METASPACE-ML: Metabolite annotation for imaging mass spectrometry using machine learning

Bishoy Wadie [1,2,*], Lachlan Stuart [1,*], Christopher M. Rath [1,3], Theodore Alexandrov [1,4,5,6,+]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; [2] Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany; [3] Present: Totus Medicines, Emeryville, CA, United States; [4] Molecular Medicine Partnership Unit, EMBL, Heidelberg, Germany; [5] Metabolomics Core Facility, EMBL, Heidelberg, Germany; [6] BioStudio, BioInnovation Institute, Copenhagen, Denmark

*: these authors equally contributed, [+]: correspondence to theodore.alexandrov@embl.de

Abstract

Imaging mass spectrometry enables spatial metabolomics, yet metabolites can be assigned only to a fraction of the data generated. METASPACE-ML is a machine learning-based approach addressing this challenge. Integrating new scores and computationally-efficient False Discovery Rate estimation, trained and evaluated on 444 representative datasets from 44 groups, METASPACE-ML surpasses the rule-based predecessor in precision, coverage, and computational efficiency. Our work helps illuminate dark matter in spatial metabolomics with machine learning.

Introduction

Imaging mass spectrometry (imaging MS) has emerged as a leading technology in spatial metabolomics, finding applications in diverse fields such as biology, medicine, and pharmacology [1–3]. However, a key challenge in this field is the accurate and confident annotation of metabolites, primarily due to limitations in data collection, existing algorithms and software [4]. Similar to bulk metabolomics [5], the vast majority of the imaging MS data, so-called "dark matter", cannot be molecularly annotated with existing tools.

We previously developed METASPACE, an engine for metabolite annotation [6] and a community-populated knowledge base [7]. METASPACE utilizes a False Discovery Rate (FDR)-controlled approach, where metabolite ions are reported at a given confidence level by ranking them against implausible generated decoy ions. However, a limitation of METASPACE is its rule-based scoring system, namely Metabolite Signal Matching (MSM), which assigns equal weights to the features and lacks adaptability to data variations. Previous attempts to employ a data-driven rescoring approach did not consistently increase the number of annotations across diverse datasets [8].

Here, we introduce METASPACE-ML, a machine learning model for FDR-controlled metabolite annotation for imaging MS, incorporating novel features and trained on 165 representative datasets from 40 groups. By evaluating METASPACE-ML on 279 datasets from 20 groups, we demonstrate its superiority over the traditional MSM rule-based approach, delivering more annotations, particularly at low FDR thresholds. We offer an efficient implementation using the Lithops serverless framework [9] This advancement enables richer molecular information from imaging MS data to be extracted with enhanced likelihood of correct assignment.

Results

Key principles of METASPACE-ML

We aimed to optimize the discrimination between target ions and implausible decoy ions [6] by employing a data-adaptive scoring method for FDR-controlled annotation (Figure 1A). We used five scores per ion, including three constituents from the MSM score [6], and two novel scores estimating the absolute and relative m/z error (Figure 1B).

Target and decoy ions were separated with a ranking-based Gradient Boosting Decision Trees [10]. An ensemble of decision trees was trained iteratively, with each subsequent tree correcting the errors of the previous ones. During each iteration, the target-decoy ion pairs from the training data were scored using a decision tree, aiming to minimize the PairLogit loss function that reflects the difference in their prediction scores (Figure 1C). Training the ensemble produced the METASPACE-ML model (Figure 1D).

Comprehensive set of representative test datasets

For training and evaluation, we selected two extensive sets from 6,478 public METASPACE datasets (as of July 12th, 2022) and EMBL datasets. The training and testing sets comprised 165 and 279 datasets (see Methods), respectively, selected to be representative of METASPACE datasets (Figure 2A). With a balanced distribution of positive and negative polarity, MALDI was the most prevalent ionization source (119 out of 165), often coupled with Orbitrap or FTICR analyzers. Approximately 65% represented common mammalian model organisms (human, mouse, and rat). Human derived samples were predominantly analyzed via FTICR analyzers, with over 90% having a high resolving power (≥ 200k @ 200 m/z). The majority of the datasets (155 out of 165) were acquired with a high upper m/z limit (maximum m/z > 500).

METASPACE-ML outperforms the rule-based approach particularly for highly-confident annotations

To assess the performance of the METASPACE-ML model, we conducted a comparative evaluation against the rule-based MSM approach [6]. The mean average precision (MAP) was used to evaluate the quality of target vs decoy ion rankings. When using cross-validation on the training datasets (see Supp Figure S1 for the cross-validation splits), METASPACE-ML demonstrated a significantly superior performance (p-value < 0.0001), with higher MAP scores (Figure 2B) and a higher median MAP value (0.36 vs 0.28). The evaluation on the 279 testing datasets highlighted the applicability and consistent improvement achieved by METASPACE-ML in annotating both metabolites and lipids, with particularly promising results for the novel CoreMetabolome database (Figure 2C, Supp Figure S2). METASPACE-ML consistently outperformed the MSM approach, enabling the discovery of substantially more annotations: At the default FDR 10%, METASPACE-ML, on average, identified approximately 15 more annotations and at least 30 more annotations in half of the datasets across all databases (Figure 2D). At FDR 5%, METASPACE-ML revealed approximately a two-fold increase in annotations compared to the MSM approach in half of the datasets across all databases (Figure 2E), indicating a greater improvement for most-confident annotations at lower FDR. This trend was consistently observed for each database, with higher relative improvements for LipidMaps and SwissLipids at lower FDRs

compared to CoreMetabolome (Supp Figure S3). METASPACE-ML also performed significantly better for data with medium-high resolving power (Wilcoxon *p*-value < 0.001) (Figure 2F, Supp Figure S4).

METASPACE-ML better separates target vs decoy with the spectral score being most useful

Attaining effective separability between target and decoy ions is crucial for FDR-controlled annotation. Upon applying UMAP (Uniform Manifold Approximation and Projection), an algorithm for visualizing multi-dimensional data, to the feature space and visually differentiating target and decoy ions, we observed that the METASPACE-ML score better reflected target-decoy separation than MSM (Figure 2G, Supp Figure S5). The SHAP (SHapley Additive exPlanations) analysis showed the impact of individual features on the machine learning output, with *rho_spectral* [6] having the highest impact on the prediction score, while *rho_chaos* [6] having the lowest impact (Figure 2H).

METASPACE-ML helps annotate low-intensity ions

Having established the improved performance of METASPACE-ML in delivering more annotations, we examined the impact of ion intensities. In nearly all test datasets with annotations at FDR 10% (221/233), compared to MSM, newly-found annotations by METASPACE-ML had significantly (Wilcoxon *p*-value<0.01) lower intensities (Figure 2I; Figure S6 for a detailed example). This property may be especially useful in annotating biologically relevant ions corresponding to low-concentration metabolites, and thus turning acquired data into actionable hypotheses.

Among molecular classes newly-found by METASPACE-ML carbonyl compounds, estrane steroids, and carboxylic acid derivatives were the most and significantly (Fisher *p*-value<0.05) overrepresented (Figure 2J).

METASPACE-ML outperforms MSM in runtime

We evaluated the computational performance of METASPACE-ML open-source implementation using the Lithops serverless framework [9], available at https://github.com/metaspace2020. Compared to MSM on the considered subset of the 50 training datasets, on average, METASPACE-ML exhibited reduced run time (Supp Figure S7) likely due to the novel way of decoys selection (see Methods). This demonstrates the readiness of METASPACE-ML for public use on the METASPACE platform, where it is already accessible to early adopters.

Discussion

Compared to our previously proposed adaptive Support Vector Machine classification [8], the Gradient Boosting Decision Trees (GBDT) approach, formulated as a ranking model, demonstrated enhanced performance across a broader range of datasets. This proved particularly valuable as GBDT is known to be robust to noisy outliers, which can significantly impact the separation of target and decoy ions. We defined ways to calculate absolute and relative m/z error for centroided data. As expected, this helped improve the accuracy of the model predictions. Given the increasing use of centroided data, these scores hold potential

for wider application beyond METASPACE-ML. We introduced a novel expert-curated metabolome database, CoreMetabolome, which resulted in better improvements compared to more general databases.

Despite the notable performance and annotation coverage achieved by METASPACE-ML compared to the rule-based MSM approach, we acknowledge its limitations. The quality of the training datasets affects the performance. We have realized it by observing an improvement when excluding non-centroided datasets submitted by mistake to METASPACE by setting a cut-off on the number of peaks (50.000). In terms of scores, *rho_chaos* [6] had only a minor impact on the model performance (Figure 2H) which demands re-defining the quantification of spatial informativeness. As discussed earlier [6], the way of producing decoy ions is key and new data shows their heterogeneity in terms of similarity to the target ions (Figure 2G). Further investigations are needed into how this affects FDR estimation and in finding the most reliable way of decoy selection. Lastly, although the training datasets were selected to be representative for a large number of public datasets from 44 labs, this selection is biased towards datasets represented on METASPACE. This warrants further work for evaluating METASPACE-ML for less common data e.g. from cultured cells in single-cell metabolomics, or data from industrial labs, where the ability to publically deposit data may be limited. Finally, while the introduced CoreMetabolome database does offer some degree of improved performance, more rigorous and automated methods for generating such databases are desired.

Due to its flexible architecture, METASPACE-ML can include additional features. Of particular interest is the integration of Collisional Cross Section (CCS) values, enabling automated use of ion mobility separation able to resolve molecular isomers and isobars [11]. Furthermore, METASPACE-ML can incorporate the Kendrick mass defect or features quantifying other phenomena e.g. co-detection of characteristic in-source fragments or biochemically-related molecules. Since the model can be trained or fine-tuned on any training imaging MS data, one can envision the development of context-specific, technology-specific, or lab-specific models as compared to generalized models.

Conclusion

Our study represents a major advancement in metabolite annotation for spatial metabolomics by introducing the first machine learning model that outperforms rule-based approaches. By providing open access to the utilized data, model, and code, we facilitate further progress in the field. By addressing a key bottleneck integrating into the widely-adopted cloud platform[7], our model has the potential to enhance imaging mass spectrometry and spatial metabolomics, and to have far-reaching implications for biology, medicine, and pharmacology.

Methods

Selection of training datasets

In order to train the model, we selected datasets from METASPACE in an unbiased way to balance maximizing diversity of exploited technologies, acquisition parameters and labs

while also staying relatively representative to all datasets submitted to METASPACE. The training datasets were selected from 6,478 public METASPACE datasets (as of July 12th, 2022) as well as 4,358 private EMBL datasets with the METASPACE API (https://metaspace2020.readthedocs.io). Each dataset was given a weight based on the values of the corresponding parameters from the metadata (polarity, ionization source, analyzer, resolving power range, submitter, group). Those parameters are ordered to highlight their respective impact on the selection, with polarity having the highest impact and group having the lowest. Datasets from the same METASPACE group with similar names (e.g. “SlideA”, “SlideB”) had proportionally decreased selection weights compared to the rest of the datasets to reduce the grouping effect. In addition, the following exclusion criteria have been applied: datasets submitted by METASPACE developers, datasets with no submitter information, and datasets with fewer than 10 HMDB annotations at 10% FDR according to the rule-based approach. After applying the exclusion criteria, 200 training datasets have been randomly selected based on their assigned weights (Table S2).

The datasets were further cleaned up to match the required input of the employed ranking model (CatBoost; see below): we considered only ions with non-zero MSM scores. Furthermore, for a dataset we considered only target adducts which had 10 or more rule-based annotations at FDR 20% and decoy adducts which had 10 or more rule-based annotations at FDR 50%. The different FDR values for target and decoy were considered to obtain a better balanced training data. In addition, we filtered out datasets with a median number of peaks higher than 50.000 to exclude likely non-centroided datasets. This provided us with 165 training datasets (Figure 1E).

Processing training data using the rule-based METASPACE and CoreMetabolome

The selected training datasets were then reprocessed on the METASPACE server using the rule-based approach as previously described [6] with minor modifications to their configuration file (Section S1). The datasets were annotated against the CoreMetabolome database. The CoreMetabolome database is a molecular database specially designed and developed by us for METASPACE annotation. CoreMetabolome was designed to be large enough for untargeted spatial metabolomics, include only chemically plausible molecules, prioritize endogenous molecules over exogenous molecules, and include the molecules from primary metabolic pathways that are commonly abundant in different types of samples. HMDB (version 4) and KEGG were used as input databases and curated manually by an experienced chemist and mass spectrometrist. Section S2 contains detailed information on the curation process.

New ion scores quantifying the m/z error from the centroided data

In addition to the scores used in the Metabolite Signal Match score (MSM) (spatial isotope rho_spatial, spectral isotope rho_spatial, and spatial chaos rho_chaos) we have introduced two new scores: *m/z error abs* (absolute m/z error) and *m/z error rel* (relative m/z error). These scores quantify the error in estimating the m/z value for an ion of interest compared to its theoretically defined value: defined as follows:

$$mz_{mean} = \frac{1}{\sum_{i=1}^{n} PxI_i} \sum_{i=1}^{n} mz_i * PxI_i,$$

$$mz\ error\ abs =- abs(mz_{mean_{j=1}} - mz_{theo_{j=1}}),$$

$$mz\ error\ rel =- abs(\frac{1}{\sum_{j=2}^{m=4} TI_j} \sum_{j=2}^{m=4} (mz_{mean_j} - mz_{theo_j} - mz\ error\ abs) * TI_j),$$

where, for a given ion, mz_i is the m/z value of that ion in pixel *i* of the respective ion image, PxI_i is the corresponding intensity in pixel i, and *n* is the total number of pixels in a given ion image. Moreover, mz_{theo_j} is the theoretical m/z value of that *j*'th isotopic peak of that ion, mz_{mean_j} is the mean m/z value of *j*'th isotope of that ion (across all pixels and centroids samples within the considered ppm window), *m* is the total number of isotopic ion images considered (equal to 4 in this study), and TI_j is the theoretical relative intensity of the j'th isotope of the ion.

FDR estimation

In the rule-based approach [6], the target-decoy strategy was formulated so that for an MSM threshold, the ratio of positive decoys to all positives provides an estimate for FDR. To reduce the variability introduced by the random choice of decoy adducts, the decoy adducts were sampled $S_D = 20$ times for each target adduct, taking the median value across S_D rankings for each formula before applying monotonicity adjustments. In METASPACE-ML, we also sample decoy adducts S_D times, yet we propose to use a single weighted ranking where decoys are weighted with $1/S_D$. First, for a database and each target adduct, both target ions and all sampled decoy ions are sorted in descending order based on the model prediction score. For each rank threshold i of sorted ions, we calculate T_i and D_i which are the numbers of targets and decoys, respectively, with ranks smaller or equal than i. The FDR value for an ion with the rank threshold i was defined as

$$FDR_i = \frac{(D_i+1) / S_D}{(T_i+1) + ((D_i+1) / S_D)}.$$

A pseudocount of 1 was added to both T_i and D_i as per the rule of succession to avoid misleading 0% FDR and for a better estimate of the mean of targets and decoys which have a binomial-like distribution.

In summary, we introduced the following changes to the FDR estimation compared to the rule-based approach [6]: 1) Using a single weighted ranking, where decoys are given $1/S_D$ of the weight, 2) Using a single selection of S_D random decoys per formula which are shared between all FDR rankings, 3) Allowing for calculation of continuous FDR values for each ion

instead of snapping FDRs to fixed thresholds (5%, 10%, 20%, 50%), and 4) Introducing a rule of succession where a pseudocount of 1 is added to the number of targets and decoys.

These changes increase the computational performance. In the rule-based approach, S_D random decoys would be sampled from a set of implausible adducts for each formula and target adduct (e.g. +H, +Na, +K). In METASPACE-ML, we propose instead to randomly sample S_D decoy adducts per target formula and share them across all possible target adducts for that formula. This change should not affect the FDR rankings as they are statistically independent. However, this allows to reduce the calculations of scores as it produces fewer decoy ions overall.

Training and cross-validating the model

We employed a ranking-based model using gradient boosting decision trees implemented using the CatBoost framework [12]. The model was trained on 200 processed training datasets. As input features, we used the original MSM features (spatial isotope, spectral isotope and spatial chaos) in addition to the newly introduced features (relative and absolute m/z error), five features in total. The CatBoost model was first initialized using the “CatBoost" method using PairLogit as the loss function and fitted on the 165 training datasets for 1000 iterations. In each iteration, the decision trees are built in such a way that it improves the previous trees’ output based on a loss function. For each dataset, a decision tree at a specific iteration scores all decoys and target ions based on their feature scores, using combinations of pairwise objects where one is considered a winner (target ion) and the other a loser (decoy ion). The pairlogit loss function [13] selects the best tree that maximizes the positive difference between the tree score for the target vs decoy ion . To evaluate the model and ensure that it was not overfitting, we performed cross validation where the training datasets were split into 5 splits, where 80% of the dataset were used for training and the remaining 20% were used for evaluation (see Evaluation metrics and benchmarking). The final prediction score of the model was scaled [0,1] so that scores closer denote high confidence annotations and vice versa.

Evaluation metrics and benchmarking

Mean average precision (MAP). In order to evaluate the performance of METASPACE-ML compared to the rule-based approach, we compared how well their respective scores were able to rank target ions relative to decoy ions. We used Mean Average Precision (MAP), a commonly used metric that provides a comprehensive evaluation of the ranking accuracy and precision. Precision is defined as the number of targets in the top “k” ions of a ranked list, divided by k. Then, the average precision for each ion in the ranked list is calculated followed by taking the mean of these average precision over all datasets. MAP scores were calculated for each cross-validated dataset for each of the 5 splits (see Training and cross validation) and for the testing datasets.

Number of annotations. While MAP and PR-AUC evaluate the ranking quality of METASPACE-ML, they do not necessarily quantify the desired increase in the number of annotations. So, we calculated the relative fold change and the absolute difference in the number of target ions captured relative to the rule-based approach at specific FDR thresholds (5%, 10%, 20%, 50%).

Selection of testing datasets and database comparison. In order to appropriately evaluate the performance of METASPACE-ML and ensure its ability to generalize to new unseen data, we have selected 300 datasets using the same selection criteria as the training datasets (see selection and pre-processing of training datasets) which is also annotated using different databases (CoreMetabolome, SwissLipids, LipidMaps) (Supp Figures 8-9). In addition, we also filtered out datasets with a median number of peaks higher than 50.000 yielding a total of 279 datasets (Table S3) for downstream evaluation.

Separation of target-decoy and feature importance

An important aspect of the ideal annotation score is its ability to optimally differentiate between target and decoys. In order to assess it, we have developed a visualization strategy where for each ion, we calculated three original rule-based scores plus the new m/z error features and projected both target and decoy ions onto a two-dimensional space using UMAP implemented in the M3C R package [14] . Additionally, to evaluate which features are most important in driving the METASPACE-ML model's prediction, we used SHAP values from the "Shap" python package [15]. SHAP values quantify the contribution of each feature to the final prediction score for a given ion in each dataset. This was done on 200 training datasets and the results were aggregated and visualized as a beeswarm plot from the Shap python package [15].

Testing for difference in intensities of annotations

To further investigate the characteristics of the newly captured ions by the METASPACE-ML model, we compared the intensity distributions of those ions to the ones captured by both METASPACE-ML and the rule-based approach across multiple datasets. Using METASPACE API (https://metaspace2020.readthedocs.io) we retrieved the ion images for each target ion in a given dataset and calculated the 99% percentile intensity across all pixels, followed by Log 10 transformation and taking the median across all ions captured at specific FDR threshold (default 10%). The distribution of the median-transformed intensities per dataset for each approach (METASPACE-ML and MSM) were compared using the Wilcoxon test. In addition, we used the 'ggbetweenstats' function in the ggstatsplot R package [16] to perform the pairwise Mann-Whitney test between the intensity distributions of ions only captured by METASPACE-ML compared to either all ions captured by METASPACE-ML or only ions captured by the rule-based approach for a given dataset. P-values were adjusted using the Benjamini-Hochberg correction [17].

Context-dependent evaluation

In addition to reporting evaluation results per dataset, we also considered grouping the testing datasets by different contexts based on their metadata (Organism, Resolving power range and Ionization source). Accordingly, we used the UpSet function in the "ComplexHeatmap" R package [18] to perform the overlapping analysis per context and add the evaluation metrics (see Evaluation and Benchmarking) as heatmap annotations.

Enrichment analysis

To learn more about the types of metabolites that were only picked up by METASPACE-ML, we performed a hypergeometric test to identify the molecular classes that were enriched in those metabolites. Accordingly, we retrieved the class and subclass information for all

annotated metabolites from the HMDB database (version 4) and used the HMDB “subclass” as background for enrichment. Then, for each dataset and subclass, we first filtered annotations with 10% FDR and then performed a two-tailed Fisher exact test where we consider the log fold enrichment as described in [19] as a proxy for the enrichment score. Finally, we filtered significantly enriched terms (p-value < 0.05) and only terms that were enriched in at least 10% of the total number of input datasets were considered for visualization purposes. The complete enrichment results per dataset and term can be found in (Table S1).

Profiling of computational performance

To assess the performance impact of METASPACE-ML we reprocessed a random selection of 50 training datasets with performance profiling. The average cloud compute resources required per dataset decreased from 3898 GB-sec to 2875 GB-sec (26% cost reduction) and the average wall-clock processing time decreased from 8.1 minutes to 5.8 minutes (28% faster). This was in part also due to more granular job subdivision affording greater parallelization, with an average dataset increasing from 91 to 117 parallel jobs and peak memory usage increasing from 20.6GB to 21.8GB.

Acknowledgments

We acknowledge software support by the METASPACE software developers: Sergii Mamedov and Lucas Maciel Vieira. The work was supported by grants from ERC (METACELL, METADRUG), NIH (KPMP, LungMAP2), and EC (CloudButton, CloudSkin, NEARDATA) and by the EMBL central funding.

Data and Software Availability

The METASPACE-ML code for inference is available at https://github.com/metaspace2020/metaspace and covers scores calculation, METASPACE-ML and FDR calculation based on the trained model. The METASPACE-ML model is available at https://github.com/metaspace2020/metaspace/tree/master/metaspace/scoring-models. The functionality is available through METASPACE in the annotations settings, see the model "v2.20230517 (METASPACE-ML)" in “Analysis version”.

The public training and testing datasets are available for download on METASPACE and more information on each dataset can be accessed from Tables S2 and S3, respectively.

Supplementary Information

The Supplementary Sections S1-S2, Tables S1-S3, and Figures S1-S9 are provided in a separate file.

Author Contributions

LS has conceived the METASPACE-ML approach, implemented it, and conceived and performed key steps of validation. BW has conceptualized substantial improvements of METASPACE-ML, trained the final METASPACE-ML model, expanded the validation methodology, performed data analysis and visualization, and wrote the manuscript. CMR has developed and curated the CoreMetabolome database. TA has obtained funding, supervised the study, and edited the manuscript.

Conflict of Interest

TA is an inventor of multiple patents in imaging mass spectrometry and is the leader of the startup-in-incubation that uses imaging mass spectrometry.

References

1. Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annu. Rev. Biomed. Data Sci.* **3**, 61–87 (2020).
2. Granborg, J. R., Handler, A. M. & Janfelt, C. Mass spectrometry imaging in drug distribution and drug metabolism studies – Principles, applications and perspectives. *Trends Analyt. Chem.* **146**, 116482 (2022).
3. Ma, X. & Fernández, F. M. Advances in mass spectrometry imaging for spatial cancer metabolomics. *Mass Spectrom. Rev.* e21804 (2022).
4. Baquer, G. *et al.* What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in mass spectrometry imaging. *Mass Spectrom. Rev.* e21794 (2022).
5. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America* vol. 112 12549–12550 (2015).
6. Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods* **14**, 57–60 (2017).
7. METASPACE: The metabolite annotation engine and community-populated knowledgebase of spatial metabolomes. https://metaspace2020.eu.
8. C Silva, A. S. *et al.* Data-driven rescoring of metabolite annotations significantly improves sensitivity. *Anal. Chem.* (2018) doi:10.1021/acs.analchem.8b03224.
9. Sampé, J., Sánchez-Artigas, M., Vernik, G., Yehekzel, I. & García-López, P. Outsourcing Data Processing Jobs With Lithops. *IEEE Transactions on Cloud Computing* **11**, 1026–1037 (2023).
10. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv [cs.LG]* (2018).
11. Sans, M., Feider, C. L. & Eberlin, L. S. Advances in mass spectrometry imaging coupled to ion mobility spectrometry for enhanced imaging of biological tissues. *Curr. Opin. Chem. Biol.* **42**, 138–146 (2018).
12. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *arXiv [cs.LG]* (2017).
13. Wang, Y., Khardon, R., Pechyony, D. & Jones, R. Online Learning with Pairwise Loss Functions. *arXiv [stat.ML]* (2013).
14. John, C. R. *et al.* M3C: Monte Carlo reference-based consensus clustering. *Sci. Rep.* **10**, 1816 (2020).
15. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**, 56–67 (2020).

16. Patil, I. Visualizations with statistical details: The ‘ggstatsplot’ approach. *J. Open Source Softw.* **6**, 3167 (2021).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
18. Gu, Z. Complex heatmap visualization. *Imeta* **1**, (2022).
19. Dalmer, T. R. A. & Clugston, R. D. Gene ontology enrichment analysis of congenital diaphragmatic hernia-associated genes. *Pediatr. Res.* **85**, 13–19 (2019).

Figures

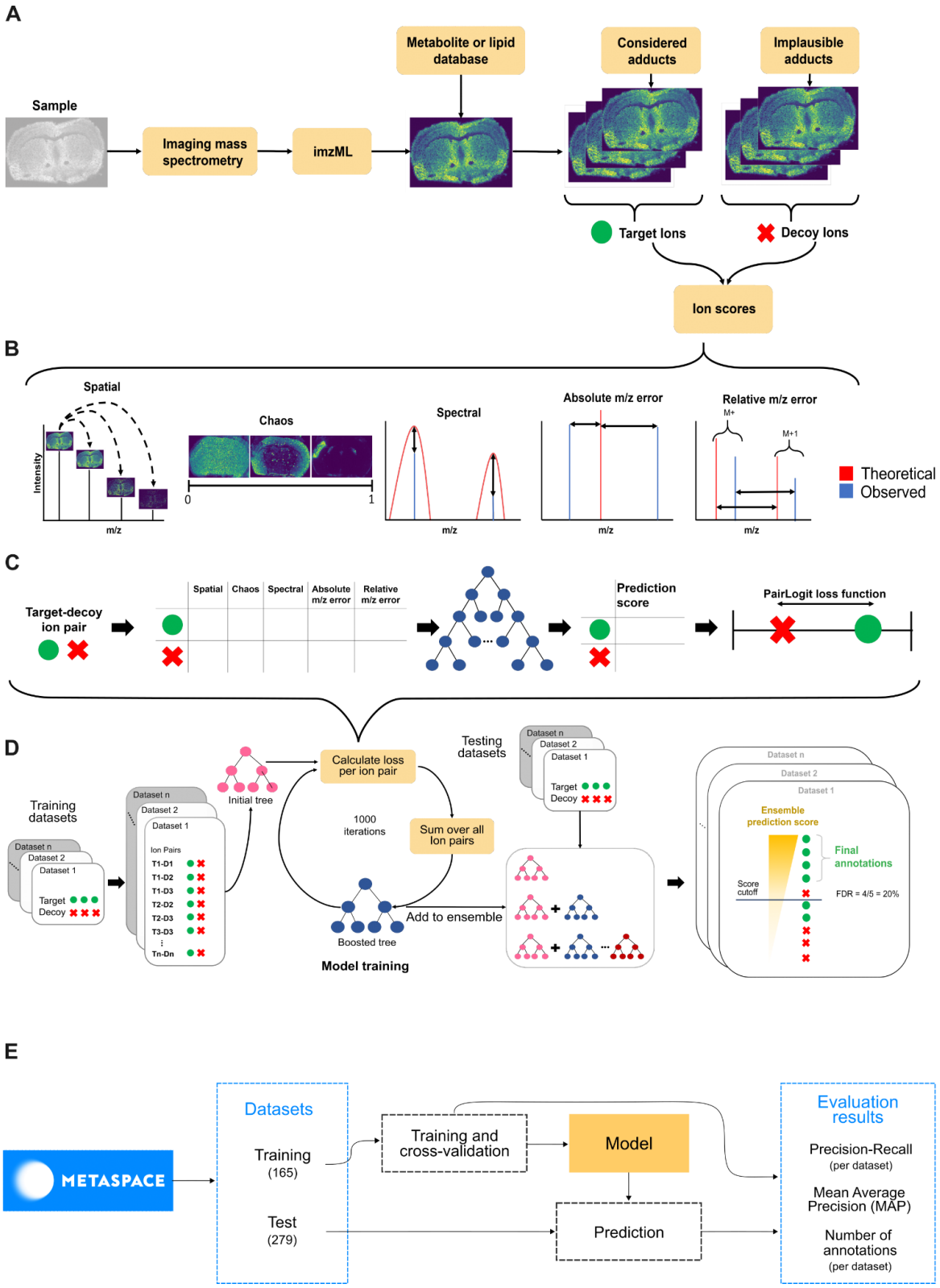

Figure 1. The data flow and key principles of METASPACE-ML. (A) The data flow. (B) Scores for an ion. (C) The principle of the target-decoy approach and scoring using a decision tree and the PairLogit loss function. (D) The workflow of model training and application to testing datasets. (E) A high-level overview of model training and evaluation.

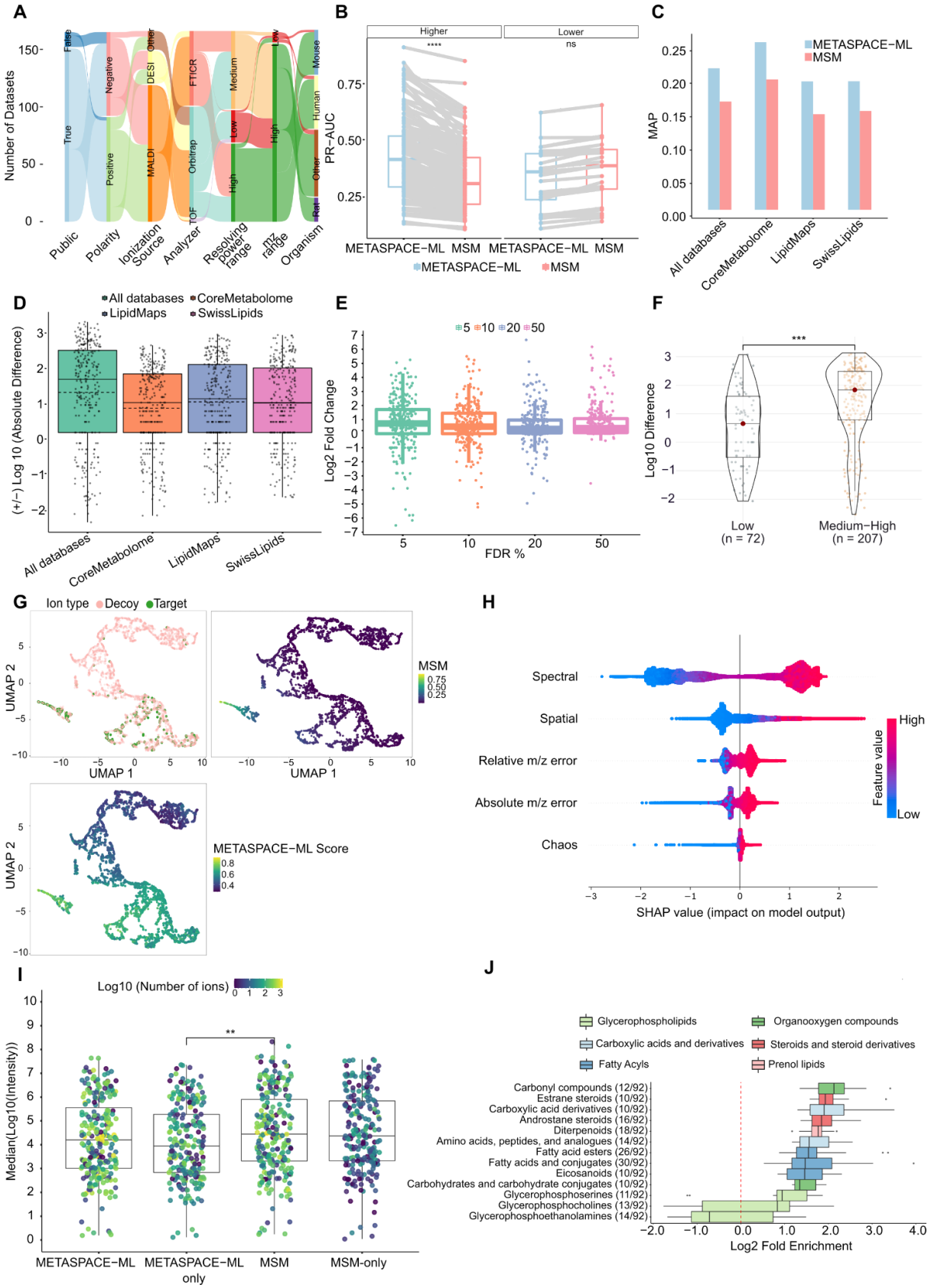

Figure 2. METASPACE-ML improves metabolite annotation compared to the rule-based MSM approach. (A) Breakdown of 200 training datasets (B) Precision-Recall Area Under Curve (PR-AUC) values for training datasets during cross validation. Each dot

represents a dataset; a gray edge connects PR-AUC values for the same dataset annotated with either METASPACE-ML or MSM. (C) MAP values for 300 test datasets annotated against either each database separately or all databases combined. (D) Increase of the number of ions annotated by METASPACE-ML relative to MSM for different databases at FDR 10%. A dot corresponds to a test dataset. Y-axis shows negative values when MSM has more annotations than METASPACE-ML. Solid and dashed lines show median and mean, respectively. (E) Fold changes of the number of annotations of METASPACE-ML relative to MSM across all databases, for different FDR thresholds. A dot corresponds to a test dataset. (F) Increase in the numbers of annotations for METASPACE-ML relative to MSM, separately low vs medium-high resolving power. A dot corresponds to a test dataset. (G) UMAP for both target and decoy ions for an example dataset (https://metaspace2020.eu/dataset/2021-03-22_15h30m15s). A dot represents an ion. (H) Features importance on the METASPACE ML model according to SHAP values. Each dot corresponds to either a decoy or a target ion across all training datasets. (I) The difference in intensities for METASPACE-ML-exclusively annotated ions compared to other annotations, at FDR 10%. A dot represents a test dataset, colored by the log10 number of annotations. (J) Metabolite classes overrepresented in the METASPACE-ML-exclusively annotated ions at FDR 10% for test datasets. ***, **, and * denote Wilcoxon p-values <0.001, <0.01 and <0.05, respectively.