for each pair of peaks was found in both atomic mass units, and in ppm relative to the average peak mass. A histogram of all mass differences in atomic mass units and in ppm

is shown in Figure S1. The nominal MS2 precision of 10 ppm, shown in red in the figure (b and d), correctly captures the boundary between non-matching and matching peaks. Using a cut-off of 10 ppm difference to define if peaks are measuring ions with the same sum formula, the standard deviation in peak mass for ions with the same formula was found to be 1.4 ppm or 0.0002 amu. No significant difference in standard deviation was found for different compounds or for different measurement runs.

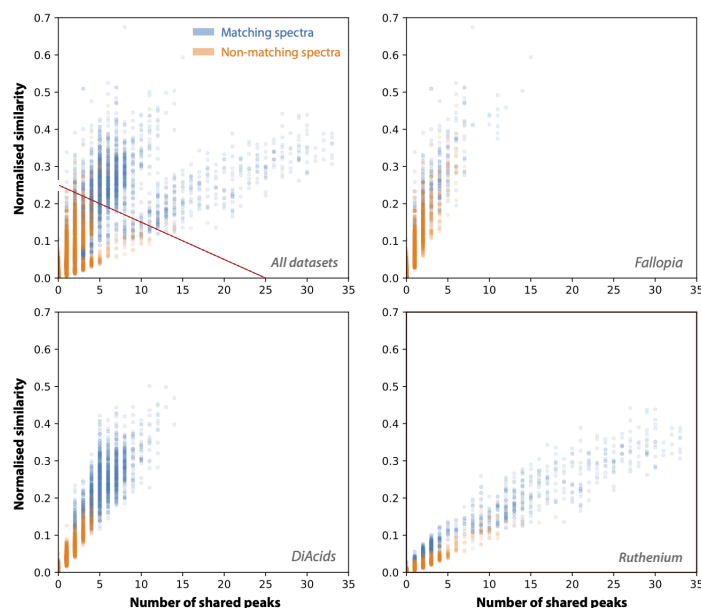
1.3.2 Distinguishing between matching and non-matching spectra

The similarity index threshold to distinguish matching (representing the same molecular structure) and non-matching spectra was investigated for the three datasets (Table 1). To limit computational time, 50 compounds identified by CompoundDiscoverer were randomly selected from the Fallopia and Ruthenium datasets, and a maximum of 200 spectra were randomly selected for each compound. Only compounds with a match quality ≥ 0.85 were selected. Selected spectra from each dataset were compared to all other selected spectra from the same dataset using Eqs. 1 and 2 with $\sigma_{mz} = 1.4$ ppm; spectra were not compared between datasets. Additionally, the Tanimoto fingerprint similarity was calculated for each pair of compounds using the function `DataStructs.FingerprintSimilarity` from the `rdkit` package, based on a MACCS fingerprint generated with `MACCSkeys.GenMACCSKeys`.



Supplementary Figure 2: a) Frequency histogram of spectral similarity index for MS2 spectra from separate measurement runs assigned to the same molecular structure (matching spectra) and spectra representing different molecular structures (non-matching spectra). Note the split scale indicated on the Y-axes. The cumulative distribution is shown in the upper panel - the intersection point at $S = 1.05$ corresponds to 9% overlap. b) Analogous for normalised spectral similarity index - the intersection point at $nS = 0.04$ corresponds to 10% overlap.

Based on similarity index or normalised similarity index, no clear cut off between matching and non-matching spectra can be found (Fig. S2). 80% of non-matching spectra show a (normalised) similarity index of 0, however normalised similarity indices between 0.01 and 0.13 can correspond to either non-matching or matched spectra. The number of peaks in common can be used as a further indicator of similarity, as shown in Fig. S3. When instrument sensitivity settings produce MS2 spectra with a large number of peaks, as in the Ruthenium dataset, spectral comparison may reveal many



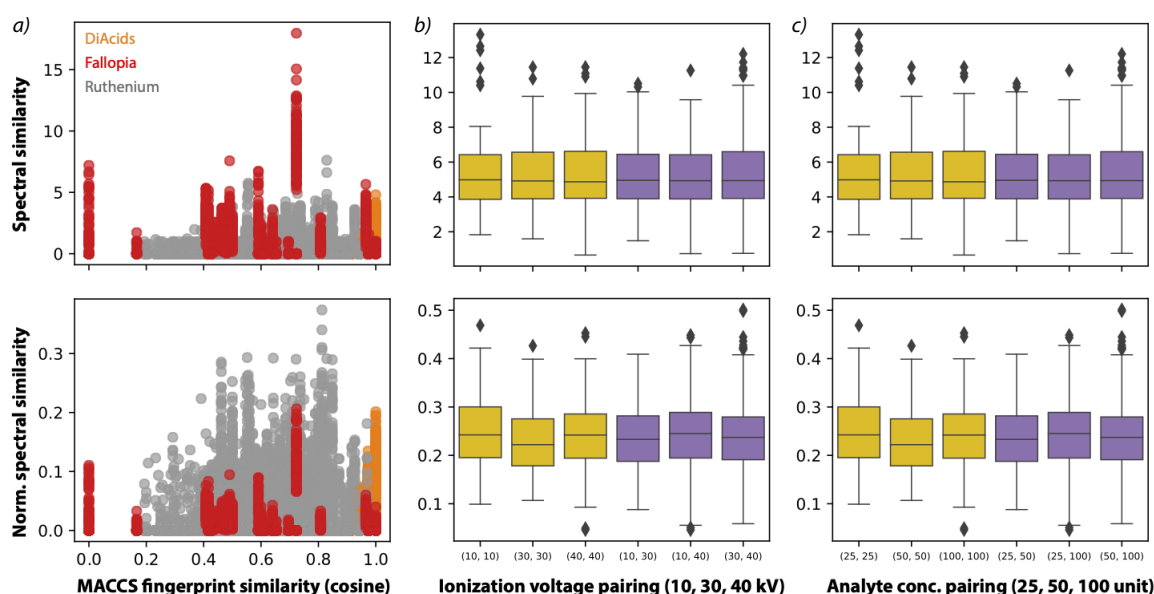
Supplementary Figure 3: Normalised similarity compared to number of peaks in common for the full dataset (top left) and for each of the datasets individually. The threshold to classify non-matching and matching spectra according to Eq. 3 ($S_{max}^{\text{norm}} = 0.25$ and $a = 0.01$) is shown in dark red.

peaks in common, but lower normalised similarity due to the large number of total peaks. A threshold to classify matching and non-matching spectra based on both the normalised similarity index and the number of peaks in common was defined as:

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

Where S_{max}^{norm} is the intercept and corresponds approximately to the maximum normalised similarity possible for non-matching spectra, and a is the slope. Precision and recall for matching and non-matching classes were investigated for a range of S_{max}^{norm} and a values, with the particular aim of minimising the number of non-matching pairs classified as matching. The threshold illustrated in Fig. S3 ($S_{max}^{\text{norm}} = 0.25$ and $a = 0.01$) is able to classify matching and non-matching spectral pairs with precision of 0.95 for both classes, and recall of 1.0 for non-matching spectra and 0.55 for matching spectra.

For most non-matching spectra, spectral similarity is 0. However, some non-matching spectra show moderate similarity, presumably due to the presence of fragments with matching sum formulas, which may or may not represent the same substructure. Non-matching spectra showing high normalised similarity are always associated with moderate (>0.4) molecular fingerprint similarity (Fig. S4). However, above fingerprint similarity of 0.4, there is no relationship between fingerprint similarity and normalised similarity, and moreover, high molecular fingerprint similarity does not necessarily correspond to high spectral similarity. The impact of measurement conditions on similarity index for matching spectra was investigated with the DiAcids dataset, where the same mixture of compounds was measured with differing ionization voltages (10, 30 and 40 kV) and at different analyte concentrations (25, 50, 100 units). Ionization voltage and analyte concentration had no effect on spectral similarity for matching spectra (Fig. S4). In summary, high spectral similarity indicates at least moderately similar molecular structures. However, high molecular similarity does not guarantee high spectral similarity. Measurement conditions do not play a clear role in determining spectral variability. Therefore, spectral similarity matching can be used to assign structures based on exact spectral matches, but it does not appear to be a useful tool to assign molecules to families or groups with similar functional groups and substructures.

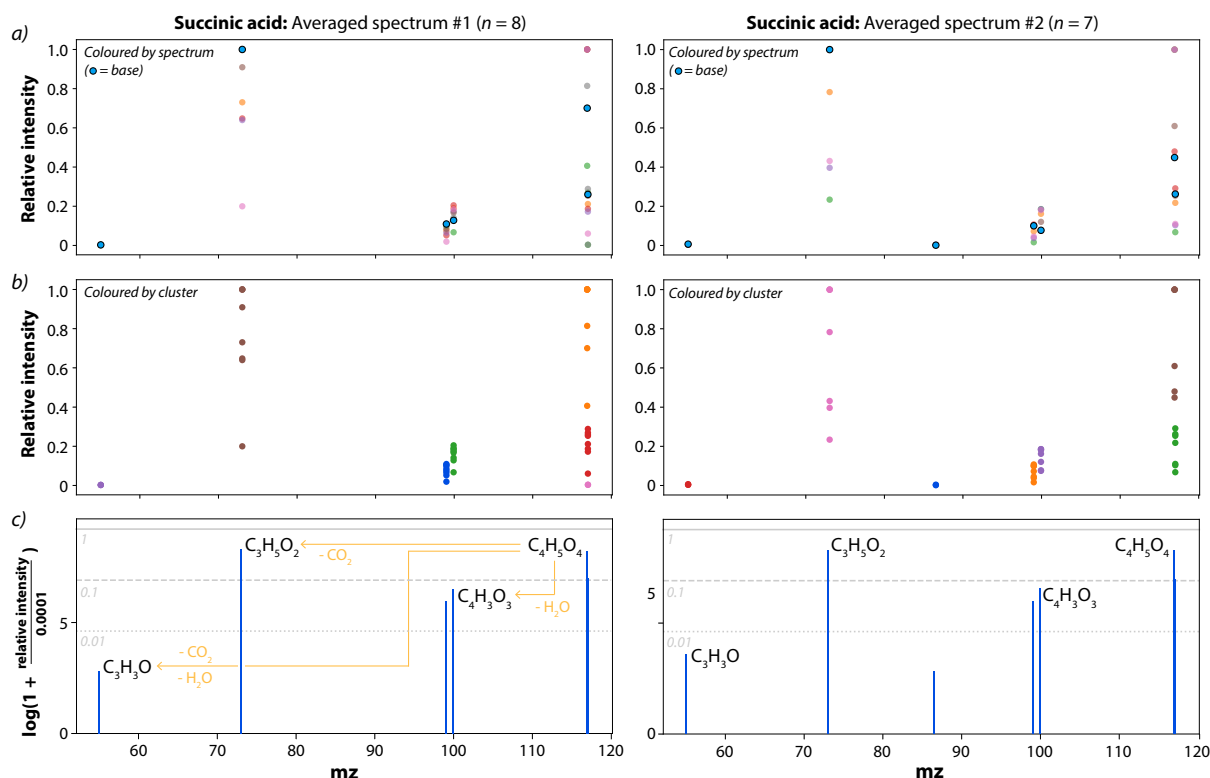


Supplementary Figure 4: *a*) Relationship between molecular similarity (cosine similarity of MACCS fingerprints) and S and S^{norm} for non-matching spectra. *b/c*) S and S^{norm} for matching spectra collected under varying ionization voltages (*b*) and with varying analyte concentrations (*c*). Yellow bars indicate where both spectra were collected with the same ionization voltage resp. analyte concentration, while purple bars indicate different measurement conditions.

1.3.3 Combining spectra with high similarity

Groups of matching spectra were found using an iterative approach, whereby a spectra was chosen at random as the ‘base’ spectrum. Potential matching spectra from all datasets were found based on retention time matching to within 5 times the nominal precision in retention time (± 50 seconds) and parent ion mass matching to within 10 times nominal precision (± 0.0001 amu). All candidate spectra were then compared to the base spectrum using Eqs. 1 and 2, and matching spectra were identified using the threshold approach in Eq. 3 ($\sigma_{mz} = 1.4$ ppm, $nS_0 = 0.25$, $a = 0.01$).

All peaks in the group of matched spectra (including the base spectrum) were clustered using agglomerative clustering (from the `sklearn` package) to recursively merge peaks into a full tree based on a similarity threshold of $5 \times \sigma_{mz}$ (Fig. S5). Peaks belonging to the same cluster were considered to represent the same molecular fragment. Clusters present in at least a quarter of matching spectra (or at least two spectra) were considered to be true peaks not representing instrumental noise. An ‘average’ spectrum representing the group of matched spectra is then constructed from the clusters representing true peaks, by finding the mean mass, the mean intensity, and the minimum and maximum relative intensity of the peaks in each cluster. Metadata such as the number and identity of the matching spectra used to produce each average spectra, the number of peaks in each cluster, the spectrum and compound ID, and the sum formulas of peaks were recorded for each averaged spectrum. The base spectrum and the matching spectra were flagged as matched and not considered in following iterations. Aggregation and 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).



Supplementary Figure 5: Averaged spectra for succinic acid (butanedioic acid), calculated based on two groups of matching spectra with 8 and 7 members respectively. *a*) and *b*) show the *m/z* and relative intensity of all individual peaks from matching spectra, coloured by spectrum (*a*) and by peak mass cluster (*b*). *c*) The averaged spectra produced by aggregating the peaks in each cluster, based on clusters with at least 2 peaks. The sum formulas are shown for peaks where sum formulas could be assigned, and the fragmentation loss pattern is indicated. Relative intensity is shown with a logarithmic scale, and lines corresponding to relative intensities of 1, 0.1 and 0.01 are shown in pale grey.

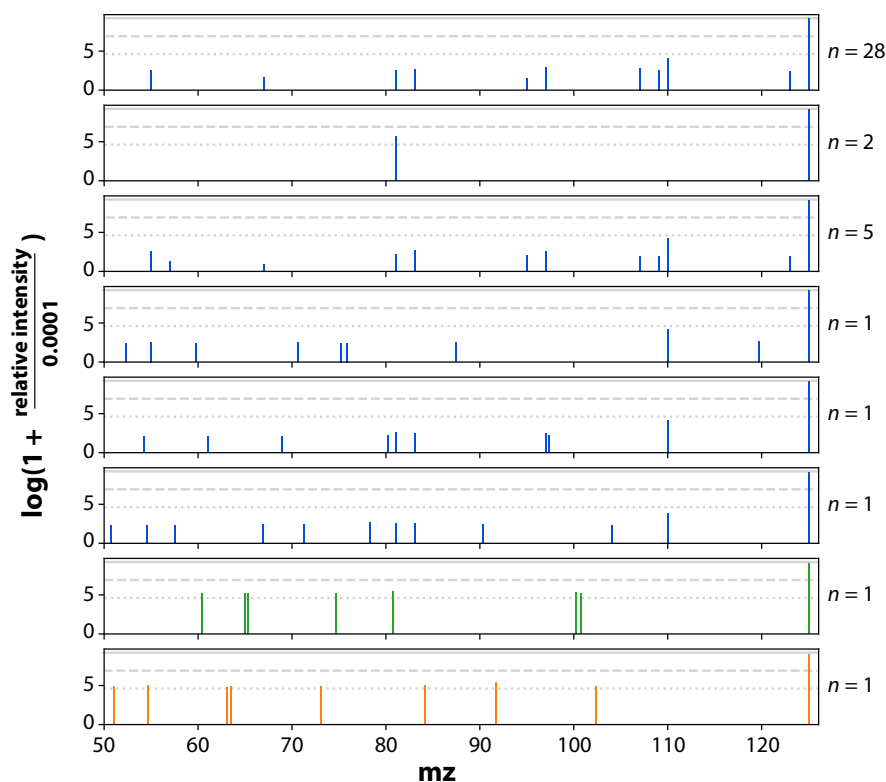
1.4 Quality control

The final part of the spectral processing involves quality checking and flagging of averaged spectra, as well as a consideration of possible isomers in CompoundDiscoverer compound identifications. MS2 spectra not containing the correct parent ion from the matching MS1 spectra are removed, and peaks larger than the parent ion are dropped from MS2 spectra, as these represent either adducts or noise (there was no evidence for doubly-charged parent ions for any spectra considered). Additionally, compounds with no assigned MS2 spectra are removed from the compound list.

All averaged spectra assigned to a particular compound by CompoundDiscoverer are intercompared and grouped using the similarity index. If all spectra are similar (based on Eqs. 1, 2, and Eq. 3 with $\sigma_{mz} = 1.4$ ppm, $nS_0 = 0.25$, $a = 0.01$) to at least one other spectra, they are considered to be matching, eg. if Compound X has 3 averaged spectra, and spectrum 1 matches spectrum 2 and spectrum 3, but spectrum 2 does not match spectrum 3, the spectra for Compound X are considered to be consistent. However, if Compound Y has 5 spectra, and spectrum 1 is similar to spectrum 2 and spectrum 3, but spectrum 4 is only similar to spectrum 5, the spectra for Compound Y will be flagged as belonging to two groups. In this case, it is likely that one group of spectra represents an isomer with similar molecular structure. An example of groupings of consistent and inconsistent averaged spectra is shown in Fig. S6.

1.4.1 Retention time prediction for quality control

The next stage of the quality control uses retention time and hydrophobicity to flag possible isomers in molecular structure assignments from CompoundDiscoverer. The retention times in the DiAcids dataset are calibrated to the Fallopi dataset



Supplementary Figure 6: Grouping of averaged spectra for Compound #26 (unnamed) in the Fallopi dataset. Bar colours indicate groups of consistently matching averaged spectra: Six of the eight averaged spectra for this compound agree (blue), whereas the spectra shown in green and orange are inconsistent. The number of MS2 spectra aggregated to produce each averaged spectrum is shown on the right. The grey dashed lines indicated relative intensities of 1, 0.1 and 0.01, as in Fig. S5.

based on eight compounds which were measured in both datasets [3]. The retention times between these two datasets were very consistent, with a slope of 0.97, an intercept of -0.19 minutes, and an R^2 of 0.96. The Ruthenium dataset had no compounds in common for calibration, however given the consistency of the other two datasets which were measured on the same set up, retention times in the Ruthenium dataset can reasonably be compared without calibration.

Hydrophobicity (logP) estimates were not available for all identified compounds, and moreover, hydrophobicity estimates for many compounds in PubChem can be highly uncertain. A range of molecular representations and machine learning methods were therefore investigated to predict hydrophobicity using literature datasets of measured hydrophobicity for training and testing [4]. The graph neural network (GNN) + multitask learning method offered the best compromise between in-sample and out-of-sample hydrophobicity prediction, and was therefore selected for hydrophobicity prediction in this study [4]. Hydrophobicity was predicted for all compounds identified by CompoundDiscoverer, and additionally for all potential compound matches (see Section 2.7 of the main article).

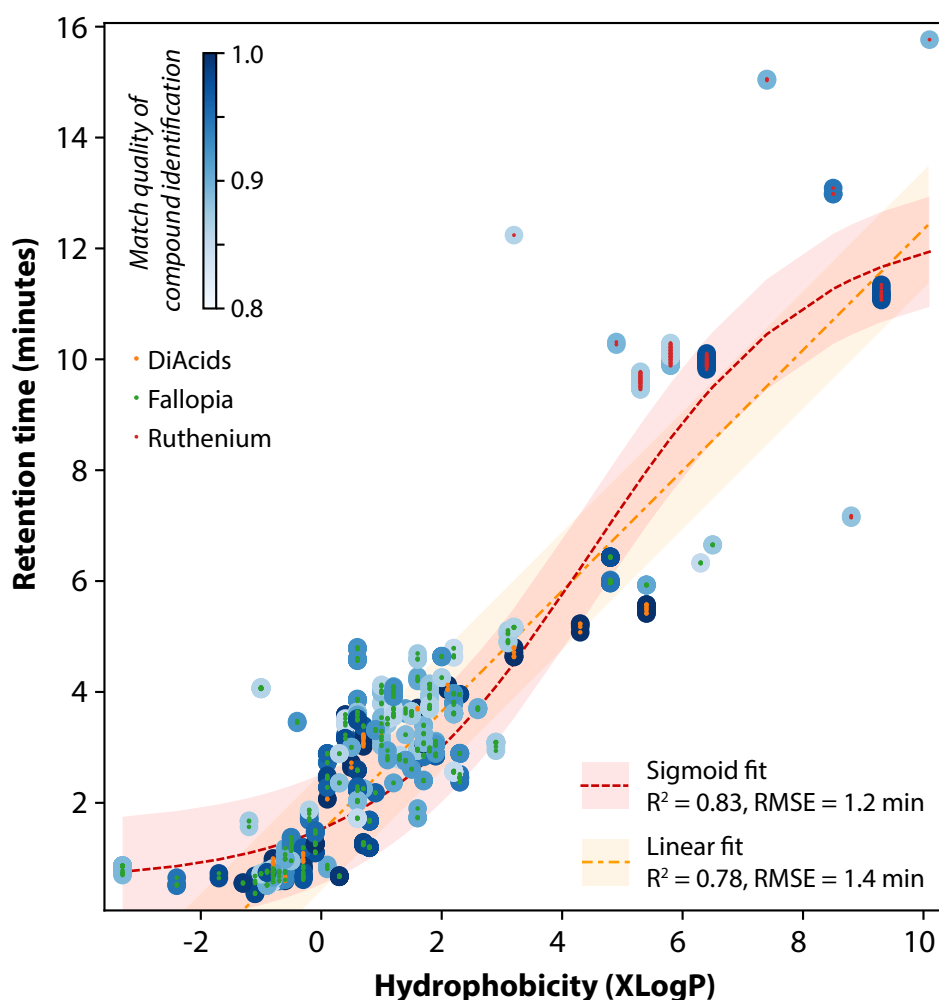
The relationship between retention time and hydrophobicity was investigated for all measurements where a structure was assigned with a match quality of ≥ 0.85 by CompoundDiscoverer. Retention time was then fit based on hydrophobicity using a linear equation ($RT = a + b \times H$) and a sigmoid equation:

$$RT = \frac{L}{1 + \exp(-k * (H - H_0))} + c \quad (4)$$

where RT is the retention time, H is the hydrophobicity, and a, b, c, H_0, k, L are constants describing the fits. A bootstrapping approach was used to find the best fits and determine uncertainty. Half of the data was used with replacement for training the

model fit in each bootstrap, and additionally, retention time values were perturbed according to the measurement uncertainty of 10 seconds. The equation was fit using the `scipy.optimize.curve_fit` function. The fit was used to predict retention time for the validation data, and the absolute difference between each measured and modelled retention times were found. Following 1000 bootstrap runs, the mean fit parameters were found, and the mean absolute difference between test data and measurements was used as an estimate of the uncertainty in retention time prediction (σ_{RT}). The linear model was fit with $a = 1.45$ and $b = 1.1$ to achieve $R^2 = 0.78$, $\sigma_{RT} = 1.1$ and RMSE = 1.4. The sigmoid model outperformed the linear model, with parameters of $L = 11.9$, $H_0 = 4.5$, $k = 0.55$ and $c = 0.6$ delivering a fit with $R^2 = 0.83$, $\sigma_{RT} = 1.0$ and RMSE = 1.2 (Figure S7).

The sigmoid model was used to predict retention time based on logP for all matched compounds. For all spectra with a compound identification, the predicted and measured retention time were compared. If they did not agree within σ_{RT} , the compound identification for the spectrum was flagged as a possible isomer of the true molecular structure. Additionally, the retention time was used to predict the expected hydrophobicity (logP) for each spectrum, for later comparison during molecular structure identification.

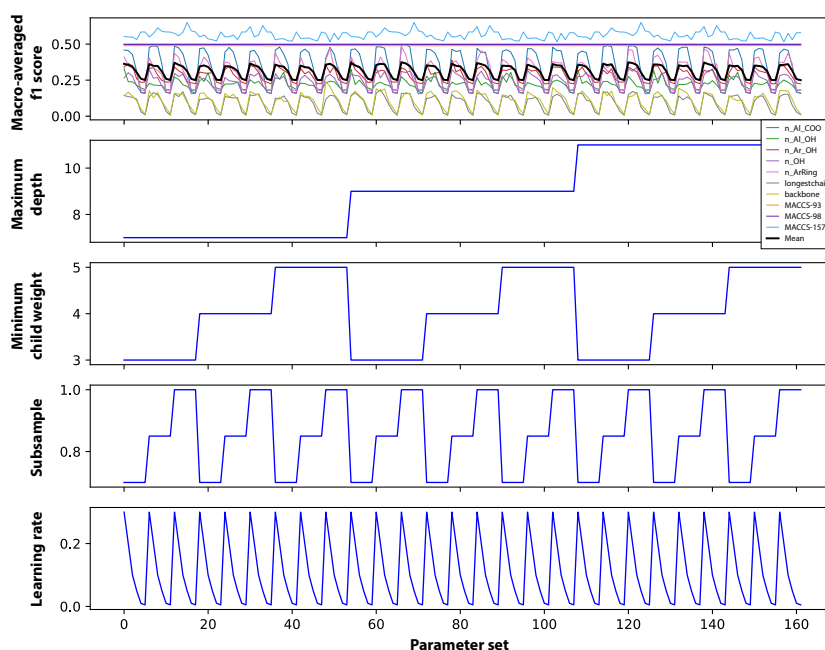


Supplementary Figure 7: Relationship between compound hydrophobicity (described using XLogP from the PubChem database) and retention time for the DiAcids, Fallopiia and Ruthenium datasets, fit using linear and sigmoid equations. The uncertainty in predicted retention time, based on the mean absolute difference between measurements and validation data in each bootstrap, is shown with shaded regions.

2 SI text: Hyperparameter tuning

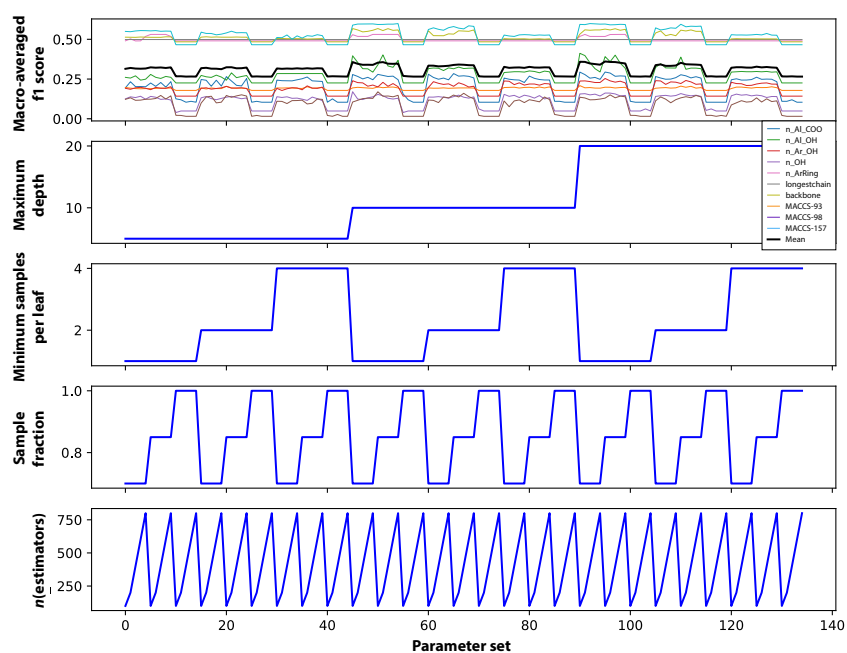
Hyperparameter tuning for the RF-G and XGB-G models was carried out using the input data features selected for prediction of each fingerprint digit in (see Section 3.2 of the main article) with the train/validate dataset. The 10 fingerprint digits with the lowest prediction quality (F1 score) were used for hyperparameter tuning: n_Al_COO, n_Al_OH, n_Ar_OH, n_OH, n_ArRing, longestchain, backbone, MACCS-93 (QCH₃ where Q is any heteroatom), MACCS-98 (heteroatom-containing ring) and MACCS-157 (C-O group). The models were trained and validated using three cross-validation folds for all combinations of hyperparameters for the 10 weakest fingerprint digits.

Hyperparameter tuning for XGB-G tested maximum depth of 7, 9, 11; minimum child weight of 3, 4, 5; subsample of 0.7, 0.85, 1; and learning rate (eta) of 0.3, 0.2, 0.1, 0.05, 0.01, 0.005. The learning rate had the largest impact on prediction quality across the fingerprint digits. Optimal model performance was found with maximum depth = 9; minimum child weight = 3; subsample = 1; and learning rate = 0.3 (Figure S8).



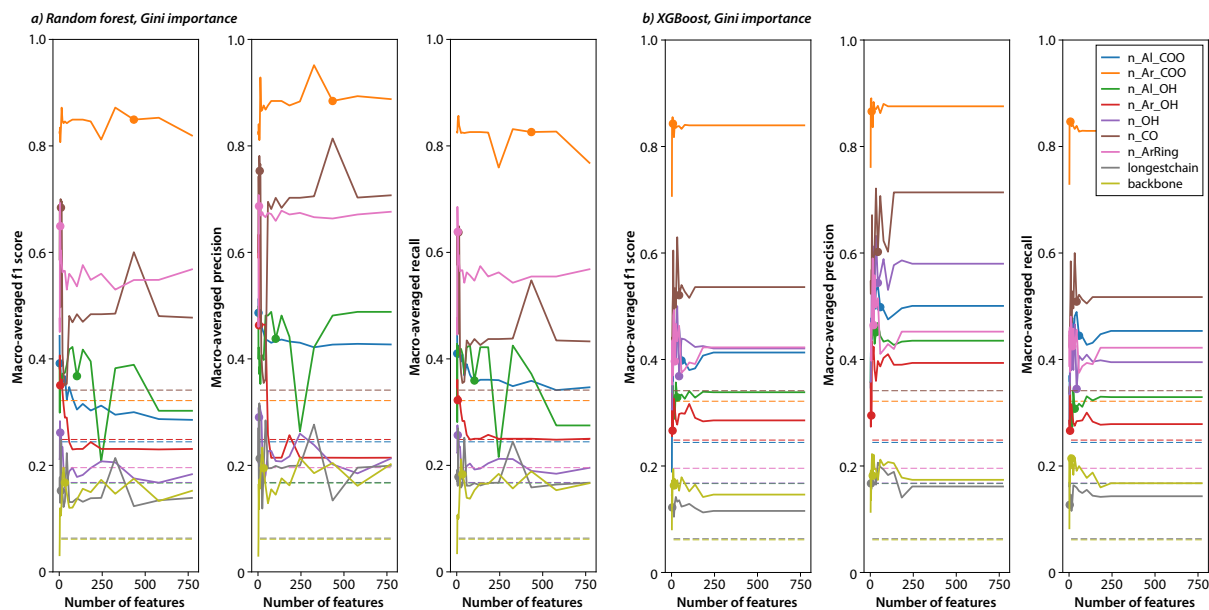
Supplementary Figure 8: Macro-averaged F1 score for each of the 10 weakest fingerprint digits using different hyperparameter combinations for the XGB-G model.

Hyperparameter tuning for RF-G tested maximum depth of 5, 10, 20; minimum samples per leaf of 1, 2, 4; sample fraction of 0.7, 0.85, 1; and $n(\text{estimators})$ of 10, 200, 400, 600, 800. The sample fraction used to train the model and the maximum depth had the largest impact on prediction quality across the fingerprint digits. Optimal model performance was found with maximum depth = 20; minimum samples per leaf = 1; sample fraction = 0.85; and $n(\text{estimators}) = 100$ (Figure S9).

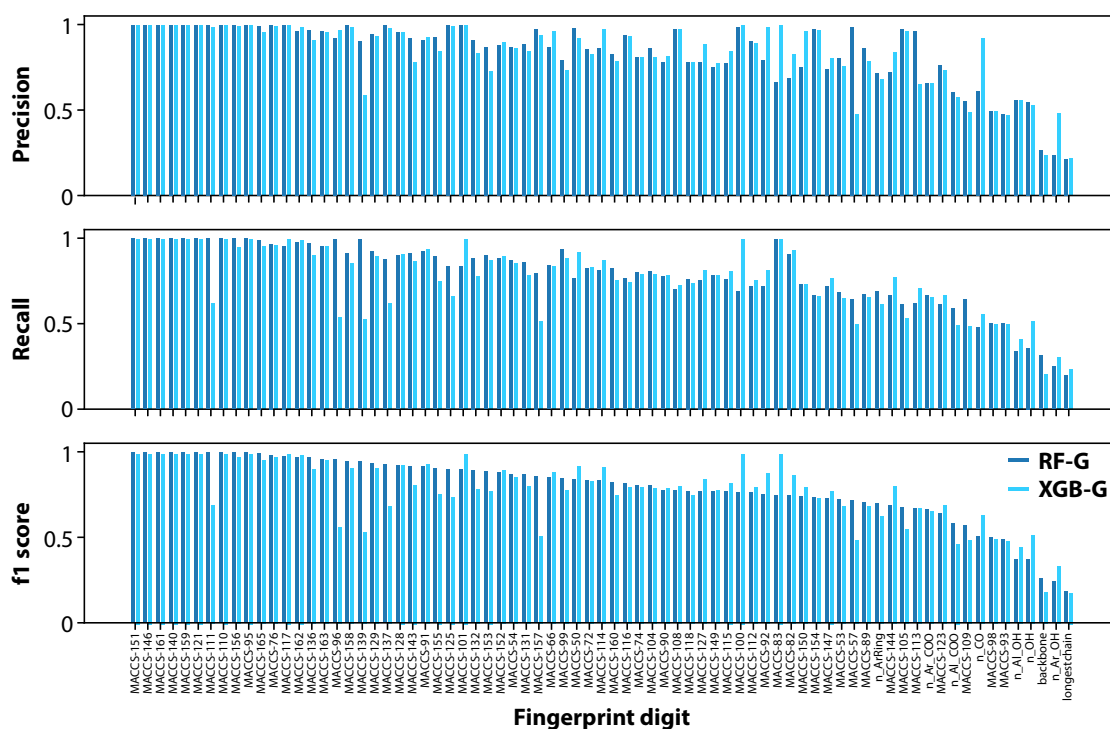


Supplementary Figure 9: Macro-averaged F1 score for each of the 10 weakest fingerprint digits using different hyperparameter combinations for the RF-G model.

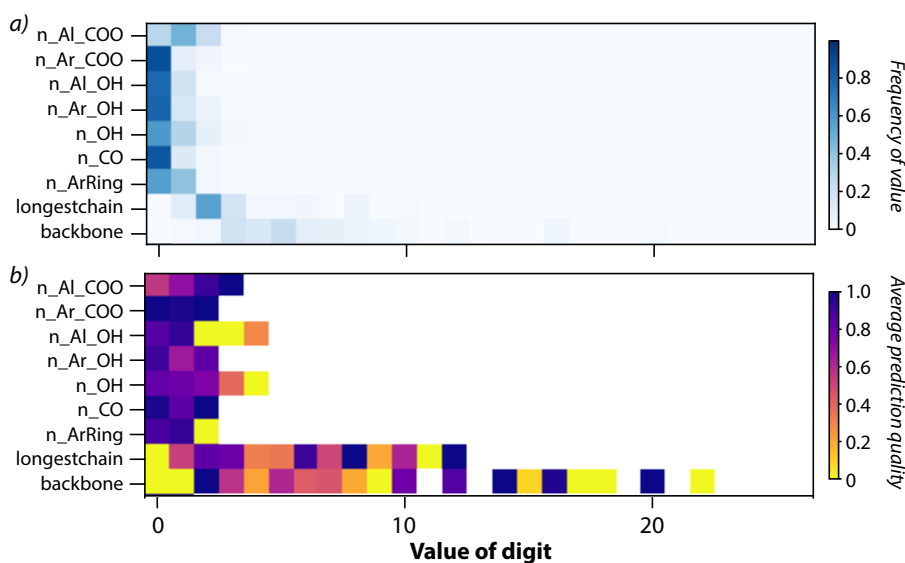
3 SI: Additional Figures



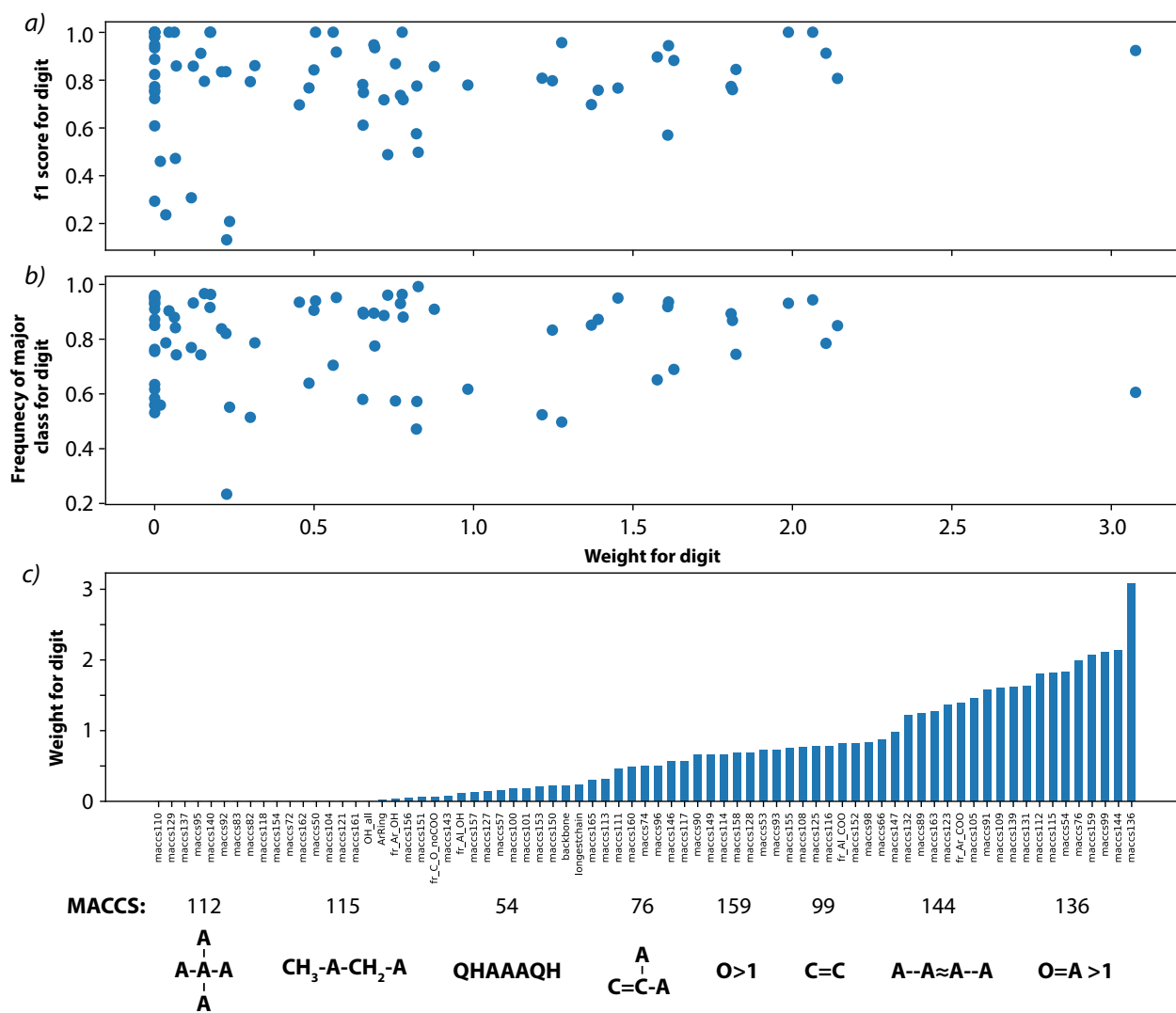
Supplementary Figure 10: Model performance with varying number of features for the first nine fingerprint digits (Table 1 (main article)) using the RF (a) and XGB (b) models with Gini feature importance. The optimum number of features for each digit is shown with a circle. The dashed lines show the random baseline prediction performance for comparison with model performance for each fingerprint digit.



Supplementary Figure 11: Macro-averaged precision, recall and F1 score for all fingerprint digits predicted for the test dataset using optimized RF-G and XGB-G models. Fingerprint digits are ordered according to the F1 score for the RF-G model.



Supplementary Figure 12: Frequency and prediction quality for non-MACCS digits of the molecular fingerprint using the RF-G model (Table 2 (main article)). The colour scale in *a*) shows the frequency of each value of each fingerprint digit. The colour scale in *b*) shows the prediction quality for each value of each digit.



Supplementary Figure 13: Fingerprint digit weights found using an SVM with expectation values. *a)* The macro-average F1 score for each digit compared to the weight; *b)* The frequency of the most common class for each digit compared to the weight; *c)* The weights for each digit. The most important eight MACCS keys are shown below panel *c*: A represents any atom, Q represents any non-C and non-H heteroatom. Single and double bonds are shown; – indicates a non-aromatic bond and ≈ indicates an aromatic bond.

4 SI: Additional Tables

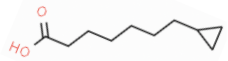
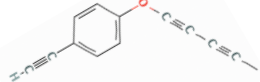
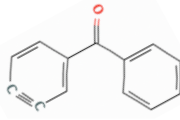
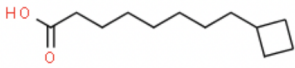
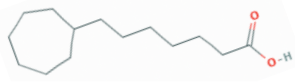
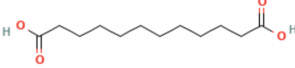
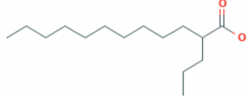
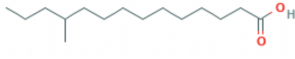
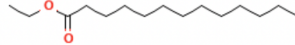
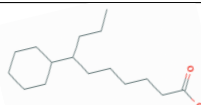
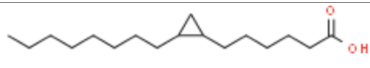
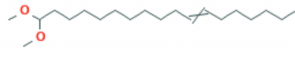
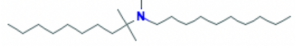
Supplementary Table 1: Example of the peak priority ranking procedure using an averaged spectra for glutaric acid taken from the DiAcids dataset (Table 1 (main article)). ‘Intensity groups’ refers to relative intensity bands of 10-100%, 1-10%, 0.1-1% etc. ‘Loss rankings within intensity groups’ ranks peaks by the mass of the lost fragment (mz, loss) within each intensity group. ‘Final priority’ describes the combined priority of each peak according to both intensity and loss rankings; the parent ion is always given priority of 1 and peaks with no assigned sum formula are not ranked.

mz	mz, loss	Intensity	Rel. intensity	Sum formula	Intensity group	Loss ranking within int. group	Final priority
59.0	72.0	2692	0.0018	C ₂ H ₃ O ₂	4	3	5
69.0	62.0	96800	0.0629	C ₄ H ₅ O	2	2	4
87.0	44.0	1051926	0.6840	C ₄ H ₇ O ₂	1	2	2
113.0	18.0	104388	0.0679	C ₅ H ₅ O ₃	2	1	3
115.9	15.1	2606	0.0017		4	2	NA
130.9	0.1	3318	0.0022		4	1	NA
131.0	0.0	276137	0.1796	C ₅ H ₇ O ₄	1	1	1

Supplementary Table 2: Comparison of feature selection for fingerprint prediction, comparing random forest (RF) and extreme gradient boosting (XGB) models with Gini (G) and permutation (P) feature importances. ‘Diff.’ refers to the percentage difference of the Model 1 results compared to Model 2, averaged across all fingerprint digits. ‘% of features chosen by both models’ is the percentage of features chosen by the model that selected less features which are also chosen by the other model.

Model 1	Model 2	Diff. in F1 score	Diff. in precision	Diff. in recall	Diff. in number of features	% of features chosen by both models
RF-G	RF-P	+8.1	+6.6	+8.0	-48	57
XGB-G	XGB-P	-0.1	+1.8	-0.8	+11	64
RF-G	XGB-G	+3.6	+3.7	+3.5	+10	62
RF-P	XGB-P	-1.7	+1.3	-3.0	+70	52

Supplementary Table 3: Compounds identified only for spectra corresponding to reactivated catalyst (see [5] for details). Structural images are taken from the PubChem database.

	Name	MW	Formula	Image
1	7-Cyclopropylheptanoic acid	170.2	C ₁₀ H ₁₈ O ₂	
2	1-ethynyl-4-penta-1,3-diynoxybenzene	180.2	C ₁₃ H ₈ O	
3	cyclohexa-1,5-dien-3-yn-1-yl(phenyl)methanone	180.2	C ₁₃ H ₈ O	
4	8-Cyclobutyloctanoic acid	198.3	C ₁₂ H ₂₂ O ₂	
5	7-cycloheptylheptanoic acid	226.4	C ₁₄ H ₂₆ O ₂	
6	Dodecanedioic acid	230.3	C ₁₂ H ₂₂ O ₄	
7	2-propyldodecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
8	11-methyltetradecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
9	5-ethyltridecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
10	7-Cyclohexyldecanoic acid	254.4	C ₁₆ H ₃₀ O ₂	
11	6-(2-octylcyclopropyl)hexanoic acid	268.4	C ₁₇ H ₃₂ O ₂	
12	18,18-Dimethoxyoctadec-7-ene	312.5	C ₂₀ H ₄₀ O ₂	
13	N-decyl-N,2-dimethyldecan-2-amine	325.6	C ₂₂ H ₄₇ N	

Supplementary Table 4: Compounds identified only for spectra corresponding to inactive catalyst (see [5] for details). Structural images are taken from the PubChem database.

	Name	MW	Formula	Image
1	7-hydroxychromen-2-one	162.1	C ₉ H ₆ O ₃	
2	8-ethyldecanoic acid	200.3	C ₁₂ H ₂₄ O ₂	
3	12-hydroxydodecanal	200.3	C ₁₂ H ₂₄ O ₂	
4	Tetracyclo[10.2.2.02,11.04,9]hexadeca-1(14),2,4,6,8,10,12,15-octaene	202.3	C ₁₆ H ₁₀	
5	Cyclopentane;decanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
6	2-ethyltridecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
7	3-tridecyloxiran-2-ol	242.4	C ₁₅ H ₃₀ O ₂	
8	pentadecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
9	13-methyltetradecanoic acid	242.4	C ₁₅ H ₃₀ O ₂	
10	cycloheptane;decanoic acid	270.5	C ₁₇ H ₃₄ O ₂	
11	methyl (E,9S)-9-hydroxyoctadec-10-en-12-ynoate	308.5	C ₁₉ H ₃₂ O ₃	
12	methyl (9E,11E)-8-hydroxyoctadeca-9,11-dienoate	310.5	C ₁₉ H ₃₄ O ₃	
13	methyl 9-hydroxyoctadeca-2,4-dienoate	310.5	C ₁₉ H ₃₄ O ₃	
14	(E)-docos-11-enamide	337.6	C ₂₂ H ₄₃ NO	

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.

Declarations

Ethical Approval

Not applicable.

Competing interests

Not applicable.

Authors' contributions

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

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.

Availability of data and materials

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 [4] can be downloaded from:

- <https://renkulab.io/gitlab/msei/capstone-project>

References

- [1] C. A. Hughey, C. L. Hendrickson, R. P. Rodgers, A. G. Marshall, and K. Qian. Kendrick mass defect spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra. *Analytical Chemistry*, 73(19):4676–4681, 2001.
- [2] Huibin Shen, Kai Duehrkop, Sebastian Boecker, and Juho Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(12), 2014.
- [3] C Lanz, A Patterson, J Slavik, K Krausz, M Ledermann, F Gonzalez, and J Idle. Radiation Metabolomics. 3. Biomarker Discovery in the Urine of Gamma-Irradiated Rats Using a Simplified Metabolomics Protocol of Gas Chromatography-Mass Spectrometry Combined with Random Forests Machine Learning Algorithm. *Radiation Research*, 172(2):198–212, 2009.
- [4] Patrik Friedlos, L Gasser, and Eliza Harris. Retention time prediction to facilitate molecular structure identification with tandem mass spectrometry. *bioRxiv*, pages 1–20, 2022.
- [5] Urs Gasser and Sasa Bjelic. Deposits of non-volatile organics in methanation catalyst (Ru/C) studied with SANS. *In preparation*, pages 1–7, 2022.