Supporting Information

DecoID improves identification rates in metabolomics through database-assisted MS/MS deconvolution

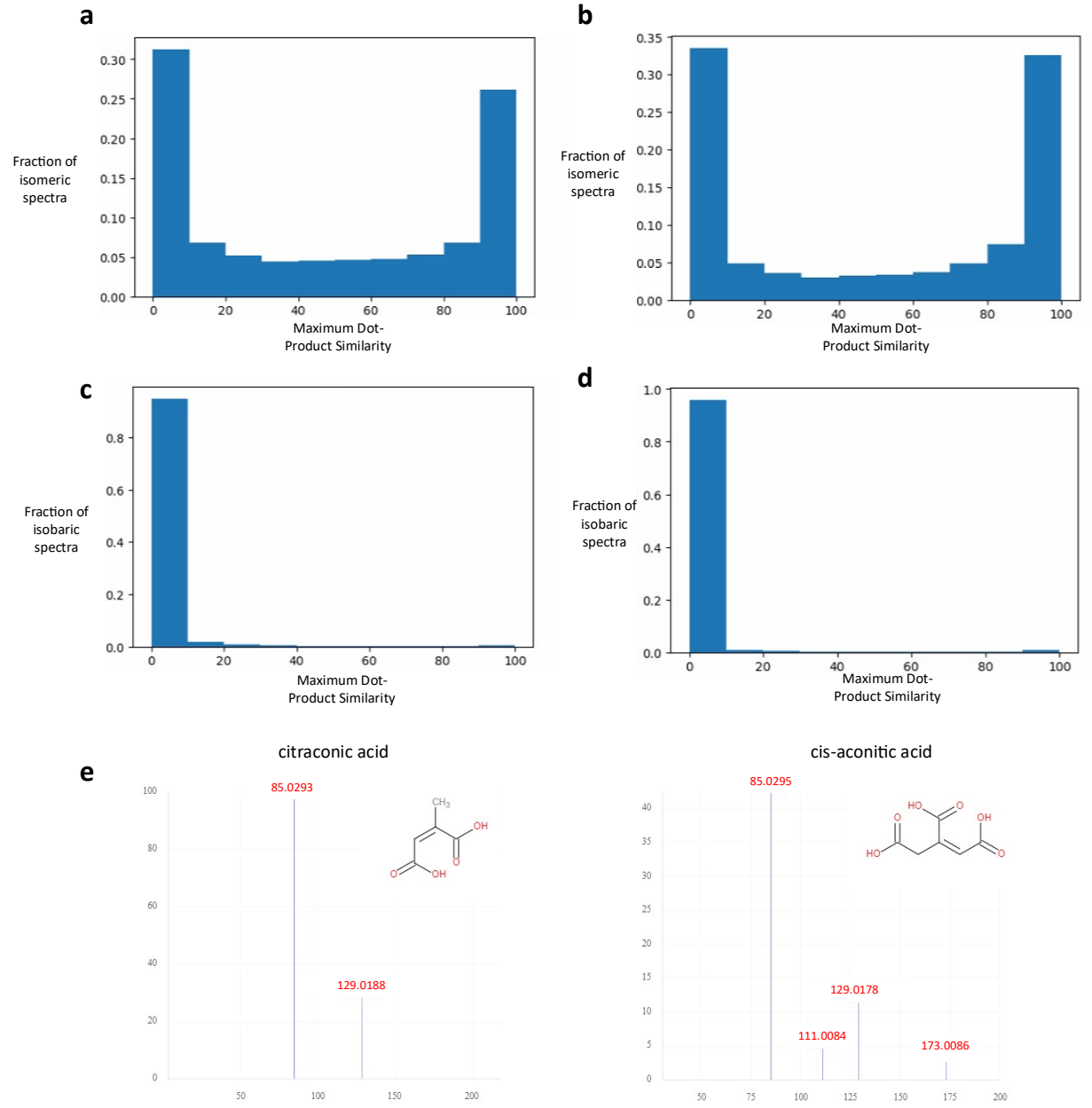
Ethan Stancliffe[1,2], Michaela Schwaiger-Haber[1,2], Miriam Sindelar[1,2], Gary J. Patti[1,2,*]

[1]Department of Chemistry, Washington University in St. Louis, St. Louis, MO, USA

[2]Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA

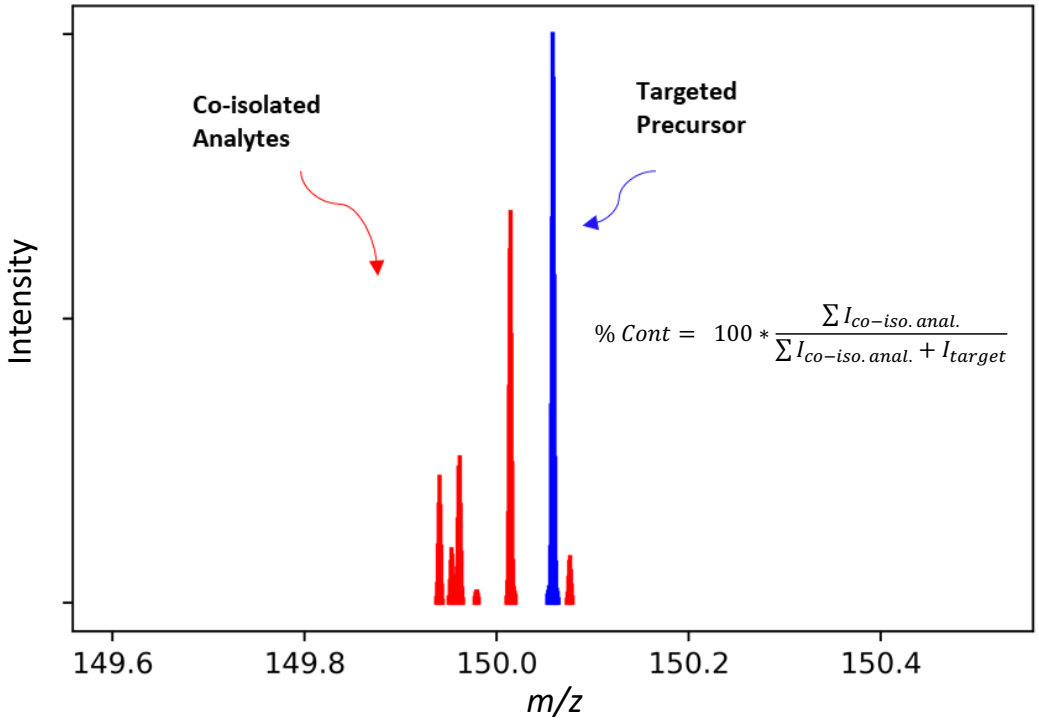
*To whom correspondence should be addressed: gjpattij@wustl.edu

Supplementary Fig. 1



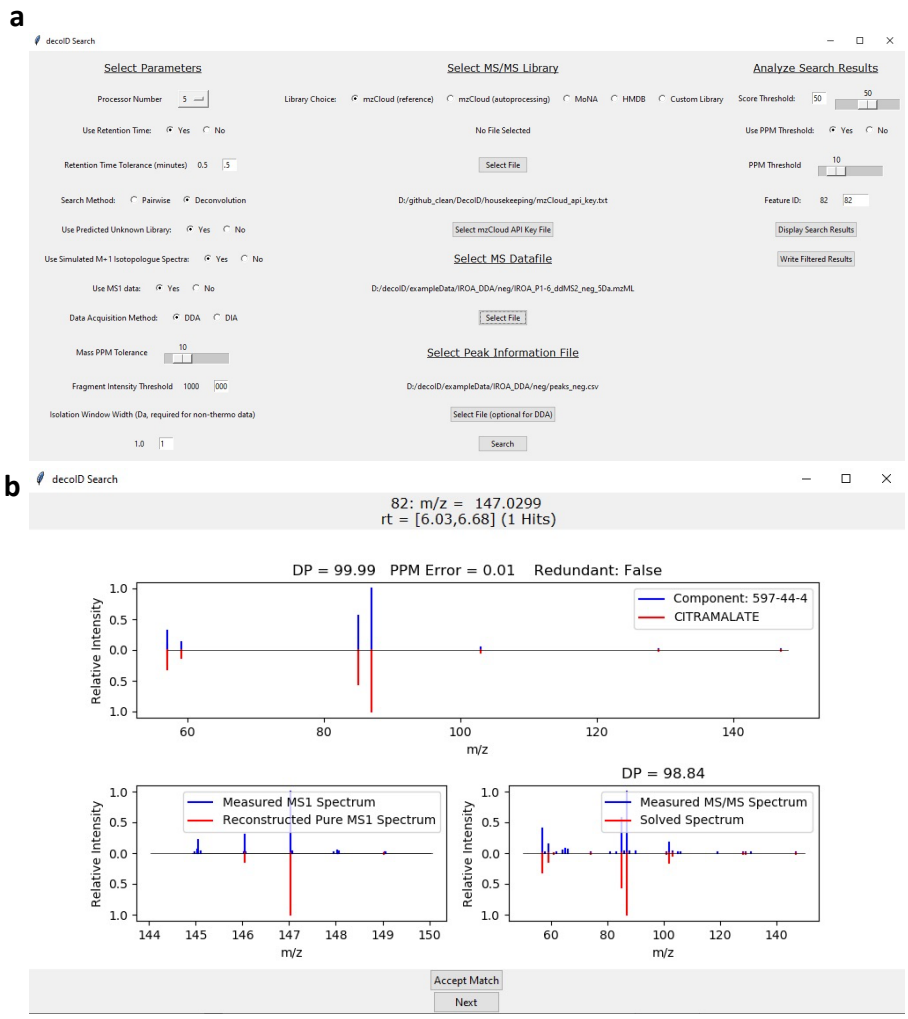
Supplementary Fig. 1 | Similarity of metabolite MS/MS spectra. a,**b**, Histograms showing the distribution of maximum dot-product similarities between experimental MS/MS spectra of isomeric compounds in the Mass Bank of North America (MoNA) database for positive mode (**a**) and negative mode (**b**). **c**,**d**, Histograms showing the distribution of maximum dot-product similarity between MS/MS spectra of compounds that are isobaric (having the same nominal mass) but not isomeric (having the same exact mass) in the MoNA database in positive mode (**c**) and negative mode (**d**). **e**, Example of two structurally distinct compounds with highly similar MS/MS spectra. Data were retrieved from the MoNA database. Only the smaller fragments at *m/z* 111 and 173 differentiate citraconic acid from cis-aconitic acid. We note that when metabolomic MS/MS spectra have a relatively small number of fragments, as shown here, high confidence identifications require high dot-product similarity scores.

Supplementary Fig. 2



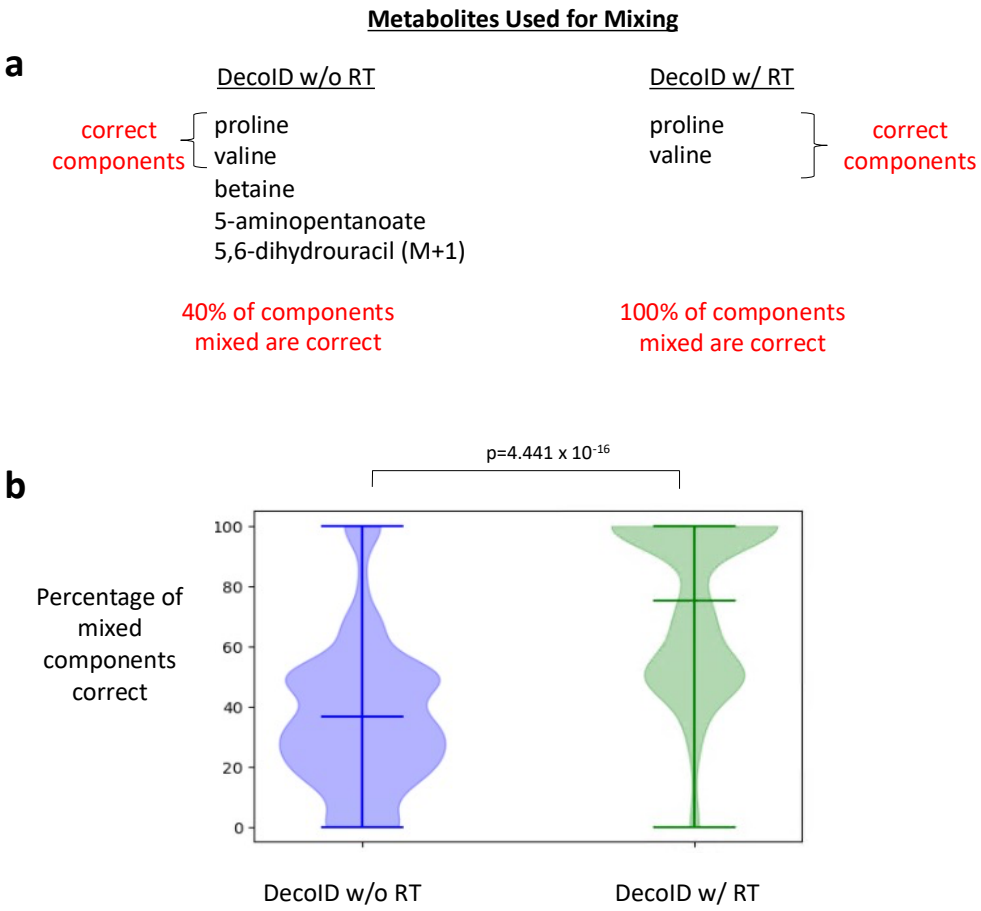
Supplementary Fig. 2 | Metric for assessing degree of contamination in an MS/MS spectrum. The nearest MS1 spectrum to the MS/MS spectrum of interest is selected. The region of the MS1 spectrum representing the isolation window of the MS/MS spectrum is examined and the signal coming from co-isolated analytes (red) is summed to give $\sum I_{co-iso.\ anal.}$. The summed intensity is then divided by the total sum of the co-isolated analytes and the targeted precursor ion (blue), I_{target}, to give an estimate of the percent contamination in the resulting MS/MS spectrum.

Supplementary Fig. 3



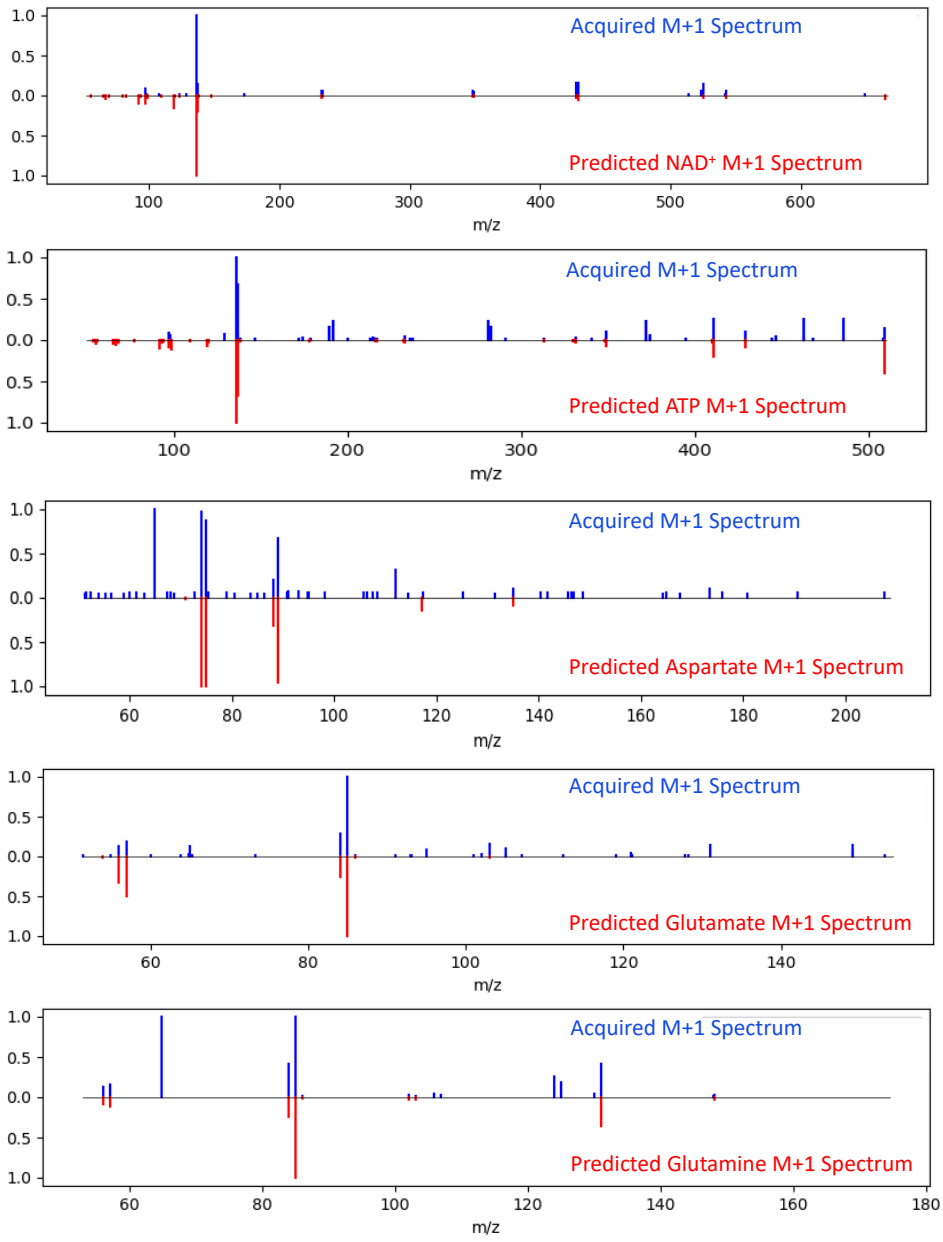
Supplementary Fig. 3 | User interface for DecoID. a, Screenshot showing user-defined entries within the DecoID interface. The "Processor Number" specifies the number of spectra to deconvolve in parallel on separate CPU cores. If one does not wish to deconvolve MS/MS spectra, the "Pairwise" search method can be selected. For increased accuracy in the deconvolution, retention time can be toggled to "yes" if the database has retention times. A tolerance can also be set to determine how far database and observed retention times can differ. The unknown library and predicted M+1 isotopologues can be toggled on or off. The acquisition method must be entered as either DDA or DIA. The mass PPM tolerance sets the internal mass tolerance of DecoID and should be set based on the mass accuracy of the instrument. An intensity threshold can also be used to remove low-intensity fragment ions. In the example setup above, only fragment ions above an intensity of 1000 will be considered. For non-Thermo datafiles, the isolation window width used to acquire the data must be entered. The choice of MS/MS database can be selected and the path to the MS data file must be provided. For the mzCloud database, the API key file must be selected. The peak information file path must be entered for DIA but is optional for DDA. After deconvolution and identification, the results can be visualized for the entire dataset or just a single feature of interest by entering a specific featureID (i.e., the row number of a feature in the peak list or the scan number if not provided). Results can be filtered by mass error and dot-product similarity. **b**, An example match visualization. The top panel shows the purified spectrum after deconvolution (labeled by the component's compoundID, Methods) in blue and its database match (citramalate) in red. The reconstructed MS1 spectrum is shown in the bottom left along with the acquired MS1 spectrum. In the bottom right, the entire reconstructed MS/MS spectrum (minus the residual noise) is shown in red with the measured MS/MS spectrum in blue. The similarity of these two spectra is also provided (98.84 in this example). Details about these reconstructed spectra can be found in the Methods. The user can accept this match. If accepted, this entry will be appended to an additional csv file.

Supplementary Fig. 4



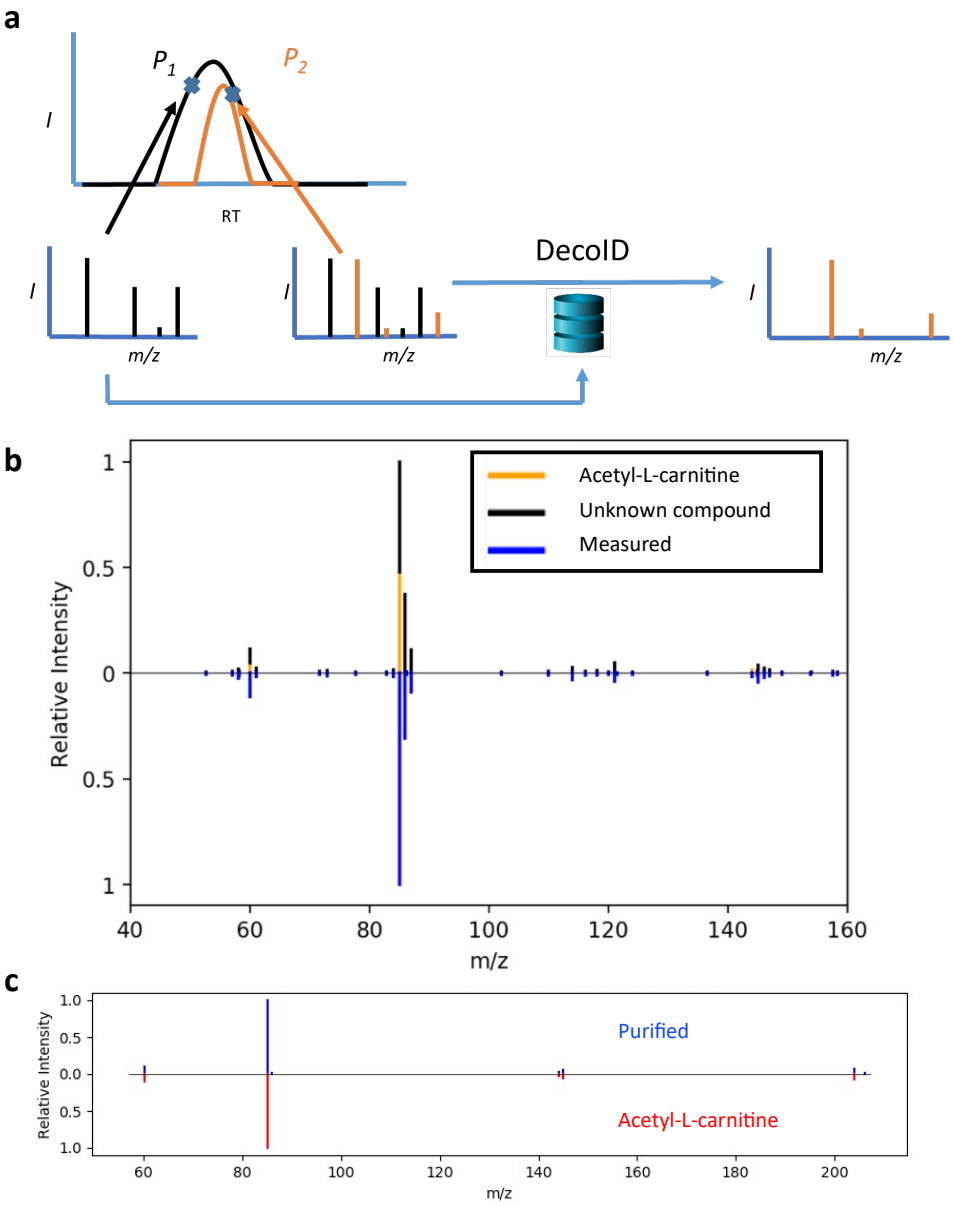
Supplementary Fig. 4 | Use of retention times from the in-house database improves the accuracy of DecoID deconvolutions on the IROA DDA dataset. **a**, Toy example of when a chimeric spectrum of proline and valine was deconvolved by using DecoID with and without retention time. When retention time was used, 100% of the components mixed were correct. Without retention time, only 40% of the components mixed were correct. **b**, Violin plots showing the distribution of the percentage of components mixed that are correct in the IROA standard mixture DDA dataset. The results were searched against the in-house database with and without retention time. When retention time was used, the mean percentage of correct components increases from 37% to 75%. Correct components were determined from the retention-time bounds of the mixture metabolites and the exact times when MS/MS data were acquired. Statistical significance was computed by using a two-sided, two-sample Kolmogorov–Smirnov test. The horizontal lines in the violin plots mark the mean, minimum, and maximum percentage.

Supplementary Fig. 5



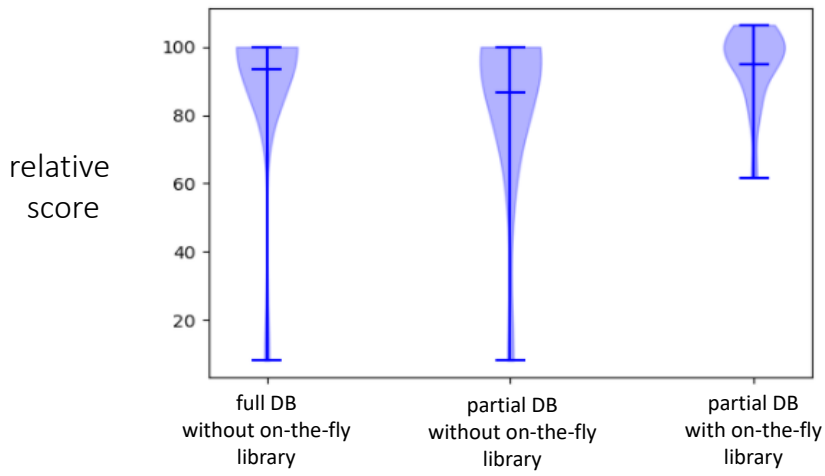
Supplementary Fig. 5 | Example M+1 spectrum predictions are highly similar to experimentally acquired M+1 spectra. M+1 spectra were predicted by using the DecoID algorithm for glutamine, glutamate, aspartate, ATP, and NAD^+ from the mzCloud M+0 spectra. The predicted spectra (shown in red) are highly similar to the M+1 spectra acquired from reference standards (shown in blue).

Supplementary Fig. 6



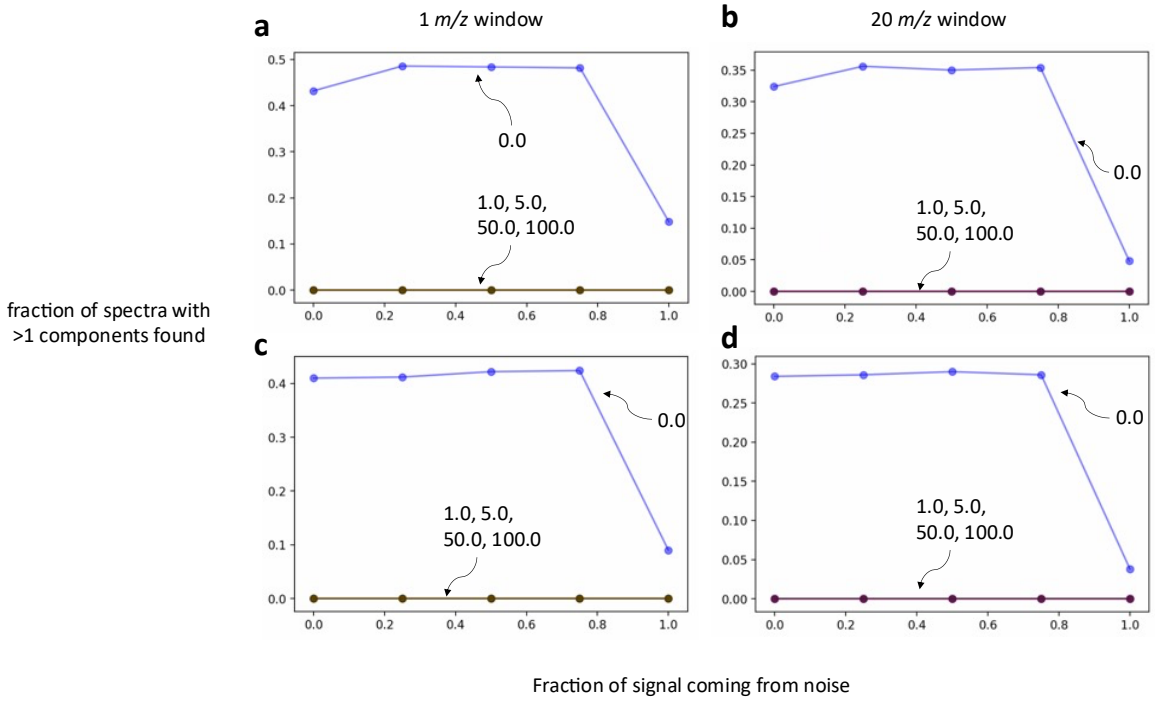
Supplementary Fig. 6 | Generation of an 'on-the-fly' library for unknowns and example deconvolution from the NIST 1950 dataset. a, Extracted ion chromatogram of two isobaric precursors P_1 and P_2. The blue "x" indicates where MS/MS data were acquired. A non-contaminated MS/MS spectrum of P_1 was acquired. However, the acquired MS/MS spectrum for P_2 is contaminated with fragments from P_1. DecoID uses the first spectrum to deconvolve the contaminated spectrum, even though P_1 does not return any library matches. The pure spectrum for P_1 is added to the library spectra used for deconvolution of acquired MS/MS spectra where P_1 is present. **b**, Deconvolution of a MS/MS spectrum of acetyl-L-carnitine from human plasma. The acetyl-L-carnitine spectrum is contaminated by an unknown metabolite whose pure MS/MS spectrum was acquired. **c**, After deconvolution, the purified spectrum is a near exact match to the reference spectrum of acetyl-L-carnitine.

Supplementary Fig. 7



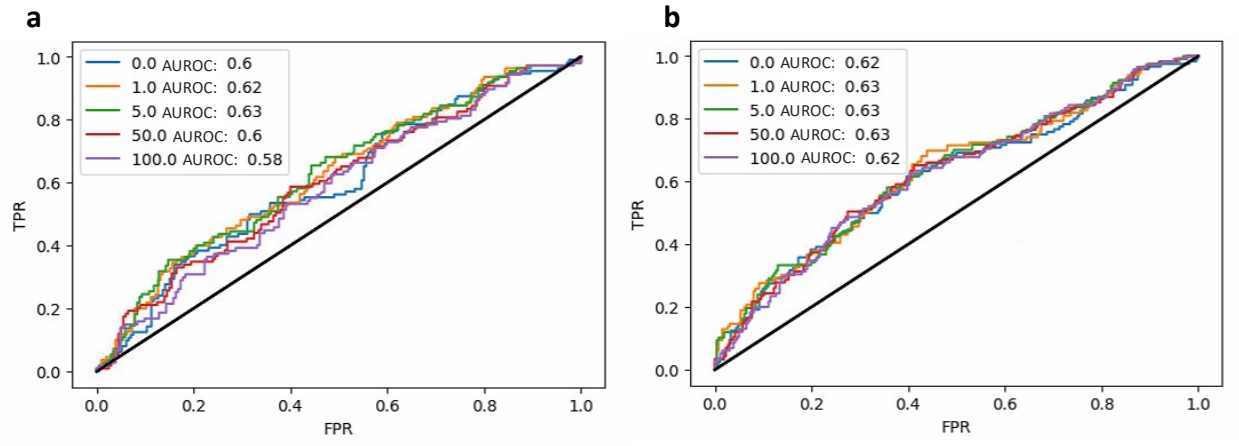
Supplementary Fig. 7 | The 'on-the-fly' library rescues DecoID performance when database coverage of the IROA DDA dataset is low. The impact of the on-the-fly unknown library was assessed when the positive-mode DDA spectra from the IROA standard mixture (1 *m/z* isolation window) were deconvolved with DecoID. First, the spectra were deconvolved by using the on-the-fly unknown library and the complete in-house database. Similarity scores for each compound were recorded. Second, the spectra were deconvolved without the on-the-fly unknown library but still using the complete in-house database, leading to a mean 7% decrease in similarity scores (left). Next, the spectra were deconvolved with a partial database where 50% of the entries in the in-house database were randomly removed. This was done without the on-the-fly library and resulted in a further reduction in database scores to a mean 87% of the original values (middle). Lastly, the spectra were deconvolved by using the partial database. However, this time the on-the-fly library was used (right), rescuing scores to greater than 96% of what was achieved when both the on-the-fly library and the complete database were used. Data shown are for compounds not removed during the database reduction but that were contaminated by one of the forgotten database compounds. The relative score is the dot-product similarity of a purified spectrum relative to the dot-product similarity of a spectrum purified by using the complete in-house database and the on-the-fly unknown library. Horizontal lines in the violin plots mark the mean, minimum, and maximum score.

Supplementary Fig. 8



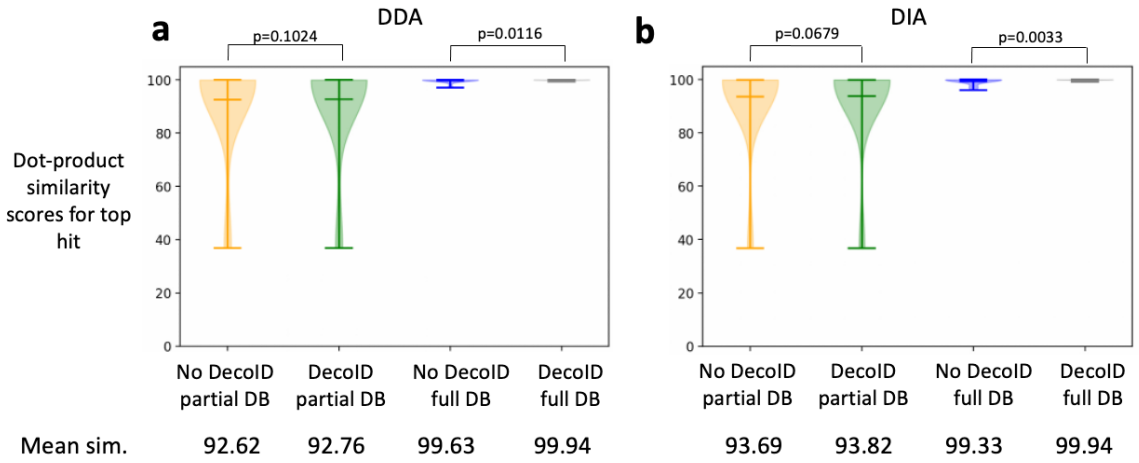
Supplementary Fig. 8 | DecoID does not falsly deconvolve synthetic datasets of non-chimeric spectra. To ensure that DecoID does not lead to unfaithful deconvolutions, two null evaluations were performed on 500 simulated DDA spectra with 1 *m/z* isolation windows (**a**, **c**) and 500 simulated DIA spectra with 20 *m/z* isolation windows (**b**, **d**). The simulated spectra were composed of randomly selected MS/MS spectra from MoNA with simulated noise added to amount to 0%, 25%, 50%, 75%, or 100% of the total MS/MS spectrum. In the first null evaluation, DecoID was applied to deconvolve the simulated noisy, but non-chimeric, spectra by using the full MoNA database. For each of the 500 simulated DDA and DIA spectra, the fraction of the spectra that we deconvolved into >1 component was calculated. In (**a-b**), the behavior of DecoID on the simulated dataset with different LASSO parameter values (0.0, 1.0, 5.0, 50.0, and 100.0) was characterized. In all cases, except when the LASSO parameter was zero (amounting to linear regression), DecoID did not falsely deconvolve the simulated dataset. In the second null evaluation (**c**-**d**), the same simulated dataset was deconvolved with DecoID at various LASSO parameters, but the spectra for the compounds in the simulated dataset were removed from the database. Again, in all cases except when the LASSO parameter was equal to zero, DecoID did not falsely deconvolve the non-chimeric spectra into linear combinations of other compounds' spectra. These results underscore the inadequacy of linear regression to deconvolve metabolomic MS/MS spectra and the importance of the LASSO penalty to the deconvolution.

Supplementary Fig. 9



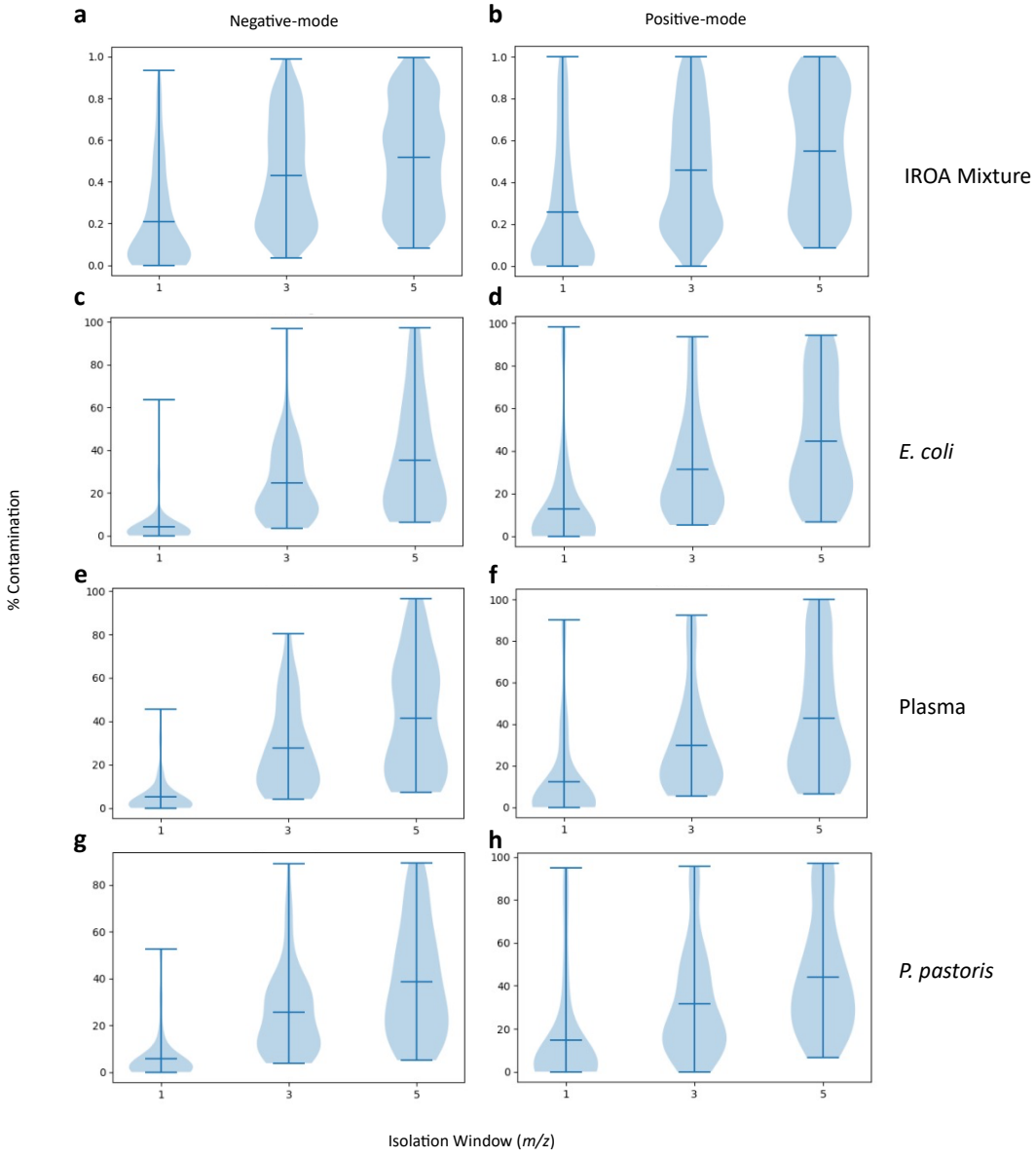
Supplementary Fig. 9 | Tuning the DecoID LASSO penalty by using the parameter optimization DDA and DIA datasets. When optimizing the LASSO parameter for DecoID to give the best area under the receiver operating characteristic curve (AUROC), five different parameter values were tested (0.0, 1.0, 5.0, 50.0, and 100.0) on a DDA and and DIA dataset from a mixture of chemical standards used only to optimize the DecoID parameters. The top three hits for each mixture compound were considered. For DDA (**a**), a LASSO parameter value of 5.0 gave the highest AUROC. For DIA (**b**), 1.0, 5.0, and 50.0 gave equally high AUROC. Although these parameters gave the highest result, DecoID's performance is robust to the choice of LASSO parameter.

Supplementary Fig. 10



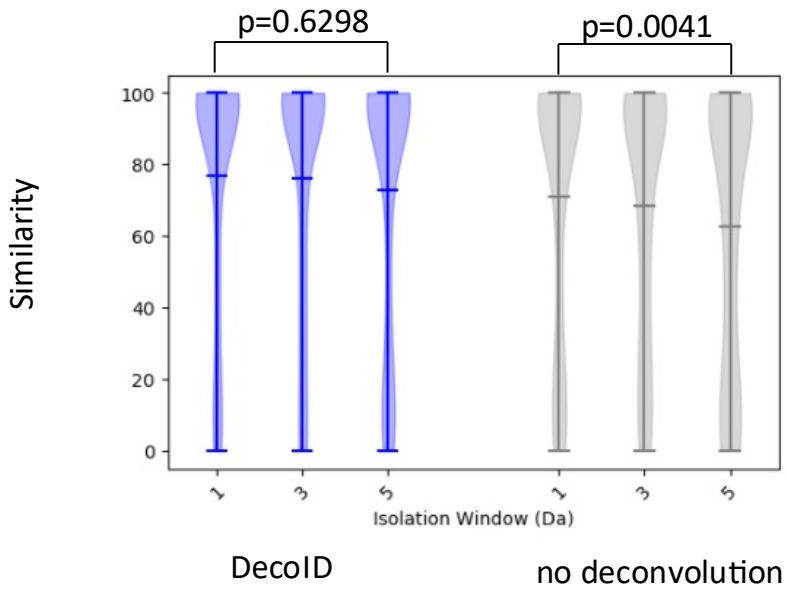
Supplementary Fig. 10 | DecoID does not artificially increase similarity scores when synthetic DDA and DIA datasets are deconvolved with a partial database. To ensure that DecoID does not artificially increase database scores, reference MS/MS spectra of the metabolite standards used to optimize the DecoID parameters were computationally mixed to generate a synthetic chimeric dataset. Reference MS/MS data for these compounds were then removed from the database before applying DecoID. Two synthetic datasets were generated to represent DDA (**a**) and DIA (**b**). The datasets were then subjected to deconvolution with and without the mixture compounds' spectra being removed from the mzCloud database. For DDA, a LASSO parameter of 5.0 was used. For DIA, a LASSO parameter of 50.0 was used. In both cases, there was no significant change in the dot-product scores for the top hit after the mixture compounds had been removed from the database. Conversely, when deconvolving with the complete database, significantly different dot-product scores were achieved. Statistical significance was assessed with a two-sided Wilcoxon signed-rank test. Horizontal lines in the violin plots mark the mean, maximum, and minimum similarity.

Supplementary Fig. 11



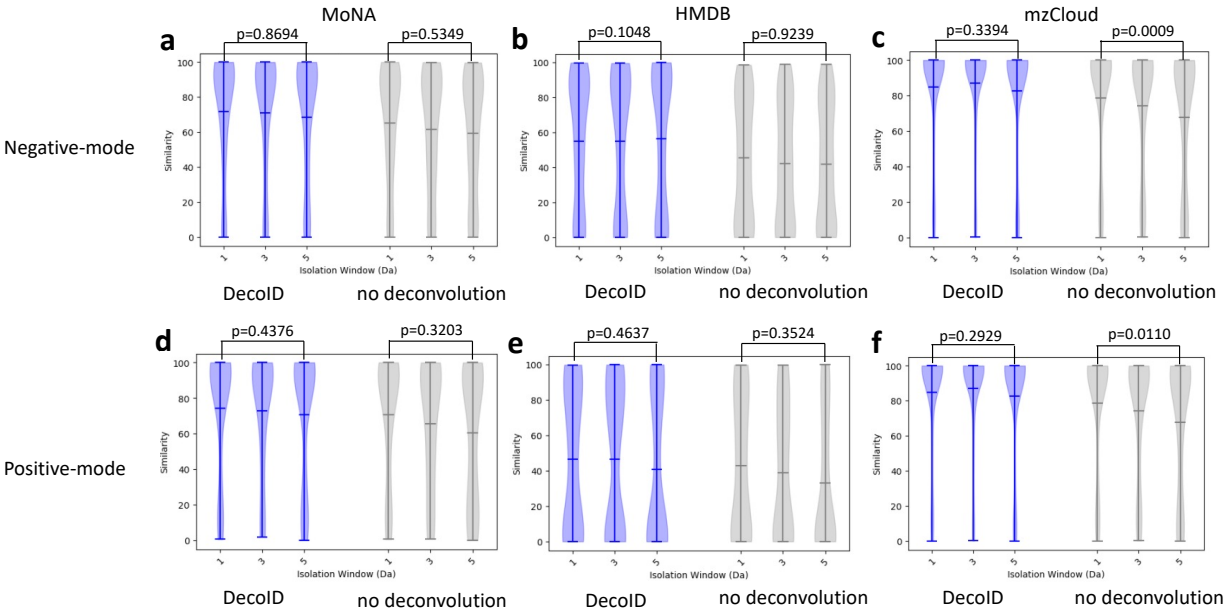
Supplementary Fig. 11 | MS/MS contamination increases with larger isolation windows in multiple sample matrices. The distribution of contamination increases as the isolation window increases in negative mode (**a**,**c**,**e**,**g**) and positive mode (**b**,**d**,**f**,**h**) for the IROA standard mixture (**a**-**b**) as well as biological extracts, *E. Coli* (**c**-**d**), plasma (**e**-**f**), and *P. pastoris* (**g**-**h**), spiked with 81 standards. Horizontal lines in the violin plots mark the mean, minimum, and maximum contamination.

Supplementary Fig. 12



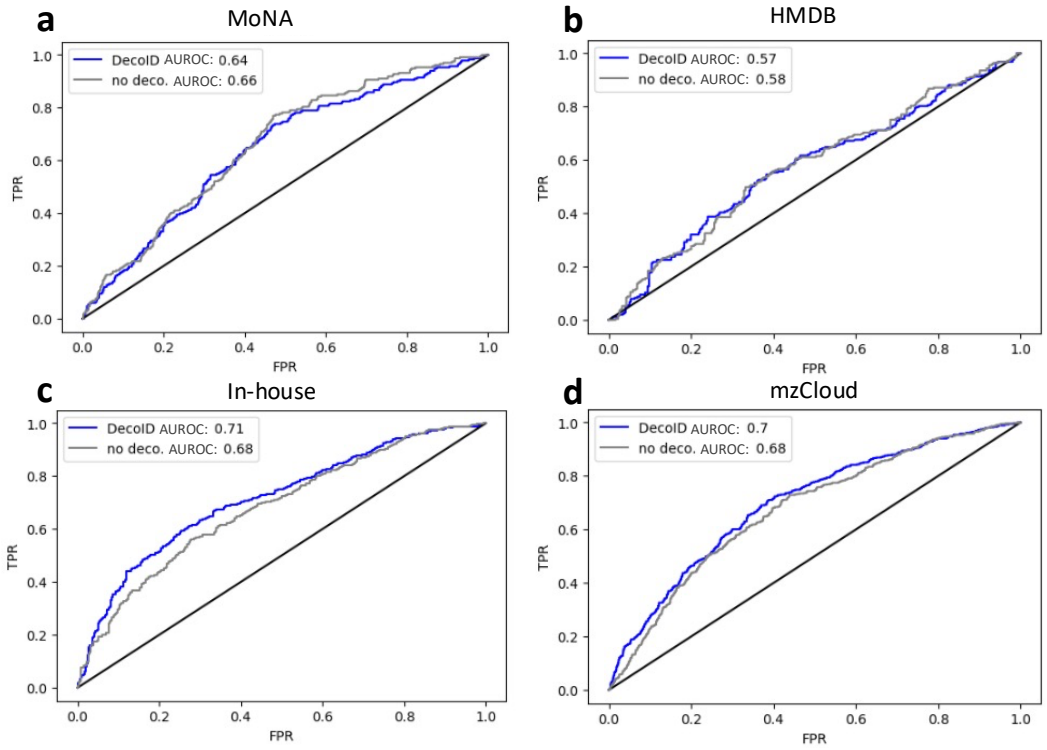
Supplementary Fig. 12 | DecoID preserves high-identification scores on the IROA DDA dataset when our in-house database is used, even as the MS/MS spectra become more chimeric. The violin plots depict the dot-product similarity distribution of identification scores for correctly identified metabolites from the IROA standard mixture when using our in-house database. The IROA standard mixture was analyzed in positive mode, and the data were processed by using DecoID with and without deconvolution. As the isolation window increases and more chimeric spectra are produced, the database similarity decreases without deconvolution. With DecoID, however, no decrease occurs. Statistical significance was assessed with a two-sided, two-sample Kolmogorov–Smirnov test. The horizontal lines in the violin plots mark the mean, maximum, and minimum similarity. The negative mode counterpart of these data is shown in Fig. 3a. Results for other databases used are provided in Supplementary Fig. 13.

Supplementary Fig. 13



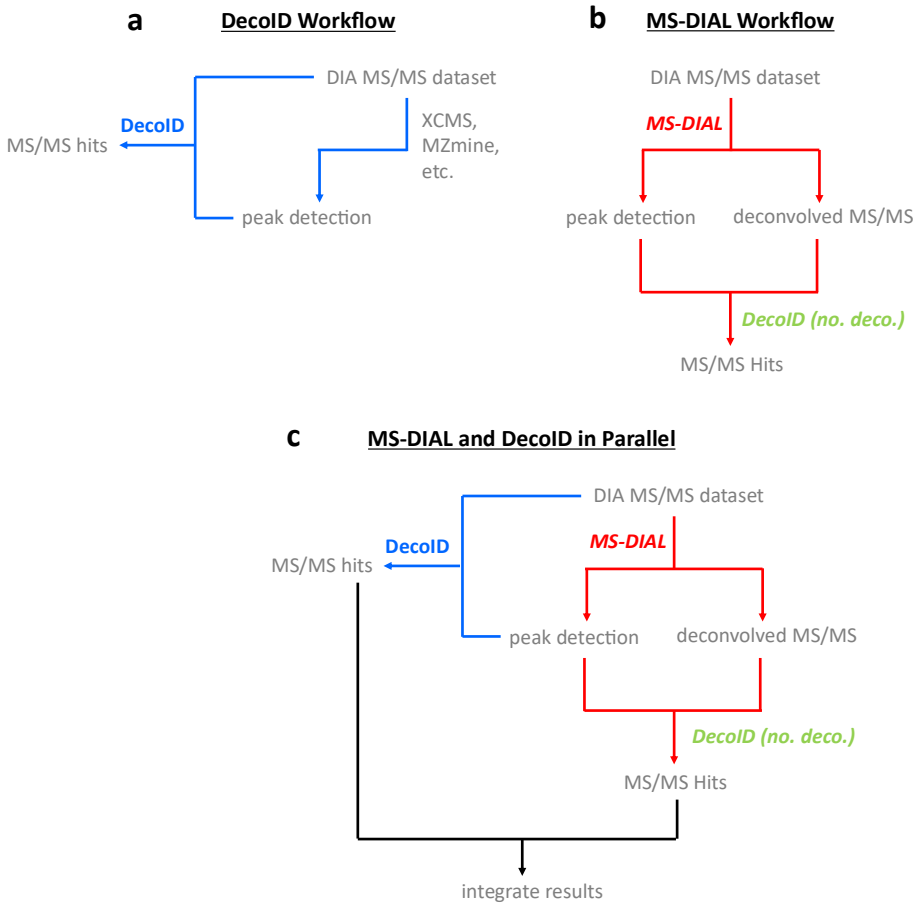
Supplementary Fig. 13 | DecoID preserves high-identification scores across multiple databases in the IROA DDA dataset, even as the MS/MS spectra become more chimeric. The violin plots depict the dot-product similarity distribution of identification scores for correctly identified metabolites in the IROA standard mixture in positive mode (**d**-**f**) and negative mode (**a**-**c**) when using DecoID to search mzCloud (**c**,**f**), MoNA (**a**,**d**), and HMDB (**b**,**e**) with and without deconvolution. As the isolation window increases and more chimeric spectra are produced, the database similarity decreases without deconvolution. With DecoID, however, no decrease occurs. Statistical significance was assessed with a two-sided, two-sample Kolmogorov–Smirnov test. The horizontal lines in the violin plots mark the mean, maximum, and minimum similarity. Results for the in-house database are given in Fig. 3a and Supplementary Fig. 12.

Supplementary Fig. 14



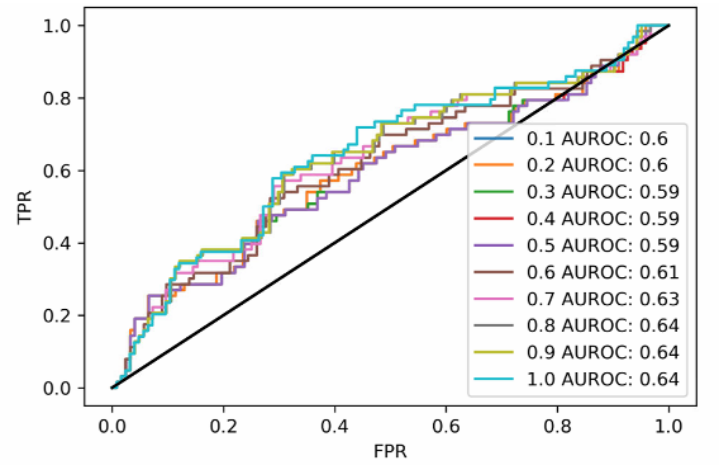
Supplementary Fig. 14 | ROC curves for DecoID results on the IROA DDA dataset. The metabolite identification accuracy of DecoID compared to directly searching the acquired spectra is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for each feature in the dataset.

Supplementary Fig. 15



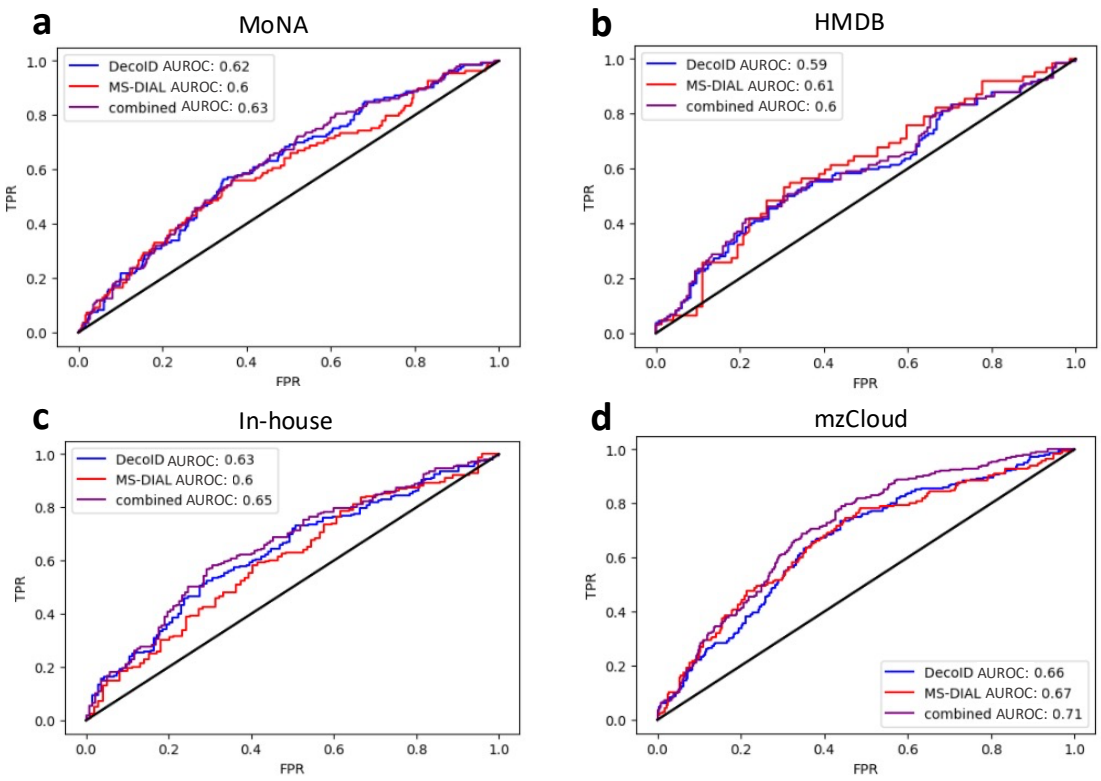
Supplementary Fig. 15 | Schematic workflow for using DecoID and/or MS-DIAL to deconvolve DIA MS/MS spectra. DIA data in this study were processed by using these three distinct deconvolution workflows. In **a**, DecoID is used to deconvolve the DIA spectra and match against reference databases. Peak detection can be performed with any software that produces a peak list (e.g., XCMS, MZmine, etc.). In **b**, both peak detection and MS/MS deconvolution are performed with the MS-DIAL software. Next, the MS-DIAL deconvolved MS/MS spectra are searched against reference databases by using DecoID. Importantly, the DecoID software is used without deconvolution. Using DecoID without deconvolution requires de-selecting the "deconvolution" box within the DecoID interface (see Supplementary Fig. 3). This ensures uniform scoring when using the DecoID workflow or the MS-DIAL workflow, thereby enabling a direct comparison. The peak list output by MS-DIAL (or any other peak list) can be used to search the MS-DIAL deconvolved spectra. In **c**, both MS-DIAL and DecoID are used in parallel. Peak detection is performed by using MS-DIAL. The MS/MS hits from DecoID and MS-DIAL are then automatically combined by the DecoID software. An example script showing the implementation of this combined workflow is available on the DecoID GitHub page.

Supplementary Fig. 16



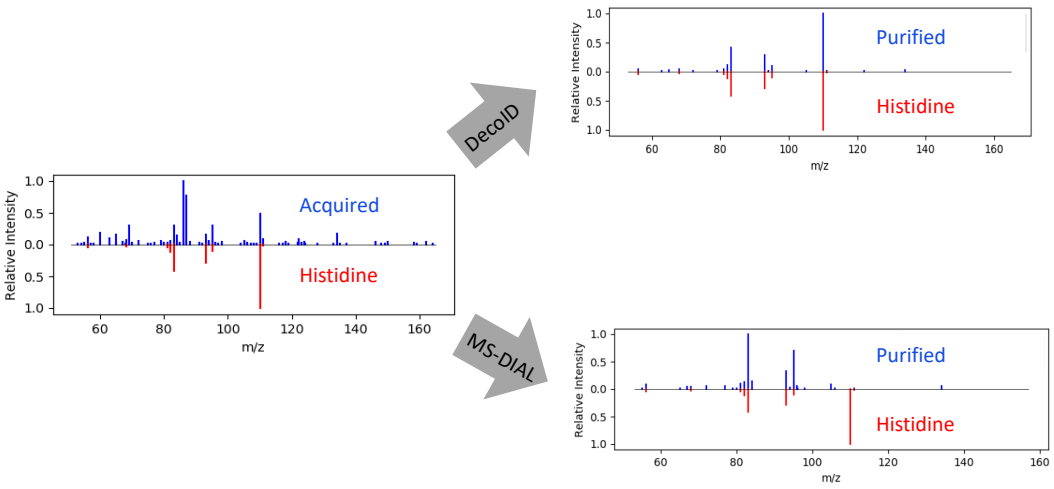
Supplementary Fig. 16 | Tuning performance of MS-DIAL on the parameter optimization DIA dataset. The MS-DIAL ”sigma” deconvolution parameter was optimized to give the highest area under the receiver operating characteristic curve (AUROC) on a training DIA dataset of mixed metabolite standards. Ten sigma values were tested to cover the recommended parameter range listed within the MS-DIAL software (0.1-1.0). On this dataset, a monotonic increase in AUROC as a function of sigma was achieved. To remain within the suggested parameter limits, a sigma value of 0.9 was selected as optimal.

Supplementary Fig. 17



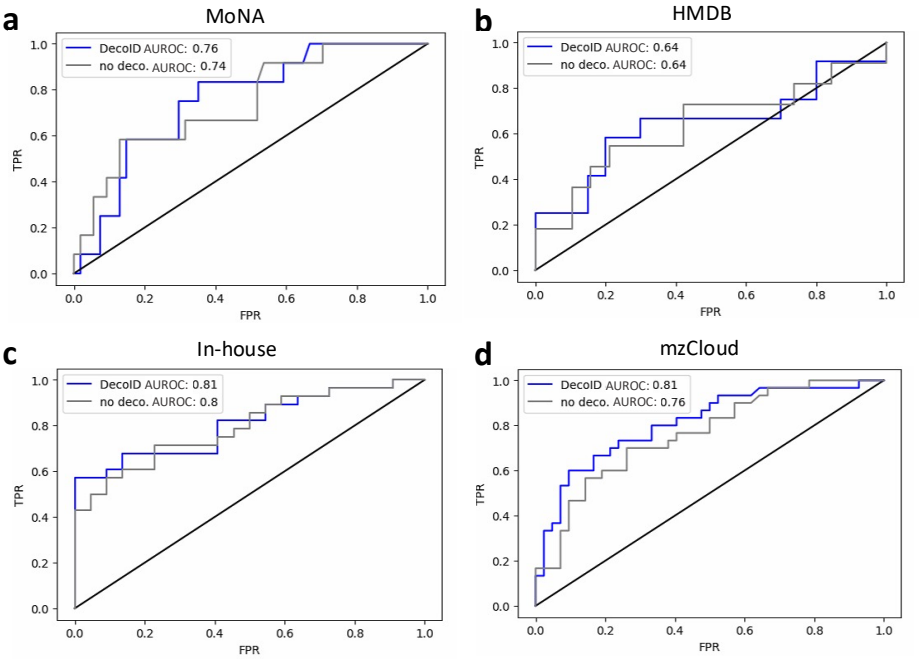
Supplementary Fig. 17 | ROC curves for DecoID, MS-DIAL, and the combined use of DecoID and MS-DIAL when analyzing the IROA DIA dataset. The metabolite identification accuracy of DecoID, MS-DIAL, and the combined use of DecoID and MS-DIAL in parallel is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). In all databases except HMDB, the area under the ROC curve (AUROC) is greatest when using the combination of DecoID and MS-DIAL. ROC curves were drawn by using the top 3 MS/MS hits for each feature in the dataset.

Supplementary Fig. 18



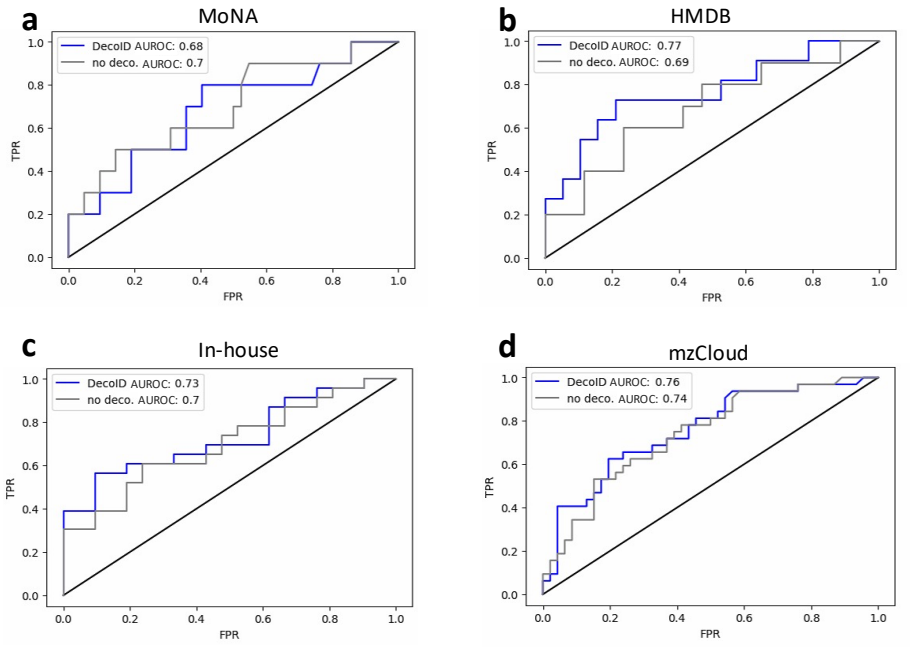
Supplementary Fig. 18 | DecoID successfully identifies histidine from a chimeric DIA MS/MS spectrum from the IROA dataset. The DIA positive-mode IROA mixture dataset was deconvolved by MS-DIAL and DecoID. Only the data deconvolved with DecoID matched the database spectrum of histidine.

Supplementary Fig. 19



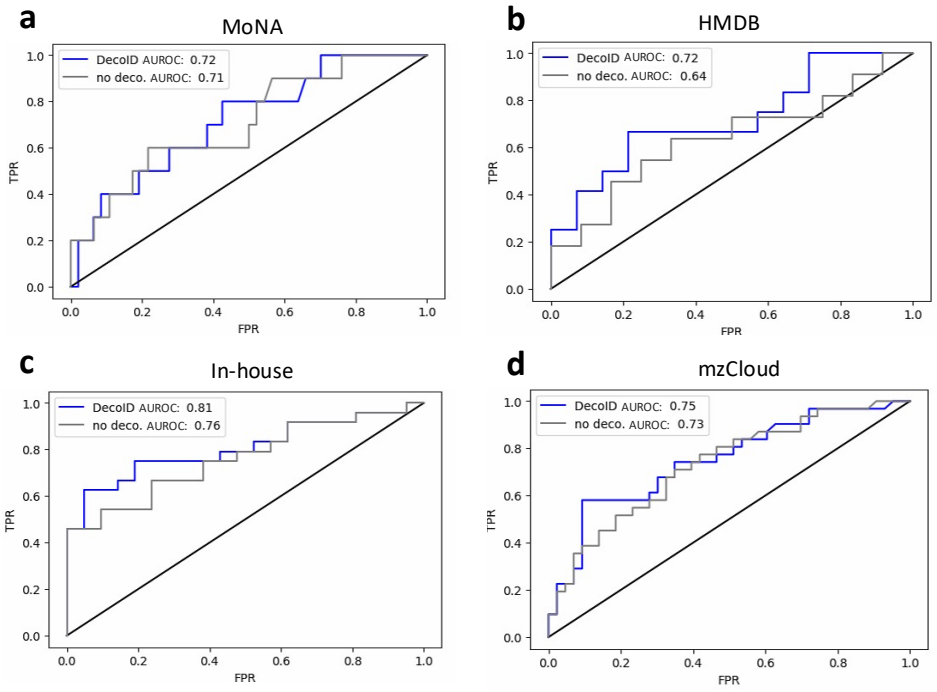
Supplementary Fig. 19 | ROC curves when using DecoID to evaluate metabolites spiked into an *E. coli* extract and analyzed with a DDA workflow. The metabolite identification accuracy of DecoID compared to directly searching the acquired spectra is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 20



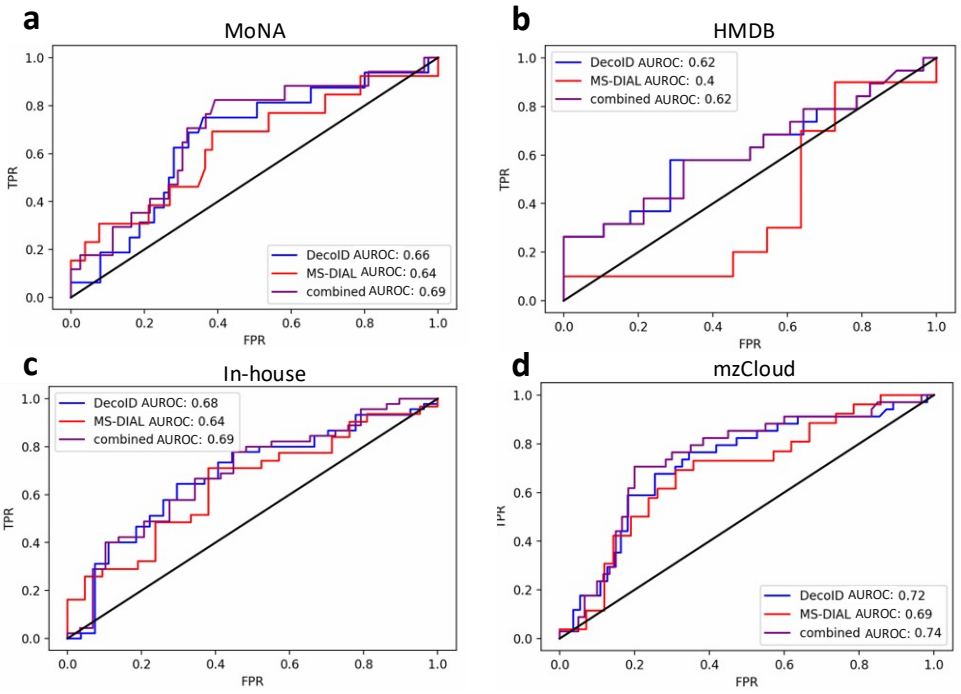
Supplementary Fig. 20 | ROC curves when using DecoID to evaluate metabolites spiked into a plasma extract and analyzed with a DDA workflow. The metabolite identification accuracy of DecoID compared to directly searching the acquired spectra is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 21



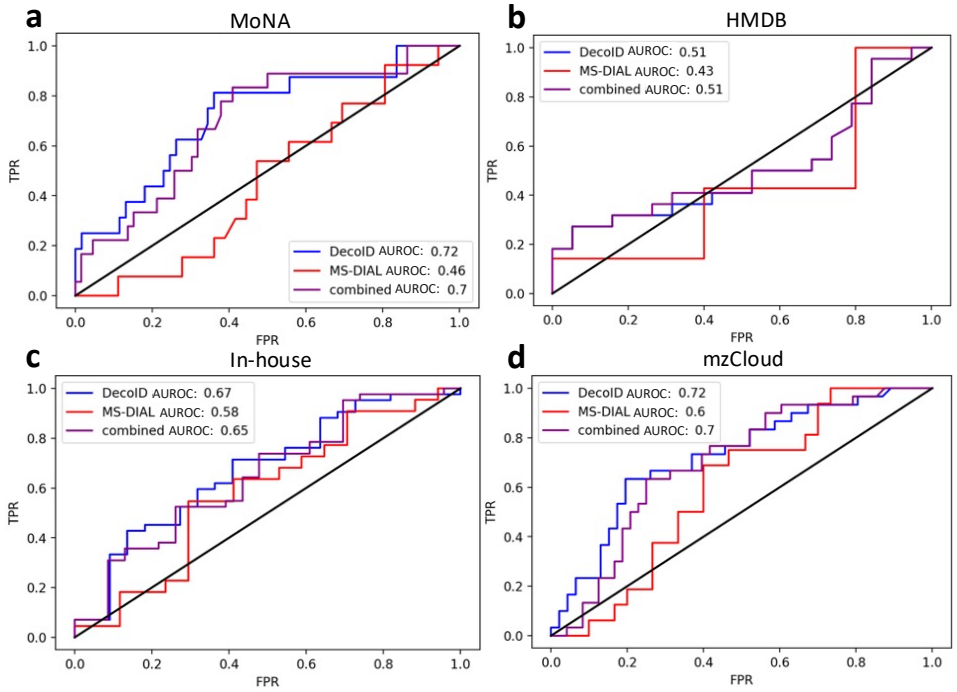
Supplementary Fig. 21 | ROC curves when using DecoID to evaluate metabolites spiked into a *P. pastoris* extract and analyzed with a DDA workflow. The metabolite identification accuracy of DecoID compared to directly searching the acquired spectra is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 22



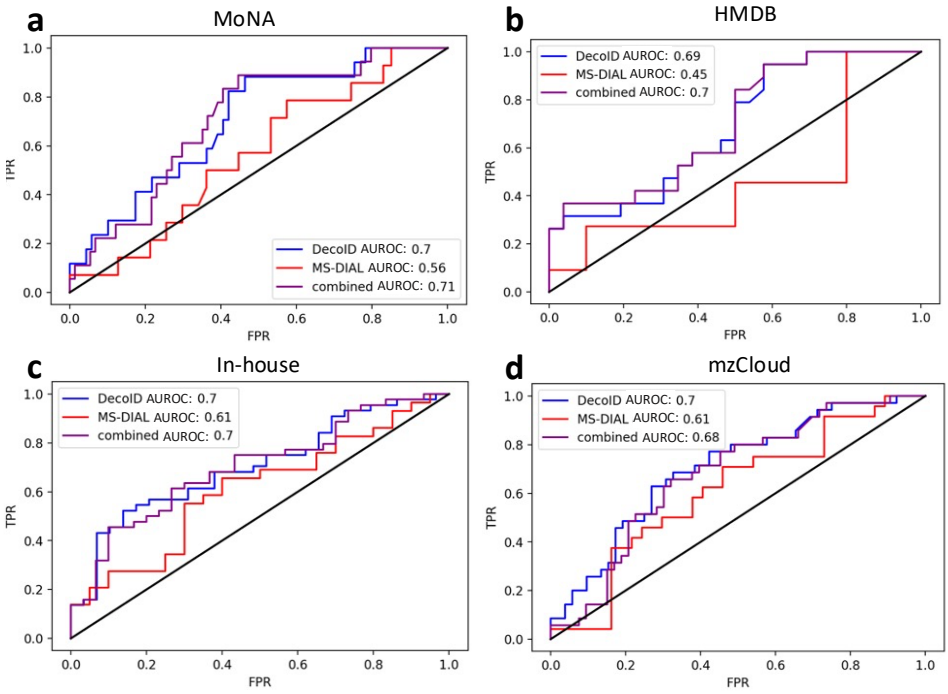
Supplementary Fig. 22 | ROC curves when using DecoID to evaluate metabolites spiked into an *E. coli* extract and analyzed with a DIA workflow. The comparative metabolite identification accuracy of DecoID, MS-DIAL, and a combined parallel usage of DecoID and MS-DIAL is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 23



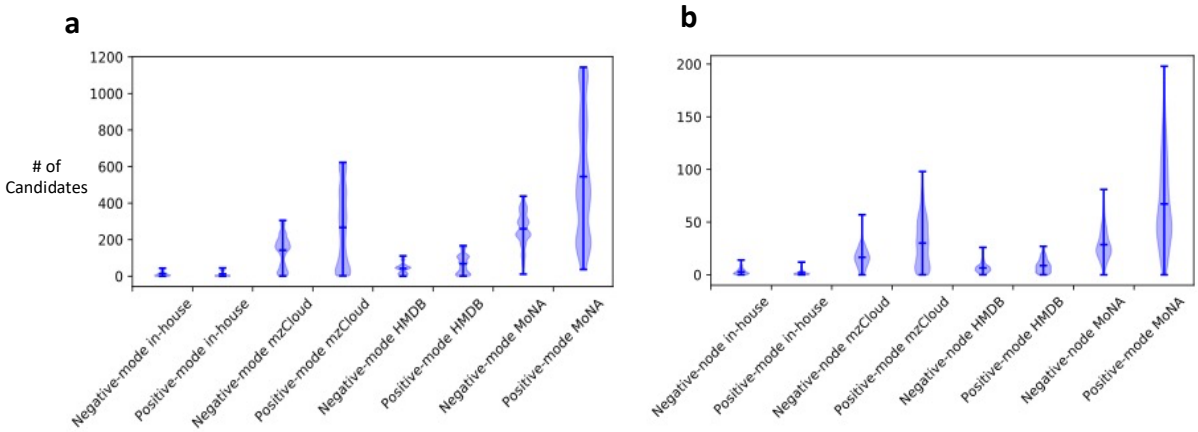
Supplementary Fig. 23 | ROC curves when using DecoID to evaluate metabolites spiked into a plasma extract and analyzed with a DIA workflow. The comparative metabolite identification accuracy of DecoID, MS-DIAL, and a combined parallel usage of DecoID and MS-DIAL is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 24



Supplementary Fig. 24 | ROC curves when using DecoID to evaluate metabolites spiked into a *P. pastoris* extract and analyzed with a DIA workflow. The comparative metabolite identification accuracy of DecoID, MS-DIAL, and a combined parallel usage of DecoID and MS-DIAL is shown by using receiver operating characteristic (ROC) curves for the four different databases: MoNA (**a**), HMDB (**b**), the in-house IROA database (**c**), and mzCloud (**d**). ROC curves were drawn by using the top 3 MS/MS hits for the spiked-in metabolites in the dataset.

Supplementary Fig. 25



Supplementary Fig. 25 | Number of candidate compounds from different databases. The violin plots show the distribution of the number of candidate compounds when positive-mode and negative-mode DIA (**a**) and DDA (**b**) data from a plasma sample are deconvolved with DecoID. The horizontal lines on the violin plots mark the mean, maximum, and minimum number of candidates.