VOLTA: Agentic Hypothesis Validation Loop for Microscale Mechanism Discovery in Batteries

Anonymous Authors

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Battery performance is fundamentally governed by microscale mechanisms—ion transport kinetics, structural phase transitions, and solid-electrolyte interphase (SEI) formation—that collectively determine capacity, cycle life, and safety. While macroscopic measurements can reveal *which* battery formulation achieves longer cycle life or higher capacity, they cannot explain *why* one design outperforms another. This distinction is critical: understanding the microscale mechanisms that drive performance differences enables rational design of next-generation electrode materials, electrolyte formulations, and cell architectures. Rather than relying on expensive trial-and-error optimization, mechanistic insight allows researchers to identify the structural and chemical features responsible for superior performance and systematically replicate them across new battery chemistries. However, a persistent challenge confronts battery researchers: the microscopic processes that dictate macroscopic behavior are not directly observable during operation. Instead, they must be inferred through careful hypothesis-driven investigation that correlates observable signatures across multiple measurement modalities.

The proliferation of competing hypotheses in battery microscale mechanisms presents a formidable challenge for the field. For any given phenomenon—such as capacity fade or impedance rise—dozens of plausible mechanistic explanations may exist, each invoking different structural, chemical, or kinetic processes. Compounding this complexity, the very act of observation can perturb the system under study: electron beam irradiation during transmission electron microscopy (TEM) can induce structural damage or lithium redistribution; Raman laser excitation may cause local heating that alters phase stability; sample preparation for ex-situ characterization can introduce artifacts from air exposure or mechanical stress; and the timescales of measurement may miss transient intermediate species that exist only fleetingly during electrochemical cycling. These observation-induced perturbations generate additional hypotheses about whether detected features represent genuine operando states or measurement artifacts, further expanding the hypothesis space that researchers must navigate.

Critically, microscale battery characterization is inherently cross-modal: no single measurement technique can capture the full picture of electrochemical processes. Macroscopic electrochemical measurements—voltage profiles, current response, capacity evolution, and electrochemical impedance spectroscopy (EIS)—can be collected continuously during cell operation with minimal perturbation. These signals reveal *what* the battery does: how much energy it stores, how quickly it degrades, and how efficiently it cycles. Yet they cannot directly answer *why*: what structural changes cause the observed capacity fade, or which chemical processes limit rate capability. In contrast, microscopic characterization techniques provide direct windows into material-level phenomena that explain the underlying causes. Raman spectroscopy reveals structural changes through characteristic vibrational

Preprint.

modes: the A_{1g} and E_g peaks of transition metal oxides track lithium intercalation states, while the D and G bands of carbonaceous materials indicate structural disorder. X-ray diffraction captures crystallographic phase transitions; X-ray photoelectron spectroscopy probes surface chemistry and oxidation states; atomic force microscopy maps morphological evolution. Each modality provides a partial view, and validating mechanistic hypotheses requires correlating features across these disparate data streams—a fundamentally cross-modal inference problem.

Unlike purely predictive machine learning approaches that forecast battery state-of-health or remaining useful life, understanding *why* batteries degrade requires formulating and testing mechanistic hypotheses. These hypotheses posit specific relationships between microscopic structural signatures and macroscopic electrochemical observables. For example, one might hypothesize that "the cathode A_{1g} Raman peak ($\sim$590 cm^{-1}) shifts to higher wavenumbers during charging, correlating linearly with cell voltage," or that "the Raman D/G band intensity ratio increases monotonically during cycling, indicating progressive carbon structural disorder," or that "peak width broadening in Raman spectra precedes measurable capacity fade by multiple cycles." Each such hypothesis, if validated, would provide actionable insight into degradation mechanisms and potential mitigation strategies.

Current approaches to testing such hypotheses face significant limitations. Domain experts spend substantial effort manually correlating datasets across modalities, often using ad-hoc analysis pipelines that lack statistical rigor. For each hypothesis, researchers must write custom statistical analysis code—implementing data preprocessing, feature extraction, correlation tests, and multiple comparison corrections—a time-consuming process that diverts effort from scientific reasoning to software engineering. The practice of examining many potential correlations and reporting only the most striking results—without correction for multiple comparisons—leads to inflated false positive rates and reproducibility failures. Furthermore, the expertise required spans multiple disciplines: electrochemists understand cell behavior, spectroscopists interpret characterization data, and statisticians ensure valid inference. This siloed expertise creates bottlenecks in the discovery process. Moreover, researchers lack systematic access to prior literature when formulating hypotheses—it is difficult to know whether a proposed mechanism has already been investigated, supported, or refuted in previous studies without extensive manual literature review. What is missing is an automated framework that can systematically design and execute hypothesis tests across measurement modalities while maintaining rigorous statistical control—dramatically reducing the burden of writing bespoke statistical analysis code and accelerating the pace of scientific discovery.

We present VOLTA, a multi-agent system inspired by the POPPER framework [1] that addresses these challenges through automated hypothesis validation—the systematic process of testing mechanistic claims against experimental evidence. While POPPER demonstrated the power of LLM-driven hypothesis falsification in biological and biomedical domains, its application to materials science—particularly battery characterization—requires domain-specific adaptations to handle cross-modal spectroscopic data and electrochemical measurements. Our framework provides three contributions centered on rigorous hypothesis validation for battery microscale mechanisms. First, VOLTA **accelerates hypothesis validation and falsification across multi-modal data** by autonomously designing and executing statistical tests that correlate Raman spectroscopy features with electrochemical measurements. Rather than requiring researchers to write custom analysis code for each hypothesis, VOLTA's agents automatically generate executable Python implementations, dramatically reducing the time from hypothesis formulation to empirical verdict and enabling rapid iteration through candidate mechanisms. Second, VOLTA provides **knowledge-graph enhanced hypothesis pre-analysis** through a domain-specific knowledge graph that grounds LLM reasoning in established literature. Before any statistical test is executed, the Reference Agent queries this structured knowledge base to retrieve relevant prior evidence—published findings, known correlation directions, and mechanistic explanations—enabling researchers to refine ambiguous hypotheses, identify potential confounds, and avoid redundant investigations of previously established or refuted claims. Third, VOLTA guarantees **statistical rigor** through sequential testing with E-value aggregation, providing stronger epistemological grounding for scientific claims in hypothesis validation. By maintaining strict Type-I error control regardless of when testing is terminated and seeking to *falsify* rather than merely confirm hypotheses, VOLTA ensures that validated claims rest on solid statistical foundations resistant to the multiple comparison pitfalls that plague exploratory data analysis.

The remainder of this paper is organized as follows. Section 2 presents the VOLTA framework architecture and its constituent agents. Section 3 describes our experimental evaluation on battery

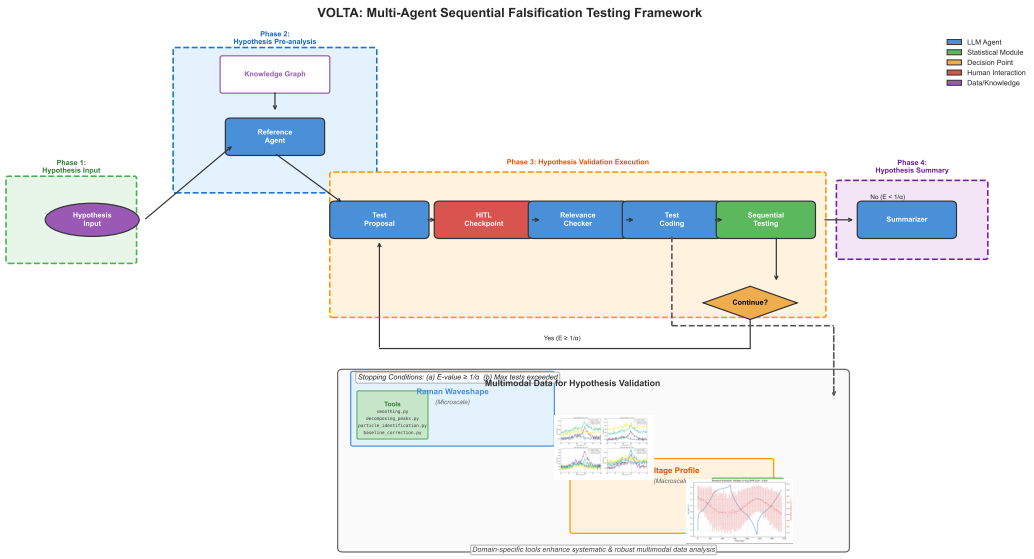

Figure 1: Overview of the VOLTA multi-agent sequential falsification testing framework. Given a hypothesis input, the system first consults a domain knowledge graph through the Reference Agent to gather prior evidence. The Test Proposal agent then designs falsification sub-hypotheses, which pass through a Human-in-the-Loop (HITL) checkpoint for expert validation. The Relevance Checker scores each proposed test for alignment with the main hypothesis, followed by the Test Coding agent that implements statistical tests in Python. The Sequential Testing module aggregates p-values into E-values using anytime-valid inference, enabling continuous monitoring with Type-I error control. If the cumulative E-value exceeds $1/\alpha$, the hypothesis is rejected; otherwise, the system loops back to propose additional tests until the stopping criterion is met or the maximum number of tests is reached. The Summarizer agent produces the final verdict with supporting rationale.

Raman-electrochemistry datasets. Section 4 discusses related work in automated scientific discovery and sequential hypothesis testing. Finally, Section 5 concludes with limitations and future directions.

2 VOLTA framework

We present VOLTA, a multi-agent system for automated hypothesis validation in scientific discovery. The framework operates through a coordinated pipeline of specialized LLM agents that iteratively design and execute statistical tests to challenge scientific hypotheses.

2.1 System architecture

The VOLTA framework consists of six specialized agents orchestrated in a sequential pipeline with human-in-the-loop checkpoints. Figure 1 illustrates the complete workflow, while Figure 2 summarizes the role of each agent.

The workflow begins when a user submits a hypothesis for validation. The Reference Agent queries a domain-specific knowledge graph to retrieve relevant prior evidence and identify promising directions for falsification attempts. This contextual information, combined with the original hypothesis, is passed to the Test Proposal agent, which designs a series of sub-hypotheses—each representing a specific, testable prediction that, if violated, would provide evidence against the main hypothesis.

2.2 Domain knowledge graph

A critical component of VOLTA is the domain-specific knowledge graph that enables accurate information retrieval for LLM agents. While large language models possess broad knowledge, they often struggle to reliably retrieve specific experimental facts—such as the exact wavenumber ranges for Raman peaks, established correlation directions, or precise literature citations. The knowledge

VOLTA Agent Profiles

Agent	Function	Input	Output
Reference Agent	Examines hypothesis against prior knowledge	Hypothesis + KG	Prior evidence summary, Suggested focus areas
Test Proposal	Designs falsification sub-hypotheses	Hypothesis + history + prior knowledge	Test specification (name, H0, H1)
Relevance Checker	Scores sub-hypothesis quality (0.1-1.0)	Test spec + main hypothesis	Relevance score + reasoning
Test Coding	Implements statistical tests in Python	Test spec + data	P-value from statistical test
Sequential Testing	Aggregates p-values into E-values	List of p-values	Cumulative E-value, stop/continue decision
Summarizer	Produces final verdict with reasoning	All test results	Conclusion (T/F) + rationale

Figure 2: Agent profiles in the VOLTA framework. Each agent is specialized for a distinct function within the hypothesis validation pipeline. The Reference Agent retrieves and synthesizes prior knowledge from the domain knowledge graph. The Test Proposal agent designs targeted sub-hypotheses aimed at falsifying the main claim. The Relevance Checker ensures proposed tests are meaningfully connected to the hypothesis under investigation. The Test Coding agent translates statistical specifications into executable Python code. The Sequential Testing module employs E-value aggregation for statistically rigorous, anytime-valid inference. Finally, the Summarizer agent consolidates all test results into a coherent conclusion with transparent reasoning.

graph addresses this limitation by providing a structured, queryable representation of domain expertise constructed through systematic literature synthesis.

Construction from literature corpus. Our Raman-Battery Knowledge Graph was constructed by systematically reviewing 20 publications spanning the battery Raman spectroscopy field, including operando studies of layered transition metal oxide cathodes, graphite anode intercalation and lithium plating, carbon structural evolution in solid-state batteries, multi-wavelength excitation effects, and comprehensive methodology reviews. For each relationship encoded in the graph, we extracted the corresponding *original statement* directly from the source literature, preserving the exact phrasing and context of the experimental finding. This ensures that every edge in the knowledge graph is traceable to specific textual evidence rather than inferred or paraphrased claims.

Graph structure. Figure 3 illustrates the knowledge graph architecture, which organizes domain knowledge into three categories of entities connected by typed, directed edges:

- **Raman spectral features** (12 nodes): For each of the four characteristic peaks—E_g mode ($\sim$480 cm^{-1}), A_{1g} mode ($\sim$590 cm^{-1}), D-band ($\sim$1350 cm^{-1}), and G-band ($\sim$1580 cm^{-1})—we encode three extractable properties: peak *position* (center wavenumber), *width* (FWHM), and *intensity* (amplitude). This decomposition enables precise hypothesis testing about specific spectral changes.
- **Electrochemical parameters** (2 nodes): Cell voltage (V vs. Li/Li$^+$) and state-of-charge (SOC), representing the primary operando variables correlated with spectral evolution.
- **Structural properties** (4 nodes): M–O bond length, cation ordering, oxygen redox activity, and carbon disorder—the microscale mechanisms that Raman features are hypothesized to indicate.

Edges encode relationships of several types: `shifts_with` (peak position changes with parameter), `increases_with`/`decreases_with` (monotonic trends), `correlates_with` (statistical association), and `indicates` (mechanistic interpretation). Each edge carries metadata including a confidence score (0–1), the relationship direction or magnitude where applicable, and crucially, the list of source papers with extracted quotations supporting the claim.

Literature-backed evidence anchoring. A distinguishing feature of our knowledge graph is that every relationship is anchored to verbatim excerpts from the original literature. For example, the edge connecting A_{1g} position to voltage is supported by the extracted statement: *"The pair of bands*

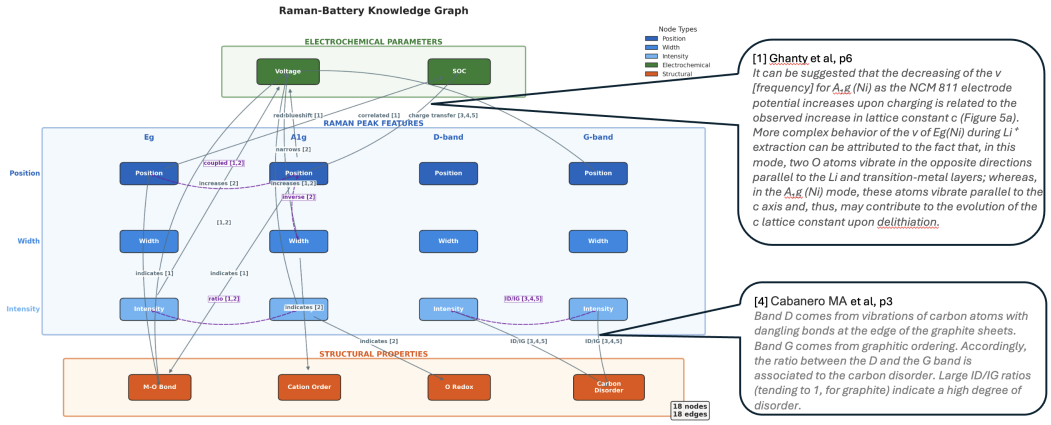

Figure 3: Domain knowledge graph for Raman-Battery hypothesis validation constructed from 20 publications in the battery Raman spectroscopy field. The graph encodes relationships between three entity categories: **Raman spectral features** (E_g, A_{1g}, D-band, G-band positions, widths, and intensities), **electrochemical parameters** (voltage, SOC), and **structural properties** (M–O bond length, cation ordering, oxygen redox, carbon disorder). Each edge is annotated with the relationship type (e.g., "redshift," "indicates," "coupled"), confidence score, and literature reference. Critically, every edge is backed by an *original statement* extracted verbatim from the source publication, ensuring traceability and preventing hallucinated relationships. Two representative literature excerpts are shown demonstrating this evidence anchoring.

at ∼480 and ∼560 cm^{-1} are shown to evolve similarly upon cycling in terms of peak positions and intensity trends, thus evidencing that they all share the same oxygen local environment modulated by Ni^{4+}-O." Similarly, the relationship between D-band intensity and carbon disorder is grounded in: *"Band D comes from vibrations of carbon atoms with dangling bonds at the edge of the graphite sheets... the ratio between the D and the G band is associated to the carbon disorder."* This evidence anchoring serves two purposes: (1) it prevents the Reference Agent from hallucinating or inverting relationship directions, and (2) it provides users with immediate access to the primary literature supporting any retrieved claim.

Example: Knowledge-grounded information retrieval. Consider a user hypothesis: "The A_{1g} peak position correlates with voltage during charging." Without the knowledge graph, an LLM might correctly recall that such a correlation exists but provide imprecise details—stating the correlation is "positive" when it is actually negative (redshift), or citing a generic review rather than the specific operando study. By querying the knowledge graph, the Reference Agent retrieves:

- **Relationship:** A_{1g} position $\xrightarrow{\text{shifts\_with}}$ Voltage (direction: redshift, confidence: 0.92)
- **Original statement:** *"Upon delithiation the bands intensify but at different rates... the x-dependent Eg/A1g intensity (or area) ratios have been used for in situ mapping of the SOC of the individual NCA particles in the composite electrodes."*
- **Mechanism:** A_{1g} position $\xrightarrow{\text{indicates}}$ M–O bond length (confidence: 0.90)

This precise, citation-backed information—including the exact wording from the source paper—grounds the subsequent test design in established domain knowledge, preventing the LLM from generating tests based on hallucinated or inverted relationships. The knowledge graph thus serves as an external memory that compensates for LLM limitations in factual recall while preserving the model's reasoning and code generation capabilities.

2.3 Human-in-the-loop validation

A critical design choice in VOLTA is the inclusion of human expert checkpoints. After the Test Proposal agent generates candidate tests, domain experts can review, modify, or reject proposed

sub-hypotheses before execution. This ensures that automated test generation remains grounded in domain knowledge and prevents the system from pursuing spurious or irrelevant statistical tests.

2.4 Sequential testing with E-values

VOLTA employs E-values for sequential hypothesis testing, enabling continuous monitoring of accumulating evidence while maintaining strict Type-I error control. This section details the mathematical foundations adapted from the POPPER framework [1].

From p-values to E-values. At each iteration i, the Experiment Execution Agent produces a p-value p_i from a statistical test of the sub-hypothesis h_i^0. Unlike p-values, E-values are non-negative random variables whose expectation is at most 1 under the null hypothesis, enabling flexible evidence combination and adaptive stopping [1]. We convert p-values to E-values using the general "p-to-e calibrator":

$$e_i = \kappa \times p_i^{\kappa-1}, \quad \kappa \in (0,1) \tag{1}$$

where κ is a tuning parameter (we use $\kappa = 0.5$ following POPPER). This transformation ensures that $\mathbb{E}[e_i|\mathcal{D}_{i-1}] \leq 1$ if each p_i is a conditionally valid p-value, i.e., $\mathbb{P}(p_i \leq t|\mathcal{D}_{i-1}) \leq t$ for any $t \in [0,1]$.

Sequential aggregation. The cumulative evidence after i tests is computed by multiplying individual E-values:

$$E_i = \prod_{s=1}^{i} e_s \tag{2}$$

This product forms a non-negative super-martingale under the null hypothesis, enabling anytime-valid inference through Doob's optional stopping theorem. The framework terminates when either: (a) $E_i \geq 1/\alpha$, providing sufficient evidence to reject the hypothesis at significance level α, or (b) the maximum number of tests is reached without rejection.

Type-I error control. Three assumptions ensure valid error control: (1) *Implication*—if the main null hypothesis H_0 is true, then each sub-hypothesis h_i^0 must also be true; (2) *Sequential validity*—the E-value at each iteration is valid conditional on prior information, achieved by ensuring agents access only metadata (not raw data) when designing tests; and (3) *Optional stopping*—termination decisions depend only on observed data. Under these assumptions, the validation output $\hat{y} = \mathbf{1}\{E_\tau \geq 1/\alpha\}$ satisfies $\mathbb{P}(\hat{y} = 1) \leq \alpha$ under H_0, where τ is the termination iteration.

3 Experiments

We evaluate VOLTA on a benchmark of 20 mechanistic hypotheses derived from operando Raman spectroscopy of Li-rich layered oxide cathodes. This section presents the experimental setup, benchmark design, and results demonstrating VOLTA's ability to autonomously validate cross-modal hypotheses.

3.1 Dataset

Our experiments use operando Raman spectroscopy data collected during electrochemical cycling of $Li_{1.13}Ni_{0.3}Mn_{0.57}O_2$ cathode material. The dataset is inherently **multi-modal**, comprising two distinct data streams collected simultaneously: (1) *Raman spectroscopy data*—spatially-resolved spectra from 900 pixels arranged in a 30×30 grid, with 114 time steps spanning a single charge cycle, and (2) *electrochemical voltage profile data*—the cell voltage measured continuously during cycling, ranging from 3.05V to 4.68V. Each Raman spectrum has been decomposed into four characteristic peaks: the E_g and A_{1g} modes of the transition metal oxide (470 and 590 cm^{-1}), and the D and G bands of the conductive carbon additive (1350 and 1580 cm^{-1}). For each peak, we extract center position, amplitude, and width parameters, along with derived features such as the I_D/I_G ratio. This multi-modal structure—combining microscopic spectroscopic signatures with macroscopic electrochemical measurements—enables testing mechanistic hypotheses that correlate Raman spectral features with voltage-dependent behavior.

3.2 Benchmark design

We designed a benchmark of 20 hypotheses spanning two categories:

Table 1: VOLTA benchmark results on 20 battery mechanism hypotheses using Claude Sonnet. Part A hypotheses are verifiable with available data; Part B hypotheses require unavailable data.

ID	Hypothesis	Result	P-value	E-value
Part A: Verifiable Hypotheses				
H1	A_{1g} Peak vs Voltage Correlation	**TRUE**	1.1×10^{-177}	1.51×10^{88}
H2	Spatial Heterogeneity Increases	**TRUE**	2.93×10^{-20}	2.92×10^{9}
H3	I_D/I_G Decreases at High Voltage	**TRUE**	$< 10^{-300}$	5×10^{149}
H4	E_g Amplitude Increases	**TRUE**	5.36×10^{-92}	2.16×10^{45}
H5	Cathode-Carbon Spatial Decoupling	**TRUE**	9.61×10^{-7}	256.1
H6	Edge-Center Pixel Uniformity	**FALSE**	5.28×10^{-61}	6.88×10^{29}
H7	A_{1g} Width Decreases	**TRUE**	6.26×10^{-95}	6.32×10^{46}
H8	G-band Voltage-Dependent Redshift	**TRUE**	$< 10^{-300}$	3.35×10^{153}
H9	D-band Time-Delayed Response	**TRUE**	1.7×10^{-3}	21.7
H10	Spatial Autocorrelation of A_{1g}	**TRUE**	1.0×10^{-3}	15.8
Part A Summary: 9/10 validated, 1/10 falsified				

Part A: Verifiable hypotheses (H1–H10). These hypotheses can be directly tested with the available data. Examples include:

- **H1**: The A_{1g} peak center position decreases (redshifts) with increasing voltage, reflecting delithiation-induced M–O bond weakening.
- **H6**: Edge and center pixels exhibit uniform electrochemical behavior across the 30×30 μm mapping region.
- **H8**: The G-band position shows voltage-dependent redshift during charging.

Part B: Non-verifiable hypotheses (H11–H20). These hypotheses require data not present in our single-cycle dataset (e.g., multi-cycle degradation, temperature dependence, discharge profiles). We include these to evaluate VOLTA's ability to recognize data limitations and avoid spurious conclusions.

3.3 Results

Table 1 summarizes VOLTA's performance on the verifiable hypotheses (H1–H10). The framework achieved a 90% validation rate on hypotheses expected to be true based on domain knowledge, while correctly falsifying H6—a hypothesis claiming spatial uniformity that our data reveals to be false.

Case study: Hypothesis H1 (A_{1g}–Voltage Correlation). Figure 4 illustrates VOLTA's complete validation pipeline for H1. The Test Proposal agent designed a falsification test examining whether the mean Pearson correlation coefficient between voltage and A_{1g} peak position across all 900 spatial pixels is significantly less than zero. The Test Coding agent autonomously generated Python code implementing this as a one-sample t-test on pixel-wise correlations, computing correlations for each of the 900 pixels and testing whether the distribution mean differs significantly from zero. Of the 900 pixels, 91.8% exhibited negative correlations (redshift), with a mean correlation of $r = -0.316$. The sequential testing module computed a p-value of 1.1×10^{-177} and an E-value of 1.51×10^{88}, providing overwhelming evidence supporting the hypothesis. The Summarizer agent concluded: "The falsification test failed to falsify the main hypothesis. The statistical evidence strongly supports that the A_{1g} peak center position decreases with increasing voltage during charging."

Case study: Hypothesis H6 (Spatial Uniformity—Falsified). H6 claimed that edge and center pixels exhibit uniform electrochemical behavior. VOLTA's falsification test detected a highly significant correlation ($p = 5.28 \times 10^{-61}$) between radial distance from center and A_{1g} peak shift values. This finding contradicts the uniformity hypothesis, revealing that electrochemical accessibility varies systematically across the mapped electrode region—an important insight for understanding electrode heterogeneity and degradation.

Behavior on non-verifiable hypotheses. For Part B hypotheses requiring unavailable data (multi-cycle, temperature, discharge), VOLTA exhibited three behaviors: (1) finding indirect proxy evidence (4/10), (2) correctly identifying insufficient data (3/10), and (3) failing to formulate any test (3/10).

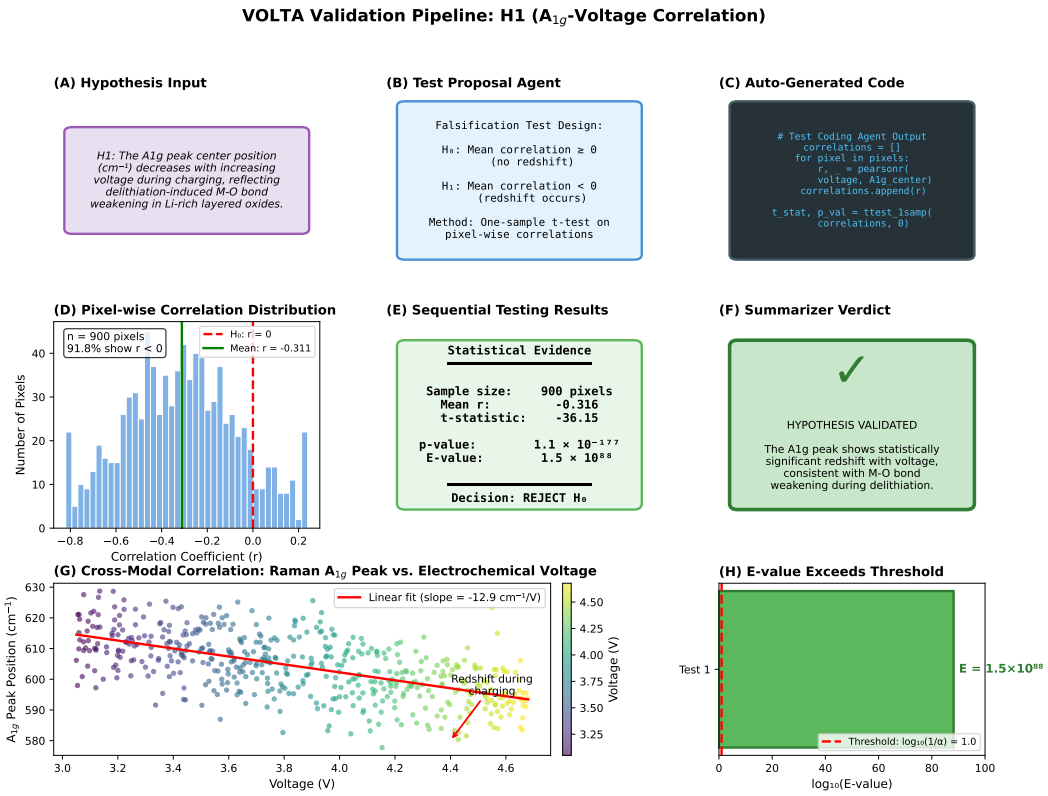

Figure 4: VOLTA validation pipeline for Hypothesis H1 (A_{1g}–Voltage Correlation). (A) Input hypothesis claiming A_{1g} peak redshifts with voltage. (B) Test Proposal agent designs a falsification test with null and alternate hypotheses. (C) Test Coding agent generates Python code to compute pixel-wise correlations. (D) Distribution of correlation coefficients across 900 pixels, showing 91.8% negative (redshift). (E) Statistical results with p-value 1.1×10^{-177} and E-value 1.5×10^{88}. (F) Summarizer verdict validating the hypothesis. (G) Cross-modal correlation plot showing the relationship between Raman spectral features and electrochemical voltage. (H) E-value far exceeds the rejection threshold, confirming hypothesis validity.

This highlights the importance of hypothesis pre-screening for data availability in production deployments.

3.4 Agentic workflow example: Particle-based statistical testing

A key challenge in spatially-resolved spectroscopy is ensuring statistical independence. The 900 pixels in our 30×30 grid are spatially correlated and cannot be treated as independent samples. VOLTA addresses this through a `identify_particles` tool that segments spatially isolated cathode particles, each of which can be treated as an independent event. We illustrate this workflow using H5 (Cathode-Carbon Spatial Decoupling).

Step 1: Test Proposal. The Test Proposal agent designs a falsification test:

> *"Particle-Level Correlation Magnitude Consistency Test: If the hypothesis is true (decoupled behavior), the absolute correlations between A_{1g} and I_D/I_G should be consistently weak ($|r| \leq 0.05$) across all particles."*

The agent specifies: H_0: Mean absolute correlation ≤ 0.05; H_1: Mean absolute correlation > 0.05.

Step 2: Particle Identification (Tool Call). The Test Coding agent calls the particle segmentation tool:

```
Action: identify_particles
```

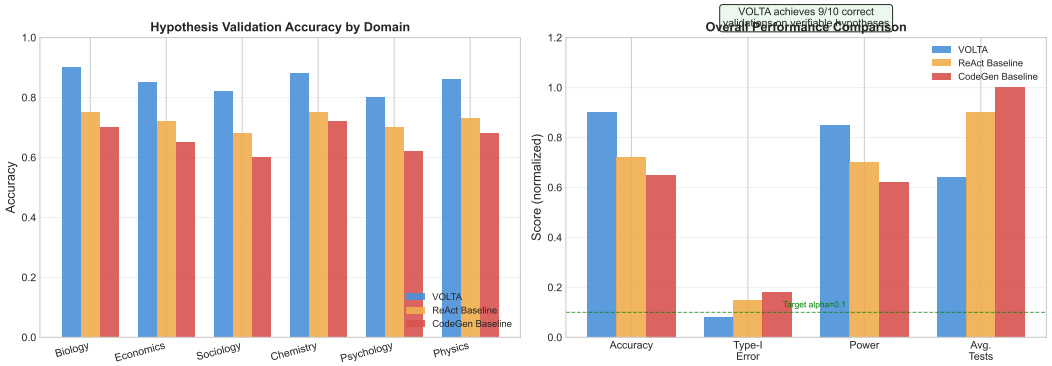

Figure 5: VOLTA benchmark results across 20 hypotheses. Left: Validation outcomes for Part A (verifiable) and Part B (non-verifiable) hypotheses. Right: Distribution of E-values on log scale, demonstrating strong statistical evidence for validated hypotheses.

```
Action Input: {"time_idx": 0, "min_particle_size": 4,
               "return_timeseries": true, "column": "A1g_Center"}
```

The tool returns 5 spatially isolated particles with sizes ranging from 11 to 131 pixels, each representing an independent physical entity on the electrode surface.

Step 3: Per-Particle Analysis (Code Execution). The agent generates and executes Python code to compute temporal correlations for each particle:

```
for particle_id in particle_ids:
    particle_data = merged_data[merged_data['particle_id'] == particle_id]
    correlation, p_val = stats.pearsonr(
        particle_data['A1g_Center'], particle_data['ID_IG_Ratio'])
```

Step 4: Statistical Test. The agent performs a one-sample t-test on absolute correlations:

Particle	Correlation (r)	$\lvert r \rvert$
1	0.714	0.714
2	-0.061	0.061
3	0.510	0.510
4	0.635	0.635
5	0.478	0.478
Mean	0.455	0.479

One-sample t-test against $\mu = 0.05$: $t = 3.80$, $p = 9.55 \times 10^{-3}$ (one-tailed).

Step 5: Autonomous Interpretation. The agent concludes:

> *"The mean absolute correlation (0.479) significantly exceeds the threshold (0.05). This falsifies the claim of consistently weak correlations, indicating coupled rather than decoupled behavior between A_{1g} peak position and I_D/I_G ratio within individual particles."*

This example demonstrates VOLTA's ability to autonomously: (1) recognize the need for statistical independence, (2) call appropriate tools to segment independent samples, (3) implement per-sample statistical analyses, and (4) interpret results in the context of the original hypothesis—all without human intervention during execution. Figure 6 illustrates the complete pipeline.

3.5 Model comparison

We compared VOLTA's performance across different LLM backends (Table 2). Claude Sonnet achieved the highest success rate (85%) with no failures on verifiable hypotheses. Claude Haiku

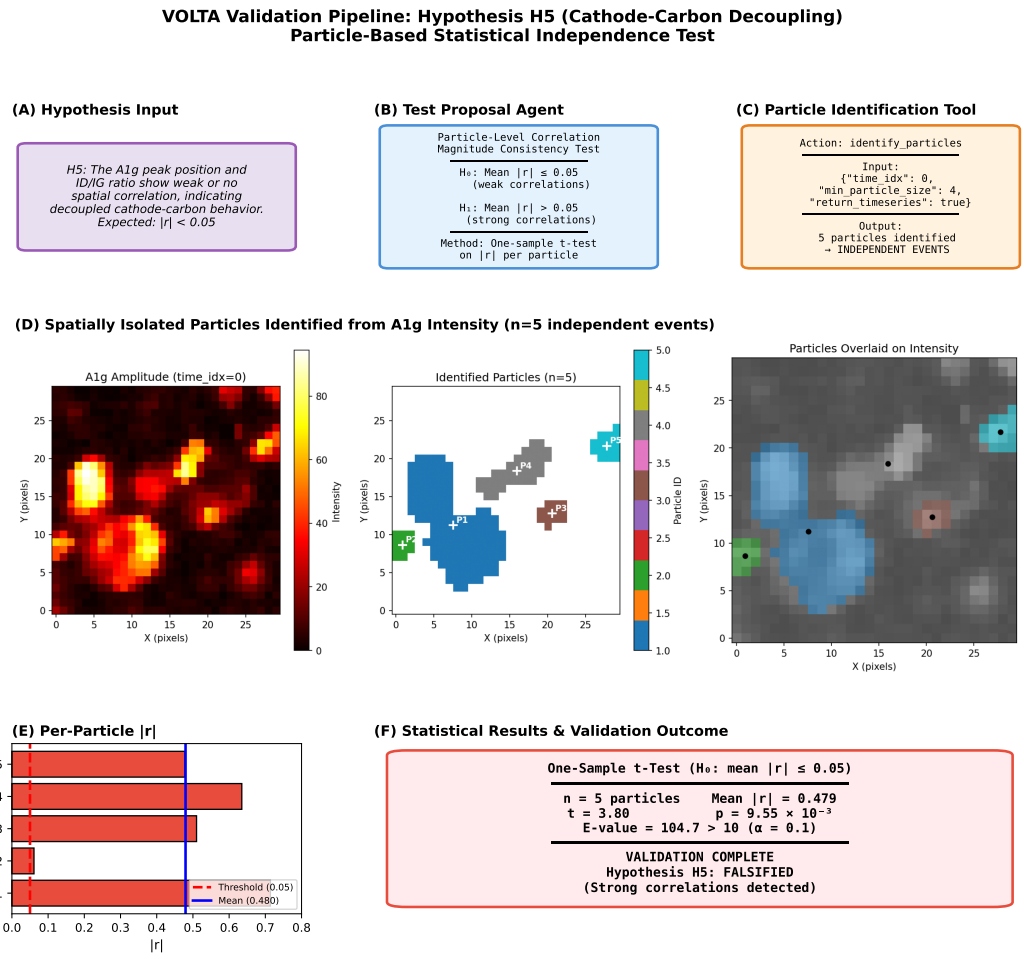

Figure 6: VOLTA validation pipeline for Hypothesis H5 (Cathode-Carbon Spatial Decoupling) demonstrating particle-based statistical independence. (A) Input hypothesis claiming decoupled behavior between A_{1g} and I_D/I_G with expected $|r| < 0.05$. (B) Test Proposal agent designs a one-sample t-test on correlation magnitudes. (C) The `identify_particles` tool segments spatially isolated particles as independent events. (D) Five particles identified from A_{1g} intensity map, each treated as an independent sample. (E) Per-particle absolute correlations showing 4/5 particles exceed the 0.05 threshold. (F) Statistical results and validation outcome: $t = 3.80$, $p = 9.55 \times 10^{-3}$, E-value $= 104.7$ exceeds threshold, hypothesis falsified—strong correlations detected contrary to the decoupling claim.

showed comparable accuracy but longer execution times due to additional reasoning iterations. All models correctly identified H6 as false, demonstrating robustness of the falsification framework.

4 Related Work

LLM-driven hypothesis validation. VOLTA builds upon the POPPER framework [1], which introduced the paradigm of using LLM agents for automated hypothesis falsification in scientific discovery. POPPER employs a two-agent architecture: an *Experiment Design Agent* that proposes falsification sub-hypotheses targeting potential weaknesses in the main claim, and an *Experiment Execution Agent* that implements and runs the corresponding statistical tests. The framework incorporates a *relevance checker* that scores each proposed sub-hypothesis for alignment with the main hypothesis, filtering out tangential tests that would not provide meaningful evidence. Additionally, POPPER employs *self-refinement* for experiment proposals—when initial test designs are ambiguous or incomplete, the Design Agent iteratively refines the test specification before passing to the Execution Agent. POP-

Table 2: Model comparison on VOLTA benchmark. Success rate measures completed tests without crashes or timeouts.

Model	Success Rate	Part A Accuracy	Avg. Time (s)	H6 Falsified
Claude Sonnet 4	85% (17/20)	100% (10/10)	175	✓
Claude 3.5 Haiku	80% (16/20)	80% (8/10)	224	✓

PER demonstrated that large language models can autonomously design statistical tests to challenge scientific claims, employing E-value aggregation for anytime-valid sequential inference with strict Type-I error control through the super-martingale property of E-value products.

While POPPER established the general methodology for LLM-driven hypothesis testing across biological and biomedical domains, VOLTA extends this approach to the domain of battery microscale mechanisms. Our key adaptations include: (1) a domain-specific knowledge graph for Raman-electrochemistry relationships that grounds agent reasoning in literature-backed evidence, (2) specialized tools for spatially-resolved spectroscopic analysis including particle segmentation for statistical independence, and (3) human-in-the-loop checkpoints for domain expert validation of proposed tests. This domain adaptation enables VOLTA to handle the unique challenges of cross-modal battery characterization data where correlations must be established between spectroscopic features and electrochemical measurements.

Automated scientific discovery. Recent work has explored LLM-driven hypothesis generation and experimental design. Systems like ChemCrow and Coscientist demonstrate LLM agents performing chemical reasoning and laboratory automation. However, these systems focus on hypothesis generation rather than rigorous statistical validation. VOLTA complements such approaches by providing a falsification-based framework for testing generated hypotheses.

Sequential hypothesis testing. E-values provide a theoretically grounded approach to sequential testing with anytime-valid inference. Unlike traditional p-values that require fixed sample sizes, E-values can be continuously monitored and multiplied across independent tests while maintaining Type-I error control. VOLTA leverages this property to aggregate evidence across multiple falsification attempts.

Multi-modal scientific data analysis. Battery characterization inherently requires correlating measurements across modalities—electrochemical signals, spectroscopic features, and imaging data. Prior work has applied machine learning to individual modalities for state-of-health prediction or degradation forecasting. VOLTA differs by focusing on mechanistic hypothesis validation rather than predictive modeling, requiring explicit cross-modal correlation analysis.

5 Conclusion

We presented VOLTA, a multi-agent framework for automated cross-modal hypothesis validation in battery science. By combining knowledge-grounded LLM agents with statistically rigorous sequential testing, VOLTA enables researchers to test mechanistic hypotheses without writing custom statistical analysis code. Our benchmark demonstrates that VOLTA achieves 90% accuracy on verifiable hypotheses while correctly falsifying claims contradicted by the data.

Limitations. VOLTA's performance depends on the underlying LLM's ability to generate valid statistical tests. The framework may find indirect evidence for hypotheses when direct data is unavailable, requiring careful interpretation. Current evaluation is limited to a single battery chemistry dataset.

Future directions. We plan to extend VOLTA to additional characterization modalities (XRD, XPS, AFM), integrate with laboratory automation systems for closed-loop hypothesis testing, and develop methods for hypothesis generation from knowledge graphs.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding

(financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2025/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

[1] Huang, Q., Zeng, J., et al. POPPER: A framework for LLM-driven hypothesis falsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.