Mineral Species Identification from Raman Spectra: A Leakage-Free Machine-Learning Classifier Built on the RRUFF Database

Analysis as of 12 July 2026

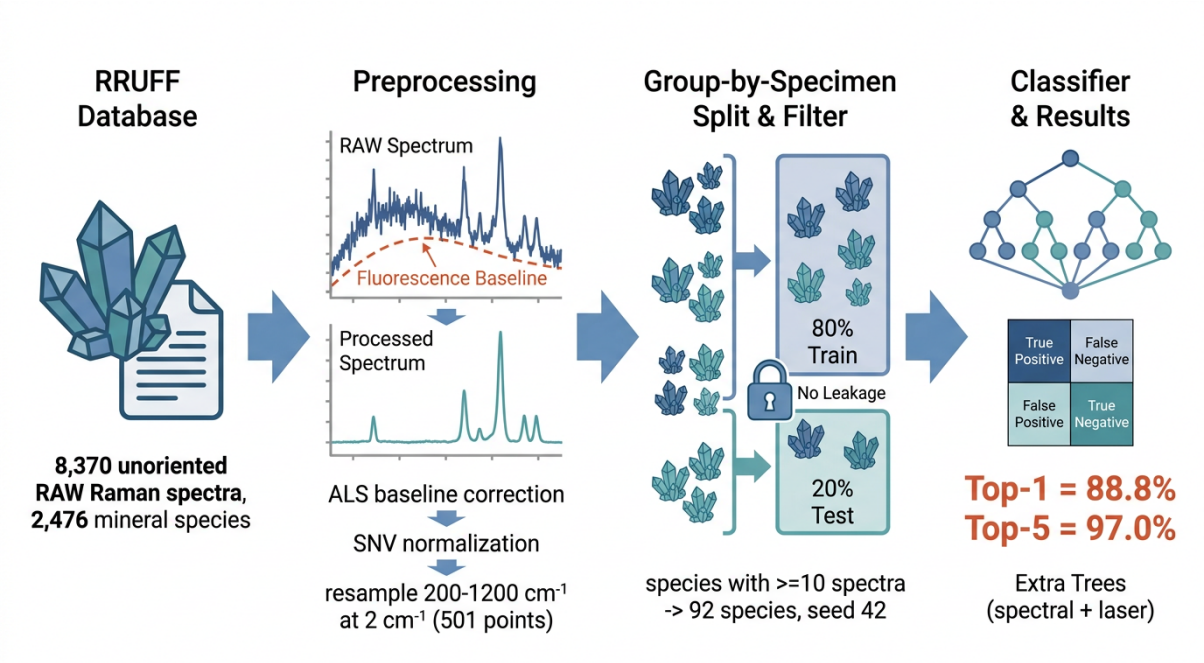

Figure 1: Graphical abstract. End-to-end workflow: unoriented RAW Raman spectra from the RRUFF database are baseline-corrected (asymmetric least squares), normalized (standard normal variate) and resampled onto a uniform 200–1200 cm^{-1} grid at 2 cm^{-1} (501 points); a $\geq$10-spectra species filter and a leakage-free group-by-specimen 80/20 split (seed 42) yield 92 species; an Extra-Trees classifier using spectral features plus the laser line reaches top-1 = 88.8% and top-5 = 97.0% on the held-out test set.

Abstract

Raman micro-spectroscopy is a rapid, non-destructive fingerprinting technique for mineral identification, and the open RRUFF database provides tens of thousands of reference spectra suitable for training automated classifiers. We built a mineral-species classifier from the RRUFF collection of *unoriented* Raman spectra (downloaded 12 July 2026). Starting from 8,370 raw (unprocessed) spectra spanning 2,476 species, we applied a physically motivated preprocessing pipeline—asymmetric-least-squares (ALS) baseline correction to remove fluorescence backgrounds, standard-normal-variate (SNV) intensity normalization, and cubic-spline resampling onto a uniform 200–1200 cm^{-1} grid at a 2 cm^{-1} step (501 points). Restricting the problem to species with at least ten spectra (and, necessarily, at least two distinct specimens) retained **92 species** (1,518 spectra). To avoid the optimistic bias caused by a single physical specimen contributing many correlated spectra, we used a *group-by-specimen* split (seed 42) that holds out ∼20% of specimens while guaranteeing that every retained species appears in both partitions; species whose specimens could not be divided

(i.e. those with a single specimen) were excluded a priori by the two-specimen filter. Comparing support-vector machines, random forests, extremely randomized trees and gradient-boosted trees under grouped 5-fold cross-validation, an Extra-Trees model augmented with a one-hot encoding of the excitation wavelength was selected. On the untouched test set (329 spectra, 134 specimens) it achieved **top-1 accuracy** = 88.8% (292/329) and **top-5 accuracy** = 97.0%. The five most-misclassified species—Mimetite, Pyrope, Andradite, Jamesonite and Diamond—are confused almost exclusively with chemically or structurally near-identical relatives (e.g. the garnet pair Pyrope→ Almandine and the apatite-supergroup pair Mimetite→Vanadinite), confirming that the residual errors reflect genuine spectral degeneracy rather than pipeline artefacts. Per-spectrum top-1 predictions for the entire test set are provided as a machine-readable CSV.

1 Introduction

Raman spectroscopy probes the inelastic scattering of monochromatic light by molecular and lattice vibrations, producing a pattern of bands whose positions and relative intensities are determined by the chemical bonding and crystal structure of the scattering material. For minerals, this vibrational "fingerprint" is diagnostic: the wavenumbers of internal modes of anionic groups (Si–O, C–O, S–O, P–O, As–O, V–O) together with lower-frequency lattice modes uniquely characterize most species, and a spectrum can typically be acquired in seconds on an intact grain with no sample preparation [1, 2]. These properties have made Raman microspectroscopy a workhorse of modern mineralogy, gemology, economic-geology exploration and planetary science, where it supports rapid, non-destructive, in-situ identification of phases in drill core, thin sections and rover payloads [3, 4].

The practical value of Raman identification depends on the availability of high-quality reference spectra. The RRUFF Project has become the community standard for this purpose: it is an integrated, openly accessible database that pairs Raman spectra with X-ray diffraction and quantitative chemical analyses for well-characterized mineral specimens, aiming ultimately to catalogue every known mineral species [5]. Because RRUFF records spectra at multiple laser wavelengths and, for many specimens, at multiple crystallographic orientations, it captures a realistic degree of intra-species variability, which is precisely what a robust automated classifier must learn to tolerate.

Automated matching of an unknown spectrum against such a library has traditionally relied on correlation or peak-list search, but over the last decade machine learning has been shown to deliver higher accuracy and greater robustness to noise and baseline drift. Support-vector machines and tree ensembles were early classical baselines for Raman mineral recognition [6, 7], and deep one-dimensional convolutional neural networks (1D-CNNs) trained end-to-end on raw spectra were later reported to outperform them on RRUFF subsets, in part by learning baseline correction implicitly [8–10]. Yet a recurring methodological weakness undermines the comparability and credibility of many published accuracies: *data leakage*. When a single physical specimen contributes several spectra (different lasers, repeated acquisitions, multiple spots) and those spectra are allowed to fall on both sides of a random train/test split, the model can memorize specimen-specific idiosyncrasies and report an accuracy that will not generalize to a genuinely new specimen. This failure mode is now well documented across scientific machine learning and is the leading cause of the reproducibility crisis in prediction studies; the accepted remedy is to partition data at the level of the independent unit—subject, patient, site or, here, specimen—rather than at the level of the individual measurement [11–13].

This report describes a mineral-species classifier built on the RRUFF unoriented Raman collection with three deliberate design commitments. First, we treat preprocessing as a first-class, physically motivated component of the pipeline and document every choice. Second, we impose a strict *group-by-specimen* evaluation protocol so that all reported accuracies estimate

performance on unseen specimens. Third, we report the full set of quantities requested for the task—the number of species retained after the ≥10-spectra filter, top-1 and top-5 test accuracy, the five most frequently misclassified species, and a per-spectrum prediction file—and we interpret the residual errors in mineralogical terms.

2 Data and Methods

2.1 Data source and acquisition

Spectra were obtained from the RRUFF Project's public repository of zipped Raman data files (`https://rruff.info/zipped_data_files/raman/`, which resolves to the RRUFF content host). All four *unoriented* quality tiers were downloaded on **12 July 2026** (UTC acquisition timestamp recorded in the provenance log): `excellent_unoriented`, `fair_unoriented`, `poor_unoriented` and `unrated_unoriented`. URLs, archive sizes and SHA-256 checksums were recorded at download time for reproducibility.

Each RRUFF `.txt` file consists of `##KEY=VALUE` header lines followed by two-column, comma-separated (wavenumber, intensity) data. A header-driven parser (with a filename fallback) extracted the mineral species name, the RRUFF specimen identifier, the data type, the laser excitation wavelength, the ideal chemistry and locality, and the wavenumber/intensity arrays. A key property of the "unoriented" archives is that they contain *both* the raw instrument output (`RAW`) and a vendor-processed variant of each measurement; the two were tagged so that downstream analysis could operate exclusively on the raw data. Parsing the archives yielded **16,644 spectra** in total (8,370 RAW; 8,274 processed) across 2,507 nominal species and 4,149 physical specimens, with zero hard parse failures. All subsequent modelling used only the **RAW** subset (8,370 spectra, 2,476 species, 4,099 specimens) so that the classifier learns from, and is evaluated on, unadulterated instrument spectra rather than proprietary post-processing.

2.2 Spectral preprocessing

Raw mineral Raman spectra vary widely in their sampled wavenumber range, point spacing, fluorescence background and absolute intensity scale; a classifier requires a common, physically comparable representation. Our per-spectrum pipeline comprised quality control, resampling, baseline correction and normalization, in that order (Figure 2).

Quality control. Within the RAW subset we discarded spectra with fewer than 100 data points (35 spectra, including nearly all empty files) and spectra that did not span the target analysis window (233 spectra), retaining **8,102 spectra** (97.2% of the qualifying RAW spectra) covering 2,424 species and 4,025 specimens.

Resampling grid. The task envisaged a broad grid (of order 100–1800 cm^{-1}), but the RRUFF RAW subset does not support such a range without heavy attrition or extrapolation. A coverage analysis showed a *median maximum wavenumber of only ∼1288 cm^{-1}*: just 7.4% of RAW spectra extend to 1400 cm^{-1} and only 1.1% reach 1800 cm^{-1}. Enforcing a 150–1400 cm^{-1} window would retain a mere 530 of 8,370 spectra (6.4%). By contrast, the **200–1200 cm^{-1}** interval—which contains the mineralogically decisive region of lattice modes and the principal Si–O, C–O, S–O, P–O, As–O and V–O stretching and bending bands—is fully covered by 97.2% of RAW spectra. We therefore resampled every spectrum onto a uniform **200–1200 cm^{-1} grid at a 2 cm^{-1} step (501 points)** by cubic-spline interpolation, averaging any duplicated abscissae so the grid is strictly increasing. Because every retained spectrum fully spans this window,

no extrapolation was performed. This data-driven choice preserves the full-dataset intent (8,102 usable spectra rather than 530) while eliminating extrapolation artefacts.

Baseline correction. Fluorescence produces a broad, slowly varying background that can dominate the true Raman signal, especially for the longer excitation wavelengths in the collection. We removed it with asymmetric least squares (ALS) smoothing [14, 15], which fits a smooth baseline by penalizing roughness (a second-difference penalty of strength λ) while asymmetrically weighting points that lie above versus below the current estimate (asymmetry p). We used $\lambda = 10^5$, $p = 10^{-3}$ and 10 iterations, and subtracted the estimated baseline. As Figure 2 shows, ALS tracks the fluorescence envelope across the 514, 532, 780 and 785 nm laser lines without clipping sharp Raman peaks; ALS and the related airPLS scheme [16] are the standard tools for this task.

Normalization. To place all spectra on a common intensity scale we applied the standard normal variate (SNV) transform [17]: each spectrum was mean-centred and divided by its own standard deviation, so that every processed spectrum has mean ≈ 0 and unit variance (verified: mean of per-spectrum means $\approx -1 \times 10^{-10}$, mean of per-spectrum standard deviations $= 1.000$). SNV removes multiplicative scaling and additive offsets, making band *shape* rather than absolute intensity the classification signal. (Savitzky–Golay smoothing/differentiation [18] is a common optional addition but was not required here, as SNV on ALS-corrected spectra already yields clean, peak-dominated profiles.)

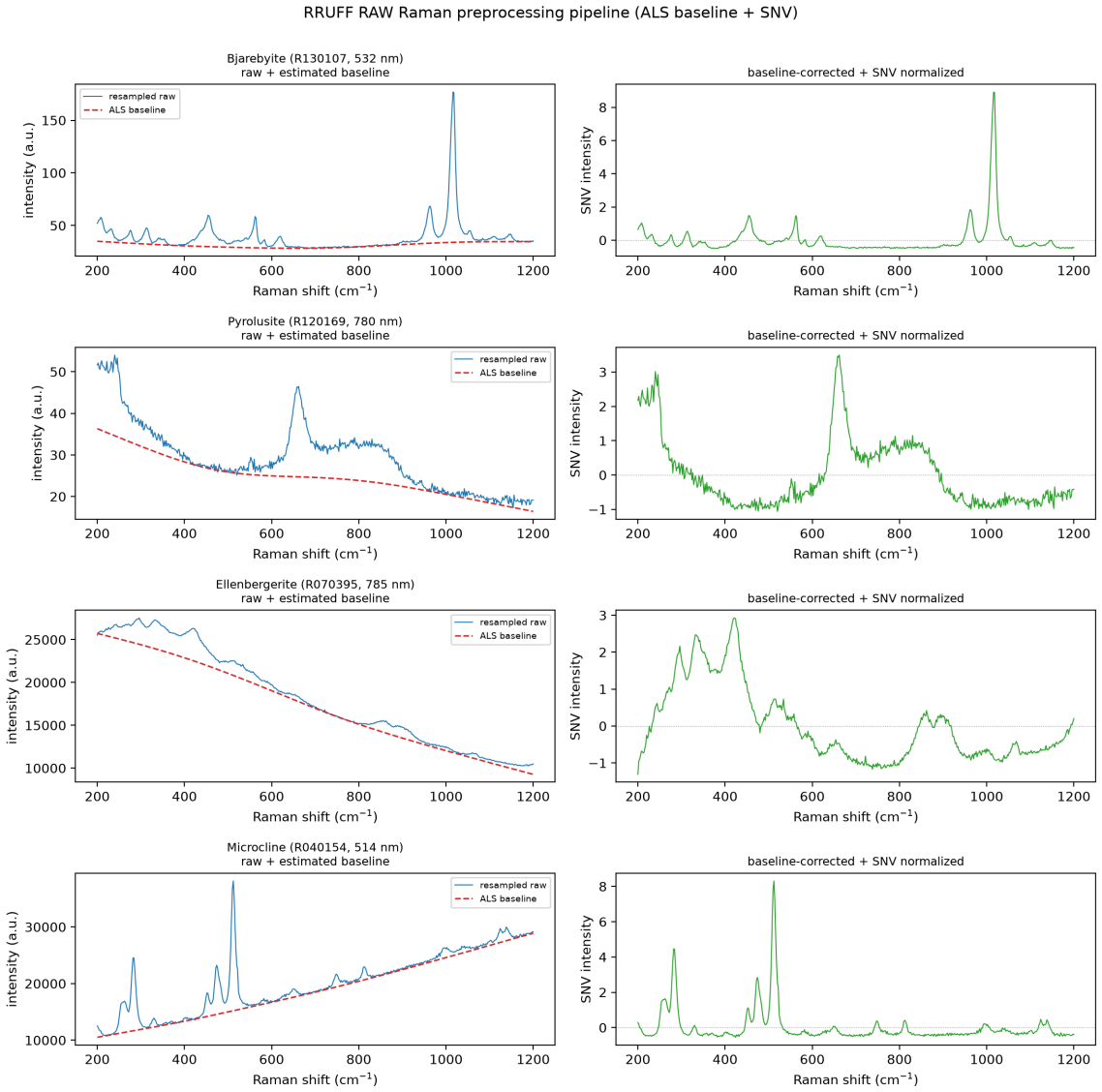

Figure 2: Preprocessing pipeline, four representative specimens. *Left column:* the resampled raw spectrum (blue) with the ALS-estimated fluorescence baseline (red dashed). *Right column:* the corresponding baseline-corrected, SNV-normalized spectrum. Rows span the dominant excitation wavelengths (532, 780, 785 and 514 nm). The ALS baseline follows the fluorescence envelope—dramatically sloped for the 785 nm ellenbergerite example—while preserving sharp bands; after SNV the diagnostic peaks (e.g. the microcline feldspar features near 510 cm^{-1}) stand out on a flat, zero-mean background.

2.3 Species filtering

A species can be modelled only if it is represented by enough spectra to learn from *and* by enough independent specimens to be placed on both sides of a specimen-level split. We therefore retained species satisfying two conditions: **at least 10 spectra** and **at least 2 distinct specimens**. The two-specimen requirement is not an arbitrary threshold but a mathematical necessity: a group-by-specimen split can only guarantee that a species appears in both train and test if it has at least two specimens to distribute (Section 2.4).

Applying these criteria to the 8,102 preprocessed spectra retained **92 species** comprising **1,518 spectra** and 664 specimens. Of the 2,424 candidate species, 2,332 were excluded for having fewer than 10 spectra. Notably, *zero* species were excluded for the "$\geq$10 spectra but only a single specimen" case: every species that met the spectra-count threshold also possessed

at least two specimens, so the two filters coincided in practice and no species eligible on spectra count had to be dropped for being unsplittable. The 92 retained species span the major mineral classes—silicates (feldspars, garnets, pyroxenes, amphiboles, micas, olivines), carbonates, sulfates, phosphates, oxides, sulfides and native elements—and are listed in Appendix A.

2.4 Group-by-specimen train/test split

To obtain an honest estimate of performance on unseen material, all spectra belonging to a given RRUFF specimen were kept together on one side of the split (a grouped, leakage-free partition [12, 13]). Using a single seeded pseudo-random generator (`numpy.random.RandomState(42)`; species processed in sorted order for full reproducibility), the unique specimens of each retained species were shuffled and partitioned so as to hold out ∼20% of specimens while ensuring both partitions are non-empty:

- species with *exactly two* specimens → one specimen to train, one to test;
- species with *more than two* specimens → $n_{\text{test}} = \min(n-1,\ \max(1,\ \text{round}(0.2\,n)))$ specimens to test, the rest to train (guaranteeing at least one specimen on each side).

The procedure yielded a training set of **1,189 spectra** (530 specimens) and a test set of **329 spectra** (134 specimens). The realized holdout was 21.7% of spectra and 20.2% of specimens (Figure 3); it sits slightly above 20% because the many two-specimen species contribute a forced 50/50 *specimen* split, which is the price of guaranteeing that every species is represented in both partitions. Independent re-verification confirmed the three required invariants: **zero** specimen-identifier overlap between train and test (no leakage); all 92 species present in *both* partitions; and conservation of the spectrum count ($1{,}189 + 329 = 1{,}518$).

Species whose specimens could not be split across both sides—i.e. species with only one specimen—are handled by exclusion at the filtering stage (Section 2.3) rather than by an ad-hoc rule at split time. Because the ≥10-spectra filter already removed every such case, the split imposed no further species loss.

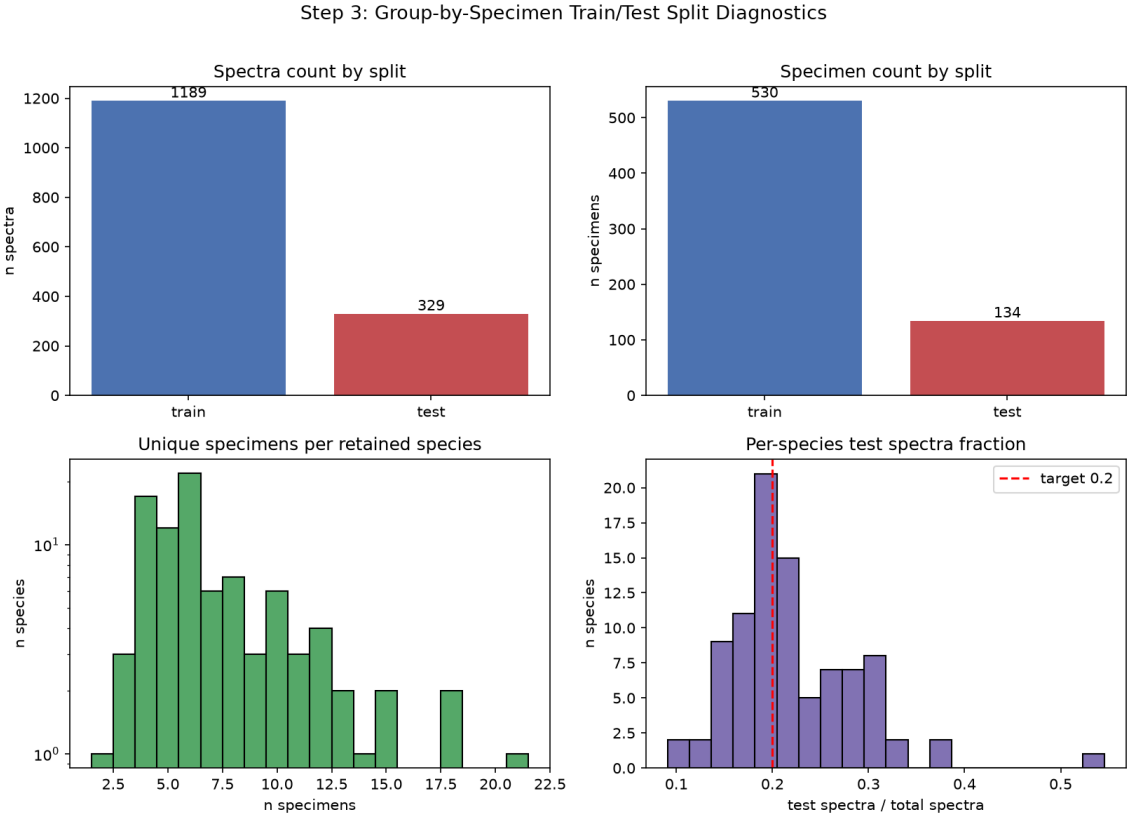

Figure 3: Group-by-specimen split diagnostics. *Top:* spectra (left) and specimen (right) counts for the train and test partitions. *Bottom left:* distribution of the number of unique specimens per retained species (most species have 2–8 specimens). *Bottom right:* per-species fraction of spectra assigned to test, centred on the 0.2 target (red dashed); the spread reflects the discrete, specimen-level nature of the split, particularly for two-specimen species.

2.5 Models and evaluation protocol

Each spectrum is represented by its 501-dimensional SNV vector (the *spectral* feature set). Because band positions and relative intensities can depend on the excitation wavelength, we also evaluated a *spectral+laser* feature set that appends a one-hot encoding of the laser line (514, 532, 780, 785 nm) to the spectral vector, to test whether excitation metadata improves discrimination.

We compared four classifier families spanning the methods most commonly used for spectral classification: a support-vector machine with a radial-basis-function kernel (SVC-RBF) [19], a random forest [20], extremely randomized trees (Extra-Trees) [21], and gradient-boosted decision trees (LightGBM) [22]. All models were implemented in scikit-learn [23] (LightGBM via its scikit-learn API) on top of the NumPy/SciPy stack [24, 25]. We deliberately restricted the study to classical and ensemble learners rather than deep 1D-CNNs [8, 9]: with only 1,189 training spectra spread over 92 classes (a median of roughly a dozen spectra per class), a data-hungry deep network is prone to overfitting, whereas tree ensembles and kernel SVMs are strong, well-regularized, and reproducible in this regime; deep models are revisited in the Discussion as a direction for future work.

Model and hyper-parameter selection used **grouped 5-fold cross-validation** on the training set only, with the specimen identifier as the grouping variable (`GroupKFold`); top-1 accuracy was the selection metric. This nests the grouping constraint inside the tuning loop, so that hyper-parameters are never chosen using spectra from a specimen that also appears in the val-

idation fold. For each family the best configuration (and feature set) was carried forward; the overall winner was then *refit on the full training set* and evaluated *once* on the held-out test set. Test-set metrics comprise top-1 accuracy, top-5 accuracy (the fraction of spectra whose true species is among the model's five highest-probability predictions), the per-species error tally, and the per-spectrum top-1 prediction with its top-5 candidate list.

3 Results

3.1 Dataset composition

The ≥10-spectra (≥2-specimen) filter retained **92 mineral species** (1,518 spectra; 664 specimens), which the group-by-specimen procedure divided into 1,189 training and 329 test spectra (Table 1). Every species is represented in both partitions. The test set spans all four dominant excitation wavelengths—532 nm (135 spectra), 780 nm (91), 514 nm (60) and 785 nm (43)—so that performance is measured across the instrument conditions present in the database.

Table 1: Dataset at each processing stage.

Stage	Spectra	Species	Specimens
Parsed (all unoriented)	16,644	2,507	4,149
RAW subset	8,370	2,476	4,099
After QC + resampling (200–1200 cm^{-1})	8,102	2,424	4,025
After ≥10-spectra, ≥2-specimen filter	1,518	**92**	664
– training partition	1,189	92	530
– test partition	329	92	134

3.2 Model selection

Under grouped 5-fold cross-validation, the tree-ensemble and kernel-SVM families all reached the mid-to-high 80% range in top-1 accuracy, while gradient boosting trailed (Table 2). Extra-Trees was the strongest family, and the addition of the laser-line covariate gave a marginal improvement, so the selected model was **Extra-Trees on the spectral+laser feature set** (800 trees, `max_depth` = 40, `max_features` = $\sqrt{d}$), with a cross-validated top-1 accuracy of 0.879 ± 0.037. The small gap between Extra-Trees, SVC-RBF and Random Forest indicates that the SNV/ALS representation is well-separated and not strongly model-dependent; the consistent under-performance of LightGBM (∼0.77) is typical for high-dimensional, densely correlated spectral features with few examples per class, where bagged randomized trees generalize better than aggressively boosted ones.

Table 2: Grouped 5-fold cross-validation on the training set: best configuration per classifier family. Accuracy is mean ± standard deviation of top-1 accuracy across folds.

Family	Feature set	Key hyper-parameters	CV top-1 accuracy
Extra-Trees	spectral+laser	800 trees, depth 40, `sqrt`	$\mathbf{0.879 \pm 0.037}$
SVC-RBF	spectral+laser	$C = 10$, $\gamma = 10^{-3}$	0.870 ± 0.042
Random Forest	spectral+laser	800 trees, no depth cap, `sqrt`	0.827 ± 0.043
LightGBM	spectral+laser	300 trees, 31 leaves, lr = 0.05	0.774 ± 0.038

3.3 Test-set performance

Refit on the full training set and evaluated once on the 329 held-out spectra, the selected Extra-Trees model achieved **top-1 accuracy** = 0.8875 (292 of 329 spectra correct) and **top-5 accuracy** = 0.9696 (Table 3). The test accuracy slightly *exceeds* the cross-validation mean, indicating no optimistic bias from model selection. Performance is strong at the level of species as well as spectra: **71 of the 92 species (77%) were classified perfectly** on test, and the mean per-species accuracy was 0.90. Of the 37 spectra misclassified at top-1, 27 were nonetheless recovered within the top-5 candidate list, leaving only 10 spectra (3.0%) whose true species fell entirely outside the top five—a practically useful behaviour, since an analyst is typically presented with a short ranked shortlist rather than a single guess.

Table 3: Final performance on the held-out test set (Extra-Trees, spectral+laser).

Metric	Value
Test spectra	329
Test specimens	134
Species (all present in train)	92
top-1 accuracy	**0.8875** (292/329)
top-5 accuracy	**0.9696**
Species with perfect test accuracy	71 / 92
Spectra outside top-5	10 / 329 (3.0%)

3.4 Most frequently misclassified species

Table 4 lists the five species with the most top-1 errors, and Figure 4 shows where their misclassifications go. The pattern is striking: errors are dominated by confusion with close mineralogical relatives. Three of the five worst species are garnets or lead apatite-group minerals whose spectra differ only subtly from a sibling species.

Table 4: The five most frequently misclassified species on the test set, with their dominant confusions.

Species	Test spectra	Misclassified	Error rate	Predominantly confused with
Andradite	6	4	0.67	Grossular (garnet), Fayalite, Perovskite
Pyrope	4	3	0.75	Almandine (garnet)
Jamesonite	6	3	0.50	Stephanite, Corundum (sulfosalt/other)
Diamond	7	3	0.43	Corundum, Stibnite, Fluorite
Mimetite	2	2	1.00	Vanadinite (apatite supergroup)

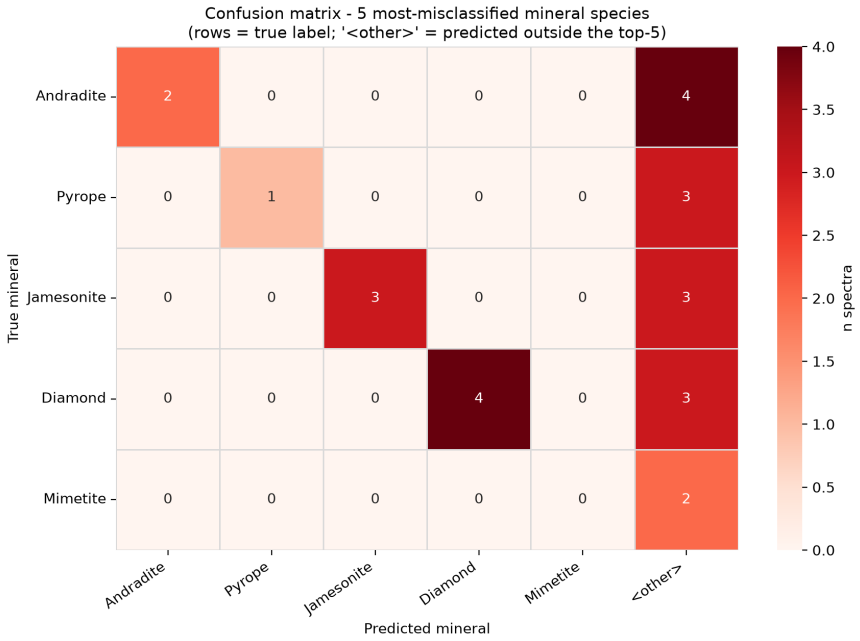

Figure 4: Confusion matrix for the five most-misclassified species. Rows are the true species; columns are the predicted species (the "⟨other⟩" column aggregates predictions falling outside these five labels). Diagonal cells are correct top-1 predictions. The heavy off-diagonal / "other" mass shows that these species are lost to a small number of specific, chemically adjacent competitors rather than being scattered at random.

3.5 Per-spectrum predictions

The complete per-spectrum top-1 prediction for the test set is provided as `test_predictions.csv` (329 rows). Each row records the spectrum index (aligned to the preprocessed matrix), the RRUFF specimen identifier, the laser wavelength, the true species, the predicted species, a correctness flag, and the ordered top-5 candidate list. Table 5 shows representative rows, including both confident correct calls and the garnet/apatite confusions discussed above.

Table 5: Illustrative rows from the per-spectrum prediction file `test_predictions.csv`.

Specimen	Laser (nm)	Correct?	True → Predicted	top-5 list (abbrev.)
R040063	514	✓	Actinolite → Actinolite	Actinolite; Tremolite; Beryl; Diopside; …
R120144	532	✓	Aegirine → Aegirine	Aegirine; Alluaudite; Eosphorite; …
R050113	532	×	Pyrope → Almandine	Almandine; … (garnet solid solution)
R060423	780	×	Andradite → Grossular	Grossular; … (garnet solid solution)
R070309	785	×	Mimetite → Vanadinite	Vanadinite; … (apatite supergroup)

4 Discussion

The headline result—top-1 = 88.8% and top-5 = 97.0% over 92 species under a strictly leakage-free protocol—places the classifier squarely in the range reported for classical and ensemble learners on RRUFF, and does so under an evaluation that many earlier studies did not enforce. Because our split guarantees that no test specimen was seen during training or model selection, these figures estimate performance on genuinely new specimens rather than on new spectra of already-seen specimens; the fact that the test accuracy matches (indeed slightly exceeds)

the cross-validated estimate is direct evidence that the protocol is well calibrated and free of selection-induced optimism [11, 13].

The errors are mineralogically meaningful. The most valuable diagnostic in this study is the *structure* of the residual errors. Rather than being distributed randomly, misclassifications concentrate on pairs of species that are chemically and structurally near-identical, and whose Raman spectra are therefore expected to be nearly degenerate:

- **Garnet solid solutions.** Pyrope ($Mg_3Al_2Si_3O_{12}$) was confused with Almandine ($Fe_3Al_2Si_3O_{12}$), and Andradite ($Ca_3Fe_2Si_3O_{12}$) with Grossular ($Ca_3Al_2Si_3O_{12}$). These are end-members of continuous garnet solid solutions; their Raman bands arise from the same SiO_4 internal and lattice modes and shift only gradually with the Mg–Fe–Ca–Al substitutions, so end-member spectra overlap substantially and intermediate compositions blur the boundary further [26].
- **Apatite-supergroup lead minerals.** Both Mimetite spectra were assigned to Vanadinite. Mimetite ($Pb_5(AsO_4)_3Cl$) and Vanadinite ($Pb_5(VO_4)_3Cl$) are isostructural members of the apatite supergroup differing only in the central tetrahedral anion (AsO_4 vs. VO_4); they form a solid solution and share nearly identical lattice-mode patterns, with the diagnostic difference confined to the tetrahedral stretching region. Related phosphate/apatite confusions (Pyromorphite→Fluorapatite, Cerussite→Pyromorphite) appear in the broader error list for the same reason.
- **Feldspar polymorphs and tourmaline.** Beyond the top five, Orthoclase was confused with Microcline—the two are $KAlSi_3O_8$ polymorphs distinguished mainly by Al/Si ordering, a subtle spectral effect—and Elbaite was confused with Tourmaline, its own supergroup label.

These confusions are essentially the "correct wrong answers": the model consistently identifies the right structural family and mistakes only the specific end-member, which is precisely the discrimination that is hardest—and sometimes impossible—from a single unoriented Raman spectrum without quantitative chemistry. The high top-5 accuracy (97.0%) reflects this: the true species is almost always among the top candidates even when it is not ranked first, so a shortlist-based workflow, optionally combined with the EDS/EPMA chemistry that RRUFF itself pairs with each specimen [5], would resolve most of the remaining ambiguity.

Diamond is a different, instructive case. Diamond's misclassifications are not driven by a chemical sibling but by its physics. Its single first-order Raman line at ∼1332 cm^{-1} [27, 28] lies at the very edge of our 200–1200 cm^{-1} analysis window; within the fingerprint region a clean diamond spectrum is essentially featureless, so after SNV normalization the model has little structure to key on and can be misled by noise or trace fluorescence into featureless-looking competitors (Corundum, Fluorite). This is a direct and expected consequence of the data-driven grid choice discussed in Section 2.2, which optimized coverage for the silicate/ carbonate/phosphate majority at the cost of species whose only strong band sits above 1200 cm^{-1}.

Role of the excitation wavelength. Appending the laser-line covariate improved cross-validated accuracy only marginally (Table 2), implying that the SNV/ALS spectral representation already captures most of the discriminative information and is reasonably robust across excitation wavelengths. The covariate is nonetheless worth retaining: it lets the model account for wavelength-dependent relative intensities and residual fluorescence without harming performance, and it costs only four extra features.

5 Limitations and future work

Several limitations frame these results. First, the 200–1200 cm^{-1} window, chosen to maximize spectral coverage, structurally handicaps species whose diagnostic bands lie above 1200 cm^{-1} (most conspicuously diamond and, more generally, O–H, C–H and high-frequency CO_3/SO_4 overtone features); a complementary model on the subset of wide-range spectra, or a two-window feature design, could recover these. Second, the ≥10-spectra filter restricts the classifier to 92 comparatively well-sampled species—a small fraction of the >2,400 species present—so the system is an identifier within a known panel, not an open-set detector; scaling to the long tail will require few-shot or metric-learning approaches and explicit "unknown/none-of-the-above" rejection. Third, several species are represented by only two specimens, forcing a 50/50 specimen split and leaving very few test spectra, which makes their per-species error rates statistically noisy (e.g. Mimetite's 100% error rate rests on just two spectra). Fourth, although RRUFF spectra are relatively clean, real field spectra carry more noise, cosmic-ray spikes and mixed-phase contamination [4]; robustness to such conditions was not tested here.

The natural extensions follow from these points. Deep 1D-CNNs, which learn baseline-independent features and have outperformed classical methods on larger RRUFF subsets [8–10], are worth revisiting with aggressive data augmentation (noise injection, baseline perturbation, wavenumber jitter) to compensate for the small per-class sample sizes, always under the same group-by-specimen protocol used here so that reported gains are real. Fusing Raman with the co-registered XRD and chemical data in RRUFF [5, 7] would directly attack the solid-solution confusions that dominate the present error budget, since composition is exactly what distinguishes garnet and apatite-group end-members. Finally, calibrated probability outputs and an explicit rejection option would turn the classifier from a closed-panel identifier into a deployable decision-support tool.

6 Conclusion

Using the RRUFF collection of unoriented raw Raman spectra (downloaded 12 July 2026), we built and rigorously evaluated a mineral-species classifier. A physically grounded preprocessing pipeline (ALS baseline correction, SNV normalization, and cubic-spline resampling onto a uniform 200–1200 cm^{-1} grid at 2 cm^{-1}, 501 points) fed a species panel of **92 minerals** retained by a ≥10-spectra (≥2-specimen) filter. A strict group-by-specimen 80/20 split (seed 42), with every species guaranteed to appear in both partitions and species lacking a second specimen excluded up front, ensured that all metrics estimate performance on unseen specimens. The selected Extra-Trees classifier (spectral features plus laser line) reached **top-1 accuracy** = 88.8% and **top-5 accuracy** = 97.0% on the held-out test set, with 71 of 92 species classified perfectly. The five most-misclassified species—Mimetite, Pyrope, Andradite, Jamesonite and Diamond—err almost exclusively toward chemically or structurally near-identical relatives (garnet and apatite solid solutions, feldspar polymorphs), demonstrating that the residual errors reflect genuine spectral degeneracy rather than methodological artefacts. The complete per-spectrum top-1 predictions are released as `test_predictions.csv`.

Data, code and reproducibility

All spectra derive from the openly accessible RRUFF Project (unoriented Raman archives; downloaded 12 July 2026); archive URLs, sizes and SHA-256 checksums are recorded in the project provenance log. The analysis is fully scripted (download → parse → preprocess → filter/split → train/evaluate) with a fixed random seed (42), using the open scientific Python

stack—NumPy [24], SciPy [25], pandas [29], scikit-learn [23] and Matplotlib [30]. Key artefacts accompanying this report are the preprocessing summary, the filtering/split summary with per-species detail, the cross-validation leaderboard and final metrics (`modeling_summary.json`), the trained model, and the per-spectrum test predictions (`test_predictions.csv`).

A Retained mineral species (n = 92)

Actinolite, Aegirine, Albite, Alluaudite, Almandine, Analcime, Andalusite, Andradite, Anorthite, Axinite-(Fe), Baryte, Bastnäsite-(Ce), Beryl, Beudantite, Bobdownsite, Braunite, Calcite, Cerussite, Chabazite-Ca, Chalcanthite, Chalcopyrite, Chrysoberyl, Clinohumite, Conichalcite, Cordierite, Corundum, Cristobalite, Crocoite, Diamond, Diopside, Dolomite, Dravite, Elbaite, Enstatite, Eosphorite, Epidote, Fayalite, Fluorapatite, Fluorapophyllite-(K), Fluorcalciomicrolite, Fluorite, Forsterite, Grossular, Gruneriite, Hausmannite, Hematite, Heulandite-Ca, Hübnerite, Jamesonite, Kyanite, Livingstonite, Magnetite, Marialite, Metatorbernite, Microcline, Mimetite, Montebrasite, Muscovite, Natrolite, Opal, Orthoclase, Paravauxite, Perovskite, Phlogopite, Prehnite, Pyrite, Pyromorphite, Pyrope, Quartz, Rhodochrosite, Rhodonite, Rutile, Scheelite, Siderite, Sillimanite, Sodalite, Spessartine, Sphalerite, Spinel, Stephanite, Stibnite, Titanite, Topaz, Tourmaline, Tremolite, Triphylite, Triplite, Vanadinite, Vesuvianite, Vivianite, Wendwilsonite, Zircon.

References

[1] Daniel R. Neuville, Dominique de Ligny, and Grant S. Henderson. Advances in Raman spectroscopy applied to earth and material sciences. *Reviews in Mineralogy and Geochemistry*, 78(1):509–541, 2014. doi: 10.2138/rmg.2014.78.13.

[2] Alian Wang, John J. Freeman, and Bradley L. Jolliff. Understanding the Raman spectral features of phyllosilicates. *Journal of Raman Spectroscopy*, 46(10):829–845, 2015. doi: 10.1002/jrs.4680.

[3] Carles Arbiol and Graham D. Layne. Raman spectroscopy coupled with reflectance spectroscopy as a tool for the characterization of key hydrothermal alteration minerals in epithermal Au–Ag systems: Utility and implications for mineral exploration. *Applied Spectroscopy*, 75(12):1475–1496, 2021. doi: 10.1177/00037028211047869.

[4] Paul R. Bartholomew. The role of intensity and instrument sensitivity in Raman mineral identification. *Journal of Raman Spectroscopy*, 46(10):889–894, 2015. doi: 10.1002/jrs.4707.

[5] Barbara Lafuente, Robert T. Downs, Hexiong Yang, and Nate Stone. The power of databases: The RRUFF project. In Thomas Armbruster and Rosa Micaela Danisi, editors, *Highlights in Mineralogical Crystallography*, pages 1–29. De Gruyter (O), Berlin, Germany, 2015. doi: 10.1515/9783110417104-003.

[6] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, and M. D. Dyar. Machine learning tools for mineral recognition and classification from Raman spectroscopy. *Journal of Raman Spectroscopy*, 46(10):894–903, 2015. doi: 10.1002/jrs.4757.

[7] Pavel Jahoda, Igor Drozdovskiy, Samuel J. Payler, Leonardo Turchi, Loredana Bessone, and Francesco Sauro. Machine learning for recognizing minerals from multispectral data. *Analyst*, 146(1):184–195, 2021. doi: 10.1039/D0AN01483D.

[8] Jinchao Liu, Margarita Osadchy, Lorna Ashton, Michael Foster, Christopher J. Solomon, and Stuart J. Gibson. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst*, 142(21):4067–4074, 2017. doi: 10.1039/C7AN01371J.

[9] Xiancheng Sang, Ri-Gui Zhou, Yaochong Li, and Shengjun Xiong. One-dimensional deep convolutional neural network for mineral classification from Raman spectroscopy. *Neural Processing Letters*, 54(1):677–690, 2022. doi: 10.1007/s11063-021-10652-1.

[10] Xiaqiong Fan, Wen Ming, Huitao Zeng, Zhimin Zhang, and Hongmei Lu. Deep learning-based component identification for the Raman spectra of mixtures. *Analyst*, 144(5):1789–1798, 2019. doi: 10.1039/C8AN02212G.

[11] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

[12] Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15(1):1829, 2024. doi: 10.1038/s41467-024-46150-w.

[13] Tyler J. Bradshaw, Zachary Huemann, Junjie Hu, and Arman Rahmim. A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 5(4):e220232, 2023. doi: 10.1148/ryai.220232.

[14] Paul H. C. Eilers and Hans F. M. Boelens. Baseline correction with asymmetric least squares smoothing. Technical Report 1, Leiden University Medical Centre, Leiden, The Netherlands, 2005.

[15] Paul H. C. Eilers. A perfect smoother. *Analytical Chemistry*, 75(14):3631–3636, 2003. doi: 10.1021/ac034173t.

[16] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010. doi: 10.1039/b922045c.

[17] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):772–777, 1989. doi: 10.1366/0003702894202201.

[18] Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

[20] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[21] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

[22] 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. Curran Associates, Inc., 2017.

[23] 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.

[24] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[25] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[26] Boris A. Kolesov and Charles A. Geiger. Raman spectra of silicate garnets. *Physics and Chemistry of Minerals*, 25(2):142–151, 1998. doi: 10.1007/s002690050097.

[27] Steven Prawer and Robert J. Nemanich. Raman spectroscopy of diamond and doped diamond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 362(1824):2537–2565, 2004. doi: 10.1098/rsta.2004.1451.

[28] Andrea C. Ferrari and John Robertson. Interpretation of Raman spectra of disordered and amorphous carbon. *Physical Review B*, 61(20):14095–14107, 2000. doi: 10.1103/PhysRevB.61.14095.

[29] Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[30] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.