Resolving the Raman crystallite size turnover in nanocrystalline graphite with a synthetic benchmark

Ryan James York

SSX360 Corp., 1413 South King Street, Unit 202A, Honolulu, HI 96814, USA

Corresponding author: mission@ssx360.com; ORCID iD 0009-0007-5979-7949

Abstract

The disorder ratio $I(\mathrm{D})/I(\mathrm{G})$ of carbon Raman spectra is non-invertible: on the Ferrari–Robertson trajectory it rises as $1/L_a$ to a turnover at 20 Å and falls as L_a^2 beyond it, so one ratio maps to two sizes. I present a deterministic, hash-pinned corpus of 3,400 synthetic spectra of disordered, amorphous and hydrogenated carbon at three wavelengths, with planted acquisition failures and five benchmark tasks; an ensemble inversion with a saturation detector as refusal signal; and the signed release gates that admitted them. On the held-out split it recovers L_a to 2.85% mean relative error and assigns amorphisation stage with macro-F1 0.986 where the physics-only baseline fails; refusal cuts corrupted-row error from 3.48 to 3.01 Å without refusing clean rows. Stage and L_a are encoded in the spectrum; sp^3 fraction leans on process metadata. All results are modelled and claim nothing about real films until measured spectra replace them.

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.

The inverse problem is the same one that two-dimensional carbon characterisation runs every time a nanocrystalline film is sized. The stage-1 branch is nanocrystalline graphite, the ordered limit of defective graphene, and the sp^2 domain size read from $I(\mathrm{D})/I(\mathrm{G})$ in defective graphene follows the same law, with multiwavelength extensions turning it into a quantitative defect metric [2–6]. The stage-2 branch is amorphous and hydrogenated carbon, the diamond-like films [7]. Both branches, and the turnover between them, therefore sit inside the scope of 2D-materials characterisation, and every pipeline that reports a crystallite size from a Raman spectrum is solving exactly the inverse problem treated here. What has been missing is a public way to test whether such a pipeline resolves the branch at all, before it is pointed at measured spectra.

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. 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. Third, the verification and provenance doctrine under which those results were admitted, including the one thing provenance cannot do.

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.

Results

The forward model

In stage 1, from graphite to nanocrystalline graphite, the disorder ratio obeys the Tuinstra–Koenig relation [8],

$$I(\mathrm{D})/I(\mathrm{G}) = C(\lambda)/L_a, \qquad C(514\ \mathrm{nm}) = 44\ \text{Å}\ [9], \tag{1}$$

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

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

The branches meet at the turnover $L_a = 20$ Å, where both give a ratio of 2.2 (Fig. 1a). The G band is given a Breit–Wigner–Fano line shape and the D band a Lorentzian, disorder-activated, as in ref. [1]. The corpus carries three additional effects. In the corpus both constants are scaled with excitation wavelength as $(\lambda/514\ \mathrm{nm})^2$,

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

a declared model choice that leaves the turnover position invariant; the measured dispersion of C has been reported as linear in λ for peak-height ratios [10] and as λ^4 for integrated-area ratios [3], and the measured phase will test the choice. Compressive stress shifts the G band upward at 5 cm^{-1} per GPa, to 8 GPa in sp^3-rich films, relaxed by hydrogen, and hydrogenated films carry a photoluminescence background whose slope rises with hydrogen content, following the trend reported by Casiraghi, Ferrari and Robertson [11], a D-band suppression of $(1 - 0.6h)$ for hydrogen fraction h, and an upward G-band shift with hydrogen. The wavelength scaling, the stress rate, the suppression factor and the shift are declared model parameters chosen for the corpus, not literature values, and the measured phase will test them.

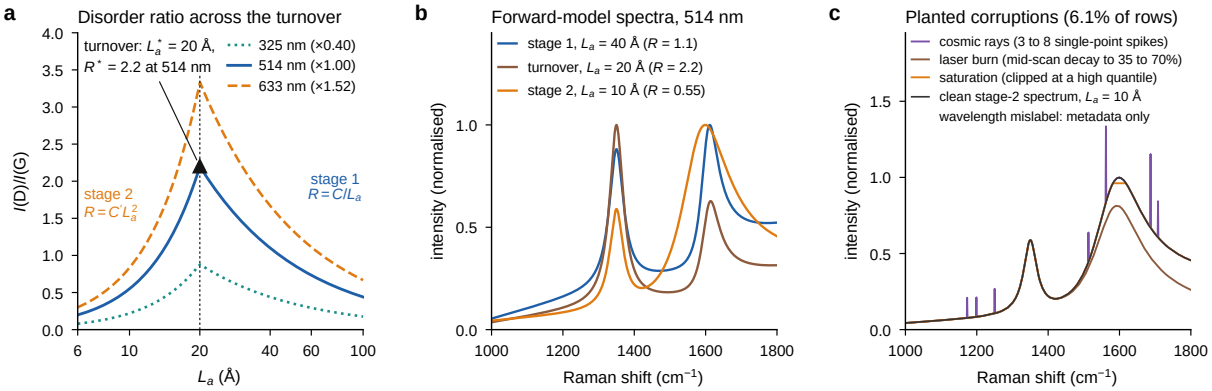

Fig. 1 | **Forward model and synthetic corpus. a**, Tuinstra–Koenig disorder ratio $R = I(\mathrm{D})/I(\mathrm{G})$ against L_a for the three excitation wavelengths used in the corpus (325 nm, teal dotted; 514 nm, blue solid; 633 nm, orange dashed). The two power-law branches meet at the turnover $L_a^* = 20$ Å, where both give $R^* = 2.2$ at 514 nm (black triangle; R^* scales as $(\lambda/514\ \mathrm{nm})^2$ with excitation wavelength, and the L_a position of the turnover is invariant). A single R below the turnover ratio maps to two crystallite sizes, one on each side of L_a^*. **b**, One noise-free forward-model spectrum per amorphisation stage at 514 nm, each normalised to its maximum: stage 1 (nanocrystalline graphite) at $L_a = 40$ Å ($R = 1.1$, blue), the turnover at $L_a = 20$ Å ($R = 2.2$, brown) and stage 2 (amorphous) at $L_a = 10$ Å ($R = 0.55$, orange), with the sp^3 fraction, hydrogen fraction and stress at the centre of the generator's distributions for that L_a. The G band is a Breit–Wigner–Fano line shape and the D band a Lorentzian, disorder-activated; the corpus carries no D′ band, and the 2D band lies outside the 1,000–1,800 cm^{-1} window. **c**, The planted corruption modes, applied by the generator's own corruption code to one clean stage-2 spectrum ($L_a = 10$ Å, 514 nm, with the generator's shot noise; black): cosmic rays (three to eight single-point spikes, purple), detector saturation (the trace clipped at a quantile drawn between 0.93 and 0.985 of its own values, orange), mid-scan laser burn (linear decay from a point in the middle third of the scan to 35–70% of the original amplitude, brown), and wavelength mislabelling, which is a metadata integrity case and leaves no trace on the spectrum. Corruption is planted on 6.1% of rows (207 of 3,400). The corpus contains 3,400 spectra of 801 points, drawn from one PCG64 stream at seed 3407; byte-identical regeneration on the pinned environment is a release gate. Panels **b** and **c** are computed by the deposited generator. All data MODELLED under the declared forward model.

The corpus and the benchmark tasks

The corpus design is given in Table 1. 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 (Fig. 1c). Five tasks are scored on the 600-row test split. T1 assigns the amorphisation stage (macro-F1 over three classes). T2 recovers L_a (mean relative error). T3 recovers the sp^3 fraction (mean absolute error). T4 scores resolution of the degenerate branch directly: the T2 error on the degenerate subset, the 100 test rows at 514 nm whose noise-free disorder ratio lies between 0.35 and 2.05, the range in which one ratio maps to a stage-1 size between 21 and 126 Å and a stage-2 size between 8 and 19 Å, so that the ratio alone cannot decide the branch. T5 scores robustness as the mean absolute L_a error on the 45 corrupted rows of the split.

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.

The 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 the subsampled spectrum together with the process metadata and the excitation wavelength; a one-dimensional convolutional network on the spectrum with a wavelength channel and a metadata branch; the tree model after a twenty-trial random hyperparameter search; and an ensemble of the tuned tree model and the network with the mixing weight fitted on the validation split. The ensemble is the pipeline's current release. Preprocessing was made robust in the same revision. A moving-median despiker replaced 205,147 points across the 3,400 spectra (11 to 152 per spectrum, median 57). The count is of points replaced, not of spike events: the rule scales every residual by one row-wide robust deviation while the shot noise grows with signal, so it also replaces ordinary positive noise excursions in the brighter parts of clean traces, each by a fraction of a per cent of its value. A flat-top detector-saturation check was added whose output is a flag, never a correction: in the test split it marks the eight saturated rows and none of the 555 clean rows. A flagged row is refused, not repaired.

Scores on the held-out split

Table 2 reports the five tasks on the 600-row held-out split. The baseline's relative error of 9.01 is not a misprint: assuming stage 1 for a stage-2 spectrum returns a crystallite size wrong by a factor of several (Fig. 2a), which is the practical content of the turnover. Applied to a stage-2 film, the baseline inverts equation (1) and returns $\hat{L}_a = C/R = (C/C')/L_a^2$, so the overestimation factor is $\hat{L}_a/L_a = (C/C')/L_a^3 = (L_a^*/L_a)^3$ with $L_a^* = (C/C')^{1/3} = 20$ Å. The overestimation therefore grows as $(L_a^*/L_a)^3$ for stage-2 films, by a factor of 8 at $L_a = 10$ Å and 37 at 6 Å. The ensemble is selected on T2 and is slightly worse than the tuned tree model on T3; that trade is recorded rather than hidden. The tuned tree model's hyperparameters were chosen by a search whose candidates were fitted on the training and validation rows together and scored on the validation rows, so the selection score was optimistic; the selected setting was then refitted on the training rows alone, and every number in Table 2 is from models that never saw the test split.

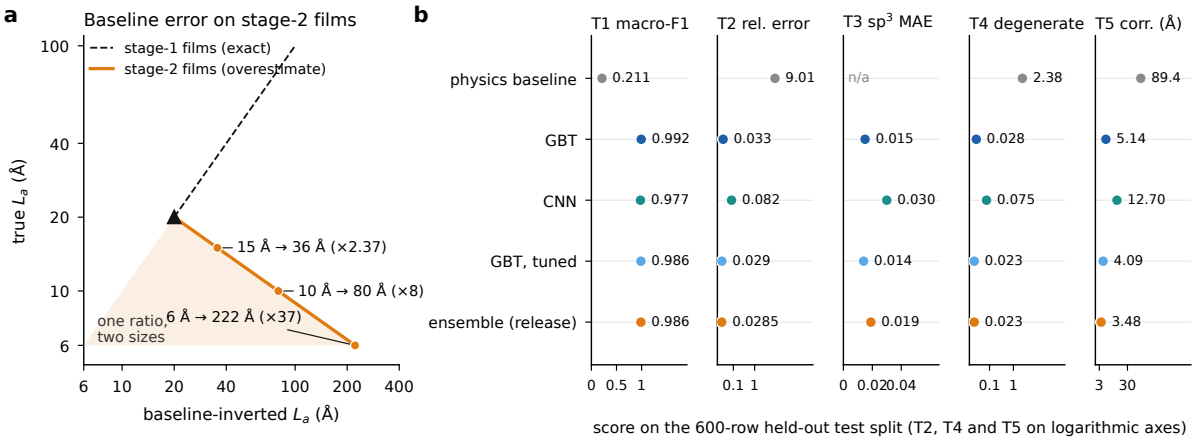

Fig. 2 | Inversion across the Tuinstra–Koenig turnover. a, A physics-only inversion that assumes stage 1 throughout, applied to stage-2 films (true L_a in $[6, 20]$ Å, the amorphous branch). The shaded wedge shows the degeneracy: the same single R value maps to two L_a sizes, and the baseline lands on the wrong one, overestimating by $(L_a^*/L_a)^3$: $\times 2.37$ at 15 Å, $\times 8$ at 10 Å and $\times 37$ at 6 Å (annotated points). On stage-1 films ($L_a \geq 20$ Å) the baseline is exact and lands on the diagonal. **b**, The five benchmark tasks (T1–T5) from Table 2, one row per pipeline, on the 600-row held-out test split; T2, T4 and T5 are on logarithmic axes and the tabulated value is printed beside each point. The physics baseline (grey, top row) fails on every task it can attempt: macro-F1 0.211, near chance for three classes, and relative error 9.01, a 901% mean relative error. The released ensemble (orange, bottom row) recovers the amorphisation stage with macro-F1 0.986 and L_a to 2.85% mean relative error, and lowers the L_a error on corrupted rows to 3.48 Å (T5). Panel **a** is MODELLED from the declared constants; panel **b** is COMPUTED from Table 2.

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; the eight refused rows are the eight saturated rows of the split, and none of the 555 clean rows is refused (Fig. 3a). A coverage probe that scores each test row's nearest-neighbour distance to the training rows flags 1.5% of the test split as out of distribution; the tree model's relative error is 0.030 in distribution and 0.097 out of it, 20% of corrupted rows are flagged and 0% of clean rows (Fig. 3b). Both signals cost nothing on good data, which is the property required of a gate that will later stand in front of measured spectra.

What the models use

Linear probes on fixed representations of the training rows test whether the quantities of interest are linearly encoded, and where (Fig. 3c). With sp^3 fraction scored by R^2 and stage by macro-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; with the labels shuffled the probes fall to chance ($R^2 \leq 0.01$ for sp^3 fraction and macro-F1 0.28–0.33 for stage, against 0.33 for three balanced classes). A nearest-neighbour attribution on one representative test query ($L_a = 14.2$ Å, stage 2, 325 nm) finds its eight nearest training rows in feature space to be same-stage neighbours at 12.9–14.7 Å, and refitting the tree model without those eight rows moves the prediction by 0.014 Å (14.02 to 14.01 Å); the probe is an illustration on one query, not a population statistic.

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 [12], and it is the kind of statement the measured phase will have to confirm or retire.

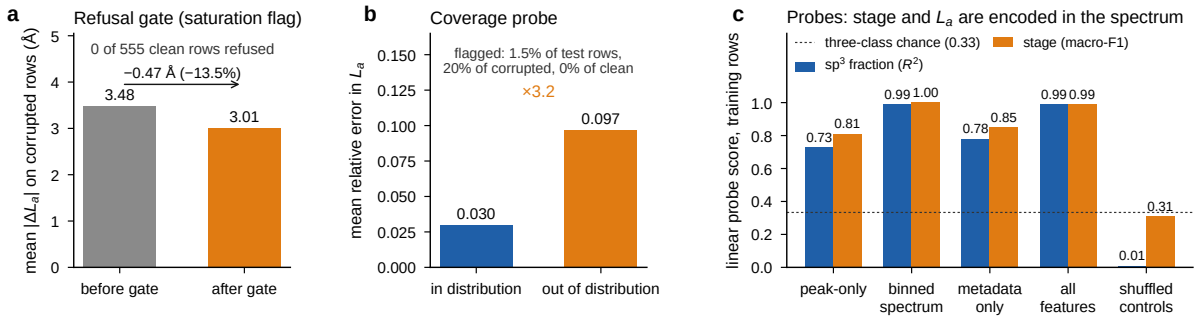

Fig. 3 | Refusal gate and model probes. **a**, The saturation-flag refusal gate lowers the ensemble's mean L_a error on the 45 corrupted test rows from 3.48 Å to 3.01 Å (-13.5%) by refusing the eight saturated rows, while refusing 0 of the 555 clean rows. **b**, A coverage probe scores each test row's distance to its tenth-nearest training row in the tree model's feature space against the 99th percentile of the same distance among training rows; 1.5% of the test split is flagged as out-of-distribution. The tree model's relative error is 0.030 in distribution versus 0.097 out of distribution ($3.2\times$); 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. **c**, Linear probes on fixed representations of the training rows (sp^3 fraction by R^2 / stage by macro-F1) show that stage and L_a are encoded in the spectrum itself (binned spectrum 0.99/1.00, all features 0.99/0.99), while the sp^3 fraction partly depends on process metadata (peak features only 0.73/0.81; process metadata only 0.78/0.85). With shuffled labels the probes fall to chance (0.01/0.31; dotted line, three-class chance 0.33). All data COMPUTED from the deposited probe records.

Verification and provenance

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.

Every artifact the platform produces, a regenerated corpus, a trained model, an evaluation record, is appended to a hash-chained ledger [13, 14] with chain head

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

where H is SHA-256 [15], and an Ed25519 signature [16] over each head (Fig. 4a). 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. The ledger deposited with this paper chains the corpus, its metadata, the evaluation records behind Table 2 and the probe records, and verifies with zero failures under the

deposited public key; an entry that fails any check fails the whole verification, and the ledger is never rewritten to make it pass.

Measurements are written by the acquisition software at the moment each is taken, as receipts

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

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 schema has been exercised on a simulated acquisition campaign that belongs to a separate programme and is not part of this deposit; it will be used unchanged when the instrument is real.

Seven release gates stand between a result and admission (Table 3, Fig. 4b). Five of them are exercised on the artifacts deposited with this paper by the deposited verification script; the two journal gates are exercised on the acquisition journal, which belongs to the separate campaign above. A result is admitted only when every gate passes, and the outcomes are printed on the document that reports the result. 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.

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.

Discussion

The practical content of this work is the number 9.01. A physics-only reading of the disorder ratio is not merely imprecise at the turnover; it is wrong by a factor of several, and the error grows as the cube of the ratio of the turnover size to the true size. Published pipelines that report L_a from $I(\mathrm{D})/I(\mathrm{G})$ on nanocrystalline graphite or defective graphene [2–6] resolve that branch implicitly, in whatever way their training data and features happen to, and nothing in the standard workflow tests whether the resolution is happening. This corpus is the test: the parameters are known exactly, the degeneracy is planted by construction, and the score on the degenerate subset separates pipelines that resolve the branch from those that average across it. To my knowledge it is the first public benchmark for this inverse problem that carries planted acquisition failures, a refusal gate scored on clean-row cost, and a signed record of the admission.

Because the corpus is a forward-model output, the scores bound the pipeline's behaviour only within the declared physics; a real film can differ from the model in ways the corpus cannot show, and the corruption catalogue is finite. The sp^3 result depends on process metadata that a measured campaign may not carry, and the stress- and hydrogen-dependent parameters are declared, not measured. None of the figures

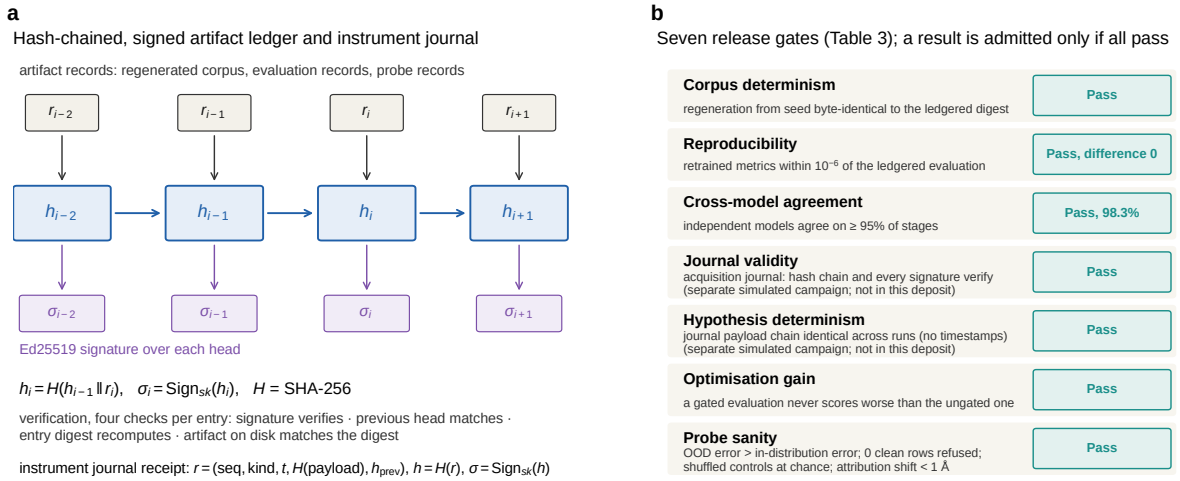

Fig. 4 | Signed release-gate provenance framework. a, Schematic of the hash-chained Ed25519 artifact ledger and the instrument journal. Each entry forms the next chain head $h_i = H(h_{i-1} \,\|\, r_i)$ and is signed $\sigma_i = \mathrm{Sign}_{sk}(h_i)$. Four checks run on every entry: the signature verifies, the previous head matches, the digest recomputes, and the artifact on disk matches the digest. The ledger deposited with this paper verifies with 0 failures. The instrument journal (receipt $r = (\mathrm{seq}, \mathrm{kind}, t, H(\mathrm{payload}), h_{\mathrm{prev}})$, $h = H(r)$, $\sigma = \mathrm{Sign}_{sk}(h)$) uses the same construction for measurements at acquisition. A mislabelled input signed at the instrument is a perfectly provenanced error; the corruptions, the refusal gate and the cross-model agreement gate are the instruments that catch it, not the ledger. **b**, The seven release gates (Table 3) that every result must pass before it is admitted. Each gate closes an independent failure mode: byte-identical corpus regeneration, retrain reproducibility (difference 0 within 10^{-6}), cross-model agreement (98.3%, criterion 95%), journal validity (hash chain and every signature verify), hypothesis determinism (payload chain identical across runs), optimisation gain (gated never worse than ungated), and probe sanity (OOD $>$ in-distribution, zero clean rows refused, shuffled controls at chance, attribution shift under 1 Å). Schematic, not a screenshot.

reported here is a measurement, and none should be quoted as one.

None of the three contributions is a new physical result, and the paper does not claim one. The Tuinstra–Koenig law and the Ferrari–Robertson trajectory are the literature's, and the learned models are standard. What is new is the integration layer: a deterministic, hash-pinned corpus, a learned inversion that refuses rather than repairs, and a signed release-gate doctrine, bound together so that every number in the paper is either the result the ledger says it is or not released at all. One verified pipeline carries a materials measurement, its inverse, and its evidence at once; a result passes every check or it is refused, and the certificate is the signed chain.

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. The forward model, the splits, the gates and the ledger schema are written to make that handoff a data replacement, not a redesign.

Methods

Forward model

Stage 1 obeys $I(\mathrm{D})/I(\mathrm{G}) = C(\lambda)/L_a$ with $C(514\ \mathrm{nm}) = 44$ Å [9]; stage 2 obeys $I(\mathrm{D})/I(\mathrm{G}) = C'(\lambda)L_a^2$ with $C'(514\ \mathrm{nm}) = 0.0055\ \text{Å}^{-2}$ [1]; both constants are scaled as $(\lambda/514\ \mathrm{nm})^2$ (equation 3), leaving the turnover at $L_a = 20$ Å and $R = 2.2$ invariant. Each spectrum is computed on the 1,000–1,800 cm^{-1} grid as follows, with E the excitation photon energy and E_{514} its value at 514 nm. The G band is a Breit–Wigner–Fano line $A_G\,(1+t/q)^2/(1+t^2)$ with $t = (\omega-\omega_G)/\Gamma_G$, $q = 4$ and $A_G = 1000$ counts. In stage 1 its centre ω_G moves linearly from 1,581 cm^{-1} at $L_a = 300$ Å to 1,600 cm^{-1} at the turnover and its width is $\Gamma_G = 18 + 8\,(1 - L_a/300\ \text{Å})$ cm^{-1}; in stage 2, ω_G moves from 1,600 cm^{-1} at the turnover to 1,510 cm^{-1} at $L_a = 2$ Å, with an added dispersion of $25\,(E - E_{514})\,\tau$ cm^{-1} eV^{-1} for fractional depth τ into the stage, and $\Gamma_G = 30 + 60\,f_{\mathrm{sp^3}}$ cm^{-1}. Compressive stress σ adds 5σ cm^{-1} per GPa to ω_G, hydrogen fraction h adds $10h$ cm^{-1} to ω_G and $15h$ cm^{-1} to Γ_G, and ω_G is clipped to 1,480–1,640 cm^{-1}. The D band, present for $L_a > 3$ Å, is a Lorentzian centred at $1{,}350 + 50\,(E - E_{514})$ cm^{-1} with half-width $22 + 0.06L_a$ cm^{-1} and peak amplitude $1.15\,R\,(1-0.6h)\,A_G$, where R is the disorder ratio of equations (1)–(3); there is no D′ band, and the 2D band lies outside the window. A linear background rises from 40 to 70 counts across the window, and hydrogenated rows carry a photoluminescence slope of $2{,}400h$ counts across the window, following the trend reported in ref. [11]. The whole trace is multiplied by an amplitude factor drawn uniformly from 0.85–1.15, given Gaussian shot noise of standard deviation $0.02\sqrt{\text{signal}}$, and clipped at zero. The wavelength scaling, the stress rate, the hydrogen suppression and shift, the line-shape parameters and the background are declared model parameters chosen for the corpus, not literature values.

Corpus generation

3,400 spectra on a 1,000–1,800 cm^{-1} grid of 801 points, at 325, 514 or 633 nm excitation drawn uniformly per row. Each row starts from five process parameters drawn independently: deposition temperature (20–

500 °C), ion energy (5–120 eV), CH_4 fraction (Beta$(2, 5)$), pressure (1–50 mTorr) and substrate bias (0–1,200 V). A disorder coordinate $x \in [0, 1]$ is the logistic function (gain 1.6) of a weighted sum of the centred and scaled ion energy, bias, CH_4 fraction and temperature (weights 1.4, 1.1, 0.7 and -0.5) plus Gaussian noise of standard deviation 0.35, so that the process metadata predicts the structure only partially. L_a falls linearly with x from 300 Å to the turnover at $x = 0.5$ (stage 1) and from the turnover to 2 Å at $x = 1$ (stage 2), with a 4% log-normal jitter and clipping to 1.5–400 Å. The sp^3 fraction is a logistic function of x centred at $x = 0.62$ with Gaussian noise of standard deviation 0.015, clipped to 0–0.95; the hydrogen fraction is $0.55(1 - x)^{1.5}$ plus a Beta-distributed term, with noise of standard deviation 0.03, clipped to 0–0.55; and the stress is $8\, f_{sp^3}^{1.5}(1 - 0.7h)$ GPa with noise of standard deviation 0.25 GPa, clipped to 0–8 GPa. The T1 label is assigned from L_a by declared thresholds, stage 1 for $L_a \geq 43$ Å, stage 2 for $6 \leq L_a < 43$ Å and stage 3 for $L_a < 6$ Å; these label boundaries are distinct from the forward model's branch point at 20 Å, so a pipeline must learn that rows between 20 and 43 Å lie on the Tuinstra–Koenig branch but carry the stage-2 label. Splits are a random permutation: 2,400 train, 400 validation, 600 test. Corruption is planted on 207 rows (6.1%) chosen without replacement, in one of four modes drawn uniformly: cosmic rays (three to eight single-point spikes, each adding 8–20 times the square root of the local signal), detector saturation (the trace clipped at a quantile drawn uniformly between 0.93 and 0.985 of its own values), mid-scan laser burn (linear decay from a point in the middle third of the scan to 35–70% of the original amplitude), and mislabelled excitation wavelength (the label switched to one of the other two wavelengths; the spectrum is untouched). All randomness is drawn from a single PCG64 stream with seed 3407. On the pinned environment regeneration is byte-identical and the SHA-256 digest of the regenerated corpus is recomputed against the artifact ledger as a release gate; an independent regeneration on a second machine with a different NumPy build reproduced every split, label, corruption assignment and count, with differences confined to the last floating-point digit of transcendental evaluations in about 3% of rows, and the deposited verification script, run on that machine, retrained the default tree model from the seed and matched the ledgered T1, T2 and T3 to every printed digit.

Preprocessing

Despiking is applied to every row. The residual of the trace against its seven-point moving median (edge-padded) is scaled by 1.4826 times the median absolute deviation of the residuals, and any point whose scaled residual exceeds 6 is replaced by the local median. Across the corpus the rule replaced 205,147 points (11 to 152 per spectrum, median 57). Because the residuals are scaled by one row-wide deviation while the shot noise grows with signal, the rule also fires on ordinary noise excursions in the brighter parts of clean traces; on the clean rows I checked, a replaced point changed by about 0.1% of its value at the median and by about 1% at most, so the count is of points replaced, not of spike events. The saturation check flags a row when at least five consecutive points lie within 10^{-4} of the trace's 98.5th percentile; in the test split it flags the eight saturated rows and no clean row, and over the whole corpus all 54 saturated rows and no clean row. The flag is a refusal signal and is never used to correct the trace.

Inversion models

Five pipelines are scored. (i) The physics baseline subtracts the fifth percentile of the trace, takes the G amplitude as the global maximum and the D amplitude as the maximum within $\pm 120\,\mathrm{cm}^{-1}$ of $1{,}350\,\mathrm{cm}^{-1}$, inverts equation (1) with the 514-nm constant at every wavelength, caps the result at 300 Å and assigns the stage from the same L_a thresholds as the labels. (ii) A gradient-boosted tree model (histogram-based gradient boosting in scikit-learn [17]: a classifier for the stage, a regressor for $\log(1+L_a)$ and a regressor for the sp^3 fraction) on every fourth grid point of the trace normalised to its maximum (201 values), the five standardised process parameters and a one-hot code of the labelled excitation wavelength; defaults 400 iterations (250 for the sp^3 regressor), learning rate 0.06, 63 leaf nodes, 12 samples per leaf and ℓ_2 regularisation 1.0. (iii) A one-dimensional convolutional network [18] whose input is the standardised trace and three constant channels coding the labelled wavelength: four convolution–batch-normalisation–GELU blocks (32, 64, 96 and 96 channels; kernels 7, 5, 5 and 3; stride 2), global average pooling, concatenation with a 32-unit branch fed by the five process parameters and the wavelength code, a 128-unit layer with dropout 0.25, and three heads for the stage logits, $\log(1 + L_a)$ and the sp^3 fraction. It is trained with AdamW (learning rate 2×10^{-3}, weight decay 5×10^{-4}) under cosine annealing for 160 epochs in batches of 64, with Gaussian input noise of standard deviation 0.03, mixup ($\alpha = 0.2$) on 60% of batches, and the loss cross-entropy $+ 2\times$ smooth-ℓ_1 on $\log(1 + L_a)$ + squared error on sp^3; the weights are averaged over the last 30% of epochs when that average is within 2% of the best validation loss, and seed 3407 fixes the initialisation and the batch order. (iv) The tree model of (ii) after a twenty-trial random search over iterations $\{400, 600, 800\}$, learning rate $\{0.03, 0.05, 0.08\}$, leaf nodes $\{63, 127, 255\}$, samples per leaf $\{6, 10, 16\}$ and ℓ_2 regularisation $\{0.3, 1, 3\}$, minimising $(1 - \mathrm{T1}) + \mathrm{T2} + 0.5\,\mathrm{T4}$; the candidates were fitted on the training and validation rows together and scored on the validation rows, which makes the selection score optimistic, and the selected setting (800, 0.08, 255, 6, 0.3) was refitted on the training rows alone before scoring. (v) The release ensemble combines (iv) and (iii): L_a is the geometric mean of the two predictions with weight w on the tree model, chosen on the validation split from a 21-point grid to minimise the relative error ($w = 0.9$); the stage is the tree model's; and the sp^3 fraction is the mean of the two. Every model is fitted on the 2,400 training rows and scored once on the 600 test rows. All models are retrained from the seed for the reproducibility gate, with a tolerance of 10^{-6} on the ledgered metrics. Library versions are pinned in the environment file deposited with the code.

Benchmark tasks and metrics

T1: amorphisation stage, macro-F1 over the three classes. T2: L_a, mean relative error $|\hat{L}_a - L_a|/L_a$ over the split; the relative error is a dimensionless fraction ($9.01 = 901\%$). T3: sp^3 fraction, mean absolute error. T4: the T2 error on the degenerate subset, the 100 test rows labelled 514 nm whose noise-free ratio, evaluated at 514 nm on the branch given by L_a, lies in $(0.35, 2.05)$. T5: mean absolute L_a error in Å on the 45 corrupted rows of the split. All scores are on the 600-row held-out test split; the baseline returns L_a only, so T3 is not defined for it.

Probes and attribution

Linear probes are ridge regressions (sp^3 fraction, scored by R^2) and multinomial logistic regressions (stage, scored by macro-F1) fitted in-sample on the 2,400 training rows for four fixed representations: five peak features (G amplitude, position and full width at half maximum, D amplitude, and their ratio), the trace binned into 120 bins, the five process parameters, and the three together, each standardised; every probe is repeated five times and the scores averaged. Shuffled-label controls repeat each probe with permuted labels and are expected at chance (R^2 near 0; macro-F1 near 0.33). The coverage probe standardises the tree model's feature vector, measures each test row's distance to its tenth-nearest training row, and flags rows beyond the 99th percentile of the same distance among training rows; the relative errors in and out of distribution are those of the default tree model fitted on the training rows. Attribution uses nearest neighbours in the tree model's feature space for one representative test query: its eight nearest training rows are listed, the tree model is refitted without them, and the change in the query's L_a prediction is reported for that query only.

Refusal gate and coverage probe

The refusal gate is the saturation flag: a flagged row is refused and excluded from the L_a score, and the score is reported with and without the gate, following the selective-prediction doctrine that a predictor should be allowed to abstain rather than err [19]. The coverage probe scores each test spectrum against the training distribution [20] and flags out-of-distribution rows.

Claim status, artifact ledger and instrument journal

Every quantitative statement carries one of three labels: MODELLED, the output of a deterministic forward model or simulation executed on the verified chain; COMPUTED, arithmetic from nameplate or catalogue values; TO-VERIFY, 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.

Artifacts are appended to a hash-chained ledger with chain head $h_i = H(h_{i-1} \,\|\, r_i)$, H = SHA-256, and an Ed25519 signature over each head (equation 4). Verification runs four checks per entry: the signature verifies under the platform key; the 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. A failing entry fails the whole verification, and the ledger is never rewritten to make it pass.

Instrument measurements are written at acquisition as receipts $r = (\text{seq}, \text{kind}, t, H(\text{payload}), h_{\text{prev}})$, $h = H(r)$, $\sigma = \text{Sign}_{sk}(h)$ (equation 5). The sequence number makes a missing measurement detectable as a gap; the previous-head field makes reordering detectable; the signature binds each receipt to the instrument's key at the time of acquisition.

Release gates

Seven gates admit a result: corpus determinism (regeneration byte-identical to the ledgered digest); reproducibility (retrained metrics within 10^{-6} of the ledgered evaluation); cross-model agreement (stage assignments of independent models agree on at least 95% of samples); journal validity (hash chain and

every signature verify); hypothesis determinism (journal payload chain identical across runs, timestamps excluded); optimisation gain (a gated evaluation never scores worse than the ungated one); probe sanity (out-of-distribution error exceeds in-distribution error, no clean row refused, shuffled controls at chance, attribution shift under 1 Å). A result is admitted only when every gate passes; the outcomes are printed on the document that reports the result. A quarterly restore drill on a clean machine (clone, check out, pull, verify with zero failures, retrain, reproduce the stage score) is itself a gate.

Statistics and reproducibility

All reported scores are on the 600-row test split ($n = 600$; 100 in the degenerate subset; 45 corrupted, of which 8 saturated); n per split is stated in Table 1. Every figure regenerates from seed 3407 and the commit identifier recorded in the code repository; Fig. 1b,c are computed by the public generator. The reproducibility gate bounds retrained metrics to within 10^{-6} of the ledgered values. The evaluation records behind Table 2 are published with the corpus; the per-row predictions of the release run, which underlie Fig. 2, are regenerated by the recorded procedure on the author's compute hardware and accompany the public record, and bootstrap 95% confidence intervals (1,000 resamples of the test rows) for the Table 2 metrics are computed from those predictions.

Use of AI tools

The digital-twin forward model and corpus generator, the inversion models, the evaluation harness and the release gates were hand-coded and run on my own offline servers and models in the United States, and the analysis was done by hand; no AI service received the corpus, the trained models or the evaluation records. Two AI systems were used outside the analysis chain, under my direction: Parallel AI for deep-research methodology at the literature stage, and Claude (Anthropic) for final checks, namely verification of references and arithmetic, an independent regeneration of the corpus and rerun of the deterministic preprocessing counts, the physics baseline and the repository's own verification script against the ledgered records, regeneration of the figures with plotting code, and language editing. No AI tool trained a model whose results are reported here, produced a result or drew a conclusion, and no generative-AI images are included; I verified every statement against the ledgered artifacts.

Data availability

The synthetic corpus analysed in this study (3,400 spectra with process metadata, ground-truth parameters, corruption flags and split assignment), the evaluation and probe records behind Tables 2 and 3, and the environment file used to re-derive them are published at `https://ryanjamesyork.com/publications` under a CC BY 4.0 licence, in the record dated 5 September 2026. The corpus is the output of the public generator at seed 3407; its SHA-256 digest is printed in that record, and an independent regeneration from the seed on a second machine reproduced every split, label, corruption assignment and count, as stated in Methods.

Code availability

The corpus generator, the inversion pipelines, the evaluation harness, the probes and the ledger verifier used in this study are available at `https://github.com/SSX360/raman-turnover-benchmark` under the Apache-2.0 licence, together with the evaluation and probe records behind Tables 2 and 3. The commit identifier that reproduces every reported number is recorded in the deposited ledger.

References

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

[2] Ferrari, A. C. & Basko, D. M. Raman spectroscopy as a versatile tool for studying the properties of graphene. *Nat. Nanotechnol.* **8**, 235–246 (2013).

[3] Cançado, L. G. *et al.* General equation for the determination of the crystallite size L_a of nanographite by Raman spectroscopy. *Appl. Phys. Lett.* **88**, 163106 (2006).

[4] Cançado, L. G. *et al.* Quantifying defects in graphene via Raman spectroscopy at different excitation energies. *Nano Lett.* **11**, 3190–3196 (2011).

[5] Lucchese, M. M. *et al.* Quantifying ion-induced defects and Raman relaxation length in graphene. *Carbon* **48**, 1592–1597 (2010).

[6] Eckmann, A. *et al.* Probing the nature of defects in graphene by Raman spectroscopy. *Nano Lett.* **12**, 3925–3930 (2012).

[7] Robertson, J. Diamond-like amorphous carbon. *Mater. Sci. Eng. R* **37**, 129–281 (2002).

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

[9] Knight, D. S. & White, W. B. Characterization of diamond films by Raman spectroscopy. *J. Mater. Res.* **4**, 385–393 (1989).

[10] Matthews, M. J., Pimenta, M. A., Dresselhaus, G., Dresselhaus, M. S. & Endo, M. Origin of dispersive effects of the Raman D band in carbon materials. *Phys. Rev. B* **59**, R6585–R6588 (1999).

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

[12] Ferrari, A. C. & Robertson, J. Resonant Raman spectroscopy of disordered, amorphous, and diamondlike carbon. *Phys. Rev. B* **64**, 075414 (2001).

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

[14] Merkle, R. C. A digital signature based on a conventional encryption function. In *Advances in Cryptology — CRYPTO '87* (ed. Pomerance, C.) Lecture Notes in Computer Science **293**, 369–378 (Springer, 1988).

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

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

[17] Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[18] Paszke, A. *et al.* PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* **32** (2019).

[19] Geifman, Y. & El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* **30** (2017).

[20] Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. International Conference on Learning Representations* (2017).

Acknowledgements

This study was funded internally by SSX360 Corp., of which the author is the founder and chief executive. No external party played any role in study design, data generation, analysis and interpretation, or the writing of this manuscript.

Author contributions

R.J.Y. conceived the study, specified the forward model and generated the corpus, built and evaluated the inversion pipelines and the refusal gate, designed and ran the ledger, journal and release gates, and wrote the manuscript.

Competing interests

The author declares the following competing interests: R.J.Y. is the founder, chief executive and a shareholder of SSX360 Corp., which develops the research platform described and funded this study.

Tables

Table 1 | Corpus design. n per split is the count of spectra; every split is drawn from the same PCG64 stream.

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	207 rows (6.1%): cosmic-ray spikes, detector saturation, mislabelled excitation wavelength, mid-scan laser burn
Metadata	Process parameters carried alongside each spectrum

Table 2 | Scores on the 600-row held-out split. $n = 600$ test rows for every entry; T4 is scored on the 100-row degenerate subset and T5 on the 45 corrupted rows. T1 is macro-F1 over three stages. T2 is the mean relative error in L_a, a dimensionless fraction (the baseline's 9.01 is a 901% error); T4 is the same error on the degenerate subset near the turnover; T3 is the mean absolute error in sp^3 fraction; T5 is the mean absolute error in L_a, in Å, on corrupted rows. A dash marks T3, which is not defined for the physics baseline, which returns no sp^3 estimate; its T5 of 89.4 Å is the error on the corrupted rows before any gate. The tuned tree model, the network and the ensemble are from the release evaluation record (pipeline revision 3); the untuned tree model is from revision 2.1, before despiking and the saturation flag were added, on the same corpus and split.

Pipeline	T1 F1	T2 rel.	T4 degen.	T3 sp^3	T5, Å
Physics baseline (Tuinstra–Koenig, stage 1 assumed)	0.211	9.01	2.38	—	89.4
Gradient-boosted trees, spectra + metadata	0.992	0.033	0.028	0.015	5.14
1-D convolutional network, spectra + metadata	0.977	0.082	0.075	0.030	12.70
Gradient-boosted trees, tuned	0.986	0.029	0.023	0.014	4.09
Ensemble of tuned trees and network (release)	0.986	0.0285	0.023	0.019	3.48

Table 3 | Release gates. A result is admitted only when every gate passes; outcomes are printed on the document that reports the result. The corpus, reproducibility, cross-model, optimisation and probe gates are run by the deposited verification script on the deposited artifacts; the two journal gates are exercised on the acquisition journal of a separate simulated campaign that is not part of this deposit.

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, 98.3%
Journal validity	Hash chain and every Ed25519 signature of the acquisition journal verify	Pass, on the separate simulated campaign; not part of this deposit
Hypothesis determinism	Journal payload chain identical across runs, timestamps excluded	Pass, on the separate simulated campaign; not part of this deposit
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 at chance; attribution shift under 1 Å	Pass