Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spec2Prop-InorgBench: An Open Raman--XRD Benchmark and AI-Assisted Screening Workflow for Inorganic Materials

Kumari Amrita
National Institute of Technology Kurukshetra
Himanshu Kumar
24sp06015@iitbbs.ac.in

Indian Institute of Technology Bhubaneswar

Research Article

Keywords: Raman spectroscopy, X-ray diffraction, Inorganic materials informatics, Spectral preprocessing, LightGBM-DART, Confidence-aware screening

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10574874/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Acknowledgments

The authors declare that no specific acknowledgments are applicable for this work. No funding was received to assist with the preparation of this manuscript.

Spec2Prop-InorgBench: An Open Raman–XRD Benchmark and AI-Assisted Screening Workflow for Inorganic Materials

Kumari Amrita[1] and Himanshu Kumar[2*]

[1]Department of Chemistry, National Institute of Technology Kurukshetra, Kurukshetra, Haryana, India.
[2*]School of Electrical and Computer Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha, India.

*Corresponding author(s). E-mail(s): 24sp06015@iitbbs.ac.in;
Contributing authors: 725101023@nitkkr.ac.in;

Abstract

Open Raman and X-ray diffraction (XRD) resources are valuable for inorganic-material characterization, but differences in file formats, spectral quality, and metadata make reproducible machine-learning evaluation difficult. We present Spec2Prop-InorgBench, an open benchmark that links raw spectra with inorganic material families and database-derived property labels. The benchmark contains 4,118 usable Raman spectra, 4,027 clean inorganic samples, 1,844 formula-matched property records, and 1,462 Raman–XRD linked samples. A runtime preprocessing pipeline converts each spectrum to a standardized 2,048-point representation. Family classification combines PCA features, chemistry-aware spectral descriptors, and train-only prototype-similarity features using a LightGBM-DART classifier, evaluated on an RRUFF-identifier-grouped 70/15/15 hold-out split. The nine-family screening model achieved 66.06% accuracy and 59.21% macro-F1 on the held-out test set, with 87.42% top-3 accuracy. Confidence thresholding at 0.8 increased accepted-sample accuracy to 89.23% at 43.05% coverage. For property prediction, descriptor-only XGBoost achieved macro-F1 scores of 75.46% for band-gap class, 83.18% for metal/non-metal classification, and 90.98% for formation-energy class. Spec2Prop-InorgBench provides reproducible benchmark tasks together with the confidence-aware Spec2Prop-Screen workflow for rapid candidate prioritization, intended to assist materials screening rather than replace phase-resolved identification or direct experimental property measurements.

Keywords: Raman spectroscopy, X-ray diffraction, Inorganic materials informatics, Spectral preprocessing, LightGBM-DART, Confidence-aware screening

1 Introduction

Raman spectroscopy is a widely used, non-destructive technique for probing the vibrational fingerprints of inorganic materials. In mineralogy and solid-state chemistry, Raman spectra reveal local bonding environments, anionic groups, lattice vibrations, hydration states, and structural disorder, making them valuable for distinguishing related inorganic families such as silicates, oxides, carbonates, sulfates, phosphates, sulfides, halides, and borates [1–4]. X-ray diffraction (XRD), in contrast, provides crystallographic and phase information through diffraction-peak positions and intensities. Together, Raman and XRD provide complementary vibrational and structural information for inorganic material characterization and phase screening [5–9].

Despite the availability of open Raman and XRD resources, converting raw instrument outputs into reproducible machine-learning-ready datasets remains challenging. Raw spectra often contain nonuniform spectral axes, missing or duplicated points, baseline drift, instrumental noise, and arbitrary intensity scaling, all of which can degrade downstream models if not handled systematically [10–13]. Inorganic materials introduce additional challenges, including polymorphism, mixed phases, hydration, compositional substitution, and overlapping vibrational bands.

Machine learning has become an important tool in materials informatics, enabling rapid screening of chemical composition, crystal structure, and materials properties from high-dimensional descriptors and database-derived representations [14–18]. Recent studies have also demonstrated its potential for Raman analysis, mineral classification, and diffraction-based structure prediction [7, 19–23]. However, many existing approaches rely on task-specific or proprietary datasets, begin with already processed spectra, focus on a single modality, or provide limited uncertainty reporting. An open benchmark that exposes the complete raw-file-to-model workflow while clearly distinguishing screening from definitive chemical identification remains valuable.

To address this need, we introduce **Spec2Prop-InorgBench**, a curated Raman–XRD benchmark for AI-assisted inorganic-material screening. The benchmark integrates open Raman and XRD resources with harmonized inorganic-family labels and formula-linked property classes. A runtime preprocessing workflow converts heterogeneous raw spectra into standardized inputs through parsing, cleaning, interpolation, baseline correction, smoothing, and normalization before feature extraction. The family-screening component, **Spec2Prop-Screen**, employs a LightGBM-DART classifier with global-shape, chemistry-aware, and train-only prototype-similarity features [24, 25]. Property-class baselines are evaluated separately using Random Forest, XGBoost, and neural descriptor/fusion models.

Spec2Prop-Screen is designed as a confidence-aware candidate-prioritization tool rather than a replacement for expert spectroscopic interpretation, crystallographic refinement, or laboratory confirmation. Likewise, the property targets are

formula-linked database classes rather than direct experimental measurements, and polymorph-specific ambiguity remains an inherent limitation.

The main contributions are:

- an open benchmark containing Raman spectra, optional XRD patterns, nine inorganic-family labels, and formula-linked property classes;
- a reproducible runtime preprocessing workflow that converts heterogeneous raw files into fixed 2,048-point representations;
- a 222-dimensional family-screening representation combining PCA global-shape features, chemistry-aware spectral descriptors, and train-only prototype-similarity features;
- grouped hold-out evaluation of classical, boosted-tree, and neural baselines for nine-family screening, including top-k, calibration, and reject-option analysis;
- descriptor, raw-spectrum, and PCA ablations for band-gap, metallicity, and formation-energy prediction; and
- an exploratory Raman–XRD dual-branch baseline reported without claiming superiority because a matched Raman-only tree baseline was unavailable on the linked subset.

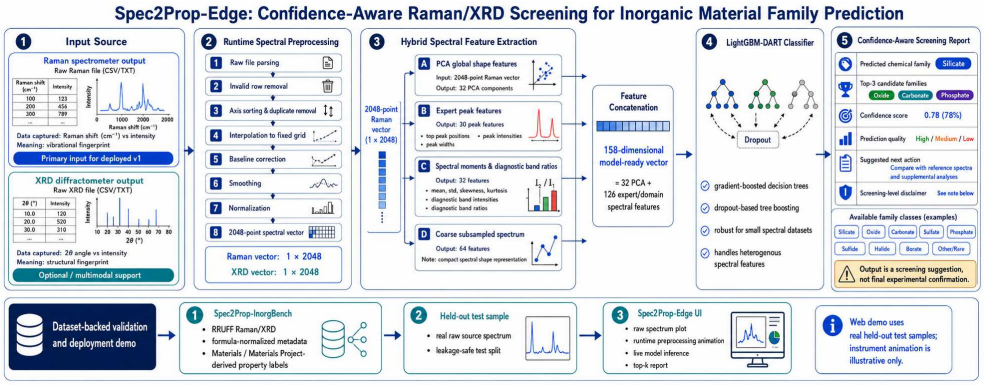


Fig. 1: Scope of Spec2Prop-InorgBench. Open Raman and XRD records are curated into benchmark subsets, standardized by a runtime preprocessing pipeline, and evaluated in three tasks: nine-family screening, formula-linked property prediction, and exploratory Raman–XRD fusion. Spec2Prop-Screen returns ranked family candidates and confidence scores for expert verification.

2 Methods

2.1 Open spectral and materials-property resources

Spec2Prop-InorgBench was constructed by integrating open spectral, crystallographic, and materials-property resources relevant to inorganic materials screening. Raman

spectra were used as the primary spectral modality because they encode vibrational information associated with local bonding environments, anionic groups, lattice modes, hydration, and structural disorder. X-ray diffraction (XRD) patterns were used as an optional complementary modality because diffraction peak positions and intensities encode crystallographic and phase-level information. The RRUFF database provided the main Raman spectral backbone and available XRD links for mineral samples [4, 26]. Crystallographic context and phase-related motivation were supported by open crystallographic resources and recent XRD machine-learning benchmarks [5–8]. Materials-property labels were linked using Materials Project and Matbench-derived annotations, including band-gap class, metal/non-metal class, and formation-energy class [14, 17, 18]. Table 1 summarizes the role of each data source.

Table 1: Open spectral, crystallographic, and materials-property resources used to construct Spec2Prop-InorgBench.

Source	**Data type**	**Main fields used**	**Linkage key**	**Role in benchmark**	**Status**
RRUFF Raman [4, 26]	Raman spectra	Raman shift, intensity, raw spectral files, and sample metadata	RRUFF ID / sample ID	Primary spectral backbone for Raman-based inorganic-family screening and feature extraction	Integrated
RRUFF XRD [4, 26]	Powder XRD patterns	2θ values, diffraction intensities, and linked sample records	RRUFF ID / sample ID	Optional multimodal branch for combining vibrational and crystallographic fingerprints	Integrated
RRUFF metadata [4]	Mineral chemistry	Mineral name, chemical formula, sample identifier, and chemistry notes	RRUFF ID / formula	Formula normalization and inorganic chemical-family label construction	Integrated
Materials Project [14]	Computed materials properties	Band gap, formation energy, metallicity-related information, and reduced formula	Normalized reduced formula	Source of materials-property annotations used for spectra-to-property classification	Integrated
Matbench [17, 18]	Benchmark property labels	Band-gap class, metal/non-metal class, and formation-energy-related labels	Normalized reduced formula	Standardized property-label source for downstream materials-informatics tasks	Integrated
MLROD	Auxiliary Raman / mineral resource	Raman/mineral records and mixture-related information	Dataset-specific sample identifiers	Auxiliary reference for robustness analysis and future extension; not used as the primary sample-level backbone	Auxiliary

Continued on next page

Table 1 (continued)

Source	Data type	Main fields used	Linkage key	Role in benchmark	Status
tmQM / tmQMg	Coordination-chemistry datasets	Transition-metal complex formulas and molecular/property metadata	Formula / complex identifier	Context resource for future transition-metal and coordination-chemistry expansion; not used as primary Raman/XRD training data	Auxiliary
Spec2Prop-InorgBench [27]	Curated benchmark	Cleaned Raman/XRD representations, family labels, property labels, and split metadata	Internal sample ID	Final released benchmark for reproducible inorganic spectra-to-property screening	Released

2.2 Construction of Spec2Prop-InorgBench

The benchmark was built through a multi-stage curation workflow designed to connect raw spectral files with chemically meaningful labels and downstream materials-property annotations. First, raw Raman records were collected and standardized at the sample level. Samples with unusable spectral files, missing essential metadata, or non-inorganic chemistry were removed. Chemical formulas were normalized to reduce inconsistencies caused by formatting differences, hydration notation, mixed occupancy, charge notation, or mineral-name duplication. Where possible, Raman records were linked to XRD patterns through shared mineral/sample identifiers. Property labels were assigned by normalized reduced-formula linkage to Materials Project/Matbench-derived entries. Because formula equality does not uniquely determine polymorph, phase, or measured property, these targets are treated as database-derived screening classes rather than phase-resolved experimental ground truth.

This curation process produced four reportable benchmark subsets. The full spectral subset contains 4,118 usable Raman spectra. After inorganic filtering and chemical-label harmonization, the clean inorganic subset contains 4,027 samples and serves as the main dataset for chemical-family classification. The property-matched subset contains 1,844 samples with usable materials-property labels. The Raman–XRD linked subset contains 1,462 samples with both Raman and XRD representations and is used for multimodal analysis. The reportable subset structure and subset definitions are summarized in Table 2.

2.3 Inorganic chemical-family label construction

Chemical-family labels were constructed from normalized chemical formulas and mineral metadata. The original chemistry labels were first organized into twelve inorganic groups, including silicates, oxides, sulfates, phosphates, sulfides, halides, carbonates, borates, elemental/native phases, telluride/selenide phases, intermetallic/alloy phases, and carbides. For model-level classification, rare or low-support groups were merged into an *Other/Rare* category to reduce label sparsity and improve screening robustness. The final fine-grained model uses nine families: Borate, Carbonate, Halide, Other/Rare, Oxide, Phosphate, Silicate, Sulfate, and Sulfide.

Table 2: Dataset subset statistics for Spec2Prop-InorgBench.

Subset	Samples	Raman	XRD	Property labels	Main use
Full spectra	4,118	Yes	Partial	Partial	Pretraining and exploration
Clean inorganic	4,027	Yes	Partial	Partial	Family classification
Property-matched	1,844	Yes	Partial	Yes	Spectra-to-property prediction
Raman–XRD linked	1,462	Yes	Yes	Partial	Multimodal fusion

This family-label design is chemically motivated. Silicates are dominated by Si–O framework and tetrahedral vibration modes; carbonates and phosphates contain characteristic anionic-group vibrations; oxides often show metal–oxygen lattice modes; sulfides and halides may exhibit lower-frequency or sparse Raman features; and borates contain B–O vibrational signatures. The *Other/Rare* class captures chemically valid but underrepresented groups that cannot support reliable independent model training. Table 3 defines the final chemical-family labels used in this study.

Table 3: Inorganic chemical-family definitions and spectral relevance.

Family	Chemical meaning	Representative spectral features	Model role
Silicate	Contains $[SiO_4]^{4-}$ units	Strong Si–O stretching (800–1200 cm^{-1})	Target class
Oxide	Contains O^{2-} with metals	M–O vibrations (broad and variable)	Target class
Carbonate	Contains $[CO_3]^{2-}$ units	Sharp C–O symmetric stretch (∼1090 cm^{-1})	Target class
Sulfate	Contains $[SO_4]^{2-}$ units	Sharp S–O symmetric stretch (∼980–1010 cm^{-1})	Target class
Phosphate	Contains $[PO_4]^{3-}$ units	P–O symmetric stretch (∼960 cm^{-1})	Target class
Sulfide	Contains S^{2-} with metals	Low-wavenumber M–S modes (< 500 cm^{-1})	Target class
Halide	Contains F^-, Cl^-, Br^-, or I^-	Very-low-wavenumber lattice modes	Target class
Borate	Contains $[BO_3]^{3-}$/$[BO_4]^{5-}$	B–O stretching (∼1300–1500 cm^{-1})	Target class
Other/Rare	Nitrates, tungstates, and other low-support groups	Family-dependent and variable	Target class

2.4 Runtime Raman and XRD spectral preprocessing

Raw Raman and XRD files are not directly suitable for machine-learning inference because different instruments and databases may use different sampling grids, axis ranges, intensity scales, and file formats. Therefore, we implemented a runtime preprocessing pipeline that converts raw instrument-style spectra into fixed-length

model-ready arrays. Raman spectra were standardized over the 100–4000 cm^{-1} range, while XRD patterns were standardized over the 5–90° 2θ range. Each spectrum was parsed into an axis–intensity representation, cleaned to remove invalid rows, sorted by spectral axis, deduplicated where repeated axis values occurred, interpolated to a fixed 2048-point grid, baseline-corrected, smoothed, and normalized.

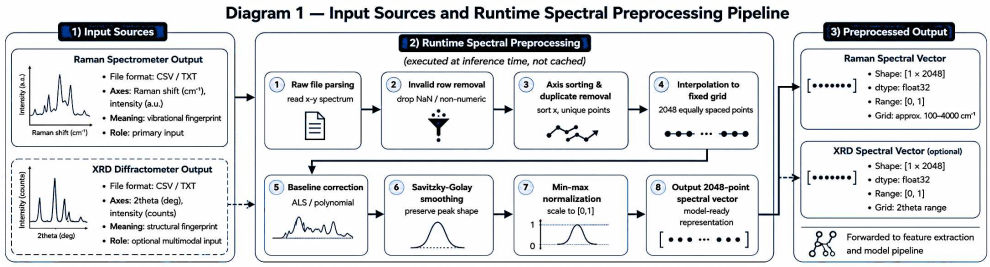


Fig. 2: Runtime Raman/XRD spectral preprocessing pipeline. Raw spectral files are parsed, cleaned, axis-sorted, deduplicated, interpolated to a 2048-point grid, baseline-corrected, smoothed, normalized, and converted into model-ready spectral vectors.

Baseline correction was used to reduce slowly varying fluorescence or background effects in Raman spectra [11, 12]. Savitzky–Golay smoothing was applied with a window length of 11 and polynomial order of 3 to suppress high-frequency noise while preserving local peak structure [10]. Intensity normalization was performed after smoothing so that models learned relative spectral patterns rather than absolute intensity scale differences. The same fixed-length representation was used for Raman and optional XRD arrays, enabling both single-modality and multimodal experiments. The preprocessing sequence is summarized in Table 4.

Table 4: Runtime preprocessing operations for raw Raman/XRD files.

Step	Operation	Chemistry/spectral rationale	Output
1	Raw-file parsing	Instruments export heterogeneous text formats	Axis–intensity array
2	Invalid-row removal	Hardware errors and malformed records can yield non-finite values	Cleaned array
3	Axis sorting	Wavenumber or 2θ axes can be reversed	Ascending spectral axis
4	Duplicate removal	Repeated detector coordinates can occur in source files	Unique axis values
5	Interpolation	Align spectra to a uniform modality-specific grid	2048-dimensional vector
6	Baseline correction	Reduce fluorescence or background drift [12]	Baseline-corrected vector
7	Savitzky–Golay smoothing	Suppress high-frequency instrumental noise while preserving peaks	Smoothed vector
8	Intensity normalization	Reduce instrument-dependent intensity-scale variation	Normalized vector

2.5 Chemistry-aware spectral feature extraction

Instead of relying only on end-to-end black-box representations, we designed chemistry-aware spectral descriptors that explicitly encode peak, band-window, shape, and similarity information. The starting representation for each Raman sample is a 2048-point normalized intensity vector. From this vector, we extracted a base 158-dimensional PCA+domain representation consisting of 32 PCA global-shape features and 126 expert spectral descriptors. The expert descriptors include peak positions, peak heights, peak widths, spectral moments, diagnostic band-window intensities and ratios, derivative statistics, spectral entropy, integrated band areas, and compact subsampled spectral-shape features.

The diagnostic band-window descriptors were designed to capture chemically meaningful spectral regions associated with inorganic families. For example, silicate-related windows encode Si–O bending and stretching behavior; carbonate and phosphate windows target characteristic anionic-group vibrations; oxide windows represent lower-frequency metal–oxygen lattice contributions; and O–H related regions help capture hydrated or hydroxylated phases. These features make the representation more interpretable than a purely learned embedding and allow tree-based models to exploit sparse but chemically meaningful spectral cues.

For confidence-aware screening, the base 158-dimensional feature vector was extended by 64 prototype-similarity features, producing a 222-dimensional PCA–domain–prototype representation. The prototypes and all related reference statistics were constructed exclusively from the training partition. This representation was used by the selected Spec2Prop-Screen classifier. Table 5 summarizes the feature groups.

Table 5: Feature groups used by the selected Spec2Prop-Screen family classifier.

Feature group	**Dimension**	**Train-only fitting**	**Role**
PCA global-shape features	32	Yes	Compact representation of broad spectral variation
Chemistry-aware domain descriptors	126	No learned fitting[a]	Peak statistics, spectral moments, diagnostic windows and ratios, derivative/entropy measures, integrated areas, and compact shape features
Prototype-similarity features	64	Yes	Similarity to reference patterns constructed exclusively from training samples
Total selected representation	**222**	–	PCA + domain + prototype representation used by LightGBM-DART

[a] Descriptor definitions are fixed, but any statistics, reference quantities, PCA transforms, scalers, and prototypes required by the pipeline are estimated from the training partition only.

2.6 Machine-learning models

The selected family-screening model is a LightGBM classifier using Dropout meets Multiple Additive Regression Trees (DART) boosting [24, 25]. LightGBM-DART was

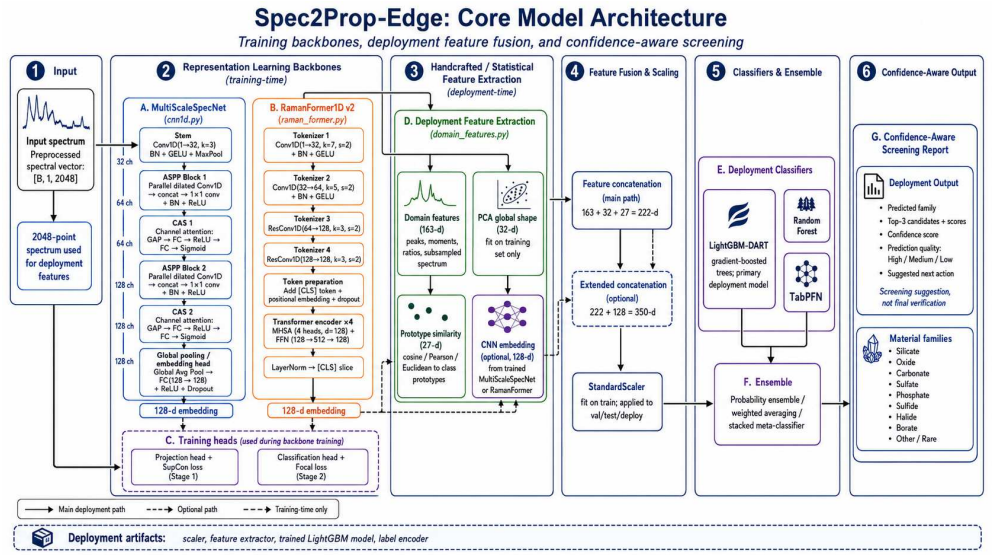


Fig. 3: Feature extraction and LightGBM-DART screening architecture. The 2048-point Raman vector is transformed into PCA global-shape features, expert spectral descriptors, and prototype-similarity features. These are concatenated into a screening-oriented 222-dimensional representation used by LightGBM-DART to predict top-k inorganic chemical-family candidates and confidence scores.

selected because gradient-boosted decision trees are well suited for heterogeneous tabular spectral descriptors, handle nonlinear interactions between sparse peak features and global-shape features, and remain effective for small-to-moderate scientific datasets where purely end-to-end neural models may overfit. The DART boosting strategy randomly drops trees during training, reducing over-specialization and improving robustness [25].

We benchmarked LightGBM-DART against Logistic Regression, support vector machines (SVM), k-nearest neighbours, Random Forest, XGBoost, and RamanFormer1D [28–32]. Property-class experiments included Random Forest and XGBoost with raw, PCA, and engineered-descriptor representations, together with DescriptorMLP, Spec2PropLite, and FusionSpec2PropNet. The exploratory multimodal experiment used a DualBranchRaman–XRDNet with separate modality encoders followed by feature fusion. Only models with final reportable test metrics are retained in the tables.

Table 6: Machine-learning models retained in the reported Spec2Prop-InorgBench evaluation.

Model	Input representation	Task	Purpose	Key notes
Logistic Regression	Raw or PCA-reduced Raman vector	Nine-family classification	Linear baseline	L2-regularized baseline for approximately linear separation
SVM	Raw or PCA-reduced Raman vector	Nine-family classification	Nonlinear baseline	RBF-kernel classifier
kNN	Raw or PCA-reduced Raman vector	Nine-family classification	Neighbourhood baseline	Tests local spectral similarity
Random Forest	Raw, PCA, or engineered descriptor representation	Family and property tasks	Non-boosted tree ensemble	Reports paired accuracy and macro-F1 from the same selected configuration
XGBoost	Raw, PCA, or engineered descriptor representation	Family and property tasks	Boosted-tree baseline	Strongest property-class baseline in the reported experiments
LightGBM-DART (PCA+Domain)	158 dimensions: 32 PCA + 126 domain descriptors	Nine-family classification	Compact ablation	Highest top-1 accuracy among the two LightGBM feature sets
LightGBM-DART (selected)	222 dimensions: PCA + domain + prototype	Ranked family screening	Spec2Prop-Screen model	Final checkpoint: DART boosting, 1,062 estimators, maximum depth 8, 66 leaves, balanced class weights, and seed 42
RamanFormer1D	2,048-point Raman vector	Nine-family classification	Neural sequence baseline	Attention-based end-to-end Raman model
DescriptorMLP	Engineered descriptor vector	Property-class prediction	Neural descriptor baseline	Multi-layer perceptron without the raw Raman vector
Spec2PropLite	Raman vector + descriptors	Property-class prediction	Lightweight fusion baseline	Compact learned spectral encoder fused with descriptors
FusionSpec2PropNet	Raman embedding + descriptor embedding	Multi-task property prediction	Neural fusion baseline	Jointly predicts band-gap, metallicity, and formation-energy classes
DualBranch Raman–XRDNet	Raman and XRD vectors, each 2,048 points	Exploratory nine-family classification	Multimodal feasibility baseline	Separate modality encoders followed by feature fusion; no matched strong tree baseline is claimed

2.7 RRUFF-ID-grouped hold-out evaluation

A fixed 70/15/15 train–validation–test split was used with random seed 42. Records were grouped by RRUFF identifier and stratified by chemical family and band-gap

class where applicable. This prevents records sharing an RRUFF identifier from crossing partitions, but it does not guarantee separation by normalized formula or mineral species; the evaluation is therefore described precisely as RRUFF-ID-grouped rather than universally leakage-free.

All PCA transformations, scalers, prototypes, and fitted model parameters were estimated using the training partition only and then applied to validation and test samples. Family classification was evaluated with accuracy, macro-F1, weighted-F1, balanced accuracy, top-k accuracy, calibration, and reject-option coverage. Property tasks were evaluated independently with accuracy and macro-F1. The manuscript reports one pre-specified grouped hold-out split rather than mean values over repeated group cross-validation, and this limitation is carried forward into the interpretation. The complete protocol is summarized in Table 7.

2.8 Mathematical formulation of spectral preprocessing

Let a raw Raman or XRD spectrum be represented as a set of axis–intensity pairs,

$$S_{\mathrm{raw}} = \{(x_i, y_i)\}_{i=1}^{M}, \tag{1}$$

where x_i denotes the Raman shift in cm^{-1} for Raman spectra or the diffraction angle 2θ in degrees for XRD patterns, and y_i is the corresponding measured intensity. The number of raw points M may vary across files because of differences in instruments, acquisition settings, or database formats. Therefore, the first objective of preprocessing is to transform each variable-length spectrum into a fixed-length model-ready vector.

Invalid entries were removed by retaining only finite axis and intensity values:

$$S_{\mathrm{clean}} = \{(x_i, y_i) \,|\, x_i, y_i \in \mathbb{R},\ \mathrm{isfinite}(x_i),\ \mathrm{isfinite}(y_i)\}\,. \tag{2}$$

This step prevents missing values, corrupted rows, or non-numeric entries from propagating into interpolation and machine-learning inference.

The cleaned spectra were sorted by the spectral axis, and duplicate axis values were merged by averaging their intensities:

$$\bar{y}(x) = \frac{1}{|\mathcal{I}(x)|} \sum_{i \in \mathcal{I}(x)} y_i, \tag{3}$$

where $\mathcal{I}(x) = i : x_i = x$ is the set of repeated measurements at the same spectral coordinate. This avoids unstable interpolation caused by duplicated x values.

Each spectrum was then interpolated onto a fixed grid of $N = 2048$ points. For Raman spectra, the grid was defined over $[100, 4000]$ cm^{-1}, while for XRD spectra it was defined over $[5, 90]^\circ$ 2θ:

$$x_j^* = x_{\min} + \frac{j}{N-1}(x_{\max} - x_{\min}), \qquad j = 0, 1, \ldots, N-1. \tag{4}$$

Linear interpolation was used to estimate the intensity at each grid point:

$$y^*(x_j^*) = y_a + \frac{y_b - y_a}{x_b - x_a}(x_j^* - x_a), \qquad x_a \le x_j^* \le x_b, \tag{5}$$

where (x_a, y_a) and (x_b, y_b) are the nearest neighboring raw spectral points surrounding x_j^*. Interpolation makes spectra comparable across samples and ensures that all models receive inputs with identical dimensionality.

To reduce slowly varying fluorescence or background contributions, baseline correction was applied. The estimated baseline b can be expressed as the solution of a penalized least-squares problem:

$$\hat{b} = \arg\min_b \left[\sum_{j=1}^{N} w_j (y_j^* - b_j)^2 + \lambda \| D^{(m)} b \|_2^2 \right], \tag{6}$$

where w_j are adaptive weights, λ controls baseline smoothness, and $D^{(m)}$ is an mth-order difference operator. The baseline-corrected spectrum is then

$$y_j^{\mathrm{bc}} = y_j^* - \hat{b}_j. \tag{7}$$

This step is especially important for Raman spectra, where fluorescence background can obscure chemically meaningful peaks.

High-frequency noise was reduced using Savitzky–Golay smoothing. Around each point, a local polynomial of degree $p = 3$ was fitted within a moving window of length $2h + 1 = 11$:

$$\hat{y}^{\mathrm{bc}}(t) = \sum_{r=0}^{p} a_r t^r, \tag{8}$$

and the smoothed value was taken at the center of the window:

$$y_j^{\mathrm{sg}} = \sum_{k=-h}^{h} c_k y_{j+k}^{\mathrm{bc}}. \tag{9}$$

Here, c_k are the Savitzky–Golay convolution coefficients. This smoothing operation suppresses random noise while preserving peak position, width, and shape better than simple moving-average smoothing.

Finally, each spectrum was intensity-normalized using maximum-intensity normalization:

$$y_j^{\mathrm{norm}} = \frac{y_j^{\mathrm{sg}}}{\max_{1 \le l \le N} |y_l^{\mathrm{sg}}| + \epsilon}, \tag{10}$$

where ϵ is a small constant to avoid division by zero for nearly flat spectra. The final preprocessed spectral vector is

$$\mathbf{s} = [y_1^{\mathrm{norm}}, y_2^{\mathrm{norm}}, \ldots, y_N^{\mathrm{norm}}]^\top \in \mathbb{R}^{2048}. \tag{11}$$

This normalization reduces the effect of arbitrary instrument-dependent intensity scaling and allows the model to focus on relative spectral shape and peak structure.

2.9 Mathematical formulation of chemistry-aware spectral descriptors

From each normalized Raman vector $\mathbf{s} \in \mathbb{R}^{2048}$, we extracted chemically interpretable descriptors representing peak structure, spectral distribution, diagnostic band responses, and reference-like similarity.

Peak-based descriptors were obtained by identifying local maxima with sufficient prominence and spacing. For the kth selected peak at index p_k, the normalized peak position, peak height, and relative width were encoded as

$$f_{p,k} = \frac{p_k}{N}, \qquad f_{h,k} = s_{p_k}, \qquad f_{w,k} = \frac{w_k}{N}, \tag{12}$$

where $N = 2048$ and w_k is the full width at half maximum estimated from the local peak profile. Peak descriptors are chemically useful because inorganic families often exhibit characteristic Raman bands associated with Si–O, M–O, CO_3^{2-}, PO_4^{3-}, SO_4^{2-}, B–O, O–H, or lattice vibration modes.

Global spectral moments were computed to summarize the intensity distribution:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} s_j, \tag{13}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(s_j - \mu)^2 + \epsilon}, \tag{14}$$

$$\text{Skew} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{s_j - \mu}{\sigma}\right)^3, \tag{15}$$

$$\text{Kurt} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{s_j - \mu}{\sigma}\right)^4. \tag{16}$$

These moments capture overall spectral sharpness, asymmetry, and intensity concentration, which are useful for distinguishing sharp crystalline spectra from broader or noisier patterns.

For each diagnostic Raman window $B_m = [l_m, u_m]$, the mean band intensity and integrated band area were computed as

$$\bar{s}_{B_m} = \frac{1}{|B_m|}\sum_{j \in B_m} s_j, \tag{17}$$

$$A_{B_m} = \sum_{j \in B_m} s_j \Delta x. \tag{18}$$

Band-ratio descriptors were then defined as

$$R_{m,n} = \frac{\bar{s}_{B_m} + \epsilon}{\bar{s}_{B_n} + \epsilon}. \tag{19}$$

These ratios are useful because relative intensity across chemically meaningful spectral windows can be more robust than absolute intensity. For example, silicate, carbonate, phosphate, sulfate, oxide, and hydrated phases can show different relative responses across characteristic Raman regions.

A compact spectral-shape descriptor was obtained by subsampling the normalized spectrum with stride $q = 32$:

$$\mathbf{s}_{\text{sub}} = [s_1, s_{1+q}, s_{1+2q}, \ldots, s_{1+63q}]^\top \in \mathbb{R}^{64}. \tag{20}$$

This retains coarse spectral morphology while reducing dimensionality, making it useful for tree-based models.

Global spectral-shape information was additionally encoded using principal component analysis (PCA). Let $\boldsymbol{\mu}_{\text{train}}$ be the training-set mean spectrum and $\mathbf{W}_{32}$ be the first 32 principal components fitted only on the training split. The PCA feature vector is

$$\mathbf{z}_{\text{PCA}} = \mathbf{W}_{32}^\top (\mathbf{s} - \boldsymbol{\mu}_{\text{train}}) \in \mathbb{R}^{32}. \tag{21}$$

Fitting PCA only on the training set avoids information leakage from validation or test samples.

For the final deployment model, prototype-similarity descriptors were added. For each chemical family c, a training-set prototype spectrum was computed as

$$\boldsymbol{\pi}_c = \frac{1}{|\mathcal{T}_c|} \sum_{i \in \mathcal{T}_c} \mathbf{s}_i, \tag{22}$$

where $\mathcal{T}_c$ is the set of training samples belonging to class c. Cosine similarity between a test spectrum and each class prototype was then computed as

$$\text{sim}_c(\mathbf{s}, \boldsymbol{\pi}_c) = \frac{\mathbf{s}^\top \boldsymbol{\pi}_c}{\|\mathbf{s}\|_2 \, \|\boldsymbol{\pi}_c\|_2 + \epsilon}. \tag{23}$$

Prototype similarity provides a reference-library-like signal that complements peak and band-window descriptors during top-k candidate screening.

The final deployment feature vector is expressed as

$$\mathbf{x}_{\text{deploy}} = [\mathbf{z}_{\text{PCA}}; \mathbf{x}_{\text{domain}}; \mathbf{x}_{\text{proto}}] \in \mathbb{R}^{222}, \tag{24}$$

where $\mathbf{z}_{\text{PCA}} \in \mathbb{R}^{32}$, $\mathbf{x}_{\text{domain}}$ contains expert spectral descriptors, and $\mathbf{x}_{\text{proto}}$ contains prototype-similarity features. This representation combines global spectral shape, chemically interpretable Raman descriptors, and reference-like class similarity for confidence-aware inorganic material screening.

Table 7: Evaluation protocol used for the reported experiments.

Task	**Subset**	**Split and fitting rule**	**Reported metrics**	**Interpretive limit**
Nine-family screening	Clean inorganic (N = 4,027)	Fixed 70/15/15 split, grouped by RRUFF identifier; PCA, scalers, prototypes, and models fitted on training data only	Accuracy, macro-F1, weighted-F1, balanced accuracy, top-k, confusion matrix	One grouped hold-out split; not repeated group cross-validation
Property-class prediction	Formula-matched (N = 1,844)	Fixed 70/15/15 split with the same train-only transformation rule	Accuracy and macro-F1 for each task and representation	Formula-linked database classes; not phase-resolved experimental targets
Raman–XRD fusion	Linked pairs (N = 1,462)	Fixed grouped split with paired Raman and XRD records kept together	Accuracy and macro-F1 for the dual-branch model	Feasibility result; no superiority claim without a matched strong comparator
Confidence-aware screening	Family-model test set (N = 604)	Thresholds applied after prediction on the untouched test set	Top-k, calibration, coverage, accepted accuracy/F1, abstention	Thresholds trade coverage for reliability and require external calibration before deployment
Demonstration interface	Held-out benchmark samples	Uses trained artifacts and selected test samples only	Ranked candidates, confidence tier, screening recommendation	Dataset-backed demonstration, not live instrument or hardware validation

2.10 Confidence-aware deployment workflow

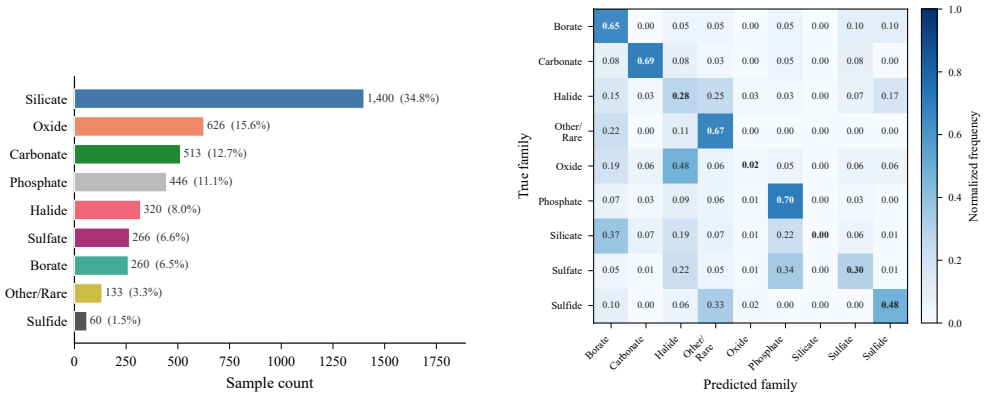
Spec2Prop-Screen was designed as a confidence-aware screening workflow rather than a final material-identification system. During inference, a raw Raman spectrum is parsed and preprocessed at runtime, transformed into a fixed 2048-point vector, converted into the 222-dimensional PCA–domain–prototype representation, and passed to the trained LightGBM-DART model. The model returns the predicted family, top-k candidate families, class probabilities, and a confidence score. Predictions are assigned to high-, medium-, or low-confidence tiers using predefined probability thresholds.

The confidence-aware design is important because spectral ambiguity, mixed phases, class imbalance, and instrument variation can produce uncertain predictions. Instead of forcing every sample into a single final class, Spec2Prop-Screen reports candidate families for scientist verification and can abstain on low-confidence samples. This makes the workflow suitable for early-stage candidate prioritization, batch triage, and preliminary material-family interpretation, while final confirmation remains dependent on expert spectroscopic analysis, crystallographic validation, and laboratory verification [33].

3 Results and Discussion

3.1 Benchmark composition and held-out class support

Spec2Prop-InorgBench contains 4,118 usable Raman spectra, 4,027 clean inorganic samples, 1,844 formula-matched property records, and 1,462 Raman–XRD linked records. The fine-family task covers Borate, Carbonate, Halide, Other/Rare, Oxide, Phosphate, Silicate, Sulfate, and Sulfide. The fixed family-model test set contains 604 samples and is imbalanced: the class supports range from 210 Silicate samples to nine Other/Rare samples. Macro-F1 is therefore more informative than accuracy alone.



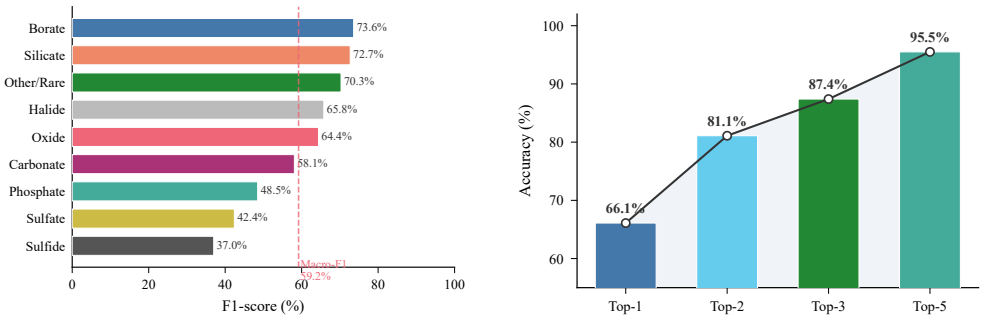
(a) Held-out class support

(b) Normalized confusion matrix

Fig. 4: Held-out evaluation data and predictions for the selected LightGBM-DART family model. (a) Exact class support in the 604-sample test set. (b) Row-normalized confusion matrix computed from the released test labels and class-probability outputs.

3.2 Raman-based inorganic-family screening

Table 8 compares paired accuracy and macro-F1 values from the same selected configuration for each model. Tree ensembles using compact spectral representations were more effective than the end-to-end RamanFormer1D baseline at the present dataset scale. XGBoost with 64 PCA features reached 64.07% accuracy and 52.55% macro-F1. The 158-dimensional LightGBM-DART ablation achieved the highest top-1 accuracy, 67.22%, with 56.57% macro-F1. The selected 222-dimensional PCA–domain–prototype model achieved a slightly lower accuracy of 66.06% but the highest macro-F1 of 59.21%; it was retained because the study emphasizes class-balanced ranking and confidence-aware screening rather than top-1 accuracy alone.



(a) Per-family F1

(b) Top-k accuracy

Fig. 5: Detailed performance of the selected family-screening model. (a) Per-family F1 mapped using the released deployment label encoder. (b) Top-k accuracy on the held-out test set.

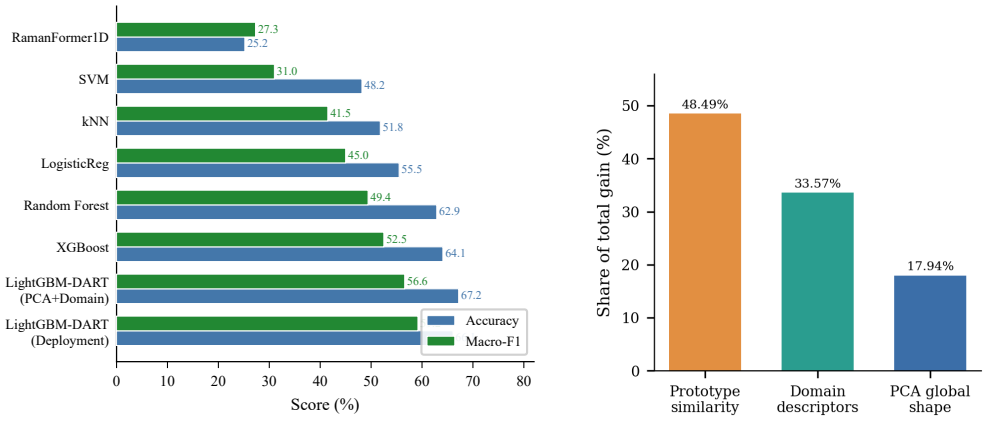
Table 8: Nine-family classification on the clean inorganic benchmark. Each accuracy/macro-F1 pair is taken from one configuration rather than combining the best values from different settings.

Model	Selected representation	Accuracy (%)	Macro-F1 (%)	Selection note
Logistic Regression	Raw Raman	55.46	45.04	Best macro-F1 setting for this model
SVM	PCA-128 Raman	48.18	31.04	Best macro-F1 setting for this model
kNN	PCA-64 Raman	51.32	41.49	Best macro-F1 setting for this model
Random Forest	PCA-64 Raman	61.92	49.43	Best macro-F1 setting for this model
XGBoost	PCA-64 Raman	64.07	52.55	Strongest non-LightGBM family baseline
LightGBM-DART	PCA+domain, 158 dimensions	**67.22**	56.57	Highest top-1 accuracy
LightGBM-DART	PCA+domain+prototype, 222 dimensions	66.06	**59.21**	Selected Spec2Prop-Screen model; best macro-F1 and available confidence/top-k evaluation
RamanFormer1D	2,048-point Raman vector	25.17	27.34	End-to-end neural sequence baseline

Note: Macro-F1 is emphasized because class support is strongly imbalanced. The selected 222-dimensional model was chosen for balanced family-level behaviour and confidence-aware ranking rather than for the highest top-1 accuracy alone.

The selected model achieved 58.74% balanced accuracy and 65.80% weighted-F1. Its candidate-ranking performance was stronger than its top-1 score: top-2, top-3, and top-5 accuracies were 81.13%, 87.42%, and 95.53%, respectively. This distinction is practically relevant for screening because a ranked shortlist can be checked against reference spectra and complementary measurements.

The strongest per-family F1 values were obtained for Silicate (73.59%), Carbonate (72.73%), and Phosphate (70.34%). The lowest values were observed for Other/Rare (37.04%), Borate (42.42%), and Halide (48.48%). The Other/Rare result must be interpreted cautiously because that class contains only nine held-out samples and is chemically heterogeneous. These results show why aggregate accuracy alone would overstate performance on low-support families.



(a) Family-model comparison

(b) LightGBM gain by feature group

Fig. 6: Benchmark comparison and model interpretation. (a) Accuracy and macro-F1 for the paired model configurations reported in Table 8. (b) Grouped LightGBM gain importance from the released 222-feature checkpoint. Prototype-similarity, domain, and PCA groups account for 48.49%, 33.57%, and 17.94% of total gain, respectively. Gain importance is associative and is not a causal explanation.

3.3 Confidence calibration and reject-option screening

The selected model produced mean and median maximum-class probabilities of 0.7159 and 0.7399. The reliability bins yield an expected calibration error of approximately 5.77%. The model is reasonably aligned at high confidence, while several middle-confidence bins are mildly overconfident. Consequently, probability thresholds should be validated again before use on a new instrument or laboratory distribution.

The reject-option analysis quantifies the coverage–reliability trade-off. At a threshold of 0.5, the model retained 79.64% of samples and achieved 73.80% accepted-sample accuracy. At 0.7, coverage decreased to 55.96%, while accuracy and macro-F1 increased to 83.14% and 74.69%. At 0.8, the model retained 43.05% of samples and achieved 89.23% accuracy and 81.46% macro-F1. These values support selective triage, but they do not remove the need for expert verification of accepted samples.

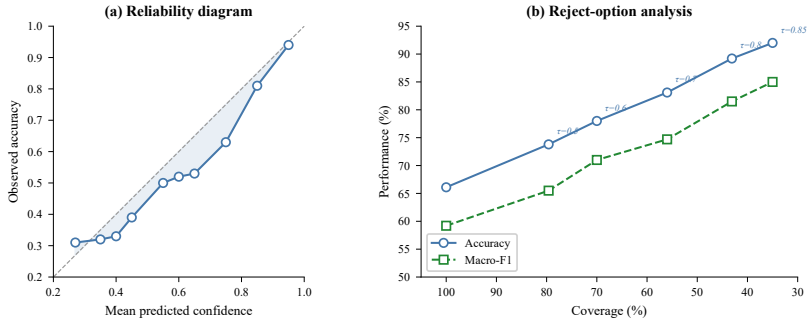


Fig. 7: Confidence analysis using the released deployment outputs. (a) Reliability diagram with an expected calibration error of approximately 5.77%. (b) Reject-option performance at thresholds 0.5–0.8, showing improved accepted-sample performance as coverage decreases.

3.4 Grouped feature importance and chemical interpretation

The released LightGBM checkpoint permits a reproducible gain-based importance analysis even though sample-level SHAP values were not included in the result package. Aggregating the 222 columns by construction shows that prototype-similarity features contribute 48.49% of total gain, chemistry-aware domain descriptors contribute 33.57%, and PCA global-shape features contribute 17.94%. This supports the design choice of combining reference-like similarity with explicit peak, band-window, moment, and shape information. It does not establish that any feature group is causally responsible for a chemical assignment.

Chemically, domain descriptors are motivated by differences in anion-related vibrations, metal–oxygen modes, hydration-related bands, peak widths, and global spectral morphology. However, the present result package does not contain sample-level explanations that would justify claims about a specific Raman band driving an individual prediction. Chemistry-reviewed case studies and local explanations are therefore reserved for future validation.

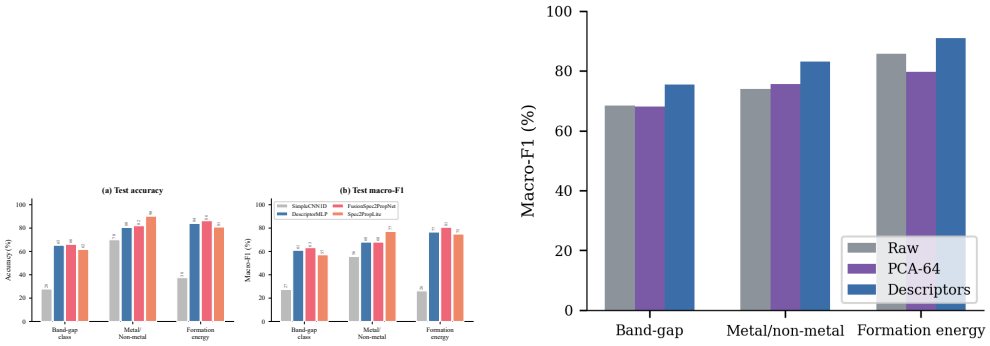
3.5 Formula-linked property-class prediction

Table 9 reports all retained property models rather than only the winning row for each task. Descriptor-only XGBoost was strongest across all three targets, achieving 79.06% accuracy and 75.46% macro-F1 for band-gap class, 93.50% accuracy and 83.18% macro-F1 for metal/non-metal classification, and 92.78% accuracy and 90.98% macro-F1 for formation-energy class. Random Forest with the same descriptor representation was the second strongest tree baseline. The neural descriptor and fusion models were competitive on some tasks but did not exceed descriptor-only XGBoost.

Table 9: Property-class prediction on the formula-matched subset ($N = 1{,}844$). Each task cell reports accuracy/macro-F1 (%).

Model	Representation	Band-gap class	Metal/non-metal	Formation-energy class
Random Forest	Engineered descriptors	76.53 / 71.93	93.14 / 82.50	92.06 / 89.76
XGBoost	Engineered descriptors	**79.06 / 75.46**	**93.50 / 83.18**	**92.78 / 90.98**
DescriptorMLP	Engineered descriptors	65.34 / 61.03	80.51 / 68.04	84.12 / 76.63
FusionSpec2PropNet	Raman + descriptors	66.06 / 63.21	81.95 / 67.97	86.28 / 80.74
Spec2PropLite	Raman + descriptors	61.73 / 57.10	90.25 / 77.08	80.87 / 74.93

Interpretation: The targets are formula-linked database classes. They should be treated as screening labels and not as direct phase-resolved experimental property measurements.



(a) Retained property models (b) XGBoost representation ablation

Fig. 8: Property-class results on the fixed formula-matched hold-out split. (a) Macro-F1 of retained tree and neural models. (b) XGBoost macro-F1 using raw Raman, PCA-64, and engineered-descriptor representations. The descriptor representation is strongest for all three tasks.

The property results should be interpreted as prediction of database-derived classes associated through normalized formula. They do not demonstrate that the model directly measures a band gap, metallic state, or formation energy from a spectrum, because one formula can correspond to multiple phases or polymorphs with different properties. A phase-resolved matching audit is required before making stronger spectra-to-property claims.

3.6 Exploratory Raman–XRD analysis

The Raman–XRD linked subset contains 1,462 samples. A DualBranchRaman–XRDNet using separate 2,048-point modality encoders achieved 44.75% accuracy and

36.60% macro-F1 on its fixed held-out split. This result demonstrates that the paired subset and dual-input training pipeline are operational. It is not presented as evidence that Raman–XRD fusion is superior, because a tuned LightGBM or XGBoost Raman-only comparator was not evaluated on the identical linked subset with the same split and feature budget.

3.7 Spec2Prop-Screen demonstration

Spec2Prop-Screen packages runtime preprocessing, 222-dimensional feature construction, LightGBM-DART inference, ranked family candidates, and confidence reporting. Figure 9 shows a dataset-backed interface using a held-out sample. The interface is a demonstration of the screening workflow, not live instrument acquisition or a hardware benchmark.

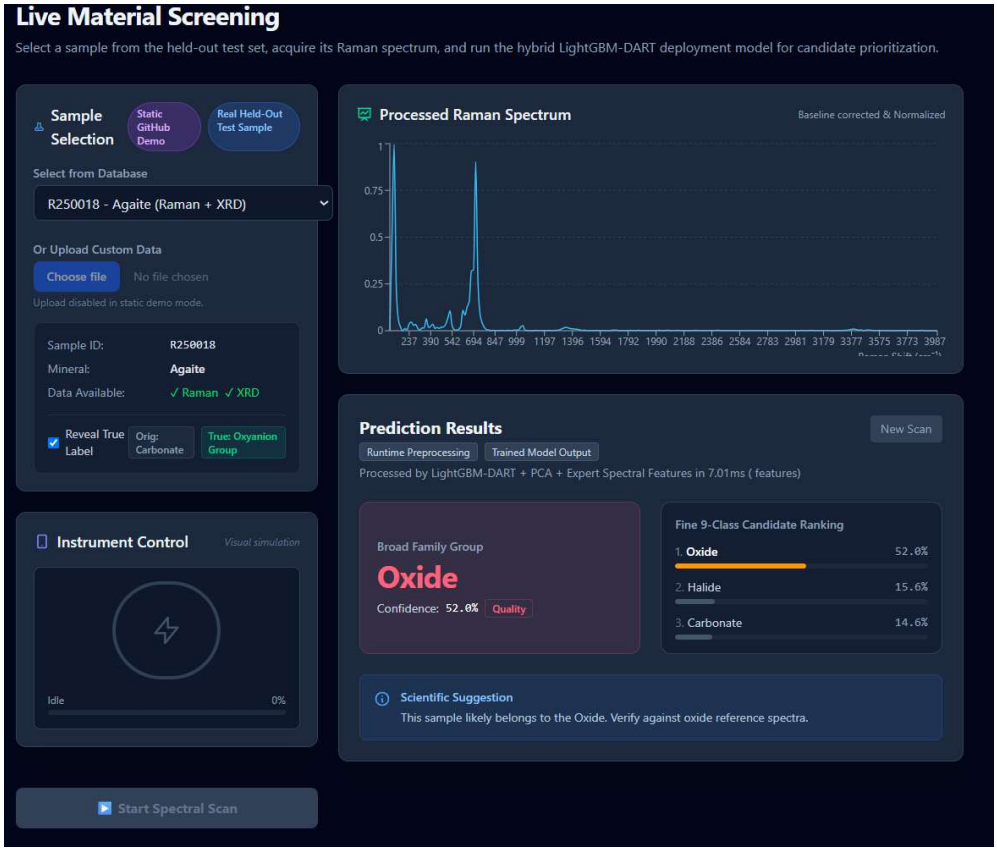


Fig. 9: Spec2Prop-Screen demonstration on a held-out benchmark sample. The interface shows the processed Raman spectrum, ranked family candidates, confidence, and a verification-oriented suggestion. The output is a screening recommendation rather than final experimental confirmation.

3.8 Limitations and future directions

The present study establishes a reproducible benchmark for screening inorganic material families and database-linked property classes; however, its conclusions are limited to the evaluated datasets and experimental protocol. The current partitioning strategy groups spectra by RRUFF identifier, thereby preventing identical records from appearing across splits. A stricter assessment based on normalized formula or mineral species, together with repeated grouped cross-validation, will be required to quantify generalization to chemically unseen materials and provide confidence intervals across multiple splits.

The property labels are obtained through normalized-formula linkage to materials databases. Because identical compositions may correspond to different polymorphs, hydration states, or substituted phases, these experiments should be interpreted as formula-linked property-class screening rather than direct experimental property determination. Future releases will incorporate phase-aware matching, candidate database identifiers, and confidence scores for ambiguous formula matches.

The Raman–XRD experiment demonstrates the feasibility of multimodal learning on the linked subset, but a matched Raman-only comparator using the same samples and feature budget is still required for a definitive assessment of multimodal benefit. Future work will therefore focus on controlled modality ablations.

4 Conclusions

Spec2Prop-InorgBench links open Raman records, optional XRD patterns, harmonized inorganic-family labels, and formula-linked materials-property classes through a reproducible raw-file-to-model workflow. On the fixed RRUFF-ID-grouped hold-out split, the selected LightGBM-DART family model achieved 66.06% accuracy, 59.21% macro-F1, and 87.42% top-3 accuracy. Confidence-based abstention increased accepted-sample accuracy to 89.23% at a threshold of 0.8, although coverage decreased to 43.05%. For the three database-derived property tasks, descriptor-only XGBoost produced the strongest reported macro-F1 values: 75.46% for band-gap class, 83.18% for metal/non-metal classification, and 90.98% for formation-energy class.

These results establish practical benchmark baselines and support ranked candidate screening, not definitive phase identification or direct experimental property measurement. Spec2Prop-Screen therefore returns top-k candidates and confidence information for expert review. The next validation priorities are formula- and mineral-grouped cross-validation, phase-resolved property matching, matched multimodal ablations, cross-instrument testing, and hardware measurements.

References

[1] Farmer, V.C. (ed.): The Infrared Spectra of Minerals. Mineralogical Society, ??? (1974)

[2] McMillan, P.: A Raman spectroscopic study of glasses in the system CaO–MgO–Al2O3–SiO2. American Mineralogist **69**, 622–644 (1984)

[3] Nakamoto, K.: Infrared and Raman Spectra of Inorganic and Coordination Compounds, 6th edn. Wiley, ??? (2009)

[4] Lafuente, B., Downs, R., Yang, H., Stone, N.: The power of databases: The RRUFF project, pp. 1–29. Walter de Gruyter GmbH, ??? (2016). https://doi.org/10.1515/9783110417104-003

[5] Gražulis, S., Chateigner, D., Downs, R.T., Yokochi, A.F.T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., Le Bail, A.: Crystallography open database – an open-access collection of crystal structures. Journal of Applied Crystallography **42**, 726–729 (2009) https://doi.org/10.1107/S0021889809016690

[6] Gražulis, S., Daskevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., Le Bail, A.: Crystallography open database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research **40**(D1), 420–427 (2012) https://doi.org/10.1093/nar/gkr900

[7] Guo, G., Goldfeder, J., Lan, L., Ray, A., Yang, A.H., Chen, B., Billinge, S.J.L., Lipson, H.: Towards end-to-end structure determination from X-ray diffraction data using deep learning. npj Computational Materials **10**, 209 (2024) https://doi.org/10.1038/s41524-024-01401-8

[8] Rincón, S., González, G., Macías, M.A., Arbeláez, P.: A new benchmark for machine learning applied to powder X-ray diffraction. Scientific Data **12**, 1186 (2025) https://doi.org/10.1038/s41597-025-05534-3

[9] Guo, G., Saidi, T., Terban, M., Valsecchi, M., Billinge, S.J.L., Lipson, H.: Ab Initio Structure Solutions from Nanocrystalline Powder Diffraction Data (2024)

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

[11] Zhang, Z., Chen, S., Liang, Y.: Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst **135**, 1138–1146 (2010) https://doi.org/10.1039/B922045C

[12] Baek, S.-J., Park, A., Ahn, Y.-J., Choo, J.: Baseline correction using asymmetrically reweighted penalized least squares smoothing. Analyst **140**, 250–257 (2015) https://doi.org/10.1039/C4AN01061B

[13] Rinnan, A., Berg, F., Engelsen, S.B.: Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry **28**(10), 1201–1222 (2009) https://doi.org/10.1016/j.trac.2009.07.007

[14] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K.A.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Materials **1**(1), 011002 (2013) https://doi.org/10.1063/1.4812323

[15] Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K., Snyder, G.J., Foster, I., Jain, A.: Matminer: An open source toolkit for materials data mining. Computational Materials Science **152**, 60–69 (2018) https://doi.org/10.1016/j.commatsci.2018.05.018

[16] Choudhary, K., Garrity, K.F., Reid, A.C.E., DeCost, B., Biacchi, A.J., Hight Walker, A.R., Trautt, Z., Hattrick-Simpers, J., Kusne, A.G., Centrone, A., Davydov, A., Jiang, J., Pachter, R., Cheon, G., Reed, E., Agrawal, A., Qian, X., Sharma, V., Zhuang, H.L., Kalish, I., Sivaraman, G., Cross, N., Dacek, S., Sorensen, C., Green, M.L., Persson, K.A., Ceder, G., Tavazza, F.: The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj Computational Materials **6**, 173 (2020) https://doi.org/10.1038/s41524-020-00440-1

[17] Dunn, A., Wang, Q., Ganose, A., Dopp, D., Jain, A.: Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Computational Materials **6**, 138 (2020) https://doi.org/10.1038/s41524-020-00406-3

[18] Riebesell, J., Goodall, R.E.A., Benner, P., Chiang, Y., Deng, B., Ceder, G., Asta, M., Lee, A.A., Jain, A., Persson, K.A.: A framework to evaluate machine learning crystal stability predictions. Nature Machine Intelligence **7**(6), 836–847 (2025) https://doi.org/10.1038/s42256-025-01055-1

[19] Dai, Y., Liu, Z., Zhao, S.: Fusion of laser-induced breakdown spectroscopy and Raman spectroscopy for mineral identification based on machine learning. Molecules **29**(14), 3317 (2024) https://doi.org/10.3390/molecules29143317

[20] Eranti, A., Tewari, Y., Palacios, R., Gupta, A.: Raman spectroscopy pre-trained encoder: A self-supervised learning approach for data-efficient domain-independent spectroscopy analysis. IEEE Access **14**, 40311–40327 (2026)

[21] Zhang, Y., Sun, J., Zhao, D., Lu, T., Liang, R., Sun, H., Liu, B., Wang, H., Zhang, H., Chen, P.: PhysRaman-Net: A physics-informed neural network for cross-device Raman spectral analysis. IEEE Photonics Journal **18**, 1–10 (2026)

[22] Liu, Y., *et al.*: Recent advances in Raman spectral classification with machine learning. Sensors **26**(1), 341 (2026) https://doi.org/10.3390/s26010341

[23] Li, Q., Jiao, R., Wu, L., Zhu, T., Huang, W., Jin, S., Liu, Y., Weng, H., Chen, X.: PXRDGen: Powder Diffraction Crystal Structure Determination Using

Generative Models (2024)

[24] 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, vol. 30, pp. 3146–3154 (2017)

[25] Rashmi, K.V., Gilad-Bachrach, R.: DART: Dropouts meet multiple additive regression trees. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pp. 489–497 (2015)

[26] Downs, R.T.: The RRUFF project: An integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. In: Program and Abstracts of the 19th General Meeting of the International Mineralogical Association, pp. 03–13 (2006)

[27] Kumar, H., Amrita, K.: Spec2Prop-InorgBench. Kaggle (2026). https://doi.org/10.34740/KAGGLE/DSV/17271313 . https://www.kaggle.com/dsv/17271313

[28] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995) https://doi.org/10.1007/BF00994018

[29] Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3), 27–12727 (2011) https://doi.org/10.1145/1961189.1961199

[30] Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[31] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[32] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[33] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 1321–1330 (2017)

Statements and Declarations

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests The authors declare no competing interests.

Author Contributions H. Kumar developed the architecture model. K. Amrita generated the idea for dataset formation and performed the dataset collection. Both authors conducted the literature survey and contributed to the preparation of the manuscript.

Data Availability The Spec2Prop-InorgBench dataset is publicly available on Kaggle at https://doi.org/10.34740/KAGGLE/DSV/17271313.

Code Availability The reproducibility package accompanying this submission contains the figure-generation scripts and machine-readable values used to generate the revised tables and figures. The complete preprocessing, fixed-split evaluation, model-training, and inference code will be archived in a versioned public repository. A browser-based demonstration of the Spec2Prop-Inorg screening workflow is publicly available at https://hkphimanshukumar321.github.io/Spec2Prop-Inorg/.