PREPRINT · VERSION 1.1 · 3 SEPTEMBER 2026 · CLAIM STATUS: MODELLED

A Deterministic Synthetic Corpus and Learned Inversion for the Amorphous-Carbon Raman Inverse Problem, Admitted Through a Signed Release Gate

Ryan James York

SSX360 Corp., Honolulu, Hawai'i, USA · mission@ssx360.com · https://ryanjamesyork.com

ORCID iD https://orcid.org/0009-0007-5979-7949

Compiled edition of Findings 2026.2, 2026.3 and 2026.4 of *ryanjamesyork.com*, whose Markdown sources are published with SHA-256 digests.

Abstract

The disorder ratio $I(\mathrm{D})/I(\mathrm{G})$ of a carbon Raman spectrum is a continuous but non-invertible function of crystallite size: on the trajectory of Ferrari and Robertson it rises as $1/L_a$ toward the turnover at $L_a = 20$ Å and falls as L_a^2 beyond it, so that a single ratio below 2.2 corresponds to two crystallite sizes. Any pipeline that inverts a spectrum must resolve that branch, and it cannot be trusted to do so on measured spectra until it has recovered known parameters from spectra whose parameters are known exactly. We describe a deterministic, seeded, hash-pinned corpus of 3,400 synthetic spectra of disordered, amorphous and hydrogenated carbon at three excitation wavelengths, with planted acquisition failures and five benchmark tasks; a learned inversion built from gradient-boosted trees and a one-dimensional convolutional network with a saturation detector used as a refusal signal; and the verification doctrine under which the results were admitted. On the held-out split, the released ensemble recovers L_a to 2.85% mean relative error and assigns the amorphisation stage with macro-F1 0.986, where the physics-only baseline fails at the turnover; refusal lowers the error on corrupted rows from 3.48 Å to 3.01 Å without refusing a clean row. Linear probes show that stage and L_a are encoded in the spectrum itself, while the sp^3 fraction is partly reached through process metadata, so that visible-excitation Raman alone should be treated as a suspect source of sp^3. Every result is a modelled result: it regression-tests the pipeline and supports no claim about physical films until measured spectra replace the synthetic corpus and the unchanged harness is rerun.

Keywords. Raman spectroscopy; amorphous carbon; inverse problem; synthetic benchmark; refusal; provenance; Ed25519; reproducibility.

1 Introduction

Raman spectroscopy is the standard non-destructive probe of sp^2 carbon, and the interpretation of Ferrari and Robertson [1] organises its first-order spectrum into a three-stage amorphisation trajectory from graphite through nanocrystalline graphite to amorphous and tetrahedral amorphous carbon. The interpretation is also a warning: the quantities most often read from a spectrum, the G-band position and the disorder ratio $I(\mathrm{D})/I(\mathrm{G})$, are not monotonic along the trajectory, and the same pair of values can occur in two stages. Resolving the ambiguity requires the width of the G band, the dispersion of its position with excitation energy, and, in practice, a model that uses all of them at once.

This paper reports three things. First, a synthetic corpus built from a declared forward model, designed so that a pipeline can be scored on whether it recovers the parameters that generated each spectrum, including under planted acquisition failures (Section 3). Second, a learned inversion that resolves the branch and refuses rather than repairs corrupted input, with results on the held-out split and probes of what the models actually use (Sections 4–5). Third, the verification and provenance doctrine under which those results were admitted, including the one thing provenance cannot do (Section 6).

The claim discipline is stated up front because it governs everything that follows. Every quantitative statement here carries the status MODELLED : it is the output of a deterministic forward model executed on a verified compute chain. No physical device, film or measurement campaign is represented. The results regression-test the pipeline and structure the measured phase; they cannot support a physical claim about real hardware until real data replaces the synthetic corpus.

2 Forward model

In stage 1, from graphite to nanocrystalline graphite, the disorder ratio obeys the Tuinstra–Koenig relation [2],

$$\frac{I(\mathrm{D})}{I(\mathrm{G})} = \frac{C(\lambda)}{L_a}, \qquad C(514\ \mathrm{nm}) = 44\ \text{Å}\ [3], \tag{1}$$

and in stage 2, from nanocrystalline graphite to amorphous carbon, the relation reverses [1],

$$\frac{I(\mathrm{D})}{I(\mathrm{G})} = C'(\lambda)\, L_a^2, \qquad C'(514\ \mathrm{nm}) = 0.0055\ \text{Å}^{-2}. \tag{2}$$

The branches meet at the turnover $L_a = 20$ Å, where both give a ratio of 2.2. The G band is given a Breit–Wigner–Fano line shape and the D band a Lorentzian, disorder-activated, as in [1].

The second corpus version adds three effects. Both constants scale with excitation wavelength as

$$C(\lambda) = C(514\ \mathrm{nm})\Big(\frac{\lambda}{514\ \mathrm{nm}}\Big)^2, \qquad C'(\lambda) = C'(514\ \mathrm{nm})\Big(\frac{\lambda}{514\ \mathrm{nm}}\Big)^2, \tag{3}$$

which leaves the turnover invariant. Compressive stress shifts the G band upward at 5 cm^{-1} per GPa, to 8 GPa in sp^3-rich films, relaxed by hydrogen; the rate is a declared model parameter. Hydrogenated films carry three effects. The photoluminescence background has a slope that rises with hydrogen content, following the trend reported by Casiraghi, Ferrari and Robertson [4]. The D band is suppressed by a factor $(1 - 0.6h)$ for hydrogen fraction h, and the G band shifts upward with hydrogen; the factor and the shift are declared model parameters chosen for the corpus, not literature values, and the measured phase will test them.

3 Corpus and benchmark tasks

Table 1. Corpus design.

Property	Value
Samples	3,400: train 2,400, validation 400, test 600
Seed	3407, fixed; regeneration is byte-identical and its SHA-256 is pinned in the ledger
Grid	1,000–1,800 cm^{-1}, 801 points
Excitations	325, 514 and 633 nm
Corruptions	6.1% of rows: cosmic-ray spikes, detector saturation, mislabelled excitation wavelength, mid-scan laser burn
Metadata	Process parameters carried alongside each spectrum

The corruptions are not noise in the ordinary sense; they are the failure modes of real acquisitions, planted so that a pipeline can be scored on whether it notices them. Five tasks are scored on the test split. T1 assigns the amorphisation stage (macro-F1). T2 recovers L_a (mean relative error), reported separately for the degenerate subset near the turnover, where the ratio alone cannot decide the branch. T3 recovers the sp^3 fraction (mean absolute error). T4 scores resolution of the degenerate branch directly. T5 scores robustness as the L_a error on corrupted rows.

The corpus answers one question: does a given pipeline recover the parameters that generated a spectrum? It cannot answer whether the forward model is complete, which is a question for measured spectra.

4 Inversion pipelines

The physics baseline inverts equation (1) and assumes stage 1 throughout; it is the honest statement of what the disorder ratio alone can do. Against it are compared a gradient-boosted tree model on binned spectra together with the process metadata; a one-dimensional convolutional network trained on the spectra alone; the tree model after a twenty-trial hyperparameter search; and an ensemble of the two learned models with weights fitted on the validation split. The ensemble is the pipeline's current release.

Preprocessing was made robust in the same revision: a median-filter despiking step removed 205,147 cosmic-ray spikes across the corpus, and a detector-saturation check was added whose output is a flag, never a correction. A flagged row is refused, not repaired.

5 Results

Table 2. Scores on the 600-sample held-out split. T2 is the mean relative error in L_a, with the degenerate subset near the turnover reported separately; T3 is the mean absolute error in sp^3 fraction; T5 is the mean absolute error in L_a, in Å, on corrupted rows.

Pipeline	T1 F1	T2 rel.	T2 degen.	T3 sp^3	T5, Å
Physics baseline (Tuinstra–Koenig, stage 1 assumed)	0.199	6.58	1.97	–	–
Gradient-boosted trees, spectra + metadata	0.992	0.033	0.028	0.015	5.14
1-D convolutional network, spectra only	0.977	0.078	0.071	0.028	14.48
Gradient-boosted trees, tuned	0.986	0.029	0.023	0.014	4.09
Ensemble, validation-fitted (release)	0.986	0.0285	0.023	0.019	3.48

The baseline's relative error of 6.58 is not a misprint: assuming stage 1 for a stage-2 spectrum returns a crystallite size wrong by a factor of several, which is the practical content of the turnover. The ensemble is selected on T2 and is slightly worse than the tuned tree model on T3; that trade is recorded rather than hidden.

5.1 Robustness and refusal

With the saturation flag applied as a refusal gate, the ensemble's L_a error on corrupted rows falls from 3.48 Å to 3.01 Å, and no clean row is refused. A coverage probe that scores each spectrum against the training distribution flags 1.5% of the test split as out of distribution; relative error is 0.030 in distribution and 0.097 out of it, 20% of corrupted rows are flagged and 0% of clean rows. Both signals cost nothing on good data, which is the property required of a gate that will later stand in front of measured spectra.

5.2 Probes: what the models use

Linear probes fitted on frozen features test whether the quantities of interest are actually encoded, and where. With sp^3 fraction scored by R^2 and stage by F1: peak-only features give 0.73 / 0.81, the binned spectrum 0.99 / 1.00, process metadata alone 0.78 / 0.85, and everything together 0.99 / 0.99; every shuffled control scores near zero. A leave-one-out nearest-neighbour attribution moves predictions by 0.014 Å, and the supporting rows for any prediction are same-stage neighbours at nearby L_a; no single cluster carries the model.

The reading is cautionary as well as reassuring. Stage and L_a are encoded in the spectrum itself. The sp^3 fraction, by contrast, is partly reached through process metadata and through the stress- and hydrogen-induced shifts of the G band, which means that a visible-excitation Raman spectrum alone should be treated as a suspect source of sp^3 fraction. That is consistent with the case for ultraviolet excitation [5], and it is the kind of statement the measured phase will have to confirm or retire.

6 Verification and provenance

6.1 Claim status

Every quantitative statement the platform emits carries one of three labels, printed inline and carried in the artifact metadata. `MODELLED` marks the output of a deterministic forward model or simulation

executed on the verified chain. COMPUTED marks arithmetic from nameplate or catalogue values. TO-VERIFY marks any site- or sample-dependent figure that has not been measured. A statement without a label is not admitted. Negative results are retained under the same labels, because positive results are only credible in that company.

6.2 Artifact ledger and instrument journal

Every artifact the platform produces — a regenerated corpus, a trained model, an evaluation record — is appended to a hash-chained ledger [6, 7] with chain head

$$h_i = H\big(h_{i-1} \,\|\, r_i\big), \tag{4}$$

where H is SHA-256 [8], and an Ed25519 signature [9] over each head. Verification runs four checks per entry: the signature verifies under the platform key; the entry's recorded previous head equals the recomputed head of the prefix; the entry's own digest recomputes; and the artifact on disk matches the digest in the entry. At the time of writing the ledger holds ten entries with zero failures; an entry that fails any check fails the whole verification, and the ledger is never rewritten to make it pass.

Measurements of a simulated campaign are written by the acquisition software at the moment each is taken, as receipts

$$r = \big(\mathit{seq},\ \mathit{kind},\ t,\ H(\mathit{payload}),\ h_{\mathrm{prev}}\big), \qquad h = H(r), \qquad \sigma = \mathrm{Sign}_{sk}(h), \tag{5}$$

so that the sequence number makes a missing measurement detectable as a gap, the previous-head field makes reordering detectable, and the signature binds each receipt to the instrument's key at the time of acquisition. The current journal verifies twenty-five of twenty-five entries with the chain intact. The same schema will be used unchanged when the instrument is real.

6.3 Release gates

Table 3. Release gates. A result is admitted only when every gate passes; outcomes are printed on the document that reports the result.

Gate	Criterion	Latest
Corpus determinism	Regeneration from seed is byte-identical to the ledgered digest	Pass
Reproducibility	Retrained metrics within 10^{-6} of the ledgered evaluation	Pass, difference 0
Cross-model agreement	Stage assignments of independent models agree on $\geq 95\%$ of samples	Pass, 98.2%
Journal validity	Hash chain and every Ed25519 signature verify	Pass, 25 of 25
Hypothesis determinism	Journal payload chain identical across runs, timestamps excluded	Pass
Optimisation gain	Each revision improves or holds the score it targets; a gated evaluation never scores worse than the ungated one	Pass
Probe sanity	OOD error exceeds in-distribution error; no clean row refused; shuffled controls near zero; attribution shift under 1 Å	Pass

The gates are deliberately mechanical. None of them asks whether a result is interesting; each asks whether it is the result the ledger says it is. Code and pointers live in version control and binaries in a content-addressed store, so that one commit identifier reproduces the corpus and every model trained on it; a quarterly restore drill on a clean machine — clone, check out, pull, verify with zero failures, retrain and reproduce the stage score — is itself a gate.

6.4 What provenance cannot do

The ledger proves that data has not changed since it was logged. It says nothing about whether the label was right at the moment of logging: a mislabelled excitation wavelength signed at the instrument is a perfectly provenanced error. That is what the planted corruptions, the refusal gate and the cross-model agreement gate are for. Provenance and verification are different instruments, and the platform keeps them separate.

7 Limitations and next steps

The corpus is a forward-model output, so the scores above bound the pipeline's behaviour only within the declared physics; a real film can differ from the model in ways the corpus cannot show. The corruption catalogue is finite. The sp^3 result depends on process metadata that a measured campaign may not carry. None of the figures reported here is a measurement, and none should be quoted as one.

The next steps are, in order: replace the synthetic spectra with measured Raman spectra from academic characterisation partners and rerun the unchanged harness; wire the coverage refusal gate in front of any measured-data deployment; and widen the synthetic cohorts, in particular the stress distributions, before the measured phase so that the harness's tolerances are set on the harder case.

Data and code availability

The corpus regenerates deterministically from its seed and its digest is recorded in the platform ledger; the findings this paper compiles are published at `https://ryanjamesyork.com/findings` with their Markdown sources and SHA-256 digests, so that any quoted number can be checked against a specific source. The generator and trained models are held pending the measured-data phase and are available from the author on request.

Competing interests

The author is the founder and chief executive of SSX360 Corp., which develops the research platform described.

References

[1] A. C. Ferrari and J. Robertson, "Interpretation of Raman spectra of disordered and amorphous carbon," *Phys. Rev. B* **61**, 14095–14107 (2000).

[2] F. Tuinstra and J. L. Koenig, "Raman spectrum of graphite," *J. Chem. Phys.* **53**, 1126–1130 (1970).

[3] D. S. Knight and W. B. White, "Characterization of diamond films by Raman spectroscopy," *J. Mater. Res.*

4, 385–393 (1989).

[4] C. Casiraghi, A. C. Ferrari, and J. Robertson, "Raman spectroscopy of hydrogenated amorphous carbons," *Phys. Rev. B* **72**, 085401 (2005).

[5] A. C. Ferrari and J. Robertson, "Resonant Raman spectroscopy of disordered, amorphous, and diamond-like carbon," *Phys. Rev. B* **64**, 075414 (2001).

[6] S. Haber and W. S. Stornetta, "How to time-stamp a digital document," *J. Cryptology* **3**, 99–111 (1991).

[7] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Advances in Cryptology – CRYPTO '87*, LNCS 293 (Springer, 1988), pp. 369–378.

[8] D. Eastlake 3rd and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)," RFC 6234 (IETF, 2011).

[9] S. Josefsson and I. Liusvaara, "Edwards-Curve Digital Signature Algorithm (EdDSA)," RFC 8032 (IETF, 2017).

R. J. York, "A deterministic synthetic corpus and learned inversion for the amorphous-carbon Raman inverse problem, admitted through a signed release gate," preprint v1.1, 3 September 2026 (v1.0 released the same day; v1.1 clarifies the attribution of two model parameters in Section 2). Compiled from Findings 2026.2–2026.4, `https://ryanjamesyork.com/findings`. All results MODELLED under declared hypotheses.