Predicting the ClassyFire Chemical Superclass of a Compound from its Tandem Mass Spectrum Alone

A Compound-Level Machine-Learning Study on the MassBank of North America (MoNA) Positive-Mode $[M+H]^+$ Library

July 12, 2026

Abstract

Tandem mass spectrometry (MS/MS) is the workhorse of untargeted metabolomics, yet the overwhelming majority of acquired fragmentation spectra remain structurally unannotated. A pragmatic intermediate goal is to assign each spectrum to a broad *chemical superclass* rather than a single structure. We ask a deliberately narrow question: how well can the ClassyFire chemical superclass of a compound be predicted from its positive-ion MS/MS fragmentation pattern *alone*, using only classical, fully reproducible machine learning? We assembled an experimental MS/MS dataset from the MassBank of North America (MoNA) "LC-MS/MS Positive Mode" export (export generated **2026-07-06**, downloaded 2026-07-12), retaining only tandem spectra acquired in positive mode with an $[M + H]^+$ precursor, a syntactically valid InChIKey, and a ClassyFire superclass annotation. Filtering 102 580 records yielded 20 403 **spectra** spanning 3695 **unique compounds** and **16 superclasses**, over a parsed collision-energy range of 0 eV to 80 eV. Spectra were noise-filtered (relative intensity $< 1\%$ of base peak removed), base-peak normalized, binned onto a 1 Da m/z grid over $[50, 1000)$ Da (950 bins), and TF-IDF weighted (fit on training data only). Compounds were split 80/20 by molecular skeleton (seed **42**), guaranteeing zero compound leakage across the split. An Extremely Randomized Trees classifier was selected by grouped cross-validation and tuned, then spectrum-level probabilities were averaged to a single per-compound prediction. On the 739 held-out test compounds the model attains an **overall accuracy of 0.503** and a **macro-averaged F_1 of 0.373** across superclasses—roughly 2.2× the majority-class baseline (0.233). Performance is strongly class dependent: phenylpropanoids/polyketides, lipids, and organic acids are recognized well ($F_1 \geq 0.55$), whereas several rare superclasses are never recovered. The two most frequently confused pairs, *Benzenoids*↔*Organoheterocyclic compounds* (50 misclassifications) and *Organic acids and derivatives*↔*Organoheterocyclic compounds* (31), are chemically adjacent, indicating that the classifier's errors are structurally sensible rather than arbitrary. We provide the complete per-compound test-set predictions. The study establishes a transparent, leakage-controlled baseline against which more sophisticated spectral models can be measured.

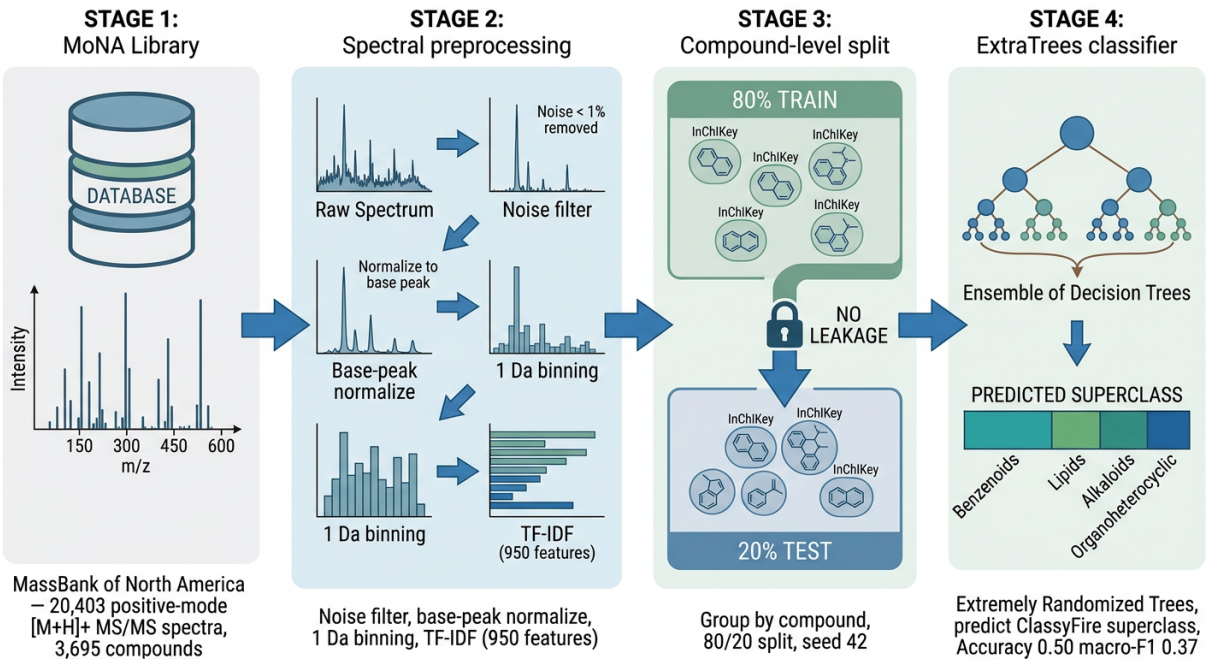


Figure 1: Graphical abstract. End-to-end workflow: (1) experimental positive-mode $[M + H]^+$ MS/MS spectra are retrieved from the MassBank of North America (MoNA); (2) each spectrum is noise-filtered, base-peak normalized, binned onto a 1 Da m/z grid, and TF-IDF weighted into a 950-dimensional feature vector; (3) compounds are split 80/20 by molecular skeleton (InChIKey), so no compound appears in both partitions (seed 42); (4) an Extremely Randomized Trees ensemble predicts the ClassyFire superclass, reaching accuracy 0.50 and macro-F_1 0.37 on held-out compounds.

1 Introduction

1.1 Small-molecule identification and the annotation bottleneck

Untargeted metabolomics and natural-product discovery both rest on tandem mass spectrometry (MS/MS), in which a precursor ion is isolated, fragmented, and its product ions recorded to produce a chemical "fingerprint" of the molecule. Despite decades of methodological progress, the single largest obstacle in the field is not data acquisition but data interpretation: in a typical liquid-chromatography MS/MS experiment only a small percentage of detected features can be confidently matched to a known structure, leaving the remainder as chemical "dark matter" [1]. Even repository-scale re-analysis of hundreds of millions of public spectra annotates only on the order of 5 % to 10 % of features by direct library search [2]. This gap is fundamental: the number of biologically and environmentally relevant small molecules vastly exceeds the number of reference spectra held in curated libraries, and rigorous statistical control is required merely to keep the false-discovery rate of library matches in check [3].

Because exact structural identification is so often unattainable, a pragmatic and increasingly popular alternative is *compound-class* annotation: rather than naming a molecule, one assigns it to a chemically meaningful category such as "lipids and lipid-like molecules" or "alkaloids and derivatives." Class-level information is frequently sufficient to interpret a metabolomic phenotype, to prioritise candidate features for follow-up, or to organise a molecular network into chemically coherent families. The landmark demonstration that compound classes can be predicted directly from fragmentation spectra—CANOPUS [4]—showed that class-level annotation remains possible even for molecules absent from every structural database, and it has become a standard component of untargeted workflows.

1.2 ClassyFire and the chemical superclass

To speak of a compound's "class" at all requires a controlled vocabulary. ClassyFire provides exactly this: a rule-based classifier that maps any chemical structure (supplied as InChIKey or SMILES) onto ChemOnt, a purely structure-based taxonomy of more than 4800 categories arranged in up to eleven hierarchical levels [5]. The taxonomy descends from broad *kingdoms* (organic vs. inorganic) through *superclass*, *class*, *subclass*, and finer levels down to the "direct parent" of a molecule (Figure 2). The *superclass* level—containing categories such as *Benzenoids*, *Lipids and lipid-like molecules*, *Organoheterocyclic compounds*, *Organic acids and derivatives*, *Alkaloids and derivatives*, and *Phenylpropanoids and polyketides*—is coarse enough to be well populated in reference libraries yet specific enough to be chemically informative. It is therefore an attractive prediction target. Crucially, ClassyFire assignments are deterministic functions of molecular structure, so a compound with a known InChIKey can be given a ground-truth superclass label without any manual curation, enabling supervised learning at scale.

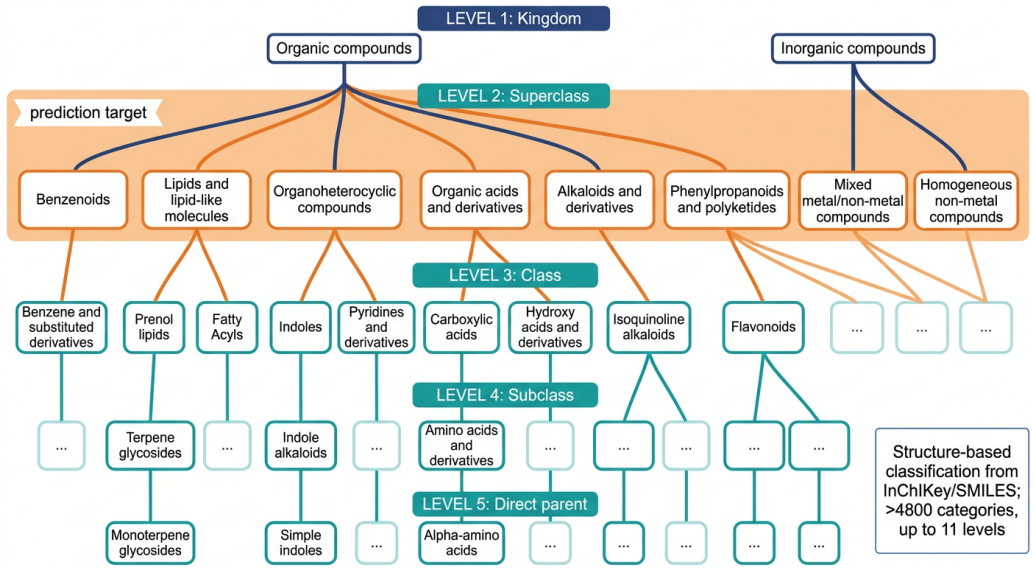


Figure 2: The ClassyFire (ChemOnt) chemical taxonomy. Structures are classified hierarchically from Kingdom down to the direct parent. This study targets the *superclass* level (highlighted), which offers a favourable trade-off between chemical specificity and per-class sample size. Assignments are computed from structure (InChIKey/SMILES), providing noise-free ground-truth labels for supervised learning [5].

Reliable structure identifiers are the second prerequisite. The IUPAC International Chemical Identifier (InChI) and its fixed-length hashed form, the InChIKey, provide a canonical, non-proprietary encoding of molecular structure that is widely used for deduplication and cross-database linking [6]. The first 14 characters of an InChIKey encode the molecular skeleton (connectivity), making it a natural unit for grouping stereoisomers and protonation states of the "same" compound—a property we exploit both for labelling and for leakage-free data splitting.

1.3 Reference spectral libraries

Supervised prediction from spectra is only possible because the community has invested in large, openly shared reference libraries. MassBank pioneered open, cross-vendor sharing of curated mass spectra [7]; the Global Natural Products Social Molecular Networking platform (GNPS) added community curation and molecular networking at repository scale [8]; and complementary resources such as the Human Metabolome Database [9] and METLIN [10] broadened chemical

coverage. The MassBank of North America (MoNA), hosted by the Fiehn Laboratory, aggregates and auto-curates spectra from many of these sources into downloadable, metadata-rich exports, and is one of the most convenient starting points for large-scale spectral machine learning. Data-processing ecosystems—MZmine [11, 12], MS-DIAL [13], and the GNPS networking suite [14, 15, 16]—feed these libraries and consume their annotations, and standardised confidence frameworks govern how the resulting identifications are reported [17, 18].

1.4 Learning from fragmentation spectra

A rich methodological toolkit now exists for turning MS/MS spectra into chemical knowledge. Fragmentation-tree modelling reconstructs the hierarchy of fragment formulae underlying a spectrum [19] and underpins the SIRIUS platform [20]; CSI:FingerID predicts molecular fingerprints from those trees and searches structure databases [21], with subsequent deep-kernel variants improving fingerprint accuracy [22]. A parallel line of work learns spectral *similarity* directly: Spec2Vec adapts word-embedding ideas to fragment peaks [23], while MS2DeepScore trains a Siamese network to predict structural (Tanimoto) similarity from spectrum pairs [24]. These state-of-the-art tools are powerful but also elaborate, often depending on formula annotation, fragmentation-tree computation, or large neural networks and their training corpora.

The purpose of the present report is complementary and deliberately minimalist. We ask how far one can get on the superclass-prediction task using only (i) the raw peak list, (ii) simple, auditable preprocessing, and (iii) a classical tree-ensemble classifier, under a strict compound-level evaluation protocol. Such a baseline is valuable for two reasons. First, it quantifies the intrinsic difficulty of the problem when no formula annotation, retention time, or precursor-mass reasoning is used. Second, it makes explicit a methodological pitfall that pervades cheminformatics machine learning: if spectra of the same molecule appear in both the training and test sets, performance is inflated and does not reflect generalisation to genuinely new chemistry [25]. We therefore split strictly by molecular skeleton and report results at the compound level.

1.5 Contributions

This report makes the following concrete contributions:

1. A transparent, reproducible pipeline that extracts experimental positive-mode $[M + H]^+$ MS/MS spectra from a dated MoNA export and restricts them to records with a valid InChIKey and a ClassyFire superclass (Section 2).
2. A leakage-controlled, compound-level 80/20 split (seed 42) with programmatic verification of zero shared compounds, together with a simple and fully specified spectral-preprocessing and feature-engineering recipe (Section 3).
3. A tuned Extremely Randomized Trees classifier, selected by grouped cross-validation, with spectrum-to-compound probability aggregation (Section 3).
4. A complete evaluation—overall accuracy, macro-averaged F_1, per-superclass F_1, and the two most-confused superclass pairs—plus the per-compound predicted superclass for every test compound (Section 4), and an interpretation of the errors in terms of chemical adjacency (Section 5).

2 Data acquisition and dataset characterisation

2.1 Source and provenance

Spectra were obtained from the MassBank of North America (MoNA; `https://mona.fiehnlab.ucdavis.edu`). The list of predefined exports was queried through the MoNA REST API, and the "LC-MS/MS Positive Mode" export (export id `7609a87b-5df1-4343-afe9-2016a3e79516`) was downloaded as a JSON archive (317 MB compressed, 1.87 GB uncompressed). This positive-mode subset was chosen in preference to the full LC-MS/MS export because the analysis targets positive-ion $[M + H]^+$ spectra, for which it is a strict, lighter-weight superset. The export's own generation timestamp defines the data vintage: **the MoNA export is as-of `2026-07-06T07:44:51Z`**; it was retrieved on 2026-07-12 and reported 102 580 spectra. The 1.87 GB JSON was parsed record-by-record with a streaming parser to remain memory-safe.

2.2 Filtering criteria

A record was retained only if it satisfied *all* of the following, mirroring the requirements of the task:

1. **Tandem MS/MS**: MS level equal to MS2.
2. **Positive ionisation mode**.
3. **Precursor adduct** $[M + H]^+$ after normalisation of formatting variants (e.g., `[M+H]`, `M+H`, `[M+H]1+`); multimers and water-loss or multiply-charged adducts such as $[2M + H]^+$, $[M + 2H]^{2+}$, and $[M + H - H_2O]^+$ were excluded.
4. **Valid InChIKey** matching the canonical `[A-Z]{14}-[A-Z]{10}-[A-Z]` skeleton-block-checksum pattern.
5. **Annotated ClassyFire superclass** present in the record's compound classification.
6. **At least two fragment peaks** in the parsed peak list.

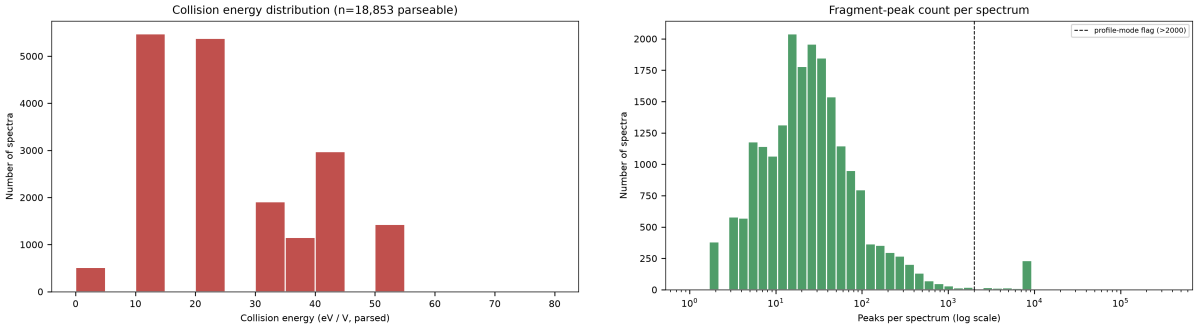
The reconciliation of dropped records is reported in Table 1. Of 102 580 scanned records, 20 403 spectra passed all criteria, corresponding to 3695 unique compounds (grouped by the 14-character InChIKey skeleton) across **16** ClassyFire superclasses.

Table 1: Record reconciliation for the MoNA LC-MS/MS Positive-Mode export (as-of 2026-07-06). Criteria are applied sequentially; each row reports the number of records removed by that criterion.

Stage	Records
Total records scanned	102,580
dropped: not MS2	0
dropped: not positive mode	33,047
dropped: adduct $\neq [M + H]^+$	31,429
dropped: invalid/absent InChIKey	17,042
dropped: no ClassyFire superclass	523
dropped: < 2 fragment peaks	136
Retained (filtered spectra)	**20,403**
unique compounds (InChIKey-14)	3,695
ClassyFire superclasses	16

2.3 Collision energy and spectral quality

Collision energy is the single most important acquisition parameter shaping a fragmentation spectrum, and its heterogeneity across a community library directly affects learnability. Of the retained spectra, 18 853 carried a numerically parseable collision energy (the remainder used non-numeric descriptors such as "ramp" or "HCD"). The parsed values span 0 eV **to** 80 eV (units eV or V as reported), with a median of 20 eV and a mean of 23.97 ± 13.48 eV; the distribution is multimodal, clustering around common instrument presets near 10, 20, and 40 eV (Figure 3a). Fragment-peak counts are likewise broadly distributed, with a median of 23 peaks per spectrum (Figure 3b). A small tail of 304 spectra (1.49%, predominantly GNPS `CCMSLIB` records) carried more than 2000 points and are best interpreted as profile/continuum traces rather than centroided peak lists; these were flagged and retained, and are implicitly centroided by the binning step described in Section 3.2.



(a) Collision-energy distribution.

(b) Fragment-peak count per spectrum (log scale).

Figure 3: Acquisition characteristics of the filtered dataset. (a) Parsed collision energies span 0 eV to 80 eV and cluster at common instrument presets. (b) Peak counts are centred near a median of 23; the dashed line marks the $>$ 2000-peak threshold used to flag likely profile-mode traces.

2.4 Class distribution and imbalance

The superclass distribution is severely imbalanced (Figure 4, Table 2), a defining and unavoidable feature of the problem. The four largest superclasses—*Organoheterocyclic compounds*, *Phenylpropanoids and polyketides*, *Organic acids and derivatives*, and *Lipids and lipid-like molecules*—together account for roughly 69 % of spectra, while the four smallest each contribute well under 1 %. Three superclasses (*Organophosphorus*, *Organic 1,3-dipolar*, and *Homogeneous non-metal compounds*) are represented by a single compound each. This imbalance motivates the use of macro-averaged F_1 as the headline metric, since macro-averaging weights every superclass equally and thereby exposes failures on rare classes that raw accuracy would mask [26].

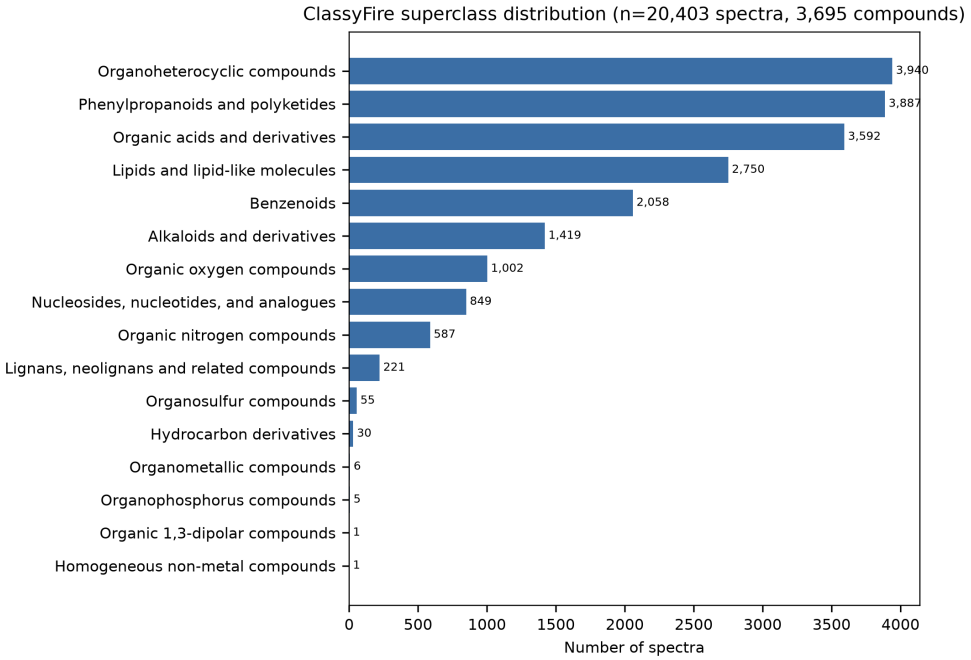


Figure 4: ClassyFire superclass distribution across the 20 403 filtered spectra (3695 compounds). The distribution is heavy-tailed: four superclasses dominate while several are represented by only a handful of spectra or a single compound.

Table 2: ClassyFire superclass distribution over the filtered dataset (spectrum counts and share). Sixteen superclasses are present.

Superclass	**Spectra**	**Share**
Organoheterocyclic compounds	3,940	19.31%
Phenylpropanoids and polyketides	3,887	19.05%
Organic acids and derivatives	3,592	17.61%
Lipids and lipid-like molecules	2,750	13.48%
Benzenoids	2,058	10.09%
Alkaloids and derivatives	1,419	6.95%
Organic oxygen compounds	1,002	4.91%
Nucleosides, nucleotides, and analogues	849	4.16%
Organic nitrogen compounds	587	2.88%
Lignans, neolignans and related compounds	221	1.08%
Organosulfur compounds	55	0.27%
Hydrocarbon derivatives	30	0.15%
Organometallic compounds	6	0.03%
Organophosphorus compounds	5	0.02%
Organic 1,3-dipolar compounds	1	0.00%
Homogeneous non-metal compounds	1	0.00%
Total	**20,403**	100%

3 Methods

3.1 Compound-level train/test split

The central methodological commitment of this study is that evaluation must reflect generalisation to *new molecules*, not merely to new spectra of molecules already seen during training.

Because the library contains many replicate spectra of the same compound acquired at different collision energies and on different instruments, a naive spectrum-level random split would place near-identical spectra of one molecule on both sides of the partition, leaking information and inflating apparent performance—a well-documented failure mode in QSAR and cheminformatics benchmarking [25].

To prevent this, the 20 403 spectra were first grouped into 3695 compounds by their 14-character InChIKey skeleton. Each compound was assigned its modal ClassyFire superclass (no compound exhibited conflicting annotations). The 80/20 split was then performed *on the compounds*, stratified by superclass, with random **seed 42**. Superclasses represented by fewer than two unique compounds (the three singletons noted above) were forced entirely into the training partition so that the test set contains only classes the model could in principle have learned. This produced **2956 training compounds** and **739 test compounds**, corresponding to 16 417 training spectra (80.5%) and 3986 test spectra (19.5%). The intersection of the training and test compound-identifier sets was verified programmatically to be *empty* at both the compound and spectrum level, and their union was confirmed to cover all 3695 compounds. The per-superclass split is reported in Table 3.

Table 3: Per-superclass compound- and spectrum-level counts for the compound-level 80/20 split (seed 42). Three singleton superclasses are forced into training and hence have no test compounds.

Superclass	**Cmp (tr)**	**Cmp (te)**	**Spec (tr)**	**Spec (te)**
Organoheterocyclic compounds	689	172	3,107	833
Phenylpropanoids and polyketides	537	135	3,156	731
Organic acids and derivatives	364	91	2,818	774
Lipids and lipid-like molecules	454	114	2,159	591
Benzenoids	433	109	1,672	386
Alkaloids and derivatives	122	30	1,176	243
Organic oxygen compounds	160	40	860	142
Nucleosides, nucleotides, and analogues	98	24	733	116
Organic nitrogen compounds	58	15	473	114
Lignans, neolignans and related compounds	24	6	184	37
Organosulfur compounds	9	2	52	3
Hydrocarbon derivatives	3	1	14	16
Organometallic compounds	2	0	6	0
Organophosphorus compounds	1	0	5	0
Organic 1,3-dipolar compounds	1	0	1	0
Homogeneous non-metal compounds	1	0	1	0
Total	**2,956**	**739**	**16,417**	**3,986**

3.2 Spectral preprocessing and feature engineering

Each spectrum was converted from a raw $(m/z, \text{intensity})$ peak list into a fixed-length numeric feature vector through the following deterministic steps.

Noise filtering. Peaks with a relative intensity below **1% of the base (most intense) peak** were discarded. This removes low-level chemical and electronic noise that would otherwise dominate the high-dimensional feature space without carrying class-discriminative information.

Base-peak normalisation. The surviving intensities of each spectrum were rescaled so that the base peak equals 1.0. This makes the representation invariant to absolute ion abundance, which varies with concentration and instrument tuning and is not chemically meaningful for

class assignment.

Binning. Normalised peaks were mapped onto a regular m/z grid spanning $[50, 1000)$ Da with a bin width of 1 Da, producing **950 bins**; the intensity assigned to each bin is the maximum normalised intensity of the peaks falling within it. Max aggregation also serves to centroid the flagged profile-mode traces onto the fragment grid without special-casing. On average 3.56 % of retained intensity mass fell outside the $[50, 1000)$ Da window and was discarded; 54 spectra became empty after noise filtering and gridding and were retained as all-zero rows.

TF-IDF weighting. Finally, term-frequency–inverse-document-frequency (TF-IDF) weighting with L_2 row normalisation was applied to the binned matrix, treating each m/z bin as a "term" and each spectrum as a "document." TF-IDF, whose inverse-document-frequency component derives from Spärck Jones's classical measure of term specificity [27], down-weights fragment m/z values that occur ubiquitously across the library (and are thus poorly discriminative) while emphasising bins that are characteristic of particular chemistries. Critically, the TF-IDF statistics were *fit on the training spectra only* and then applied to the test spectra, so that no corpus-level information from the test set leaks into the representation. The final feature matrices are $\mathbf{X}_{\text{train}} \in \mathbb{R}^{16417\times 950}$ and $\mathbf{X}_{\text{test}} \in \mathbb{R}^{3986\times 950}$.

Precursor m/z and (median-imputed) collision energy were additionally computed and stored as optional auxiliary features, but the classifier reported here uses the *spectral features alone*, in keeping with the task of predicting superclass "from the MS/MS spectrum alone."

3.3 Classifier, model selection and tuning

Two classical tree ensembles were considered as candidate classifiers: the Random Forest [28] and Extremely Randomized Trees (Extra-Trees) [29], both implemented in scikit-learn [30]. Tree ensembles are well suited to sparse, high-dimensional, heterogeneous features, are robust to irrelevant inputs, natively support multi-class problems, and—unlike large neural models—train quickly and deterministically, which suits a reproducible baseline. Extra-Trees differs from the Random Forest by choosing split thresholds at random rather than optimising them and by growing trees on the full sample, which further decorrelates the ensemble and often reduces variance [29]. Class imbalance was addressed with balanced-subsample class weighting.

Model selection used **5-fold Stratified Group k-fold** cross-validation on the training set, with the *groups* defined by compound identifier so that no compound appears in both a training and a validation fold; the absence of within-fold compound leakage was asserted for every fold. Folds were additionally stratified by superclass. Under this protocol the Random Forest achieved a cross-validated macro-F_1 of 0.355 (accuracy 0.546) and Extra-Trees a macro-F_1 of 0.356 (accuracy 0.539); Extra-Trees was selected on the primary macro-F_1 metric. A grid search over the number of trees ($\{300, 600\}$), maximum depth ($\{\text{None}, 40\}$), and the fraction of features considered per split ($\{\sqrt{\cdot}, 0.1\}$) identified $\{n_{\text{estimators}} = 600,\ \text{max\_depth} = \text{None},\ \text{max\_features} = 0.1\}$ as best (cross-validated macro-F_1 0.359). The final model was retrained on the full training set with these hyperparameters and random state 42.

3.4 From spectra to compounds: prediction aggregation and metrics

Because the evaluation unit is the compound, spectrum-level predictions must be aggregated. For each test spectrum the model produces a probability distribution over the 16 superclasses via `predict_proba`. For each of the 739 test compounds, the class probabilities of *all* of its spectra are **averaged**, and the compound is assigned the arg-max class of this mean distribution; the mean of the maximum probability serves as a prediction-confidence score. This soft-voting

aggregation is more robust than hard majority voting because it pools evidence across replicate spectra acquired at different collision energies.

We report **overall accuracy** and **macro-averaged** F_1 as the two headline metrics, the full **per-superclass** F_1 (with precision and recall), the **confusion matrix**, and the two **most-confused superclass pairs**, defined as the pairs (i, j) maximising the symmetric off-diagonal count $C_{ij}+C_{ji}$ of the compound-level confusion matrix. Spectrum-level metrics are also reported for comparison.

4 Results

4.1 Overall performance

On the 739 held-out test compounds, the tuned Extra-Trees classifier attains an **overall accuracy of 0.503** and a **macro-averaged** F_1 **of 0.373** (weighted-average F_1 0.493). For context, always predicting the largest test superclass (*Organoheterocyclic compounds*) would achieve an accuracy of only 0.233; the model therefore recovers the correct superclass more than twice as often as the majority-class baseline, using nothing but the fragmentation pattern. At the spectrum level the same model reaches an accuracy of 0.555 and a macro-F_1 of 0.412. That spectrum-level numbers exceed compound-level numbers is expected and reassuring: it reflects the fact that abundant, easily-classified compounds contribute many spectra, whereas the compound-level metric gives each molecule equal weight and thus more faithfully measures generalisation across chemical space. The predicted-class distribution is itself informative—the model predicts only 10 of the 12 test superclasses, essentially never emitting the rarest labels—and confidence is modestly calibrated, averaging 0.55 on correct versus 0.41 on incorrect compound predictions.

4.2 Per-superclass performance

Performance varies dramatically across superclasses (Figure 5, Table 4). The best-recognised well-populated classes are *Phenylpropanoids and polyketides* ($F_1 = 0.65$), *Lipids and lipid-like molecules* ($F_1 = 0.62$), and *Organic acids and derivatives* ($F_1 = 0.55$), followed by *Nucleosides, nucleotides and analogues* ($F_1 = 0.53$) and the largest class, *Organoheterocyclic compounds* ($F_1 = 0.47$). These are precisely the categories whose members share pronounced, recurring fragmentation motifs—fatty-acyl and glycerophospholipid losses for lipids, sugar and nucleobase fragments for nucleosides, and characteristic small-molecule neutral losses for organic acids. At the other extreme, three test superclasses (*Lignans*, *Organic nitrogen compounds*, and *Organosulfur compounds*) are never correctly recovered ($F_1 = 0$); each is represented by very few test compounds and even fewer training compounds, and the model has evidently not acquired a stable fragmentation signature for them. The single-compound *Hydrocarbon derivatives* test case is correctly classified, but its F_1 of 0.67 should be read as anecdotal given the support of one.

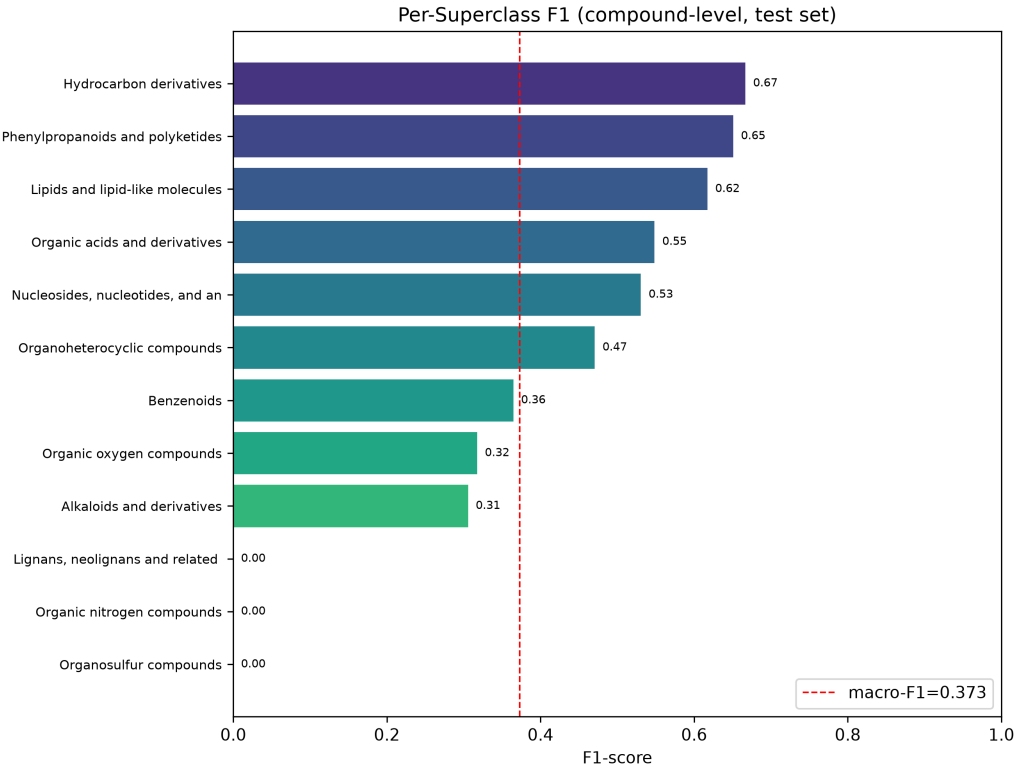


Figure 5: Per-superclass F_1 (compound level, test set). Bars are sorted by F_1; the dashed line marks the macro-average of 0.373. Well-populated classes with distinctive fragmentation (phenylpropanoids/polyketides, lipids, organic acids) are recognised well, whereas rare classes collapse to $F_1 = 0$.

Table 4: Per-superclass precision, recall, and F_1 at the compound level (739 test compounds), sorted by F_1. Support is the number of test compounds.

Superclass	F_1	**Precision**	**Recall**	**Support**
Hydrocarbon derivatives[†]	0.667	0.500	1.000	1
Phenylpropanoids and polyketides	0.651	0.600	0.711	135
Lipids and lipid-like molecules	0.617	0.581	0.658	114
Organic acids and derivatives	0.548	0.537	0.560	91
Nucleosides, nucleotides, and analogues	0.531	0.520	0.542	24
Organoheterocyclic compounds	0.471	0.476	0.465	172
Benzenoids	0.365	0.422	0.321	109
Organic oxygen compounds	0.317	0.435	0.250	40
Alkaloids and derivatives	0.306	0.262	0.367	30
Lignans, neolignans and related compounds	0.000	0.000	0.000	6
Organic nitrogen compounds	0.000	0.000	0.000	15
Organosulfur compounds	0.000	0.000	0.000	2
Macro average	**0.373**	**0.361**	**0.406**	739
Weighted average	**0.493**	**0.490**	**0.503**	739

[†]Support of 1; value is anecdotal.

4.3 Confusion structure and the most-confused pairs

The compound-level confusion matrix (Figure 6) reveals a clear and chemically interpretable error structure. Errors are not spread uniformly; instead, mass is concentrated around a few adjacent classes, and *Organoheterocyclic compounds*—the largest and most chemically diffuse

superclass—acts as an attractor into which many other classes are misclassified. Computing the symmetric off-diagonal counts identifies the **two most-confused superclass pairs**:

1. **Benzenoids ↔ Organoheterocyclic compounds**, with **50** total misclassifications (30 true Benzenoids predicted as Organoheterocyclic, 20 in the reverse direction);
2. **Organic acids and derivatives ↔ Organoheterocyclic compounds**, with **31** total misclassifications (17 and 14 in the two directions).

Both pairs involve *Organoheterocyclic compounds*, and both are chemically adjacent, as discussed in Section 5.

Figure 6: Compound-level confusion matrix (rows: true superclass; columns: predicted). Colour encodes the number of test compounds. The diagonal dominates for well-populated, fragmentation-distinctive classes, while off-diagonal mass concentrates around chemically adjacent categories—most conspicuously the Benzenoids/Organoheterocyclic and Organic-acids/Organoheterocyclic confusions.

4.4 Per-compound predictions

The complete per-compound output—the true superclass, the predicted superclass, the prediction confidence, and the full 16-way probability vector for each of the 739 test compounds—is provided in the accompanying file `test_compound_predictions.csv`. A representative excerpt

is shown in Table 5. The excerpt illustrates the qualitative behaviour discussed above: confident, correct calls for prototypical lipids and organic acids (e.g., confidence 0.89 for a correctly classified organic acid), alongside lower-confidence errors that scatter into the neighbouring Organoheterocyclic, Lipid, or Benzenoid classes.

Table 5: Excerpt from the per-compound test-set predictions (first 15 compounds by InChIKey-14). The full table with 16-way class probabilities is provided as `test_compound_predictions.csv`.

Compound (InChIKey-14)	True superclass	Predicted superclass	Conf.	Correct
AAOFJKLTRKOQTQ	Organic oxygen compounds	Phenylpropanoids and polyketides	0.451	✗
ABBPFXQJIWUCKF	Organoheterocyclic compounds	Lipids and lipid-like molecules	0.605	✗
ABHHIGWFFMCQOC	Phenylpropanoids and polyketides	Organoheterocyclic compounds	0.197	✗
ABLACSIRCKEUOB	Benzenoids	Lipids and lipid-like molecules	0.376	✗
ADDLXGZBTXNVDS	Lipids and lipid-like molecules	Lipids and lipid-like molecules	0.602	✓
AEQDJSLRWYMAQI	Alkaloids and derivatives	Benzenoids	0.287	✗
AEUTYOVWOVBAKS	Organic nitrogen compounds	Organic acids and derivatives	0.427	✗
AFUDERYHOUWFDT	Organic acids and derivatives	Alkaloids and derivatives	0.532	✗
AHLPHDHHMVZTML	Organic acids and derivatives	Organic acids and derivatives	0.893	✓
AHTPATJNIAFOLR	Organoheterocyclic compounds	Organoheterocyclic compounds	0.330	✓
AHTRGGWSBFOEEG	Lipids and lipid-like molecules	Benzenoids	0.403	✗
AHUAGKUYEMQHTH	Phenylpropanoids and polyketides	Organoheterocyclic compounds	0.298	✗
AJLFOPYRIVGYMJ	Organoheterocyclic compounds	Lipids and lipid-like molecules	0.256	✗
ALEXXDVDDISNDU	Lipids and lipid-like molecules	Lipids and lipid-like molecules	0.507	✓
ARQXEQLMMNGFDU	Phenylpropanoids and polyketides	Phenylpropanoids and polyketides	0.553	✓

5 Discussion

5.1 The errors are chemically sensible

The most important qualitative finding is that the classifier's confusions track genuine chemical adjacency rather than random noise. Consider the top confused pair, *Benzenoids↔Organoheterocyclic compounds*. Benzenoids are carbocyclic aromatic systems, whereas organoheterocyclics contain at least one ring heteroatom; enormous numbers of real metabolites sit at this boundary—a benzene ring fused or substituted with a nitrogen-, oxygen-, or sulfur-containing ring, or a molecule whose dominant fragments are aromatic regardless of a peripheral heterocycle. In positive-mode CID/HCD, such molecules produce very similar aromatic fragment ions (e.g., tropylium- and phenyl-type cations, aromatic neutral losses), so the fragmentation fingerprints genuinely overlap. The second pair, *Organic acids and derivatives↔Organoheterocyclic compounds*, is analogous: many organic acids (amino acids, nucleotide-associated acids, heterocyclic carboxylic acids) contain heterocyclic cores, and small carboxyl-related neutral losses (e.g., −44 Da, −46 Da) are shared broadly. That *Organoheterocyclic compounds* is the common attractor in both pairs reflects its status as the largest and structurally broadest superclass—a catch-all with high prior probability and diffuse fragmentation behaviour. In short, the model fails where a human mass spectrometrist would also hesitate, which is precisely the behaviour one wants from a spectrum-only baseline.

5.2 Why the absolute numbers are what they are

An overall accuracy of ∼0.50 and a macro-F_1 of ∼0.37 should be read against the intrinsic difficulty of the task rather than against near-perfect structure-identification tools built on richer inputs. Three factors bound performance. First, the input is deliberately impoverished: only the peak list is used, with no molecular formula, no precursor-mass reasoning, no fragmentation tree, and no retention information. Methods such as CSI:FingerID and CANOPUS achieve far higher class-level accuracy precisely because they first infer molecular formulae and fragmentation trees and predict fingerprints before assigning classes [21, 4, 20]. Second, the

collision-energy heterogeneity documented in Figure 3a—spectra of the same compound acquired anywhere from 0 to 80 eV—means the same molecule can present markedly different fragmentation patterns, a source of irreducible within-class variance for a model that ignores the collision-energy covariate. Third, the compound-level split is a demanding test: because no test compound (and no stereoisomer or protonation variant of it) is ever seen in training, the model must generalise to unseen chemistry, exactly the regime in which cheminformatics models are known to degrade relative to random-split estimates [25]. Reported spectrum-level accuracy is correspondingly higher (0.555), and would be higher still under a naive random split—which is exactly why we do not use one.

5.3 Class imbalance and the rare-class failures

The zero-F_1 superclasses (*Lignans*, *Organic nitrogen*, *Organosulfur*) are a direct consequence of extreme imbalance (Table 2). With only a handful of training compounds, a decorrelated tree ensemble cannot carve out a reliable decision region, and balanced class weighting is insufficient to overcome the sheer scarcity of examples. This is a data problem as much as a modelling one: it mirrors the broader "dark matter" asymmetry of metabolomics, in which reference coverage is concentrated on a few well-studied chemotypes [1, 2]. Reporting macro-F_1 alongside accuracy makes this failure visible rather than letting the abundant classes hide it, consistent with best-practice guidance for imbalanced multi-class evaluation [26].

5.4 Limitations

Several limitations should temper interpretation. (i) *Spectrum-only inputs*: by construction we exclude precursor m/z and collision energy, both of which are informative and readily available; the auxiliary features were computed but not used here. (ii) *Label granularity*: ClassyFire superclass is coarse, and some molecules legitimately straddle superclass boundaries; the "ground truth" itself embeds the aromatic/heterocyclic ambiguity that drives the top confusions. (iii) *Library biases*: MoNA aggregates spectra from heterogeneous instruments and laboratories, and the flagged profile-mode traces, ramped/HCD collision-energy descriptors, and uneven per-compound replication all inject noise. (iv) *Binning resolution*: the 1 Da grid discards the sub-Dalton accurate-mass information that high-resolution instruments provide and that formula-based methods exploit. (v) *Classical model family*: gradient-boosted trees and neural approaches were not evaluated here (LightGBM/XGBoost were unavailable in the execution environment), so the reported numbers are a baseline rather than a ceiling.

5.5 Future directions

The most promising and low-cost improvement is to concatenate the precursor m/z and collision energy to the spectral features, and to model collision energy explicitly so that replicate spectra are no longer treated as noisy repeats of one another. Higher-resolution binning (e.g., 0.01–0.1 Da) or peak-based rather than grid-based features would preserve accurate-mass information; sublinear term-frequency scaling in the TF-IDF step is a further cheap refinement. On the modelling side, gradient-boosted decision trees [31, 32] and, given enough data, message-passing or embedding-based neural models [33, 34, 23, 24] are the natural next steps. Most fundamentally, incorporating molecular-formula and fragmentation-tree information—the ingredients that make SIRIUS/CSI:FingerID and CANOPUS so effective [19, 20, 21, 4, 22]—would be expected to close much of the gap between this transparent baseline and state-of-the-art class prediction, at the cost of the simplicity and full reproducibility that motivated the present study.

6 Conclusion

Using only the fragmentation peak list of a positive-mode $[M+H]^+$ tandem mass spectrum, a tuned Extremely Randomized Trees classifier predicts the ClassyFire chemical superclass of an unseen compound with an overall accuracy of 0.503 and a macro-averaged F_1 of 0.373 on 739 held-out compounds drawn from the MoNA library (export as-of 2026-07-06), under a strict, leakage-free, compound-level split (seed 42). Performance is strong for well-populated, fragmentation-distinctive superclasses (phenylpropanoids/polyketides, lipids, organic acids) and collapses for rare ones, while the dominant errors—Benzenoids and Organic acids each confused with the broad Organoheterocyclic superclass—are chemically adjacent and therefore interpretable. The result quantifies both the promise and the limits of purely spectral, structure-agnostic class prediction, and provides a clean, fully reproducible baseline and a complete set of per-compound predictions against which richer models can be measured.

Data, code, and reproducibility

All randomness is controlled by seed 42 (compound split, cross-validation shuffling, and model `random_state`); re-running the pipeline reproduces the identical split, model, and metrics. The data provenance is fixed to the MoNA "LC-MS/MS Positive Mode" export (id `7609a87b-5df1-4343-afe9-2016a3e79516`), generated `2026-07-06T07:44:51Z`. Deliverables accompanying this report include: the per-compound test-set predictions with full class-probability vectors (`test_compound_predictions.csv`); the evaluation, split, and data summaries (`evaluation_summary.txt`, `split_summary.txt`, `data_summary.txt`); and the figures reproduced herein. Preprocessing used a 1 Da m/z grid over $[50, 1000)$ Da with a 1%-of-base-peak noise floor, base-peak normalisation, and train-only TF-IDF; classification used scikit-learn's Extremely Randomized Trees ($n_{\text{estimators}} = 600$, `max_features`= 0.1, `max_depth`=None, balanced-subsample weights).

References

[1] Ricardo R. da Silva, Pieter C. Dorrestein, and Robert A. Quinn. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41):12549–12550, 2015.

[2] Wout Bittremieux, Nicole E. Avalon, Sydney P. Thomas, Sarvar A. Kakhkhorov, Alexander A. Aksenov, Paulo Wender P. Gomes, Christine M. Aceves, Andrés Mauricio Caraballo-Rodríguez, Julia M. Gauglitz, William H. Gerwick, et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nature Communications*, 14:8488, 2023.

[3] Kerstin Scheubert, Franziska Hufsky, Daniel Petras, Mingxun Wang, Louis-Félix Nothias, Kai Dührkop, Nuno Bandeira, Pieter C. Dorrestein, and Sebastian Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nature Communications*, 8:1494, 2017.

[4] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A. Hoffmann, Daniel Petras, William H. Gerwick, Juho Rousu, Pieter C. Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39(4):462–471, 2021.

[5] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton,

Russell Greiner, and David S. Wishart. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8:61, 2016.

[6] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7:23, 2015.

[7] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010.

[8] Mingxun Wang, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A. Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V. Melnik, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34(8):828–837, 2016.

[9] David S. Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L. Lee, et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, 2022.

[10] Carlos Guijas, J. Rafael Montenegro-Burke, Xavier Domingo-Almenara, Amelia Palermo, Benedikt Warth, Gerrit Hermann, Gunda Koellensperger, Tao Huan, Winnie Uritboonthai, Aries E. Aisporna, Dennis W. Wolan, Mary E. Spilker, H. Paul Benton, and Gary Siuzdak. METLIN: a technology platform for identifying knowns and unknowns. *Analytical Chemistry*, 90(5):3156–3164, 2018.

[11] Mikko Katajamaa, Jarkko Miettinen, and Matej Orešič. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636, 2006.

[12] Robin Schmid, Steffen Heuckeroth, Ansgar Korf, Aleksandr Smirnov, Owen Myers, Thomas S. Dyrlund, Roman Bushuiev, Kevin J. Murray, Nils Hoffmann, Miaoshan Lu, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology*, 41(4):447–449, 2023.

[13] Hiroshi Tsugawa, Kazutaka Ikeda, Mikiko Takahashi, Aya Satoh, Yoshifumi Mori, Haruki Uchino, Nobuyuki Okahashi, Yutaka Yamada, Ipputa Tada, Paolo Bonini, et al. A lipidome atlas in MS-DIAL 4. *Nature Biotechnology*, 38(10):1159–1163, 2020.

[14] Louis-Félix Nothias, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, Madeleine Ernst, Hiroshi Tsugawa, Markus Fleischauer, Fabian Aicheler, Alexander A. Aksenov, et al. Feature-based molecular networking in the GNPS analysis environment. *Nature Methods*, 17(9):905–908, 2020.

[15] Allegra T. Aron, Emily C. Gentry, Kerry L. McPhail, Louis-Félix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, Julia M. Gauglitz, Nicole Sikora, Fernando Vargas, Justin J. J. van der Hooft, Madeleine Ernst, Kyo Bin Kang, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols*, 15(6):1954–1991, 2020.

[16] Robin Schmid, Daniel Petras, Louis-Félix Nothias, Mingxun Wang, Allegra T. Aron, Annika Jagels, Hiroshi Tsugawa, Johannes Rainer, Mar Garcia-Aloy, Kai Dührkop, Ansgar Korf, Tomáš Pluskal, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nature Communications*, 12:3832, 2021.

[17] Lloyd W. Sumner, Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, Teresa W.-M. Fan, Oliver Fiehn, Royston Goodacre, Julian L. Griffin, Thomas Hankemeier, Nigel Hardy, James Harnly, Richard Higashi, Joachim Kopka, et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3):211–221, 2007.

[18] Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science & Technology*, 48(4):2097–2098, 2014.

[19] Florian Rasche, Aleš Svatoš, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Analytical Chemistry*, 83(4):1243–1251, 2011.

[20] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, 2019.

[21] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

[22] Kai Dührkop. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. *Bioinformatics*, 38(Supplement_1):i342–i349, 2022.

[23] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, 2021.

[24] Florian Huber, Sven van der Burg, Justin J. J. van der Hooft, and Lars Ridder. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13:84, 2021.

[25] Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alán Aspuru-Guzik, et al. QSAR without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

[26] Sarah Farhadpour, Timothy A. Warner, and Aaron E. Maxwell. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: guidance and best practices. *Remote Sensing*, 16(3):533, 2024.

[27] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[28] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[29] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Tianqi Chen and Carlos Guestrin. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[32] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 3146–3154, 2017.

[33] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1263–1272, 2017.

[34] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.