Enhanced identity spectrum search with AI/ML confidence scoring for HRAM data

Gábor Zsemlye[1], Rajesh k. Jha [2], Maria Falaq [2], Juraj Lutišan[1], Marynka Ulaszewska[3], Samuel Benkovič[1], Tim Stratton[4], Michal Raab[1]
[1] Thermo Fisher Scientific, Bratislava, Slovakia; [2] Thermo Fisher Scientific, India; [3] Thermo Fisher Scientific, Milano, Italy; [4] Thermo Fisher Scientific, Austin, USA

Abstract

Purpose: Spectra annotation is a major challenge in untargeted small molecule analysis. Initial steps involve comparing unknown spectra with experimental spectral libraries. We propose a novel AI/ML confidence scoring system for definitive Orbitrap data identification using the mzCloud library.

Methods: A histogram gradient boosting model was employed to mimic scientists' evaluation of library search results. Multiple searches were simulated, learning from both true and false hits using data from the mzCloud curated library. The model utilized 170 input features to capture fragmentation spectra complexity, metadata, and query-hit matching scores, including Cosine, NIST, and HighChem-HighRes scores.

Results: The AI/ML model was validated with mzCloud standards, pending real sample validation. It was compared against NIST, Cosine, HighChem-HighRes, and a 2016 confidence score.

Introduction

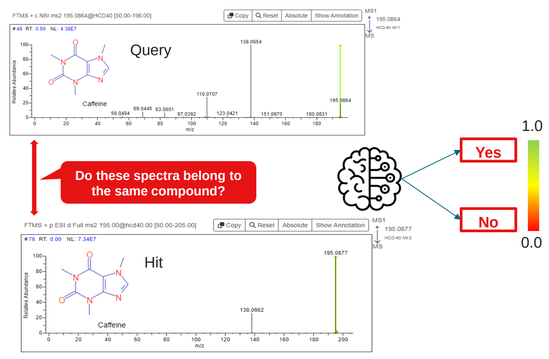
Spectra annotation is a primary challenge in untargeted analysis within small molecules applications. Comparison of unknown spectra against experimental spectral libraries is usually the first step of every annotation workflow. Ideally such identification algorithm provides list of best possible hits, where high ranking score unequivocally indicates, which library hit corresponds to unknown spectra. In real life, however the ambiguous results are unfortunately a daily routine, as isomeric and/or isobaric species may produce similar spectra, or because some regions of collision energies produce poorly specific spectra. This will result in multiple high scored hits, what does not help in taking decisions and driving conclusions. To answer that challenge, we propose a new AI/ML confidence scoring system for unequivocal spectra identification of Orbitrap data against Thermo Scientific™ mzCloud™ mass spectral library.

Materials and methods

Model Selection

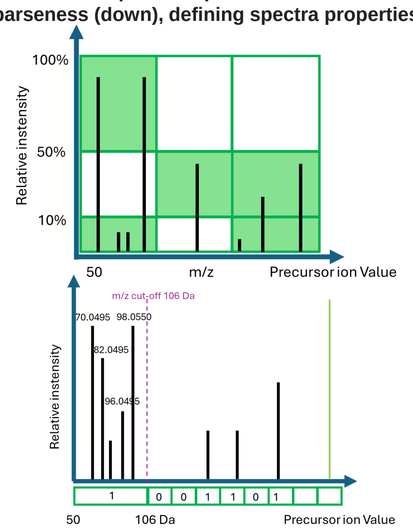
A machine learning model aims to replicate the scientist's behavior in observing hit results during the compound identification through library searches. This is achieved by simulating multiple searches and learning from true and false hits. During the training the ML model receives a pair of query and hit spectra (along with metadata and other calculated features), and the label if the two spectra belong to the same compound or not. This learned information can then be used at real spectrum search: for each query and hit candidate a confidence can be predicted, if the spectrum pair belongs to the same compound. See Figure 1

Figure 1. Input and output of the ML model



The histogram gradient boosting model was selected for this research due to its robustness and flexibility. While any regression model could be applicable, this particular model is favored for its ability to handle missing features, a common occurrence in real-world datasets, as facilitated by the scikit-learn implementation. The 170 model input features were created to account for a variety of parameters reflecting complexity of fragmentation spectra (such as sparseness, balancenes, etc), its metadata (such as analyzer, isolation width or precursor mass and its accuracy) and matching pairs query-hit including ranking scores form Cosine, NIST and HighChem-HighRes matchings from symmetric search. See Figure 2 for some examples.

Figure 2. Two examples of input features: Balanceness (up) and Sparseness (down), defining spectra properties.

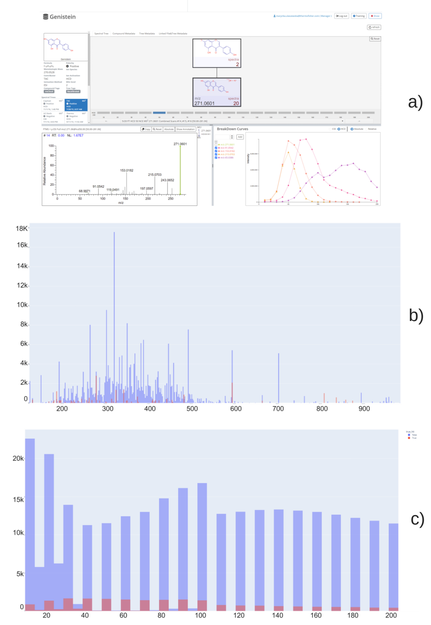


Data selection

Data for model training and validation were selected from mzCloud curated spectral library, which offers 34,000 compounds belonging to different compound classes with 3,462,578 MS2 spectra belonging to CID and HCD activation types, across 10-100 and 10-200 NCE levels, respectively. All compounds used for the model training and validation were known analytical standards defined by InChI and InChIKeys. For the search simulation, the spectra were selected from the Autoprocessed library and it was searched in the Reference library. To grant better variability and distribution of query and true hits, an intersection of 1600 compounds existing in both Autoprocessed and Reference library were chosen, with additional 1600 compounds accounting for false hits. During the validation phase, an additional 3600 compounds were incorporated into the dataset to more accurately simulate real-world conditions. This adjustment reflects the scenario in which scientists query an unknown spectrum against a spectral library containing thousands of compounds, thereby reducing the probability of accurately identifying the correct match.

The search process is emulated by invoking the mzCloud APIs via a Python script, followed by the conversion of the resulting data into feature sets. This procedure is computationally intensive but amenable to parallelization. When executed with 80 parallel workers, the task completes in 10 hours on an ml.r6i.32xlarge AWS instance.

Figure 3. Example of data in mzCloud mass spectral library: a) data available for one compound entry; b) *m/z* precursor ion distribution in training dataset; c) NCE distributions in training dataset

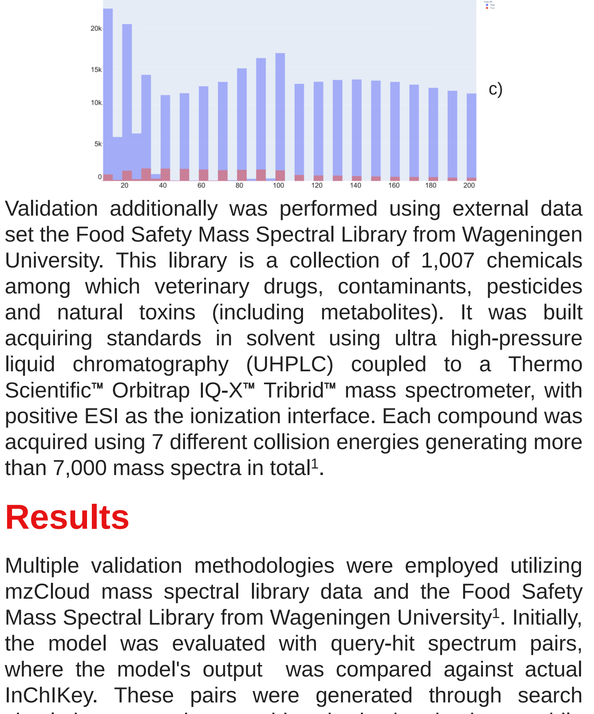


Validation additionally was performed using external data set the Food Safety Mass Spectral Library from Wageningen University. This library is a collection of 1,007 chemicals among which veterinary drugs, contaminants, pesticides and natural toxins (including metabolites). It was built acquiring standards in solvent using ultra high-pressure liquid chromatography (UHPLC) coupled to a Thermo Scientific™ Orbitrap IQ-X™ Tribrid™ mass spectrometer, with positive ESI as the ionization interface. Each compound was acquired using 7 different collision energies generating more than 7,000 mass spectra in total[1].

Results

Multiple validation methodologies were employed utilizing mzCloud mass spectral library data and the Food Safety Mass Spectral Library from Wageningen University[1]. Initially, the model was evaluated with query-hit spectrum pairs, where the model's output was compared against actual InChIKey. These pairs were generated through search simulations, ensuring true hit exist in the database, while including also false hits. The model's accuracy was determined by counting the number of correct predictions it made for a given spectrum pair out of all possible spectrum pairs. A prediction is considered correct if the model accurately identifies both true positives and true negatives. The model achieved an accuracy of 89.2%. Additionally, the ROC AUC for AI/ML model was 0.95. For comparison, the ROC AUC for other scoring methods were as follows: 0.66 for Cosine, 0.68 for HighChem-HighRes, 0.67 for NIST, and 0.58 for the legacy confidence method.

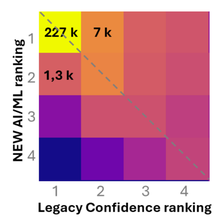
Figure 4. AUC scores at the compound level. Searching executed in mzCloud mass spectral library



A similar validation was conducted at the compound level, acknowledging that each compound is characterized by multiple spectra acquired at various CID and HCD NCE levels. Here, the model's input was a query spectrum, and a hit was defined as a compound, with the model determining if the query spectrum could belong to that compound. Using mzCloud mass spectral library data for this validation, the ROC AUC for the model was 0.99. Traditional match scores yielded significantly lower AUC values, as illustrated in Figure 4 (upper chart). The legacy confidence scoring model based on Bayesian Networks available currently in Thermo Scientific™ Compound Discoverer™ software achieved an ROC AUC of only 0.92. When using data from the Food Safety Mass Spectral Library, the AUC was slightly lower at 0.97 but still outperformed other scoring methods.

The ranking capabilities of the model were also evaluated. When a spectrum of a certain compound is searched in the spectral library, the search yields a list of hit candidates. The true hit compound should be ranked high in that list, ideally on the first position. The ranking was compared with the legacy confidence score system, in the following way: for each query-hit spectrum pair the ranking in the hit result list was calculated, once with the new AI/ML confidence, once with the legacy one. Then the ranks were counted in a two-dimensional heatmap, with the legacy confidence on the X axis and new AI/ML confidence on Y (See the Figure 5). Ideally both models should rank the true compound on the first rank, and it happened in majority of cases (yellow cell of the heatmap). Cells on diagonal represent cases where both models ranked compounds in equal way. Above this diagonal the new AI/ML confidence model performs better, below this diagonal the legacy model correctly ranked spectra pairs. It can be seen on the Figure 5 (upper heatmap) that for approx. 7k spectra pairs new AI/ML model ranked better respect to legacy one, while for 1.3k spectra pairs, legacy confidence model performed better. The spectra pairs correspond to a specific number of compounds, here we achieved improvement for 104 compounds, while model struggled with proper ranking for 48 compounds, see tables in Figure 5. Such ranking evaluation was performed for both mzCloud mass spectral library data and Food Safety Mass Spectral Library

Figure 5. Ranking capabilities of the new AI/ML confidence model, compared with the legacy confidence model. Left: Validation on mzCloud dataset. Right: Validation on Food Safety Mass Spectral Library.

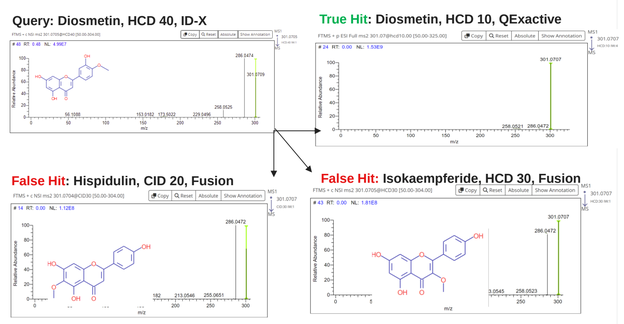


mzCloud Validation 400 compounds	
Compounds with improved ranking	104
Compounds with wrong ranking	41
Compounds with equal ranking	255

Wageningen Validation 560 compounds	
Compounds with improved ranking	48
Compounds with wrong ranking	28
Compounds with equal ranking	484

In the Figure 6 a real case is shown, where the new AI/ML confidence helps to distinguish true and false hits for the isomeric compound species. The HDC 40 spectrum of Diosmetin was searched in the mzCloud Reference library. The expectation was to achieve the correct ranking for hits and score values that will clearly distinguish between the right and false hits. The new AI/ML model achieved the goal, while legacy confidence score was not able to rank hits, nor properly distinguish between true and false hits. All scores were very low, suggesting none of hits was probable. The HighChem-HighRes scoring algorithm returned similarly high values for all hits, suggesting it can be any of three, but unable to indicate the true hit. Other two traditional scoring algorithms NIST and Cosine assigned higher value to one of the hits, but unfortunately both were false hits.

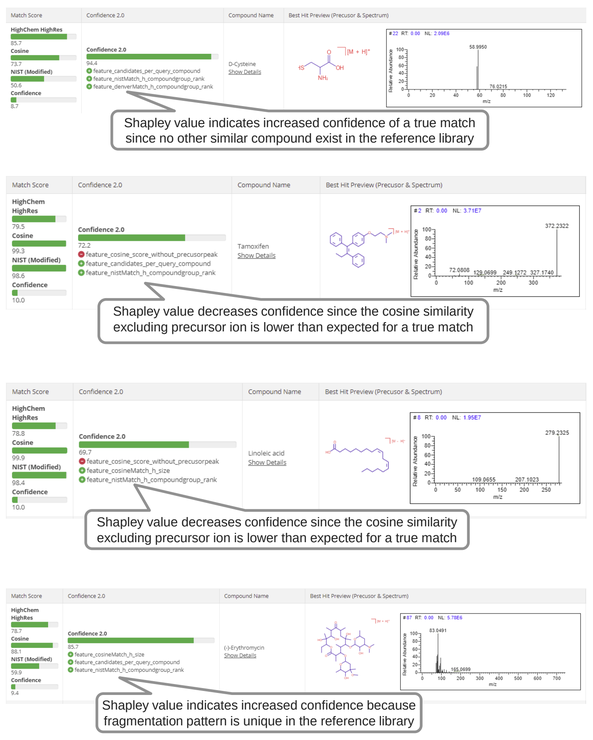
Figure 6. Search results for Diosmetin and different scoring algorithms



Hits	Metadata	AI/ML Confidence	Legacy Confidence	HighChem-HighRes	NIST	Cosine
Diosmetin	HCD 10, QE	**78.6**	8.1	89.1	47.1	61.1
Hispidulin	CID 20, Fusion	**11.4**	8.1	84.3	46.3	61
Isokaempferide	HCD 30, Fusion	**4.6**	9.8	81	74.9	96.5

Understanding the reasoning behind the outputs of an AI/ML model is essential. Shapley values serve this purpose by quantifying the contribution of each input feature to the model's final prediction, see Figure 7. These contributions can be displayed to the user, emphasizing the most influential input features for each query spectrum-hit spectrum pair, or aggregated at the compound level if multiple spectra are available for each compound in the library. In addition to the input feature name, a numerical value ranging from 0 to 1 can indicate the magnitude of its influence, while a directional sign (+/-) can show whether the feature positively or negatively affects the outcome.

Figure 7. Different examples of ranked feature contributions with directional impact for model predictions



Conclusions

The new AI/ML Confidence Score model demonstrates superior accuracy and classification capabilities for identity searches against an MS2 spectral database, surpassing traditional deterministic spectral similarity calculations used in mzCloud mass spectral library, including HighChem-HighRes, NIST, Cosine, and the previous Confidence Score of Compound Discoverer software. Over 170 features were engineered to incorporate various data and metadata clues analyzed by subject matter experts during library search candidate evaluations. Training on a substantial portion of the mzCloud mass spectral library enhances the understanding of scoring rationale in individual cases, allowing users to adjust specific scan conditions, such as varying collision energy, to improve confident identification. The new model will be available on the updated mzCloud mass spectral library site alongside the existing scoring methods.

Acknowledgements

Acknowledgement to WFSR (Wageningen Food Safety Research) for providing the open access Food Safety Mass Spectral Library

References

1. Food Safety Mass Spectral Library from Wageningen University, accessed in March 2025, https://www.wur.nl/en/show/food-safety-mass-spectral-library.htm.
QR-code for downloading the library:

Trademarks/licensing