analytical chemistry

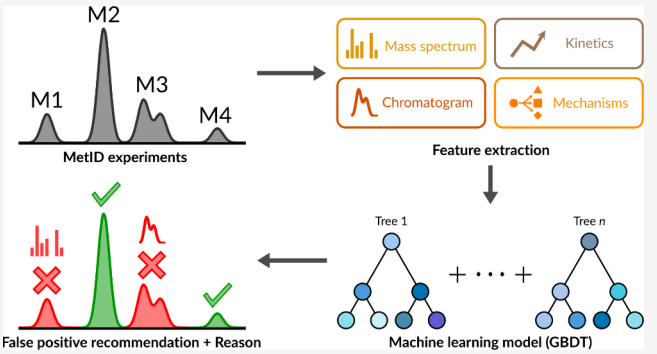
pubs.acs.org/ac
Article

Machine Learning-Assisted False Positive Detection in Metabolite Identification Workflows

Ramon Adàlia,*,⊥ Paula Cifuentes,⊥ Joyce Liu, Lionel Cheruzel, Gemma Sanjuan, Tomàs Margalef, and Ismael Zamora

Cite This: https://doi.org/10.1021/acs.analchem.5c02745
Read Online

ABSTRACT: Metabolite identification is a pivotal step in drug discovery and development, enabling the comprehensive analysis of drug-derived compounds within biological systems. However, the complexity of liquid chromatography−mass spectrometry data often results in numerous false positives, complicating the identification of true metabolites. This study introduces a machine-learning-based approach to improve the accuracy of false positive detection in metabolite identification workflows. By incorporating expert knowledge, we develop a feature set for metabolite-related chromatographic peaks that characterizes true and false positives with high accuracy, integrating data from mass spectra, chromatographic signals, and kinetic profiles. We validate this method via gradient boosting decision tree classifiers on both publicly available and proprietary "real-world" data sets, including small molecules and new modalities. Our findings demonstrate that machine learning-assisted techniques significantly reduce false positive identifications, thereby increasing the efficiency and accuracy of metabolite identification processes.



INTRODUCTION

Metabolite identification (MetID) is a critical part of drug discovery and development, providing insights into metabolic liabilities and pathways of drug candidates and aiding the identification of lead compounds for safe and effective medicines. Accurate metabolite profiling is essential not only for understanding metabolic challenges and pharmacokinetic and pharmacodynamic properties but also for meeting regulatory standards.

Liquid chromatography−mass spectrometry (LC-MS) is the predominant analytical technique for MetID in the pharmaceutical industry, valued for its speed, stability, sensitivity, and automation potential. It detects a wide range of metabolites in complex samples, producing large data sets visualized as chromatogram peaks. However, not all peaks represent true metabolites; false positives can arise from contamination, noise, processing errors, or even variations in LC-MS setups such as chromatography conditions, ionization methods, or mass analyzers.

Automatic software tools efficiently identify major metabolite peaks based on chromatographic and spectral features,[1] typically when signal intensity is sufficient. Lower-abundance metabolites may be discarded if quality thresholds are applied, risking missed detections, since MS signal intensity does not always correlate with concentration. Consequently, experts often configure software to reveal broad candidate peaks and manually approve or reject them based on criteria including cross-peak parameters, fragments shared with parent compounds, or trends across multiple samples (e.g., incubation time series).

Manual review is time-consuming and prone to error, introducing variability and limiting scalability as the LC-MS data volume grows with high-throughput screening.

To address these limitations, we developed a machine learning framework for automatic detection and reduction of false positives in MetID. Combining advanced classification models with domain-specific feature engineering, our approach improves the reliability of software-generated annotations, while reducing manual review needs. This relies on standardized and consistent MetID data, achievable via platforms like Oniro,[2] though the method is platform-agnostic.

Labels for training and evaluation derive from manual expert review, informed by chromatographic, mass spectrometry, and metabolic knowledge. While not infallible or a perfect gold

Received: May 8, 2025
Revised: November 4, 2025
Accepted: November 4, 2025

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

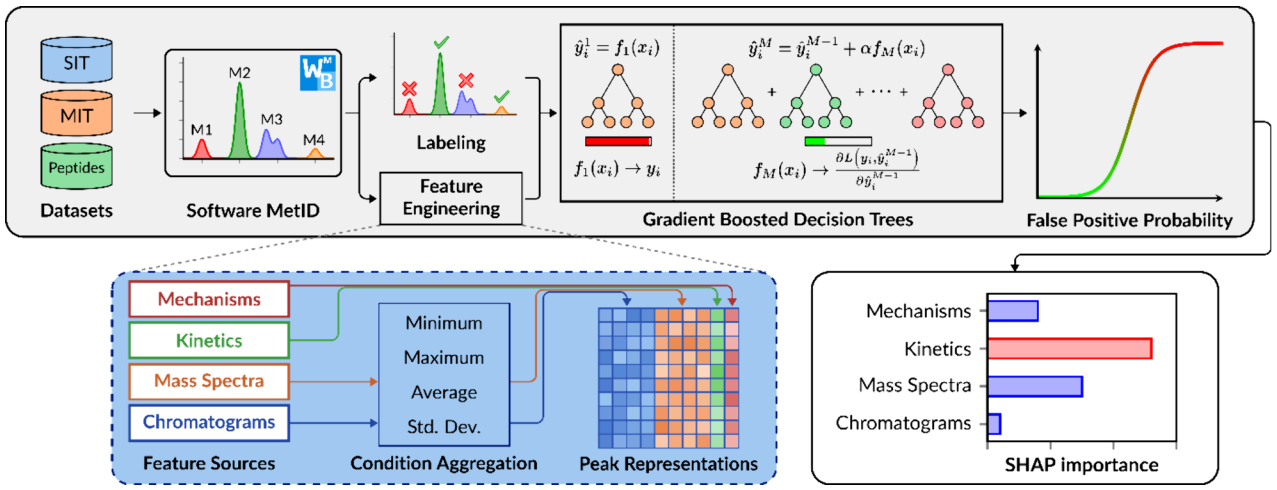


Figure 1. Steps involved in the development of a false positive detection model.

standard, this aligns with current drug discovery practice, where expert annotation remains the validation standard.

Our models use features similar to those considered by experts to approximate their decision-making. Crucially, models are trained only on expert-annotated data and do not autonomously generate labels. Thus, they are designed to assist and accelerate expert evaluation by prioritizing likely candidates and reducing manual effort, not to replace it.

We evaluated our framework on multiple drug discovery MetID data sets, including small molecules and macromolecules, some generated via Oniro by a MetID expert group.

Results show improved metabolite identification accuracy and the potential to make metabolite profiling more efficient, reliable, and scalable, supporting automation trends in drug discovery workflows.[3,4] The following sections detail our methodology, key results, and application to real-world data sets and discuss future impacts.

RELATED WORK

Metabolomics aims for comprehensive detection and quantitation of endogenous small molecules, using peak-quality assessment to distinguish the true signal from noise across thousands of features. In contrast, MetID, a targeted subdomain of drug discovery, focuses on precise confirmation of drug-derived metabolites, often relying on expert-driven workflows that incorporate chromatographic and reaction-specific cues.

Earlier metabolomics studies addressed peak quality through machine learning and heuristics. Yu et al.[5] extracted over 100 features from ion chromatograms, using m/z database matching as a proxy for true peaks, achieving scalability at the expense of indirect labeling. Gloaguen et al.[6] and Chetnik et al.[7] used expert-labeled peaks to train classifiers on chromatographic shape features, while Ju et al.[8] applied entropy- and correlation-based rules to filter noise.

Manual review of candidate metabolite peaks remains common in MetID, a time-consuming, variable process exacerbated by large LC-MS data sets from high-throughput screening. This reliance on manual or semiautomated methods slows discovery and creates bottlenecks.

To streamline these workflows, software tools automate key steps from raw LC-MS processing to structural elucidation and interpretation. Some are vendor-specific, optimized for data from the same manufacturer, such as MetaboLynx (Waters) and MetabolitePilot (SCIEX),[9] while vendor-neutral tools like MassMetaSite[10] support several instruments and formats, enhancing flexibility.

Many tools assist metabolite identification via external chemical databases,[11] but this is limited for newly synthesized drug candidates absent from public repositories. MassMetaSite overcomes this by generating metabolite structures from the parent compound using user-defined biotransformation reactions, enabling the elucidation of structure for novel compounds.

Existing MetID studies have improved data processing[12] and LC-MS parameter prediction,[13] but to our knowledge, no prior work has used machine learning to replicate expert curation logic based on the same features experts apply when flagging false positives.

Our approach differs from those in metabolomics by expanding beyond chromatographic peak quality and incorporating drug metabolism-specific information: MS fragmentation patterns, kinetic behavior, and reaction context are key elements in expert MetID review. Integrating these directly targets false positives in MetID pipelines, improving throughput, reproducibility, and standardization and supporting more robust decision-making in drug discovery.

METHODS

To enable our approach, we first created an explainable machine learning classifier capable of identifying false positives in the MetID data. The methodology, depicted in Figure 1, involved several key steps. We compiled five MetID data sets, processed them using specialized metabolite identification software[10] with settings that maximize the number of potential metabolite-related chromatographic peaks, and manually labeled the resulting data. For each identified peak, we developed a comprehensive description based on the available information. These descriptions, along with the labels, were used to train Gradient Boosting Decision Trees (GBDT)[14] models. To interpret the model's predictions, we employed explainability techniques based on SHapley Additive exPlanations (SHAP)[15] values, which provide a unified measure of feature importance by quantifying each feature's contribution to a model's prediction.

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

Table 1. Summary Statistics of the Datasets Used in This Study

		Peak count				Discard rate		
Data set	Experiments	Min	Median	Max	Total	Min	Median	Max
SIT	86	3	26	75	2485	5.71%	70.59%	100.00%
MIT	27	9	41	160	1461	22.22%	76.19%	95.12%
Peptides	123	5	30	283	5573	0.00%	90.79%	100.00%

DATA

To assess the effectiveness of our proposed approach across various MetID workflows, data were initially evaluated using publicly available data sets for both peptides and small molecules. The peptide data are divided into five distinct data sets, each characterized by different incubation conditions. These data sets included a total of 123 LC−MS/MS experiments, encompassing 47 linear and cyclic peptides with molecular weights ranging from 708 to 4184 Da. Data acquisition for three of the data sets was conducted using a Thermo Orbitrap instrument in data-dependent acquisition (DDA) mode, whereas the remaining two data sets were acquired using a Waters Q-TOF instrument in data-independent acquisition (DIA) mode. These compiled data are collectively referred to as the peptide data set. The small molecule data were compiled from two distinct data sets. The first data set, referred to as the single incubation time data set (SIT), comprised 86 LC−MS/MS experiments. These experiments used an AB Sciex TripleTOF 5600+ Mass Spectrometer in DDA mode on human liver microsomes incubated for 60 min, with substrates ranging in molecular weight from 144 to 733 Da. The second data set, referred to as the multiple incubation time data set (MIT), consisted of 27 LC−MS/MS experiments. These experiments employed a combination of Thermo Orbitrap and Agilent 6550 iFunnel Q-TOF instruments in both DDA and DIA modes on human hepatocytes and human liver microsomes, with incubation times ranging from 0 to 140 min. The substrates in the MIT data set ranged in molecular weight from 261 to 670 Da.

The raw data were visualized using the WebMetabase module in the Oniro software suite, which relies on MassMetaSite[10] to process the data. For more details on how the software works, see.[2,16,17] Detailed information on the data preprocessing steps is provided in the Supporting Information S1 and S2 for the SIT and MIT data sets, and in[18] for the peptide data set. Summary statistics of the data sets are presented in Table 1, revealing some variability in the number of peaks per experiment, although no significant differences were observed between the data sets.

LABELING

The criteria employed by the experts for MetID included but were not limited to the following:

- Mass Spectra: Mass spectral analysis involves evaluating several key parameters to assess the quality and reliability of metabolite identification, such as the following:
  - m/z Difference: The deviation between the observed and calculated m/z values for the metabolite peak, expressed as parts per million (ppm) or milliDaltons (mDa). A deviation exceeding a threshold (e.g., > 10 ppm) suggests a potential false positive.
  - Isotopic Pattern: The alignment between observed and theoretical isotopic patterns for the metabolite peak. For example, a small molecule containing chlorine should have a 3:1 ratio between the monoisotopic peak and the +1 Da peak and a 3:2 ratio for the +2 Da peak. A poor match in these patterns may indicate a false positive. The MassMetaSite software computes a similarity score as the mean squared error between the observed and theoretical patterns, normalized to the range [0, 1].
  - MS Area: The total ion count of the metabolite across the peak's retention time range. A low MS area relative to the parent compound can suggest a false positive.
  - Negative Control Area Ratio: The ratio between MS areas in the incubation sample and the blank. The signals present in both are considered nonspecific. Ideally, a specific signal unique to the sample is "Not In the Blank" (NIB). If the signal appears in both the sample and the blank (In Blank, IB), but with high area values in the latter, it may indicate background noise or system artifacts.
  - Fragmentation Pattern: The consistency of the metabolite's fragmentation pattern with that of the parent compound, accounting for expected modifications. Missing fragments or excessive noise can suggest a false positive.
- Chromatographic peak shape: The ideal chromatographic peak shape is characterized by a symmetrical narrow peak with a Gaussian shape. Tailing or fronting can occur for several reasons, including column overloading or deterioration, but irregular shapes can indicate a false positive.
- Kinetics: The relationship between incubation time and MS area for the metabolite peak. First-generation metabolites typically exhibit exponential kinetics at initial time points, whereas second-generation metabolites often show a sigmoidal shape. Deviation from these patterns can signal an unexpected reaction mechanism or a false positive.
- Reaction Mechanism: The expected reactions based on the experimental conditions. For example, glucuronides or GSH conjugates should not appear in experiments with human liver microsomes. Metabolites requiring multiple, improbable reactions for formation may be coincidental matches based on mass shifts.

For the training set, two experienced analysts manually labeled false positives in the data sets. One analyst focused on peptides, whereas the other reviewed small molecules. They evaluated each metabolite peak using all available data to determine whether it was a true or false positive. The selected metabolites for each experiment are shown in Supporting Information S3 and S4 for the SIT and MIT data sets,

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

Table 2. Description of All Features Used in the Model

Feature set	Feature name	Description
Mass spectra	Peak count	Total number of peaks detected in the spectrum
	Assignment count	Number of assigned peaks in the spectrum
	Assignment ratio	Assignment count/peak count
	Intensity sum	Sum of all relative intensities of peaks in the spectrum
	Assigned intensity sum	Sum of relative intensities for assigned peaks in the spectrum
	Assigned intensity ratio	Assigned intensity sum/Intensity sum
	Max assigned intensity	Highest relative intensity among the assigned peaks
	Peaks above thresholds	Number of peaks with intensities above specified thresholds
	Assignments above thresholds	Assigned peaks with intensities above specified thresholds
	Assignment ratio above thresholds	Assignments above thresholds/Peak count above thresholds
	Metabolite intensity	Intensity of the peak corresponding to the metabolite
	Metabolite position	m/z of the metabolite divided by highest m/z in the spectrum
	Isotopic similarity	Similarity between observed and calculated isotopic distribution
	m/z difference (ppm)	Relative difference between calculated and measured m/z
	Negative control ratio	MS area in the sample/MS area in the blank
Chromatographic peak shape	Variance	Centered moment of order 2, representing variance
	Skewness	Centered moment of order 3, indicating asymmetry
	Kurtosis	Centered moment of order 4, indicating peakedness
	EMG fit	Goodness of fit to exponentially modified Gaussian (R^2)
	Smoothness	Pearson's correlation between curve and 1-point-shifted curve
	Total variation	Sum of the absolute height differences between consecutive points
Kinetics	Nonzero count	Number of nonzero values in the kinetics
	Starting height	The height of the curve at the first incubation time point
	Peak count	Total number of peaks in the kinetic profile
	Valley count	Total number of valleys in the kinetic profile
	Is increasing	Boolean indicating if the curve is nondecreasing overall
	Is decreasing	Boolean indicating if the curve is nonincreasing overall
	Is strictly increasing	Boolean indicating if the curve is strictly increasing
	Is strictly decreasing	Boolean indicating if the curve is strictly decreasing
	Total variation	Sum of the absolute height differences between consecutive points
	Mean height	Average height of the curve across all time points
Mechanism	Is phase I	Whether the metabolite came from phase I metabolism
	Is phase II	Whether the metabolite came from phase II metabolism
	Adduct	One-hot encoded: [M + H]+, [M + Na]+, [M+K]+, [M+NH4]+ or other
	Compatible mechanisms	Multilabel binary indicator of compatible reactions by mass shift

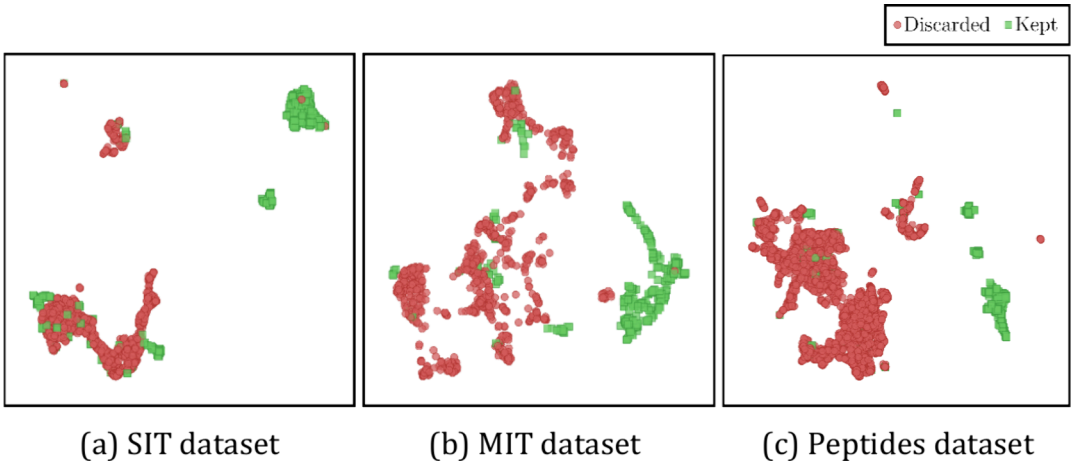


Figure 2. UMAP projection of the feature set showing separation between true positives and false positives in the three data sets. (a) SIT data set, (b) MIT data set, (c) Peptides data set.

respectively, and in[18] for the peptide data set. Summary statistics for the manual labeling process (Table 1) show a high discard rate across all three data sets, with a median exceeding 70%. Notably, the peptide data set reached a median discard rate of 90.79%. This high discard rate reflects the purpose of the analysis, as the user settings can be adjusted to either prioritize the identification of major metabolites or capture all potential metabolites by detecting the maximum number of peaks with minimal filtering. The latter approach was the intended strategy in this case. The differences in the discard rates between small molecules and peptides may also arise from variations in experimental conditions.

FEATURE ENGINEERING

A key contribution of this work is the development of a comprehensive feature set that effectively characterizes

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

Table 3. Cross-Validation Performance Metrics for the Three Datasets

Data set	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
SIT	84.36%	89.44%	88.82%	88.25%	90.58%	88.29%	2.11%
MIT	89.16%	95.85%	96.61%	94.36%	94.65%	94.13%	2.61%
Peptide	90.17%	92.44%	92.63%	93.24%	92.58%	92.21%	1.06%

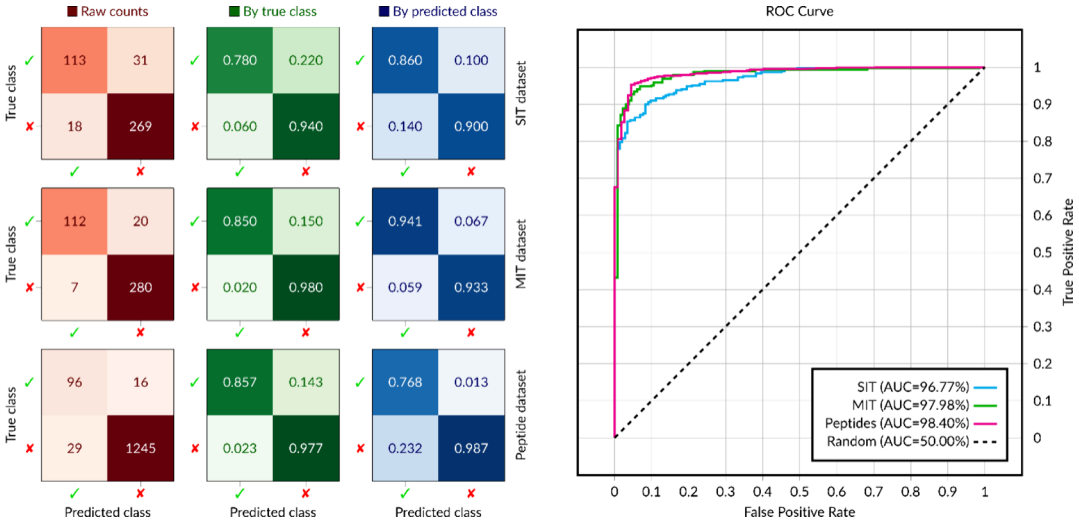


Figure 3. Test set performance metrics for different data sets. Left: Confusion matrices for the test sets of the SIT, MIT, and peptide data sets, showing results without normalization, normalized by true label, and normalized by predicted label. Right: ROC curves for the test sets of the SIT, MIT, and peptide data sets, with a reference line for a random classifier. AUC values are shown in the legend with the discarded class taken as positive.

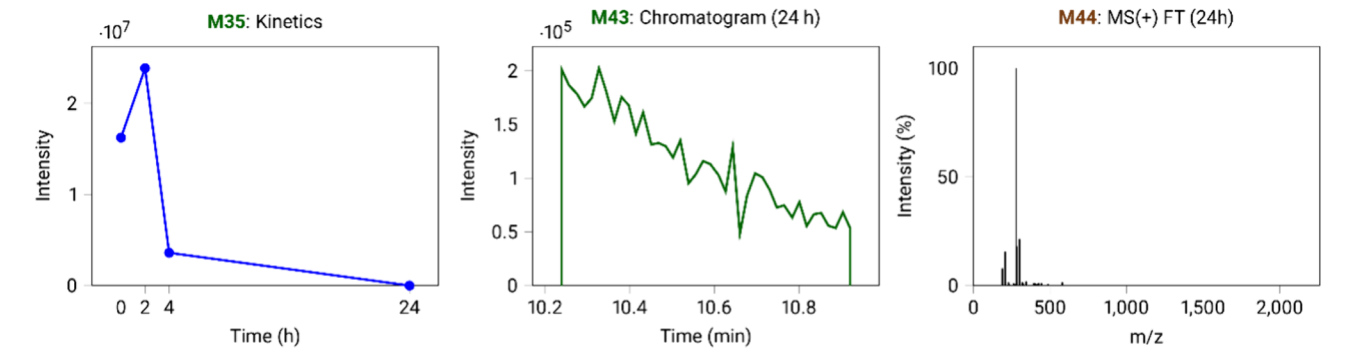
metabolite peaks, enabling the distinction between true and false positives. This feature set captures critical information from mass spectra, chromatographic peaks, kinetic data, and reaction mechanisms. The engineered features are specifically tailored to represent the quality and reliability of the data, playing a crucial role in enhancing the performance of the trained models. For each software-identified metabolite peak, features were derived from four sources: mass spectra, chromatographic peaks, kinetic data, and reaction mechanisms. The complete list of features is presented in Table 2. More details about how these features are computed are presented in Supporting Information S5. The result of this process is a feature vector for each metabolite peak, with 87 components corresponding to mass spectra, 6 components corresponding to chromatographic peaks, 10 components corresponding to kinetics, and 32 components corresponding to reaction mechanisms, totaling 135 features. Whenever multiple conditions were present in the data, multiple mass spectra and chromatogram peaks were available for each peak. To account for this in a way such that the number of conditions does not change the size of the final feature set, we computed the mean, standard deviation, minimum, and maximum of each feature across all conditions. This resulted in a feature set with 4 times the number of features derived from the mass spectra and chromatographic peaks described above, yielding a total of 414 features.

Uniform Manifold Approximation and Projection (UMAP)[19] is an unsupervised dimensionality reduction technique widely used for data visualization. Figure 2 presents a UMAP projection of the feature set into two dimensions for each data set, where noticeable separation between the two classes is observed despite UMAP not leveraging labels during projection. This highlights the feature set's high capability in distinguishing false positive identifications, suggesting that models trained on this feature set are likely to achieve high classification accuracy.

MODELING METHODS

Gradient Boosted Decision Trees (GBDTs)[14] were chosen as the method for the machine learning model. GBDTs are a powerful and flexible machine learning technique that has been shown to perform well in a wide range of applications and are considered the state-of-the-art method for tabular data.[20] They offer several advantages for our current application, including robustness to highly correlated features, invariance to monotonic transformations of the input features, ease of hyperparameter tuning, the ability to handle missing data, and interpretability potential. The GBDT models were trained using the LightGBM[21] and XGBoost[22] libraries. These libraries are widely used for training GBDT models and offer high-performance flexible implementations that can handle large data sets efficiently. The main hyperparameters of these GBDT models are the number of trees, the maximum number of leaves per tree, the maximum depth of the trees, the learning rate, and the regularization parameters. These hyperparameters were tuned via the FLAML[23] library, which is a fast and lightweight automatic machine learning library that provides state-of-the-art hyperparameter optimization algorithms. The default hyperparameter ranges and sampling distributions provided by FLAML were used for the hyperparameter tuning process. Alternative models to GBDTs were also evaluated and tuned using similar procedures, but GBDTs consistently achieved superior performance; see Supporting Information S6 for details.

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX



Name	RT	m/z	Mass shift	m/z diff [ppm]	m/z diff [mDa]	MiM	Z	MS Area	Area %	AI Suggested Action
M35 -408 RT=6.70	6.68 → 6.70	733.6396 → 733.6403	-408.1757	1.54 → 2.51	1.13 → 1.84	2929.5338	4	3.60E6 → 2.38E7	2.82 → 3.38	Hide (Kinetics)
M43 -2223 RT=10.17	10.17 → 10.33	186.9561 → 186.9562	-2222.0128	3.13 → 3.42	0.59 → 0.64	1115.6968	6	4.73E6 → 6.06E6	0.72 → 53.37	Hide (Chromatographic peak shape)
M44 -2315 RT=10.65	10.65 → 10.66	342.1671	-2314.2222	7.59 → 7.62	2.60 → 2.61	1023.4873	3	3.88E6 → 4.35E6	0.75 → 43.85	Hide (Mass spectrum)

Figure 4. Examples of false positive predictions for metabolites M35−408, M43−2223, and M44−2315. Each subfigure shows a specific feature (chromatographic peak, mass spectrum, or kinetics) that led the model to classify these as false positives, highlighting the interpretability of SHAP-based feature analysis.

EXPLAINING PREDICTIONS

To extract feature importance from the trained GBDT models, we used SHapley Additive exPlanations (SHAP).[15] SHAP values, derived from Shapley values in cooperative game theory, quantify each feature's contribution to a model's prediction. The sign of the SHAP values in a classification model indicates whether a feature contributes positively or negatively to the prediction of a specific class. In our analysis, a positive SHAP value for a feature means that the feature influences the model to predict that the peak should be discarded. We grouped features by their source, i.e., mass spectra, chromatographic peaks, kinetic data, and reaction mechanisms, to interpret the SHAP values. The combined SHAP values for each group were then analyzed to determine their relative importance in model predictions, considering only features with positive SHAP values. The feature group with the highest aggregated SHAP value was identified as the most significant contributor to the model's decision to discard a given peak.

RESULTS AND DISCUSSION

Each data set was split into training and test sets, containing 80% and 20% of the samples, respectively. Stratified sampling was applied to maintain the distribution of the target variables in both sets. To ensure rigorous model development, we performed 5-fold cross-validation within the training set for hyperparameter tuning and validation. The test sets were reserved exclusively for the final evaluation of the model generalization. Performance during cross-validation was assessed by using the balanced accuracy metric, defined as the average recall across classes. The optimal hyperparameters were selected based on the highest average balanced accuracy score across folds, employing the FLAML[23] library with a maximum of 100 trials. The balanced accuracy metrics across different folds in the cross-validation process are presented in Table 3.

Table 3 presents the cross-validation performance metrics for the three data sets, reporting the mean and standard deviation of balanced accuracy. All data sets demonstrate high cross-validation performance, with mean balanced accuracies exceeding 88%. Low standard deviations, peaking at 2.61% for the MIT data set, suggest model stability and minimal performance variability across folds. After the hyperparameters were optimized, the models were retrained on the full training sets and evaluated on the test sets. The test results are summarized in Figure 3, which includes confusion matrices and Receiver operating characteristic (ROC) curves for each data set. The balanced accuracies on the test sets were 86.0%, 91.5%, and 91.7% for the SIT, MIT, and peptide data sets, respectively. These values are slightly below the cross-validation scores but remain well within acceptable ranges, indicating strong generalizability to new data.

For the discarded class (false positives), the precision exceeds 90%, reaching 90.0%, 93.3%, and 98.7% on the SIT, MIT, and peptide data sets, respectively. This high precision indicates that few expert-retained metabolites were discarded by the model. The sensitivity for the discarded class is even higher at 94%, 98%, and 97.7% for the SIT, MIT, and peptide data sets, respectively, clearly demonstrating the model's exceptional efficacy in identifying false positives while ensuring that valid metabolites are not mistakenly discarded. The ROC curves corroborate these results as the true positive rate quickly approaches 100% at low false positive rates. Overall, these findings highlight the approach's promise in reducing false positive identifications in MetID. By flagging likely false positives, the models assist experts in prioritizing complex cases, enabling more efficient resource allocation.

EXPLAINING PREDICTIONS: APPLICATION EXAMPLE

We demonstrated our proposed approach for explaining predictions with SHAP values using taspoglutide, a 30-amino acid drug for type 2 diabetes, in a prospective experiment involving incubation alongside protease dipeptidyl peptidase-4 (DPP4) at four time points. MetID was initially performed manually by the experts according to the criteria described in the Supporting Information Section 2.2. The results were then compared with the predictions made by the peak selection model. Both approaches identified 15 metabolites and 28 false positives, showing excellent agreement. In Figure 4, we present three examples of metabolites predicted as false positives by the model, with the rationale for each exclusion noted as an additional column in Oniro's interface. The leftmost example shows the kinetics of metabolite M35−408, where signals are present at time zero of incubation, when no biotransformation

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

is expected to occur. Based on these kinetics, the model appropriately marked this metabolite as a false positive. The center example shows the chromatographic peak shape of metabolite M43−2223, where an irregular peak shape suggests that the signal could stem from noise, interference, or contamination. The model accurately flagged this metabolite as a false positive based on peak shape. Finally, the right-most example shows the mass spectrum of the metabolite M44−2315, which lacks a primary molecular ion peak and has a low ratio of assigned to total peaks, indicating a poor-quality spectrum and justifying the model's exclusion of this metabolite.

MODEL UPDATE

One key aspect to highlight is the model's ability to be updated, with its performance continuously improving as new MetID data from ongoing experiments are analyzed and curated within the Oniro platform. The model undergoes automatic updates when the following conditions are met:

- A minimum of five new experiments were conducted.
- The number of new experiments constitutes at least 10% of the current total number of experiments in the model.

These criteria ensure that updates occur only when a sufficient amount of new data is available, striking a balance between model responsiveness and stability. Frequent updates with minimal data could lead to noise-driven fluctuations, whereas infrequent updates might delay potential improvements. Data from new experiments are split using an 80/20 training/test ratio, where the new training data are added to the training set and the new test data are assigned to the test set. This process ensures that the model is retrained with the updated training data while maintaining the existing data in its respective training and test sets. To illustrate how this update mechanism functions in practice, we simulate the process using the SIT data set. Initially, the model is built using data from only five experiments. As new experiments are added, the model is updated whenever both updated conditions are met. Up to 55 experiments, and updates occur every time five new experiments are introduced. However, beyond this point, the 10% threshold becomes the dominant factor, requiring a greater number of new experiments before subsequent updates can take place.

Figure 5 illustrates the evolution of precision, recall, and ROC AUC on the test set for the SIT data set as model updates are performed. Initially, when data are scarce, the model's ability to generalize to unseen data is limited, with precision being particularly low, indicating a high rate of incorrectly discarded metabolites by the model. With only a small number of experiments, the model may capture only a narrow range of patterns, leading to a suboptimal performance and a greater risk of overfitting. However, as more experiments are conducted and additional data become available, the model consistently improves, demonstrating its ability to learn from the accumulated data. Over time, this leads to enhanced generalization, as reflected in better predictive metrics, especially in the precision metric. This process highlights the effectiveness of the update strategy in ensuring that the model adapts dynamically while maintaining robustness.

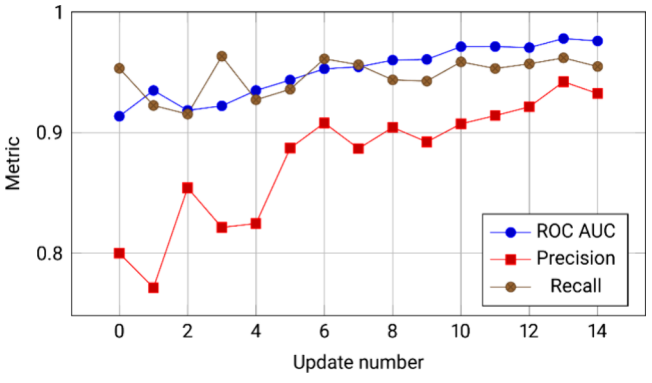


Figure 5. Change in test set metrics across updates in the SIT data set.

DATA SET-SPECIFIC GENERALIZATION

We investigated the impact of data set-specific experimental conditions on model generalization using the MIT data set, which includes two subsets with different kinetic sampling protocols. The first subset (Protocol A, 13 experiments) sampled at 5, 15, 40, 80, and 140 min, while the second subset (Protocol B, 14 experiments) sampled at 0, 2, 5, 10, 15, and 30 min. Notably, Protocol B includes a predose measurement at 0 min, which is absent in Protocol A.

We trained separate models on each subset following the described procedures with 80/20 training/test splits and evaluated them on both matching and mismatched test sets. A combined model trained on the combined training data from both protocols was also included for comparison.

As shown in Table 4, models achieved the highest performance when trained and tested on the same protocol (Local BA) but maintained respectable performance when evaluated on the other protocol (Cross-protocol BA). These results indicate that while data set-specific features influence predictive accuracy, generalization across similar experimental conditions remains viable.

Table 4. Balanced Accuracy (BA) of Models Trained on Protocol-Specific Subsets of the MIT Dataset[a]

Training Protocol	Global BA	Local BA	Cross-protocol BA
Protocol A	0.9402	0.9521	0.8824
Protocol B	0.9387	0.9458	0.8749

[a]Global BA refers to a model trained on the combined dataset (both protocols). Local BA indicates the performance of a model trained and tested on the same protocol. Cross-protocol BA reports the performance of a model trained on the other protocol and tested on the protocol it was not trained on.

More broadly, a model's ability to generalize depends on the compatibility of input features and labeling criteria. For instance, if one data set lacks kinetic information altogether (e.g., using only a single incubation time), transferability from or to time-resolved data sets would be infeasible due to structural mismatches in the feature space. Similarly, if data sets are constructed using different expert criteria to define what constitutes a false positive, the resulting decision boundaries may be incompatible. However, in the present case, both protocols provide sufficient temporal structure and were labeled using consistent criteria, which enables meaningful generalization across them and supports the broader applicability of the proposed approach.

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

Table 5. Test Performance Metrics for Data Sets 1 and Dataset 2

Data set	Support	Balanced Accuracy	Precision	Recall	ROC AUC	FN	FP	TN	TP
Data set 1	2157	86.66	98.71	97.17	96.08	58	26	83	1990
Data set 2	107	64.79	72.37	82.09	76.62	12	21	19	55

Table 6. Test Performance Metrics for Dataset 1 and Dataset 2 before and after Model Updates

	Balanced accuracy		Precision		Recall		ROC-AUC	
Data set	Original	Updated	Original	Updated	Original	Updated	Original	Updated
Data set 1	86.66	90.06	98.71	99.18	97.17	97.63	96.80	97.93
Data set 2	64.79	66.60	72.37	76.23	82.09	94.90	76.62	80.03

CASE STUDY: VALIDATION ON PHARMACEUTICAL DATA

To assess the real-world applicability of the proposed false positive detection method, we applied it to two distinct data sets from a leading pharmaceutical company. This evaluation was made possible by the standardization of the data within the Oniro platform, highlighting the essential role of consistent and well-curated data in enabling the effective application of the developed approach.

These data sets were selected to reflect several use cases from the early phase of the drug discovery process, including both small molecules and peptides and data analyzed using differing selection criteria and various false-positive rates. Data set 1 includes small molecule MetID from data generated in an automated hepatocyte incubation assay, which is highly suitable for this model, because of the standardized HPLC chromatography and HRMS conditions. The selection criterion for the metabolites in this data set was the three most abundant metabolites. The data set included 125 experiments with 13,552 software-identified metabolites with 10,510 false positives (77.55%) identified manually. Data set 2 comprises macrocyclic peptide experiments obtained from a standardized simulated intestinal fluid stability assay. Metabolite selection was guided by criteria favoring those with high molecular weights or elevated relative abundances to support the MetID assignment. Additionally, the overall charge of each metabolite relative to that of its parent compound was considered. A total of 1,225 software-identified metabolites were assessed, with 690 confirmed as false positives (56.33%).

Using these data sets, models were trained based on the described methodology, with an 80/20 training/test split. The test set performance metrics for both data sets are summarized in Table 5. Data set 1 shows a strong performance with high balanced accuracy (86.66%), precision (98.71%), and recall (97.17%). In contrast, Data set 2 demonstrates lower balanced accuracy (64.79%) and precision (72.37%), possibly because of the greater complexity of macrocyclic peptide data and the more intricate selection criteria, which may present challenges for accurate classification.

As previously discussed, the model has the capability to be updated. This has been demonstrated in the models developed for both data sets, where new data have been incorporated over several months. The performance metrics are provided in Table 6 for reference.

An increase in balanced accuracy (from 86.66% to 90.06% in Data set 1 and from 64.79% to 66.60% in Data set 2) and precision (from 98.71% to 99.18% in Data set 1 and from 72.37% to 76.23% in Data set 2) can be observed in the updated models, which translates into improved overall prediction performance and more reliable identification of false positives. These improvements reflect the model's ability to adapt to new data, ensuring that it remains relevant and robust in dynamic environments.

CONCLUSIONS

This study presents a novel machine learning approach to address the critical challenge of false positive identifications in MetID studies. Using GBDT models, the methodology demonstrated high precision and recall rates, exceeding 90% in detecting false positives across various data sets, including both small molecule and peptide data. The developed algorithm is platform- and user-agnostic, enabling its implementation on any system that supports the standardized collection of MetID data, such as the Oniro platform used in this study. Importantly, the model is designed as a decision-support tool to assist, rather than replace, expert judgment, and it does not apply hard thresholds or filtering criteria that might discard potential true metabolites.

The key findings and contributions of this research include the following:

- Comprehensive Feature Engineering: By developing an extensive feature set that effectively characterizes metabolite peaks, this study enhances the accuracy of distinguishing between true and false positives. The features derived from mass spectra, chromatographic peaks, kinetic data, and reaction mechanisms collectively contribute to the model's high performance.
- High Model Performance: The GBDT models trained on the constructed feature set exhibited robust performance across cross-validation and test sets, with balanced accuracies consistently above 86%. This finding indicates that the models have strong generalizability to unseen data.
- Reduction in Manual Effort: The proposed machine learning approach significantly reduces the reliance on manual data interpretation, which is traditionally both tedious and time-consuming. By prioritizing candidates based on likelihood of being false positives, the efficiency of MetID workflows can be improved without compromising sensitivity.
- Explainability and Adoption: Employing SHAP values to interpret the model's predictions provides transparency and aids in understanding the decision-making process of the model. This explainability is crucial for the adoption of the proposed approach in practical MetID workflows.
- Versatility Across Data sets: The methodology was validated using diverse data sets, including different incubation conditions and acquisition methods, demon-

https://doi.org/10.1021/acs.analchem.5c02745
Anal. Chem. XXXX, XXX, XXX−XXX

strating its versatility and applicability to various MetID scenarios.

While our approach helps reduce the burden of manual review by ranking and contextualizing likely false positives, it does not eliminate human oversight. The model is intended to accelerate expert workflows, not replace them. By keeping experts in control of the final decisions, the risk of discarding true metabolites due to false negatives is mitigated. Looking ahead, future work could focus on refining the feature set and model architecture as well as developing hybrid strategies that combine predictive modeling with uncertainty quantification. These improvements may further enhance prioritization while preserving the reliability and safety critical to drug metabolism studies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.5c02745.

S1. SIT-settings.pdf: Detailed information on the data preprocessing steps for the SIT data set (PDF)

S2. MIT-settings.pdf: Detailed information on the data preprocessing steps for the MIT data set(PDF)

S3. SIT-metabolites.pdf: Reports of selected metabolites for the SIT data set (PDF)

S4. MIT-metabolites.pdf: Reports of selected metabolites for the MIT data set (PDF)

S5. SI_Features.pdf: Additional details on the feature engineering process (PDF)

S6. supplementary-algorithms.pdf: Performance comparison of alternative models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ramon Adàlia − *Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain; Lead Molecular Design, S.L., Sant Cugat del Vallès 08173, Spain;* orcid.org/0009-0004-9458-1922; Email: Ramon.Adalia@autonoma.cat

Authors

Paula Cifuentes − *Universitat Pompeu Fabra, Barcelona 08003, Spain; Lead Molecular Design, S.L., Sant Cugat del Vallès 08173, Spain;* orcid.org/0009-0007-8181-8822

Joyce Liu − *Genentech, South San Francisco, California 94080, United States*

Lionel Cheruzel − *Genentech, South San Francisco, California 94080, United States*

Gemma Sanjuan − *Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain*

Tomàs Margalef − *Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain*

Ismael Zamora − *Lead Molecular Design, S.L., Sant Cugat del Vallès 08173, Spain;* orcid.org/0000-0002-7700-0354

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.5c02745

Author Contributions

⊥(R.A., P.C.) Contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was conducted with the support of Pla de Doctorats Industrials del Departament de Recerca i Universitats de la Generalitat de Catalunya (grant numbers 2023-DI-00006 and 2023-DI-00002).

■ REFERENCES

(1) Zamora, I.; Fontaine, F.; Serra, B.; Plasencia, G. *Drug Discovery Today: Technologies* **2013**, *10*, No. e199.

(2) Goracci, L.; Desantis, J.; Valeri, A.; Castellani, B.; Eleuteri, M.; Cruciani, G. *J. Med. Chem.* **2020**, *63*, 11615−11638.

(3) Taneja, A. *Technology (IJARET)* **2024**, *15*, 344−352.

(4) Leung, C.; Liu, J.; Cunico, K.; Johnson, K.; Yan, Z.; Cai, J. *Drug Metab. Dipos.* **2024**, *52*, 377−389.

(5) Yu, T.; Jones, D. P. *Bioinformatics* **2014**, *30*, 2941−2948.

(6) Gloaguen, Y.; Kirwan, J. A.; Beule, D. *Anal. Chem.* **2022**, *94*, 4930−4937.

(7) Chetnik, K.; Petrick, L.; Pandey, G. MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC-MS metabolomics data. *Metabolomics* **2020**, *16*. DOI: 10.1007/s11306-020-01738-3

(8) Ju, R.; Liu, X.; Zheng, F.; Zhao, X.; Lu, X.; Zeng, Z.; Lin, X.; Xu, G. *Anal. Chim. Acta* **2019**, *1067*, 79−87.

(9) Cuyckens, F.; Dillen, L.; Cools, W.; Bockx, M.; Vereyken, L.; de Vries, R.; Mortishire- Smith, R. J. *Bioanalysis* **2012**, *4*, 595−604.

(10) Bonn, B.; Leandersson, C.; Fontaine, F.; Zamora, I. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 3127−3138.

(11) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148.

(12) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. *Anal. Chem.* **2017**, *89*, 8696−8703.

(13) Lenski, M.; Maallem, S.; Zarcone, G.; Garcon, G.; Lo-Guidice, J.-M.; Antherieu, S.; Allorge, D. *Metabolites.* **2023**, *13*, 282.

(14) Friedman, J. H. *Annals of Statistics* **2001**, *29*, 1189−1232.

(15) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765−4774.

(16) Radchenko, T.; Kochansky, C. J.; Cancilla, M.; Wrona, M. D.; Mortishire-Smith, R. J.; Kirk, J.; Murray, G.; Fontaine, F.; Zamora, I. Metabolite identification using an ion mobility enhanced data-independent acquisition strategy and automated data processing. *Rapid Commun. Mass Spectrom.* **2020**, *34*. DOI: 10.1002/rcm.8792

(17) Paiva, A. A.; Klakouski, C.; Li, S.; Johnson, B. M.; Shu, Y.-Z.; Josephs, J.; Zvyaga, T.; Zamora, I.; Shou, W. Z. *Bioanalysis* **2017**, *9*, 541−552.

(18) Cifuentes López, P.; Zamora, I.; Radchenko, T.; Fontaine, F.; Garriga, A.; Morettoni, L.; Kammersgaard Christensen, J.; Helleberg, H.; Becker, B. A. *bioRxiv* **2025**, DOI: 10.1101/2025.05.01.651614.

(19) Sainburg, T.; McInnes, L.; Gentner, T. Q. *Neural Computation* **2021**, *33*, 2881−2907.

(20) McElfresh, D.; Khandagale, S.; Valverde, J.; C, V. P.; Feuer, B.; Hegde, C.; Ramakrishnan, G.; Goldblum, M.; White, C. *arXiv* **2023**, DOI: 10.48550/arXiv.2305.02997.

(21) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp 3149−3157.

(22) Chen, T.; Guestrin, C. *arXiv* **2016**, DOI: 10.1145/2939672.2939785.

(23) Wang, C.; Wu, Q.; Weimer, M.; Zhu, E. *arXiv* **2021**, DOI: 10.48550/arXiv.1911.04706.