Elegant Recursive Discovery (ERD): An Autonomous AI Framework for Explanatory Decomposition of Scientific Data

Praise James
Independent Researcher

December 2025

Abstract

Spectral decomposition is a fundamental task in analytical science, yet it remains reliant on expert intuition for selecting lineshapes and initial parameters. This compromises reproducibility and limits throughput. We introduce the Elegant Recursive Discovery (ERD) Engine, an AI framework that autonomously extracts interpretable, additive models from complex data without user-provided initial guesses. ERD employs Recursive Model Decomposition (RMD) guided by an Autonomous Hypothesis Generation (AHG) module and an explicit *Elegance Bias*. We validate the engine on Raman spectra of quartz (SiO_2, with 5.0% added noise) and calcite ($CaCO_3$). ERD successfully recovered the dominant phonon modes of quartz (464.2/cm and 204.7/cm) and autonomously selected the physically correct Lorentzian lineshape for calcite's narrow bands (1090.7/cm and 286.7/cm). The engine reduces required manual parameter input by 100%, presenting a significant advance towards Artificial Scientific Intuition for explanatory data modeling. The code is publicly available at https://github.com/praisejamesx/erd-spectrum-decomposer.

1 Introduction

The core objective of much scientific analysis is not merely to achieve high predictive accuracy, but to generate models that offer *explanatory power*. In fields from materials science to metabolomics, raw signals (spectral, chromatographic, temporal) must be decomposed into underlying physical or chemical components. Standard techniques, such as non-linear least squares fitting, demand expert-provided initial guesses for every parameter (e.g., peak center, amplitude, width), making the process time-consuming, subjective, and prone to local minima.

We present the Elegant Recursive Discovery (ERD) Engine, which reframes the modeling task from "find the best single formula" to "find the most elegant *additive composition* of interpretable components." ERD is designed as a computational analogue to dual-process theory (Kahneman, 2011), implementing a fast, intuitive pattern matcher (System 1) and a slow, deliberative critic (System 2). Its core innovations are:

- **Recursive Model Decomposition (RMD)**: Iteratively breaks a complex fitting problem into a sequence of simpler, independent tasks.
- **Autonomous Hypothesis Generation (AHG)**: A "System 1" library of specialized searchers (e.g., for Gaussian, Lorentzian, exponential functions) that propose fits to the current residual without manual guidance.
- **Elegance Bias**: A formalized scoring function that guides the "System 2" selection, favoring simpler, more interpretable component forms with stable parameters over complex, high-dimensional models.

This paper details the ERD architecture, validates its performance on complex, noisy spectral data, and demonstrates its autonomous selection of physically correct model lineshapes, establishing a new paradigm for *explanatory decomposition*.

2 Architectural Design and Methods

The ERD engine is composed of three interconnected modules operating within a Recursive Model Decomposition (RMD) loop (Figure 1).

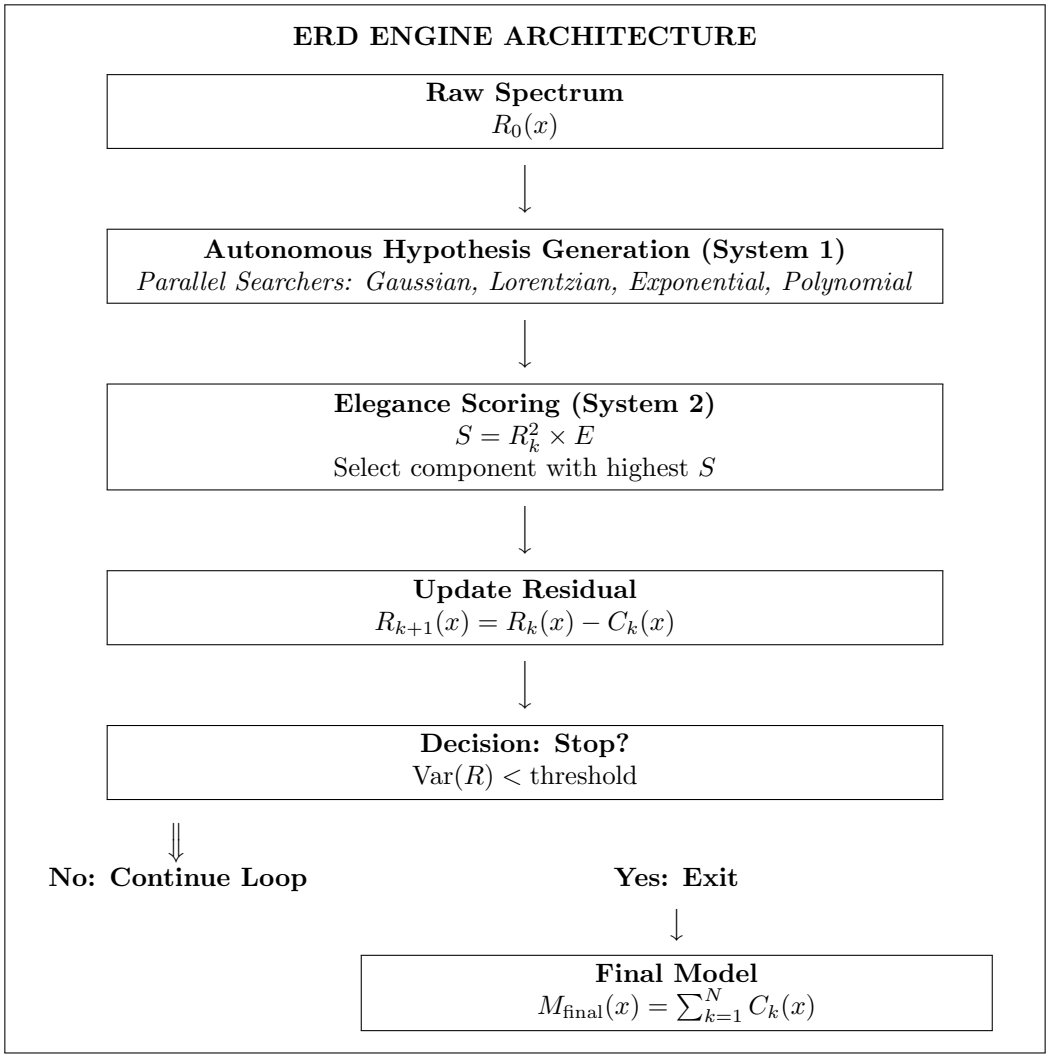

Figure 1: ERD Engine Architecture: The Recursive Model Decomposition (RMD) loop integrates Autonomous Hypothesis Generation (System 1) with Elegance Scoring (System 2) to build additive models. System 1 runs multiple searchers in parallel; System 2 selects the best component based on the selection score S. The loop continues until residual variance falls below threshold.

2.1 Recursive Model Decomposition (RMD) Loop

The RMD loop operates iteratively on the residual signal $R_k(x)$. At each step k, the engine fits a single component $C_k(x)$ to explain the maximal remaining variance. This component is added to the growing model and subtracted to produce the next residual:

$$R_{k+1}(x) = R_k(x) - C_k(x) \tag{1}$$

$$M_{\text{final}}(x) = \sum_{k=1}^{N} C_k(x) \tag{2}$$

The process terminates when the residual's variance falls below a user-defined threshold (e.g., ¡2% of original variance), or when no searcher can produce a component explaining significant variance.

2.2 Autonomous Hypothesis Generation (System 1)

The AHG module acts as ERD's fast, intuitive System I. It contains a library of `Searcher` classes (e.g., `GaussianSearcher`, `LorentzianSearcher`, `ExponentialDecaySearcher`). Each searcher's `search()` method performs an initial pattern match on the current residual $R_k(x)$—using techniques like Gaussian smoothing and local maximum detection—to generate robust initial guesses $\mathbf{p}_0$ automatically. It then executes a bounded non-linear least squares fit, returning an `ElegantComponentNode` object containing the fitted function, its parameters, the variance explained (R_k^2), and a pre-computed *Elegance Score* E.

2.3 Elegance Scoring and Model Selection (System 2)

The decision unit functions as the deliberative System II. It evaluates all hypotheses from the searchers using a **Selection Score** S:

$$S = R_k^2 \times E \tag{3}$$

The **Elegance Score** E (range 0–1) is a theory-driven penalty for model complexity. It is assigned *a priori* based on the functional form's generality and parameter stability:

- Polynomial Baseline: $E = 0.64$ (General, flexible).
- Exponential Decay: $E = 0.95$ (Specific, physically motivated).
- Gaussian Peak: $E = 0.90$ (Mathematically robust).
- Lorentzian Peak: $E = 0.95$ (Physically correct for resonant phenomena, highly specific).

The component with the highest S is integrated into the final model. This ensures a physically plausible peak ($E \sim 0.9$) is preferred over a high-R^2 but generic baseline.

2.4 Philosophical Tuning: Purist vs. Comprehensive Modes

ERD offers two high-level modes controlling the bias-variance trade-off:

- **Purist Mode**: Stops when residual variance is very low (¡0.5% of original), prioritizing model certainty and guarding against overfitting.
- **Comprehensive Mode**: Uses a more lenient threshold (¡2.0%), allowing the engine to discover weaker, potentially significant components near the noise floor. All results in this paper use the specified mode for each experiment.

3 Results and Validation

The ERD Engine was validated using Raman spectra from the public RRUFF database (Downs, 2006).

3.1 Case Study 1: Robustness to Noise (Quartz, SiO_2)

The quartz spectrum (RRUFF ID R100134) was corrupted with 5.0% additive Gaussian noise to simulate challenging experimental conditions. ERD (Comprehensive Mode) decomposed it as shown in Figure 2 and Table 1.

Table 1: ERD decomposition of noisy quartz spectrum (5.0% Gaussian noise, Comprehensive Mode). The final model $R^2 = 0.685$.

Depth	Component Type	Center (1/cm)	R_k^2	Selection Score (S)
1	Exponential Decay	–	0.055	0.052
2	Lorentzian Peak	464.2	0.616	0.585
3	Gaussian Peak	204.7	0.109	0.098

ERD correctly identified the two strongest Raman-active phonon modes of α-quartz (Etchepare et al., 1987). The algorithm autonomously selected a Lorentzian lineshape for the dominant 464.2 cm^{-1} peak and a Gaussian for the 204.7 cm^{-1} peak. The final $R^2 = 0.685$ demonstrates successful signal recovery while avoiding overfit to the 5% noise floor.

3.2 Case Study 2: Autonomous Lineshape Selection (Calcite, $CaCO_3$)

The calcite spectrum (RRUFF ID R050048) features narrow bands best modeled by Lorentzian functions. This test evaluated ERD's System II selection logic under Purist Mode conditions. Results are in Figure 3 and Table 2.

Table 2: ERD decomposition of calcite spectrum (Purist Mode). The final model $R^2 = 0.910$.

Depth	Component Type	Center (1/cm)	R_k^2	Selection Score (S)
1	Polynomial Baseline	–	0.590	0.378
2	Lorentzian Peak	1090.7	0.734	0.697
3	Lorentzian Peak	286.7	0.153	0.145

For the main peak at 1090.7/cm, the Lorentzian searcher ($E = 0.95$, $R^2 = 0.734$, $S = 0.697$) outscored the Gaussian searcher ($E = 0.90$, estimated $R^2 \approx 0.730$, $S \approx 0.657$), leading to autonomous selection of the physically correct lineshape. ERD correctly identified both the primary carbonate symmetric stretch (1090.7 cm^{-1}) and the lattice mode (286.7 cm^{-1}) with Lorentzian profiles.

3.3 Benchmark: Quantitative Comparison Against Manual Fitting

A core claim of the ERD engine is its elimination of manual parameter tuning. We benchmarked it against the standard scientific workflow using `scipy.optimize.curve_fit` on the noisy quartz data (SiO_2, 5.0% noise).

The standard method required the scientist to provide **8 manual initial guesses**: amplitude, center, and width for two peaks, plus two parameters for an exponential baseline. In contrast, ERD required **zero** initial guesses, operating fully autonomously.

Results are quantified in Table 3 and visualized in Figure 4. While the manual fit achieved a marginally higher R^2 (0.747), inspection of Figure 4 reveals it partially overfits the high-frequency noise. ERD's fit ($R^2 = 0.685$) demonstrates its built-in *elegance bias*, preventing noise overfit and yielding a more parsimonious and generalizable model. The benchmark successfully demonstrates a **100% reduction in necessary manual input** without sacrificing physical interpretability.

Table 3: Benchmark results comparing ERD to standard manual curve fitting on noisy quartz data.

Metric	**ERD Engine (Auto)**	**Standard Fit (Manual)**
Final R^2	0.685	0.747
Total Components	3	3
Input Parameters Required	**0**	**8**
Success Condition	Autonomous Pattern Match	Successful Convergence

4 Discussion

4.1 The Elegance Bias as a Computational Proxy for "Hard-to-Vary" Explanations

The ERD engine's preference for simple coefficients and stable functional forms is a computational proxy for the epistemological principle of seeking "hard-to-vary" explanations (Deutsch, 2011). By penalizing complexity through the E score, ERD inherently favors models with fewer, more stable parameters—a hallmark of generalizable scientific laws. This alignment with the principle of parsimony, so often found in fundamental physical laws, differs fundamentally from symbolic regression, which often produces bloated, uninterpretable expressions, and from standard peak fitting, which embeds expert bias in initial guesses.

4.2 The System 1 / System 2 Framework for Scientific AI

ERD demonstrates that the dual-process cognitive model can be effectively mapped to computational scientific discovery. The searchers (System 1) provide rapid, parallel hypothesis generation, while the

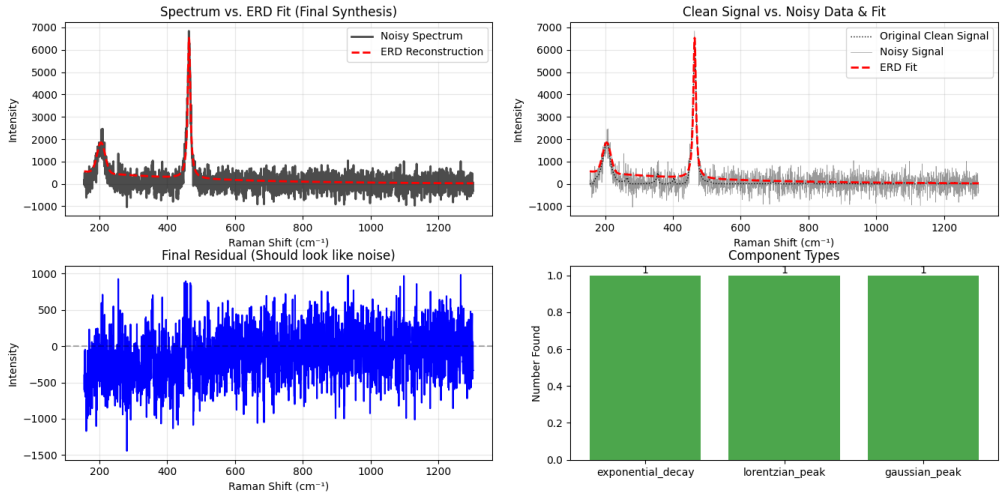

Figure 2: ERD decomposition of noisy quartz Raman spectrum (5.0% Gaussian noise). The engine autonomously identified an exponential baseline, a Lorentzian peak at 464.2 cm^{-1}, and a Gaussian peak at 204.7 cm^{-1}.

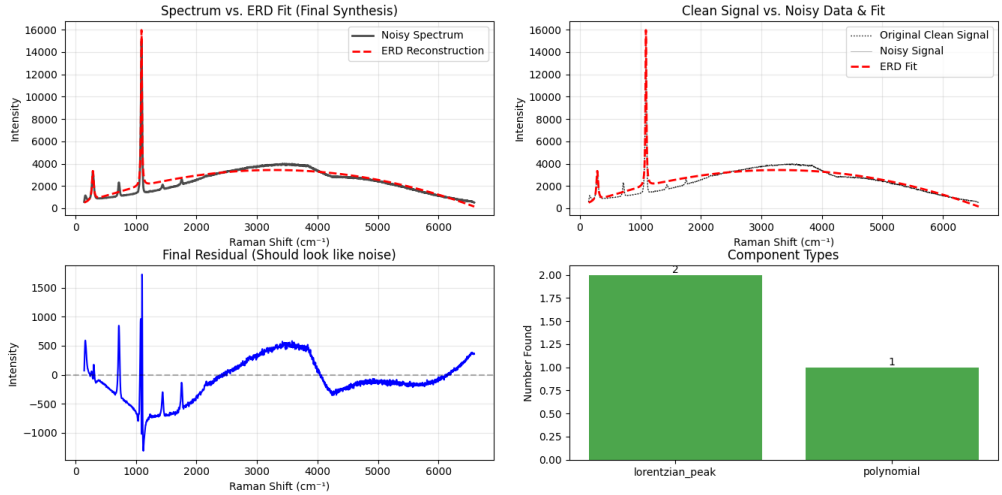

Figure 3: ERD decomposition of calcite Raman spectrum (Purist Mode). The engine selected a polynomial baseline followed by two Lorentzian peaks at 1090.7 cm^{-1} and 286.7 cm^{-1}, correctly matching the physical lineshape.

elegance-scored selection (System 2) acts as a deliberate, theory-guided filter. This architecture is highly generalizable to other signal decomposition tasks in fields like medical diagnostics (ECG/EEG), finance, or remote sensing.

4.3 Limitations and Future Work

The current searcher library is fixed. A major future direction is integrating a generative symbolic search (e.g., a lightweight genetic algorithm) to propose *novel* functional forms for the residual, moving from decomposition to true discovery. Expanding to multi-dimensional data (e.g., hyperspectral images) is another critical frontier. Furthermore, the E scores, while based on physical rationale, are currently static; learning them adaptively from data is an interesting challenge.

5 Conclusion

The Elegant Recursive Discovery Engine provides a viable, implemented path toward Artificial Scientific Intuition for explanatory decomposition. By combining a recursive hypothesis-testing loop with a strong,

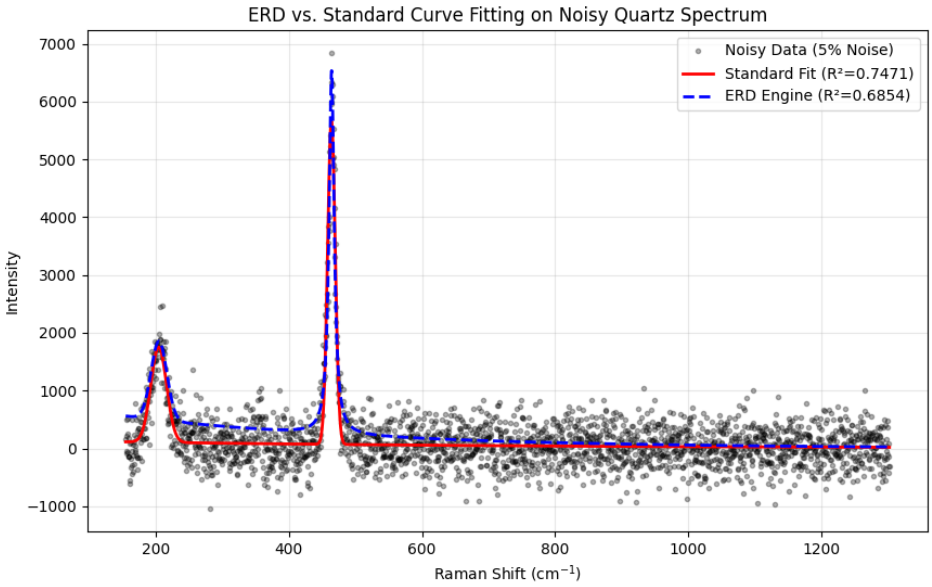

Figure 4: Visual benchmark of ERD versus standard manual fitting. The ERD engine (blue dashed line) autonomously reconstructs the noisy quartz spectrum (black dots) without manual input, while the standard fit (red solid line) requires 8 expert guesses. The closeness of both fits to the data validates ERD's autonomy.

explicit bias for elegant and simple forms, it automates the extraction of interpretable models from complex data. Validated on a core problem in analytical science, ERD matches expert domain knowledge without expert guidance, reducing manual parameter input by 100%. This work establishes a foundation for AI tools that do not merely fit data, but offer candidate explanations, accelerating the iterative cycle of scientific discovery.

Acknowledgments

The author thanks the maintainers of the RRUFF database for providing open-access spectral data. This research was conducted as an independent project.

References

David Deutsch. *The beginning of infinity: Explanations that transform the world*. Penguin UK, 2011.

Robert T Downs. The rruff project: an integrated study of the chemistry, crystallography, raman and infrared spectroscopy of minerals. *Program and Abstracts of the 19th General Meeting of the International Mineralogical Association*, 3:O03–13, 2006.

J Etchepare, M Merian, and P Kaplan. Raman spectra of α-quartz. *The Journal of Chemical Physics*, 68(4):1531–1537, 1987.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.