

Article

PAID: An AI-Ready LC-MS/MS Dataset for Pesticide Residue Analysis

Jihang Zhang ¹, Qianjin Li ², Heng Zhou ^{3,4,5}, Lin Guo ^{3,4,5}  and Xingchuang Xiong ^{3,4,5,*} 

¹ School of Information Engineering, China Jiliang University, Hangzhou 310018, China; s24030812011@cjlu.edu.cn

² School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; xc122625@gs.zzu.edu.cn

³ Center for Metrology Scientific Data, National Institute of Metrology, Beijing 100029, China; zhouheng@nim.ac.cn (H.Z.); guolin@nim.ac.cn (L.G.)

⁴ National Metrology Data Center, Beijing 100029, China

⁵ Key Laboratory of Metrology Digitalization and Digital Metrology, State Administration for Market Regulation, Beijing 100029, China

* Correspondence: xiongch@nim.ac.cn

Abstract

Liquid chromatography–tandem mass spectrometry (LC-MS/MS) is widely employed in pesticide residue analysis. Machine learning methods for automated spectral interpretation depend on large, well-curated training datasets; however, publicly available pesticide mass spectrometry data are fragmented across heterogeneous repositories, lack standardized preprocessing, and suffer from incomplete metadata. We introduce PAID (Pesticide AI-ready Dataset), comprising two curated LC-MS/MS spectral collections derived from 15 public sources (GNPS, MassIVE, MoNA, MassBank). Starting from 91,420 raw spectra, after initial pesticide-directed screening, a seven-step reproducible pipeline—spanning multi-source integration, spectral cleaning, deduplication, metadata standardization, quality scoring, stratified splitting, and feature engineering—yields PAID-Strict (7527 spectra; 3197 compounds) and PAID-Extended (21,292 spectra; 3224 compounds). Both versions cover eight pesticide categories across QTOF and Orbitrap platforms, with core metadata fields (SMILES, InChIKey, molecular formula) exceeding 98% completeness. A feature suite of 32 chemoinformatic descriptors, 2214 molecular fingerprints, and 30 spectral features is provided alongside the spectra. Benchmark classification of the eight pesticide categories using XGBoost and LightGBM achieved 81.5% accuracy. The dataset, code, and pre-computed features are publicly available under CC BY 4.0 and MIT licenses.

Keywords: pesticide residue; liquid chromatography–tandem mass spectrometry; spectral preprocessing; molecular fingerprint; machine learning



Academic Editor: Yongqing Cai

Received: 18 May 2026

Revised: 1 July 2026

Accepted: 7 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Summary

Liquid chromatography–tandem mass spectrometry (LC-MS/MS) is a core technology for pesticide residue analysis [1]. Machine learning methods for mass spectrometry interpretation require training data that are high in quality, standardized in format, and complete in metadata. However, existing public pesticide mass spectrometry data are distributed across heterogeneous formats, exhibit variable spectral quality, and frequently lack essential metadata and standardized preprocessing. This paper describes PAID (Pesticide AI-ready Dataset) and the accompanying construction workflow (Figure 1). After initial

pesticide-directed screening, the pipeline consists of seven modular processing steps: multi-source data integration, spectrum cleaning and deduplication, metadata standardization, quality filtering, stratified splitting, multi-dimensional feature engineering, and baseline model validation.

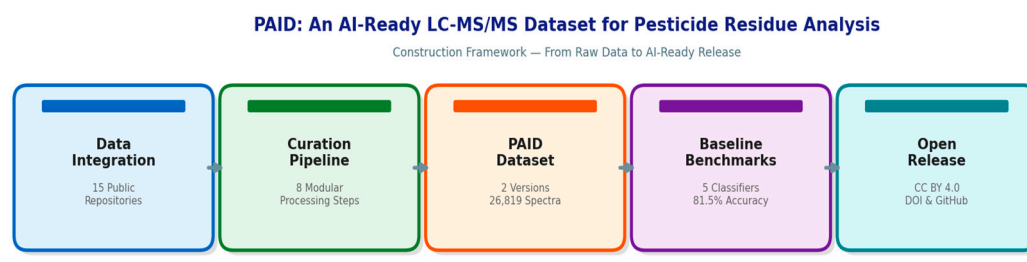


Figure 1. Overall framework of the PAID dataset construction and validation workflow. The pipeline integrates 15 public data sources through a seven-step processing workflow to produce two curated dataset versions (PAID-Strict and PAID-Extended), accompanied by pre-computed features and baseline model benchmarks.

We integrated 15 public data sources including GNPS, MassIVE, MoNA, and MassBank, processing 91,420 raw spectra through multi-level quality control to produce two dataset versions: PAID-Strict (7527 spectra, 3197 unique compounds), designed for high-precision model training with minimal redundancy; and PAID-Extended (21,292 spectra, 3224 compounds), which retains greater experimental condition diversity and is suited to data augmentation and large-scale pretraining. The dataset spans eight pesticide functional categories, supports both QTOF and Orbitrap high-resolution platforms, and exceeds 98% completeness for SMILES, InChIKey, and molecular formula.

We constructed a feature system incorporating 32 physicochemical descriptors, 2214-dimensional molecular fingerprints (ECFP4 + MACCS), and 30-dimensional spectral features. Baseline evaluation with five classifiers yielded 81.5% accuracy for XGBoost and LightGBM on the eight-category task. All processing scripts, curated datasets, and pre-computed feature matrices are publicly released (CC BY 4.0 for data, MIT for code) and deposited at Science Data Bank (DOI: 10.57760/sciencedb.35414).

2. Data Description

2.1. Dataset Overview

PAID-Strict v1.0 contains 7527 high-quality LC-MS/MS spectra covering 3197 unique compounds (by InChIKey), with an average of 2.35 spectra per compound. This ratio balances per-compound condition diversity characterization with data redundancy control. PAID-Extended v1.0 contains 21,292 spectra covering 3224 unique compounds, retaining more experimental condition variations at the cost of slightly relaxed quality thresholds. Core statistics for PAID-Strict are summarized in Table 1, and a comparison between the two versions is provided in Table 2.

PAID-Strict achieves significantly higher coverage of core molecular identifier fields compared to raw public data sources. InChIKey coverage reaches 100%, as spectra lacking InChIKey were automatically filtered during quality control. SMILES and molecular formula coverage both reach 98.4% (7407/7527 spectra). Instrument-type coverage is 82.3% (6195/7527), with missing values mainly from early community-contributed GNPS entries. Ion mode coverage reaches 99.1%, reflecting the effectiveness of automated ion mode derivation implemented in the processing pipeline.

The molecular weight range spans 100–1500 Da, covering the molecular weight range of the vast majority of pesticide active ingredients. The distribution is unimodal and

right-skewed, with a peak concentration in the 200–500 Da range, consistent with the distribution characteristics of mainstream small-molecule pesticides in the global market.

Table 1. Core statistics of PAID-Strict v1.0.

Statistical Indicator	Value
Total spectra	7527
Unique compounds (InChIKey)	3197
Pesticide functional categories	8
SMILES coverage	98.4%
InChIKey coverage	100%
Molecular formula coverage	98.4%
Instrument-type coverage	82.3%
Ion mode coverage	99.1%
Mean quality score	84.7 +/− 5.2 (100-point scale)
Molecular weight range	100–1500 Da

Table 2. Specifications comparison of PAID-Strict and PAID-Extended.

Specification	PAID-Strict	PAID-Extended
Quality threshold	≥ 70	≥ 50
Max spectra per compound	3	10
Total spectra	7527	21,292
Unique compounds	3197	3224
Mean quality score	84.7	N/A
Use case	High-precision	Data augmentation

PAID-Strict covers two mainstream high-resolution MS platforms: Orbitrap (including Q Exactive, Q-Exactive HF, Orbitrap Fusion sub-models, accounting for ~68% of spectra with instrument annotations) and QTOF (~29%). The coexistence of both platforms enables cross-instrument model generalization assessment.

Positive ion mode (ESI+) spectra account for approximately 75%, negative ion mode (ESI-) for ~25%. This distribution relates to the physicochemical properties of pesticide compounds—most pesticide molecules contain protonatable basic nitrogen atoms, exhibiting higher ionization efficiency in positive mode.

The dataset is split using 8:1:1 InChIKey-stratified partitioning: training set (~6022 spectra, 2558 compounds), validation set (~753 spectra, 320 compounds), and test set (~752 spectra, 319 compounds). The three subsets have completely disjoint InChIKey sets with consistent class proportions across splits.

To assess the extent of structural analog leakage under the InChIKey-stratified split (8:1:1), we computed the maximum pairwise ECFP4 (radius = 2, 2048 bits) Tanimoto similarity between all training-set and test-set compounds. Across the 2294×307 cross-set compound pairs, the median of the per-compound maximum Tanimoto was 0.65, with 45.0% of test compounds having a training-set analog exceeding 0.7 (Figure S1). Upon closer inspection, 53 of the 307 test compounds (17.3%) had a Tanimoto = 1.0 match in the training set—all of which were stereoisomer pairs sharing identical InChIKey connectivity layers but differing in stereochemistry (ECFP4 does not encode stereochemistry). Excluding these stereoisomer pairs, the median drops to 0.61, only 5.5% exceed 0.9, and 33.5% exceed 0.7. At the scaffold level, 176 of 214 test compounds (82.2%) share a Murcko scaffold with at least one training compound—a consequence of the inherent structural homogeneity of pesticide chemical space rather than data leakage. For users requiring strict scaffold-level independence, a Murcko scaffold-based OOD split achieving zero train-test scaffold overlap is provided in the code repository.

2.2. Pesticide Categories

PAID covers eight pesticide functional categories. The three largest categories—plant growth regulators (36.8%), herbicides (23.5%), and fungicides (21.9%)—collectively account for 82.2% of the dataset. Insecticides account for 10.4%, rodenticides for 6.2%, while acaricides (0.5%) and nematocides (<0.1%) are minority classes reflecting their limited availability in public databases. Detailed category statistics are provided in Table 3.

Table 3. Pesticide category statistics in PAID-Strict.

Category	Spectra Count	Proportion
Plant growth regulators	2770	36.8%
Herbicides	1769	23.5%
Fungicides	1650	21.9%
Insecticides	784	10.4%
Rodenticides	469	6.2%
Acaricides	39	0.5%
Nematicides	1	<0.1%
Unclassified	45	0.6%

The category distribution pattern is influenced by: (1) natural category bias in public sources, where plant growth regulators (especially plant hormone analogs) are extensively measured as reference standards in metabolomics studies; (2) the structural diversity of plant growth regulators spanning auxins, gibberellins, cytokinins, abscisic acid analogs, and ethylene precursors, leading to higher spectral accumulation. The PAID-Extended version maintains similar relative category proportions to PAID-Strict, and users requiring category-balanced datasets can perform custom subsampling based on the standardized metadata.

2.3. Data Format

PAID is released in two standardized formats. MGF format (Mascot Generic Format) provides complete spectral data with precursor m/z , charge, retention time metadata, and full m/z -intensity peak lists, compatible with matchms, MS-DIAL, MZmine [2], and OpenMS. TSV format (Tab-Separated Values) provides structured tabular data with spectrum ID, SMILES, InChIKey, molecular formula, instrument type, ion mode, and collision energy metadata fields, directly loadable with pandas. Additionally, pre-computed feature matrices in NPY format (32 physicochemical descriptors + 2214 molecular fingerprints + 30 spectral features) are provided for immediate model training.

2.4. FAIR Compliance

PAID is designed from inception to comply with the FAIR data principles [3]. Findability is ensured through a permanent DOI (10.57760/sciencedb.35414) assigned by Science Data Bank with comprehensive metadata enabling multi-dimensional retrieval. Accessibility is provided via CC BY 4.0 open-source licensing with unrestricted public download. Interoperability is achieved through standard open formats (MGF, TSV) with InChIKey as the universal compound identifier and computable SMILES structural representations. Reusability is supported by pre-computed feature matrices, pre-trained baseline models, complete usage tutorials, API documentation, and MIT-licensed code [4].

2.5. Instrument Platforms and Experimental Conditions

PAID aggregates spectra from diverse instrument platforms and experimental conditions. The instrument-type distribution (among spectra with instrument annotations) is summarized in Table 4. Orbitrap-platform spectra dominate (~68%), primarily from Q Exactive series instruments in MassIVE and Orbitrap data in MoNA. QTOF platforms

contribute ~29%, mainly from early high-resolution data in MassBank and GNPS. Systematic studies on Orbitrap platform optimization for non-targeted pesticide analysis have been reported [5].

Table 4. Instrument platform distribution in PAID-Strict.

Instrument Platform	Spectra Count	Proportion
Orbitrap (Q Exactive series, Orbitrap Fusion, etc.)	~4200	~68%
QTOF (various quadrupole-time-of-flight instruments)	~1800	~29%
Other/unspecified	~195	~3%

Collision energy, the key parameter determining MS/MS fragmentation degree, ranges from 10 to 60% (normalized collision energy, NCE, peaking at 25–40%) and 5–80 eV (absolute, mostly 10–40 eV). GNPS and MassIVE data predominantly use stepped collision energy (e.g., 20–40% NCE), while some MoNA and MassBank spectra use a single fixed collision energy.

The data source contribution distribution is summarized in Table 5. Source diversity is a key strength of PAID—different sources represent varied experimental conditions, sample preparation methods, and instrument configurations, enhancing model robustness and generalization capability.

Table 5. Data source contribution to PAID-Strict.

Original Data Source	Contributed Spectra	Description
MoNA	~3200	General LC-MS spectra after pesticide-directed screening
MassBank	~1500	MassBank reference spectra (MassBank_NIST combined file) after screening
GNPS	~1800	COLLECTIONS-PESTICIDES; published GNPS datasets
MassIVE	~900	High-resolution data from fruit/vegetable matrix samples and standards
Others	~127	Other specialized small-scale datasets

The parsed collision energy is provided as a numeric field (eV) in the TSV distribution, enabling users to filter spectra by CE condition or incorporate CE as a model input feature (see Section 3.7 for the detailed parsing methodology and Figure S2 for the CE value distribution).

2.6. Dataset Quality Assessment

PAID dataset quality is quantitatively assessed across three dimensions. First, meta-data completeness: core molecular identifiers (SMILES, InChIKey, molecular formula) exceed 98% coverage, significantly higher than typical coverage levels of comparable public datasets. Second, quality score distribution: the mean quality score of PAID-Strict is 84.7 ± 5.2 (100-point scale), with over 89% of spectra scoring ≥ 70 . Among 52,410 candidate spectra, only 7527 (~14.4%) were selected for the strict version, reflecting the stringency of quality control. Third, compound isolation verification: InChIKey-based stratified splitting ensures zero compound overlap across train/validation/test sets.

2.7. Baseline Model Performance

Five classification models were evaluated on the eight-category pesticide recognition task. XGBoost and LightGBM achieved the best performance with 81.5% accuracy and 81.1% weighted F1-score. Detailed performance metrics are provided in Table 6. Notably, this is a challenging classification task given the presence of extreme minority classes (acaricides:

39 samples; nematocides: 1 sample), and partial chemical structure overlap between certain categories. The tree ensemble models achieving > 81% accuracy under these conditions validate the feature quality and usability of the dataset. To address class imbalance concerns, Table 7 provides per-class precision, recall, F1-score, and training/test sample counts for all eight categories under the best-performing model (XGBoost). A normalized row-wise confusion matrix (Figure S3) is also provided. The per-class metrics reveal a 13-percentage-point gap between macro-averaged F1 (0.68) and weighted F1 (0.81), quantifying the performance penalty of the long-tailed class distribution. Minority classes with fewer than 35 training samples (acaricide, nematocide) exhibit the lowest F1-scores, while majority classes such as plant growth regulators (n = 1896 training samples) achieve F1 > 0.92. (Nematocide, with only 1 sample in the entire dataset, was excluded from classification model training and evaluation as it could not be stratified into the train/test split.)

Table 6. Pesticide classification model performance.

Model	Accuracy	Weighted F1-Score
XGBoost	0.815	0.811
LightGBM	0.815	0.810
Random Forest	0.794	0.787
SVM	0.690	0.674
Logistic Regression	0.589	0.560

Table 7. Per-class precision, recall, F1-score, and sample counts for the best-performing model (XGBoost) on the eight-category pesticide classification task.

Class	Train N	Test N	Precision	Recall	F1-Score
acaricide	31	8	0.29	0.25	0.27
fungicide	1206	301	0.75	0.80	0.78
herbicide	1321	331	0.79	0.84	0.82
insecticide	614	153	0.71	0.56	0.62
Plant growth regulator	1896	474	0.91	0.93	0.92
rodenticide	331	83	0.80	0.59	0.68
macro avg	5399	1350	0.71	0.66	0.68
weighted avg	5399	1350	0.81	0.81	0.81
accuracy	5399	1350	—	—	0.8148

2.8. Spectrum Similarity Retrieval

Spectrum similarity retrieval is a core task in mass spectrometry data analysis [6]. We computed cosine similarity between spectra based on ECFP4 molecular fingerprints and systematically compared similarity distributions for same-compound (identical InChIKey) versus different-compound spectrum pairs. Results demonstrate that same-compound spectral cosine similarities concentrate in the 0.85–1.0 range, while different-compound similarities primarily distribute in the 0.1–0.4 range, exhibiting a clear separation boundary that quantitatively validates the intra-class consistency and inter-class discriminability of PAID.

3. Methods

3.1. Data Sources

This study integrates four categories of authoritative public mass spectrometry data sources, totaling 15 data files, all publicly accessible under CC BY 4.0 (MassBank [7], GNPS [8], MoNA [9]) or equivalent open-access terms (MassIVE, NIST SRD where applicable). Data sources are summarized in Table 8.

Table 8. Integrated data source inventory.

Repository	Dataset/File Identifier	Format	Size (MB)	Data Type
MoNA	MoNA-export-LC-MS_Spectra	MSP	1105	General LC-MS spectra
MassBank	MassBank_NIST	MSP	111	MassBank + NIST combined reference spectra
GNPS	COLLECTIONS-PESTICIDES (positive/negative)	MGF	3.8	Pesticide-specific spectra
GNPS	HMDB, MSMLS	MGF	2.7	Metabolite reference spectra
MassIVE	MSV000099284 (orange/tomato samples + standards)	MSP/MGF	N/A	Food matrix + standards
MassIVE	MSV000080604/05/16	MGF	N/A	Q Exactive high-resolution data

A detailed inventory of the 15 integrated data files, including repository sources, file identifiers, formats, sizes, and data types, is provided in Table 7.

3.2. Experimental Environment

All processing is based on the Python 3.8+ ecosystem. Core dependencies include: *matchms* $\geq 0.21.0$ (Netherlands eScience Center, Amsterdam, The Netherlands) for mass spectrometry data processing [10], *RDKit* $\geq 2023.03.1$ (open-source, <http://www.rdkit.org/>) for chemoinformatics [11,12], leveraging compound reference data from PubChem [13], *scikit-learn* $\geq 1.2.2$ (INRIA, Saclay, France) for machine learning [14], *XGBoost* $\geq 1.7.5$ (University of Washington, Seattle, WA, USA) [15], and *LightGBM* $\geq 3.3.5$ (Microsoft Research, Redmond, WA, USA) [16]. The complete environment configuration file is available in the open-source code repository.

3.3. Workflow Overview

We designed a seven-step modular data-processing workflow (Figure 2), with each step implementing explicit quality-control gating conditions. The entire pipeline is parameterized and fully automated. The workflow design references FAIR practices established by the reproducible Metabolome Annotation Workflow (MAW) [17]. A brief overview of each step is provided below, followed by detailed descriptions of key steps.

3.4. Step 1: Multi-Source Data Integration

This step unifies 15 heterogeneous data files into a standardized in-memory representation. Core processing includes (1) multi-format unified parsing—MSP and MGF formats are parsed via *matchms*, with tolerance for non-standard source files; (2) meta-data field standardization—a field name mapping table resolves naming conflicts across data sources (e.g., “PrecursorMZ”, “Precursor_MZ”, “precursormz” all map to the same attribute); (3) data cleaning—invalid charge values (zero or out-of-range) are filtered. The innovation lies in establishing a unified field mapping and format parsing framework across 15 heterogeneous data sources, resolving naming discrepancies into a standardized internal representation.

After merging, 91,420 raw spectra were obtained. Source distribution and metadata completeness are shown in Figures 3 and 4, respectively.



Figure 2. End-to-end construction workflow of the pesticide LC-MS/MS dataset (Steps 1–7), showing input/output data volumes and quality-control checkpoints at each step.

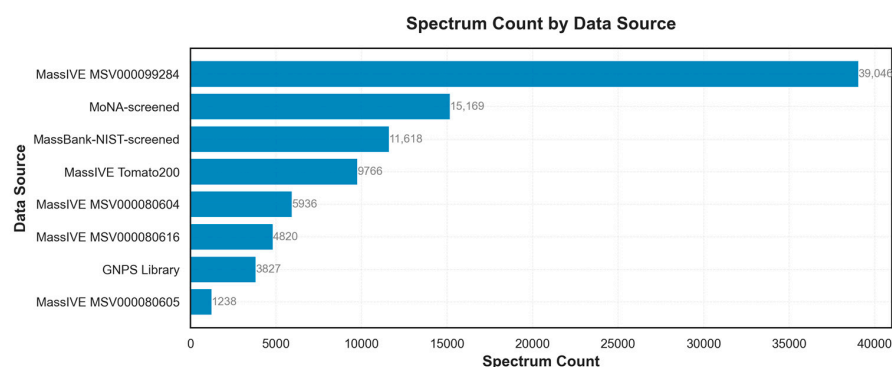


Figure 3. Spectral contribution distribution across 15 data sources (91,420 total raw spectra).

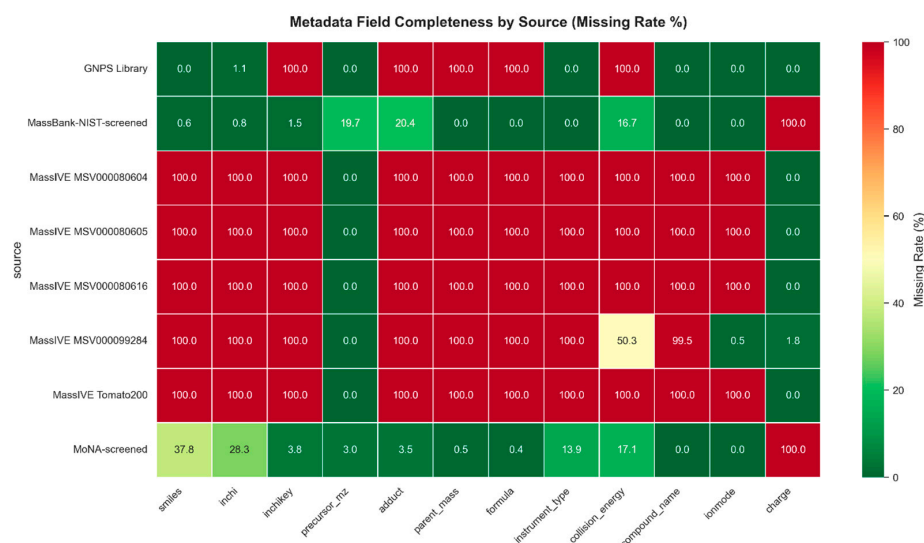


Figure 4. Core metadata field completeness heatmap by data source (after Step 1 merging).

3.5. Step 2: Spectrum Cleaning

This step ensures spectral data correctness through a composable, standardized cleaning pipeline built on matchms. Five sub-steps are executed sequentially: (1) structural identifier consistency repair—cross-validating SMILES, InChI, and InChIKey using RDKit, correcting inconsistencies with SMILES as the trusted source; (2) ion mode derivation and adduct correction—automatically deriving precursor ion types from meta-data and correcting adduct annotations based on common adduct rules; (3) precursor m/z annotation—calculating precursor m/z for spectra missing this field; (4) peak list standardization—normalizing m/z and intensity values, filtering zero-intensity and out-of-range peaks; (5) GC-MS data removal—identifying and removing all GC-MS spectra based on instrument-type metadata and spectral characteristics, retaining only LC-MS/MS data.

The core innovation of this step is the composable pipeline architecture that modularizes structural verification, ion derivation, and peak filtering operations. Each module can be independently replaced or extended, providing a flexible framework for custom cleaning workflows across different data sources.

After cleaning, the data volume decreased from 91,420 to 56,550 spectra (~38% filtered due to structural inconsistencies, non-LC-MS data, or peak list anomalies). The per-filter removal statistics are shown in Figure 5.

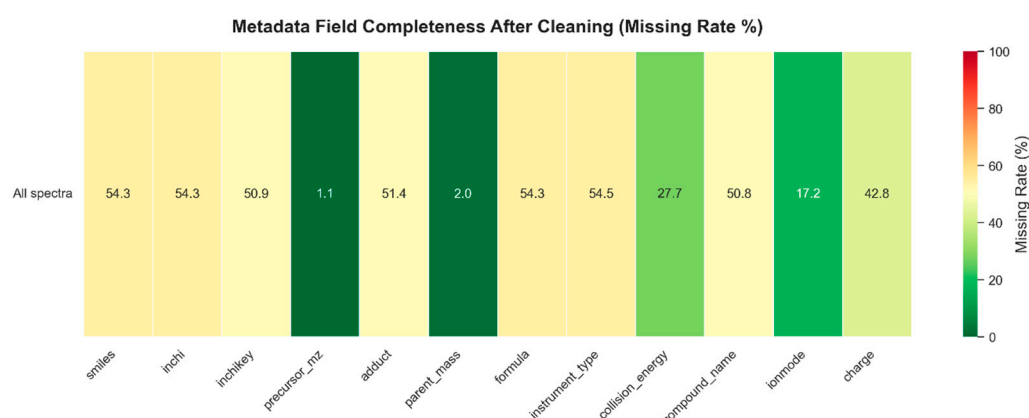


Figure 5. Step 2: Spectrum cleaning data flow diagram—spectrum removal counts per filtering sub-step.

3.6. Step 3: Deduplication and Quality Filtering

Step 2 output still contains substantial redundancy and quality anomalies. Three-dimensional filtering is applied: (1) Profile mode removal—automatically identifying and removing profile-mode spectra based on peak density and shape characteristics, retaining only centroid-mode data. (2) Abnormal peak count filtering—removing spectra with >300 peaks (typically profile-mode residuals or high-noise fusion data). (3) CosineGreedy deduplication—within each InChIKey group, pairwise cosine similarity is computed using the CosineGreedy algorithm. Spectrum pairs with similarity ≥ 0.99 are judged as complete duplicates, retaining only one. This strategy removes identical records while preserving spectral diversity from different experimental conditions (varying collision energies, different instruments).

After deduplication, 52,410 spectra remained. The innovation lies in the use of the CosineGreedy algorithm for within-compound spectrum deduplication—achieving a balance between redundancy control and diversity preservation.

3.7. Step 4: Metadata Standardization

This step standardizes metadata for retained spectra: (1) SMILES regularization—all SMILES are converted to canonical aromatic form using RDKit, ensuring identical com-

pounds map to the same SMILES string; (2) instrument-type unified classification—a classification hierarchy is established: Fourier transform instruments (ITFT, QFT) are unified as Orbitrap; QTOF remains independently classified. (3) Collision energy parsing—structured numeric fields are derived from diverse source formats (numeric, percentage NCE, string descriptions). (4) Automatic molecular formula derivation using RDKit from SMILES. The innovation lies in establishing an instrument-type classification hierarchy and collision-energy standardization rules, providing the foundation for cross-instrument and cross-energy model generalization assessment. In detail, the pipeline handles four CE format categories via regular expression matching: (i) plain numeric values (e.g., ‘35’); (ii) values with explicit unit labels (e.g., ‘35 eV’); (iii) normalized collision energy indicated by a percent sign (e.g., ‘35%’); and (iv) ramped or stepped CE (e.g., ‘20–40’, ‘Ramp 20–40’, ‘20 ≥ 40’). For ramped CE, the midpoint (min + max)/2 is recorded. NCE values are converted to approximate absolute CE using the formula $ACE = NCE \times \text{precursor_m/z} / 500$, following the standard linear relationship for Orbitrap-type instruments. The parsed numeric CE value (eV) is provided as the collision_energy column in the distributed TSV, enabling users to filter spectra by CE condition or to incorporate CE as a model input feature. The distribution of parsed CE values (Figure S2) shows a median of 40.0 eV (IQR 20–60 eV), consistent with typical LC-MS/MS pesticide acquisition parameters. For QTOF platforms, collision energy is typically reported as absolute eV values (formats i and ii), while Orbitrap platforms commonly use normalized collision energy (NCE) percentages (format iii), which are subsequently converted via the ACE formula.

3.8. Step 5: Quality Scoring and Dataset Export

This is the core step determining final dataset quality. Three-tier quality control is implemented: (1) 100-point quality scoring—a four-dimensional weighted scoring system (Table 9) quantitatively evaluates each spectrum: metadata completeness (40%), spectral quality (30%), instrument metadata (15%), and structural consistency (15%). (2) InChIKey deduplication—within each InChIKey group, spectra are ranked by quality score, retaining the top N (PAID-Strict: N = 3; PAID-Extended: N = 10). (3) Stratified split and export—the dataset is partitioned into train/validation/test sets (8:1:1) by InChIKey stratification, ensuring all spectra of the same compound are assigned to the same split, eliminating compound leakage. Both MGF and TSV formats are exported.

Table 9. Spectrum quality scoring system.

Evaluation Dimension	Weight	Scoring Rule
Metadata completeness	40%	SMILES + InChIKey + molecular formula, 13.3 pts each
Spectral quality	30%	Full score for 20–100 peaks; 20 pts for 10–19 or 101–200 peaks
Instrument metadata	15%	Instrument type + ion mode + collision energy
Structural consistency	15%	Precursor m/z within a reasonable range (50–2000 Da)

The core innovations of this step are (1) the four-dimensional weighted quality scoring system that integrates metadata completeness, spectral quality, instrument metadata, and structural consistency into a single comparable score; (2) the dual-version export strategy (PAID-Strict and PAID-Extended) that adjusts quality thresholds and deduplication aggressiveness to meet different application requirements.

Quality thresholds were determined through iterative experiments: a 70-point threshold for the strict version ensures high usability while retaining sufficient diversity, with

an average quality score of 84.7 and 89% of spectra scoring ≥ 70 . The overall processing funnel is shown in Figure 6, and the quality dashboard in Figure 7.

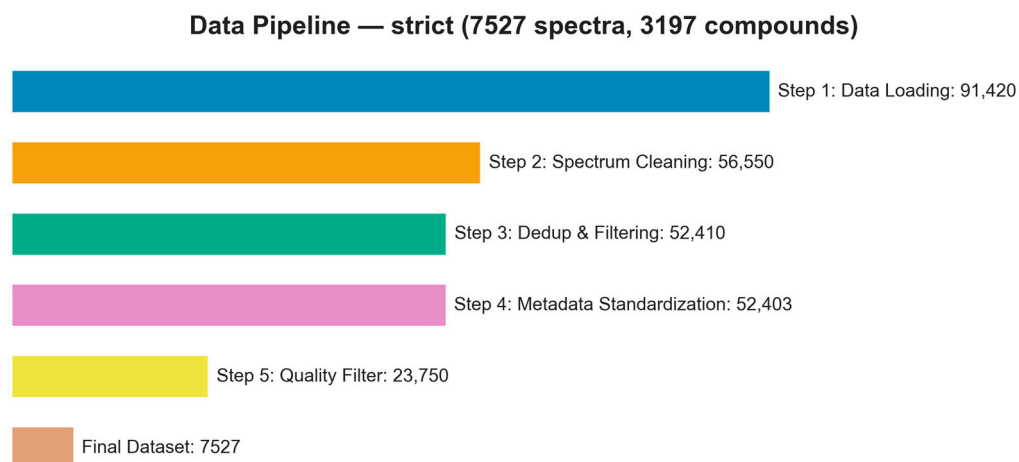


Figure 6. Full processing pipeline funnel chart (Steps 1–5), illustrating the stepwise filtering process from 91,420 raw spectra to 7527 PAID-Strict spectra.

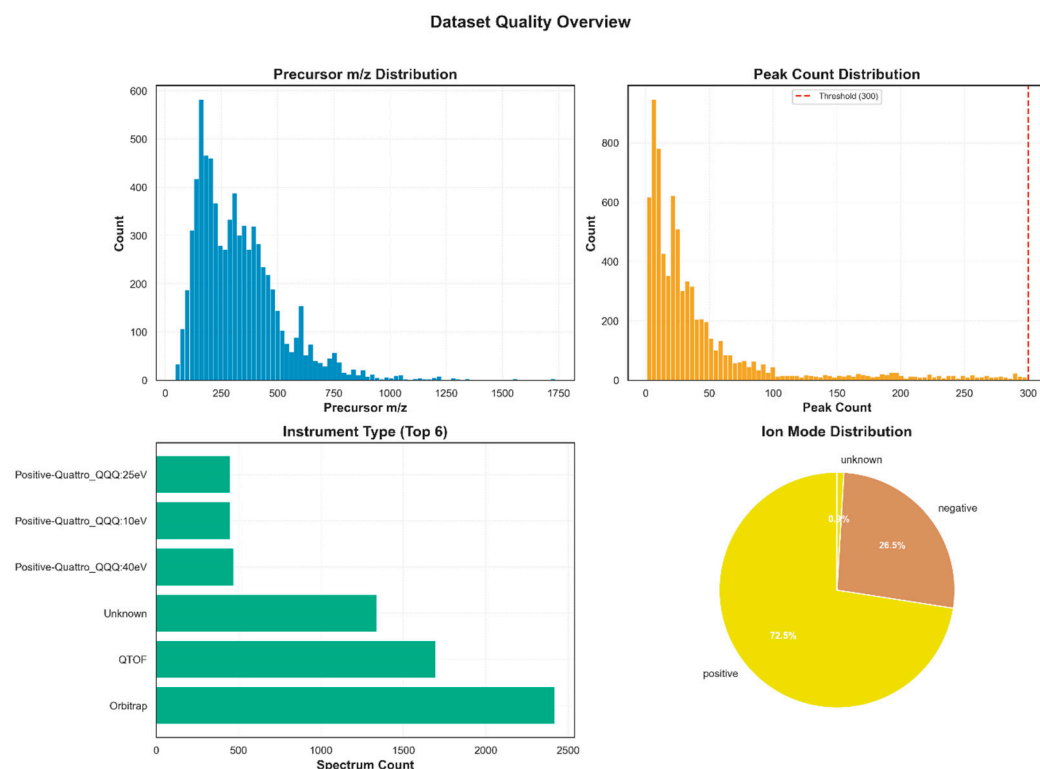


Figure 7. PAID-Strict dataset quality overview dashboard: quality score distribution, metadata completeness statistics, molecular weight distribution, and peak count distribution.

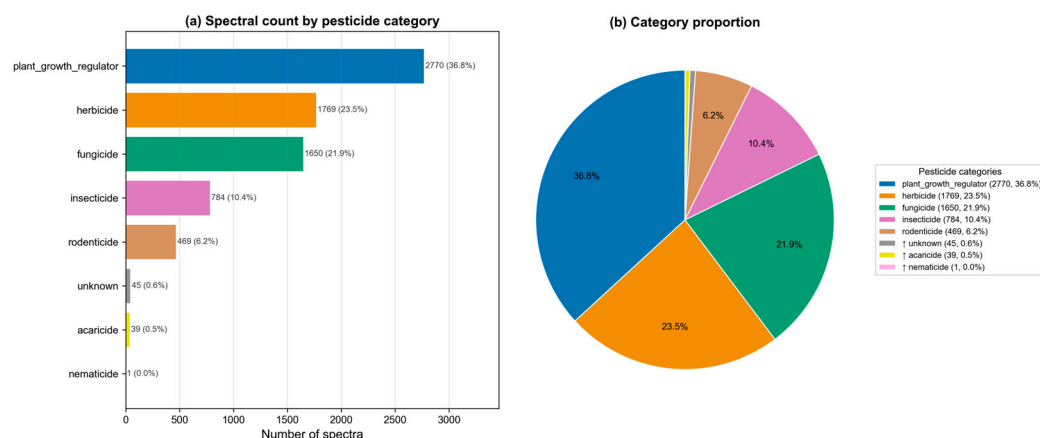
3.9. Step 6: Multi-Dimensional Feature Engineering

This step achieves AI-readiness—users can directly use the dataset for machine learning model training without secondary preprocessing. The innovation is the three-tier cascaded feature system [molecular descriptors—molecular fingerprints—spectral features] that systematically integrates chemical prior knowledge with experimental observation data. The feature system is summarized in Table 10.

Table 10. Dataset feature dictionary.

Feature Category	Dimensions	Tool	Representative Features
Physicochemical descriptors	32	RDKit	MolWt, LogP, TPSA, NumHAcceptors, NumHDonors, RingCount
ECFP4 fingerprints	2048	RDKit	Morgan fingerprints (radius = 2)
MACCS fingerprints	166	RDKit	166 predefined structural keys
Spectral statistical features	8	matchms/custom	Peak count, m/z mean/SD, intensity mean/SD
Spectral distribution features	4	custom	Intensity skewness/kurtosis, m/z skewness/kurtosis
Spectral quality features	8	custom	Base peak ratio, top-10 peak concentration, S/N estimate
Other features	10	custom	Precursor m/z related features

- (1) Molecular descriptor computation (32 dimensions): basic physicochemical properties (molecular weight, LogP, TPSA, hydrogen bond donor/acceptor counts, rotatable bond count, ring count), topological indices (BertzCT, Chi connectivity indices), and shape descriptors, computed via RDKit.
- (2) Molecular fingerprint generation (2214 dimensions): a dual-representation strategy employing ECFP4 fingerprints (2048 dimensions, radius-2 Morgan circular substructure fingerprints) and MACCS fingerprints (166 dimensions, 166 predefined structural keys). ECFP4 provides high structural resolution, capturing subtle differences in local chemical environments; MACCS fingerprints offer inherent interpretability for analyzing model decision rationales.
- (3) Spectral feature extraction (30 dimensions): basic statistical features, distribution features (skewness/kurtosis), and signal quality features (base peak ratio, top-10 peak concentration, signal-to-noise ratio estimation). Spectral features complement molecular structure representations by capturing fragmentation pattern specificity from the experimental observation dimension.
- (4) Pesticide category annotation: combining RDKit SMARTS substructure matching with compound name keyword matching to automatically annotate spectra into seven functional pesticide categories plus an unclassified category. The category distribution overview is shown in Figure 8.

**Figure 8.** PAID-Strict pesticide functional category distribution. (a) Spectral count bar chart by category; (b) Category proportion pie chart.

3.10. Step 7: Baseline Model Validation

To quantify PAID dataset usability in AI training, we established pesticide functional category recognition as a benchmark task and systematically evaluated five classical classification models. Core innovations include: (1) an InChIKey-stratified splitting strategy that definitively eliminates compound leakage—ensuring all spectra of the same compound

appear exclusively in one of the train/validation/test subsets; (2) a multi-tier benchmark framework spanning from linear models (Logistic Regression) to tree ensemble models (XGBoost, LightGBM), providing standardized performance references for downstream model selection at varying complexity levels.

Experimental setup: Input features concatenate ECFP4 fingerprints (2048 dimensions) with spectral features (30 dimensions) into a 2078-dimensional feature vector; prediction target is eight pesticide functional categories; evaluation metrics are accuracy and weighted F1-score; all experiments use a fixed random seed (random_state = 42) for full reproducibility.

Models evaluated include Random Forest, XGBoost, LightGBM, SVM, and Logistic Regression, covering the complete complexity spectrum from linear classifiers to tree ensemble models. All hyperparameters were optimized via GridSearchCV under 5-fold cross-validation; complete hyperparameter configurations are provided in Supplementary Material Table S1. A stratified 5-fold cross-validation strategy was employed with class proportions maintained in each fold.

Results overview: Tree ensemble models (XGBoost, LightGBM, Random Forest) significantly outperformed SVM and Logistic Regression. XGBoost and LightGBM both achieved 81.5% accuracy. Model performance comparison is shown in Figure 9. The baseline models were intentionally trained without class weighting or oversampling, establishing a reference performance that faithfully reflects the native class distribution of PAID. We further evaluated a balanced variant of XGBoost using sample_weight = 'balanced'; while the acaricide F1-score improved from 0.27 to 0.32 (+18.5%), the overall accuracy slightly decreased from 81.5% to 80.5%, illustrating the inherent accuracy–fairness trade-off of class re-weighting under extreme sample scarcity (acaricide: 31 training samples). Users targeting minority classes may enable class_weight = 'balanced' in scikit-learn classifiers, scale_pos_weight in XGBoost/LightGBM, or apply SMOTE oversampling, as discussed in Section 4.2.

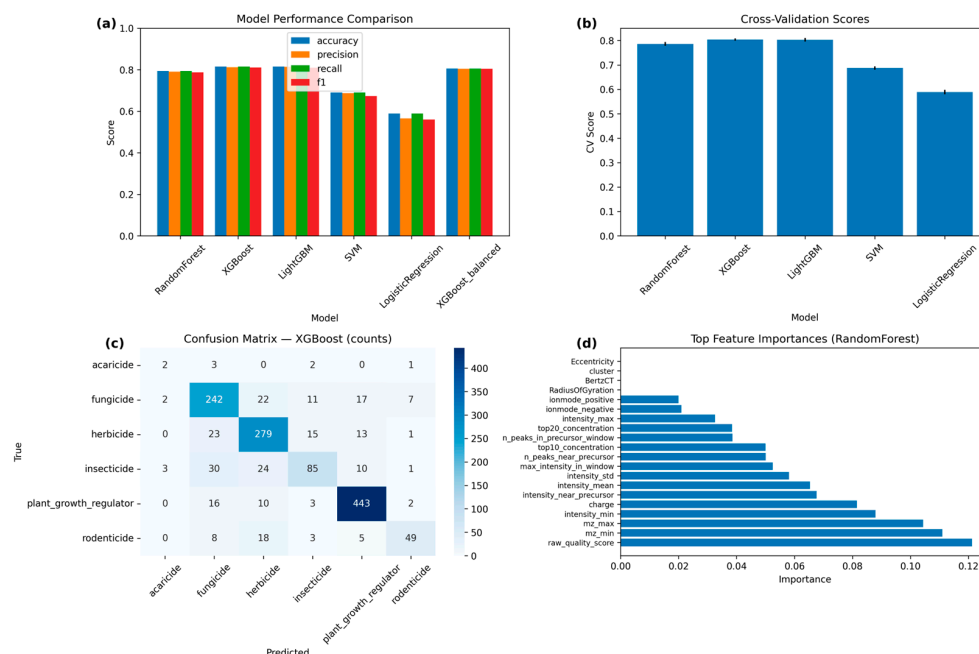


Figure 9. Performance comparison of five classification models on the eight-category pesticide recognition task. (a) Accuracy, precision, recall, and weighted F1-score for each model; (b) cross-validation scores with standard deviation; (c) confusion matrix of the best-performing model (XGBoost); (d) top 20 feature importances from Random Forest.

4. Discussion

4.1. Application Scenarios

PAID can be widely applied to the following scenarios: (1) Food safety testing—AI-driven rapid screening of pesticide residues in fruits, vegetables, and grains [18]. (2) Environmental monitoring—rapid identification and source tracing of pesticide contaminants in water and soil. (3) Method development—serving as a standardized benchmark dataset for mass spectrometry data analysis algorithms (spectral deconvolution, peak detection, retention time prediction). (4) AI research—training and evaluation of cutting-edge models for molecular structure prediction and spectrum-structure translation. (5) Education—serving as a teaching dataset at the intersection of analytical chemistry and chemoinformatics. Deep learning combined with portable mass spectrometry for on-site rapid detection has been preliminarily validated [19,20].

Beyond the baseline classification tasks demonstrated in Section 2.7, PAID supports a broad range of analytical workflows. (i) Machine learning-based pesticide annotation: users can train custom classifiers or deep learning models on the 2214-dimensional fingerprint + 30-dimensional spectral feature matrix to achieve task-specific compound identification. (ii) Suspect screening: the comprehensive metadata (SMILES, InChIKey, formula, precursor m/z) enables direct integration with suspect lists in tools such as MS-DIAL or MZmine, facilitating high-throughput screening of known pesticides in food and environmental samples. (iii) Non-target analysis (NTA): spectrum similarity retrieval using CosineGreedy on the standardized MGF library supports feature annotation in NTA workflows where authentic standards are unavailable. (iv) Confidence improvement in compound identification: the quality-scored reference spectra (0–100 scale) can serve as a gold-standard library to improve identification confidence levels (Level 1–5) in accordance with the Metabolomics Standards Initiative (MSI) guidelines [21]. All workflows are supported by the open-source code and pre-trained models released alongside PAID.

4.2. Usage Guide

The code repository (<https://github.com/jijic137/paid-pesticide-ms> (accessed on 1 April 2025)) provides complete usage tutorials and API examples covering the following typical workflows: (1) Dataset loading and exploration—loading MGF spectra via `matchms` and TSV metadata via `pandas` for rapid understanding of dataset structure, category distribution, and metadata statistics. (2) Feature extraction and model training—using pre-computed feature matrices (NPY format) directly as input with `scikit-learn`, `XGBoost`, or `LightGBM` for rapid classification, regression, or clustering model construction. (3) Spectrum similarity retrieval—based on ECFP4 molecular fingerprints or spectral cosine similarity for fast matching and retrieval of query spectra against the reference dataset. (4) Custom model training—users can seamlessly integrate their own model code and training pipelines. (5) Statistical visualization—visualization scripts for category distribution, quality score distribution, molecular property distribution, and dimensionality reduction clustering are provided.

4.3. Limitations and Future Work

Despite PAID achieving high levels of scale, quality, and standardization, the following limitations remain for continuous improvement in subsequent versions:

- (1) Class imbalance: Plant growth regulators (36.8%) and herbicides (23.5%) dominate, while acaricides (0.5%) and nematicides (<0.1%) are severely underrepresented. This long-tailed distribution reflects real-world pesticide market structure but leads to significantly lower classification performance for minority classes. Future work will address this through targeted collection of real environmental sample spectra, data augmentation techniques, and few-shot learning methods.

- (2) **Limited matrix coverage:** The current dataset is predominantly composed of pure standard spectra, with a low proportion of real sample spectra collected from food/environmental matrices. Matrix effects on ionization efficiency and fragmentation patterns are not adequately characterized. Future versions will prioritize supplementation with spiked sample spectra from typical matrices, including fruits, vegetables, tea, soil, and water. Specifically, matrix-induced ion suppression or enhancement can alter peak intensity distributions, and co-eluting matrix interferences can introduce extraneous peaks, creating a distribution shift between training (standard spectra) and deployment (real-sample extracts). This may lead to reduced classification confidence, lower recall for minority classes (e.g., acaricide F1 dropping below 0.2), and increased false positive rates. Practical mitigation strategies include: (a) transfer learning by fine-tuning on a small set of matrix-annotated spectra; (b) intensity-robust spectral features such as binarized peak presence/absence or rank-based top-N peak encoding; and (c) domain adaptation techniques. Incorporating matrix-rich spectra remains a high priority for PAID v2.
- (3) **Instrument-type limitation:** The current dataset primarily covers high-resolution MS platforms (QTOF and Orbitrap), lacking low-resolution triple quadrupole (QQQ) mass spectrometry data widely used in routine testing laboratories. Future work will incorporate QQQ platform data through public data curation or collaboration with reference laboratories.
- (4) **Feature system extensibility:** Current feature engineering is based on RDKit molecular descriptors and traditional molecular fingerprints. Recent advances in deep learning-based spectral representation learning, including graph neural networks (GrAFF-MS [22], QC-GN2oMS2 [23], MoMS-Net [24]) and Graph Transformer architectures (MassFormer [25]) for mass spectrum prediction, have demonstrated significant improvements. Future versions will integrate pre-trained spectral embedding features to construct a hybrid feature system combining chemical prior knowledge with data-driven representations [26].
- (5) **Scaffold-level analog leakage:** The InChIKey-stratified split guarantees zero compound leakage, but does not explicitly control for scaffold-level structural similarity between splits. Under the original split, 176 of 214 test compounds (82.2%) share a Murcko scaffold with at least one training compound, reflecting the inherent structural homogeneity of pesticide chemical space (triazoles, neonicotinoids, etc., sharing core scaffolds). Cross-set ECFP4 Tanimoto analysis confirms this: median max similarity 0.65 (0.61 excluding 53 stereoisomer pairs where ECFP4 does not encode stereochemistry), with only 5.5% exceeding 0.9. While limited, users requiring strict scaffold independence may use the Murcko scaffold-based OOD split (`ood_scaffold_split.py`) provided in the code repository, which achieves zero train-test scaffold overlap.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/data11070177/s1>, Figure S1: Distribution of maximum pairwise ECFP4 Tanimoto similarity between training-set and test-set compounds (InChIKey-stratified split); Figure S2: Distribution of parsed collision energy values across PAID-Strict spectra; Figure S3: Normalized confusion matrix (row-wise) for XGBoost on the eight-category pesticide classification task; Table S1: Hyperparameter configurations for all baseline classification models.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z.; software, J.Z. and Q.L.; validation, Q.L. and J.Z.; formal analysis, J.Z.; investigation, L.G. and H.Z.; resources, H.Z. and L.G.; data curation, J.Z. and H.Z.; writing—original draft preparation, J.Z.; writing—review and editing, X.X. and J.Z.; visualization, J.Z.; supervision, X.X.; project administration, X.X.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Resources Investigation Program (Grant No. 2022FY101200).

Data Availability Statement: The PAID dataset is publicly hosted at Science Data Bank (DOI: 10.57760/sciencedb.35414) under a CC BY 4.0 license. Both PAID-Strict (7527 spectra) and PAID-Extended (21,292 spectra) versions are provided, each including (1) finalized spectral files (MGF); (2) structured metadata and feature matrices (TSV/NPY). Code Availability: All processing code is hosted at GitHub (<https://github.com/jijic137/paid-pesticide-ms> (accessed on 1 April 2025)) under the MIT license, including eight modular Python scripts for each processing step, environment configuration files, and usage tutorials.

Acknowledgments: We thank the GNPS, MassIVE, MoNA, and MassBank consortia for providing public mass spectrometry data, and the matchms and RDKit open-source communities for providing core chemoinformatics tools.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fiehn, O. Metabolomics—The link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–169. [CrossRef] [PubMed]
2. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, 395. [CrossRef]
3. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
4. Du, X.; Dastmalchi, F.; Ye, H.; Garrett, T.J.; Diller, M.A.; Liu, M.; Hogan, W.R.; Brochhausen, M.; Lemas, D.J. Evaluating LC-HRMS metabolomics data processing software using FAIR principles for research software. *Metabolomics* **2023**, *19*, 11. [CrossRef] [PubMed]
5. Jimenez-Carvelo, A.M.; Lopez-Ruiz, R.; Romero-Gonzalez, R.; Garrido-Frenich, A. Critical evaluation of MS acquisition conditions and identification process in LC-Q-Orbitrap-MS for non-targeted analysis: Pesticide residues as case of study. *Anal. Bioanal. Chem.* **2023**, *415*, 6269–6277. [CrossRef] [PubMed]
6. Wang, M.; Jarmusch, A.K.; Vargas, F.; Aksenov, A.A.; Gauglitz, J.M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A.V.; et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **2020**, *38*, 23–26. [CrossRef] [PubMed]
7. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714. [CrossRef] [PubMed]
8. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kaponov, C.A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837. [CrossRef] [PubMed]
9. MoNA: MassBank of North America. Available online: <https://mona.fiehnlab.ucdavis.edu/> (accessed on 1 April 2025).
10. Huber, F.; Verhoeven, S.; Meijer, C.; Spreeuw, H.; Castilla, E.M.V.; Geng, C.; van der Hooft, J.J.J.; Rogers, S.; Belloum, A.; Diblen, F.; et al. Matchms—Processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* **2020**, *5*, 2411. [CrossRef]
11. RDKit: Open-Source Cheminformatics Software. Available online: <https://www.rdkit.org/> (accessed on 1 April 2025).
12. Landrum, G. RDKit: A Software Suite for Chemoinformatics, Computational Chemistry, and Predictive Modeling. 2013. Available online: http://www.rdkit.org/RDKit_Overview.pdf (accessed on 1 April 2025).
13. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380. [PubMed]
14. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
15. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
16. 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. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; Volume 30, pp. 3146–3154.
17. Zulfiqar, M.; Gadelha, L.; Steinbeck, C.; Sorokina, M.; Peters, K. MAW: The reproducible Metabolome Annotation Workflow for untargeted tandem mass spectrometry. *J. Cheminform.* **2023**, *15*, 32. [CrossRef] [PubMed]

18. Chen, T.; Liang, W.; Zhang, X.; Wang, Y.; Lu, X.; Zhang, Y.; Zhang, Z.; You, L.; Liu, X.; Zhao, C.; et al. Screening and identification of unknown chemical contaminants in food based on liquid chromatography-high-resolution mass spectrometry and machine learning. *Anal. Chim. Acta* **2024**, *1287*, 342116. [[PubMed](#)]
19. Zhou, Y.; Ai, J.; Cai, L.; Yan, Y.; Wang, B.; Ma, H.; Yu, Q.; Zhou, J.; Huo, X. Deep learning enabled miniature mass spectrometer for rapid qualitative and quantitative analysis of pesticides on vegetable surfaces. *Food Chem. Toxicol.* **2023**, *180*, 114023. [[CrossRef](#)]
20. Taheri, S.; de Andrade, J.C.; Conte-Junior, C.A. Emerging perspectives on analytical techniques and machine learning for food metabolomics in the era of industry 4.0: A systematic review. *Crit. Rev. Food Sci. Nutr.* **2025**, *65*, 6045–6071. [[PubMed](#)]
21. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A.A.; Melnik, A.V.; Meusel, M.; Dorrestein, P.C.; Rousu, J.; Böcker, S. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16*, 299–302. [[CrossRef](#)] [[PubMed](#)]
22. Murphy, M.; Jegelka, S.; Fraenkel, E.; Kind, T.; Healey, D.; Butler, T. Efficiently predicting high resolution mass spectra with graph neural networks. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 25549–25562.
23. Overstreet, R.; King, E.; Clopton, G.; Nguyen, J.; Ciesielski, D. QC-GN2oMS2: A graph neural net for high resolution mass spectra prediction. *J. Chem. Inf. Model.* **2024**, *64*, 5806–5816. [[CrossRef](#)] [[PubMed](#)]
24. Park, J.; Jo, J.; Yoon, S. Mass spectra prediction with structural motif-based graph neural networks. *Sci. Rep.* **2024**, *14*, 1400. [[CrossRef](#)] [[PubMed](#)]
25. Young, A.; Rost, H.; Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat. Mach. Intell.* **2024**, *6*, 404–416. [[CrossRef](#)]
26. Hupatz, H.; Rahu, I.; Wang, W.C.; Peets, P.; Palm, E.H.; Krueve, A. Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening. *Anal. Bioanal. Chem.* **2024**, *417*, 473–493. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.