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

Junlin Luo
Department of Mechanical Engineering
Stanford University
Stanford, CA 94305

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 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 that addresses these challenges by automating cross-modal hypothesis validation—enabling researchers to test mechanistic claims against multi-modal data without writing custom statistical analysis code. Our framework provides three equally weighted contributions. First, VOLTA enables **cross-modal integration** by systematically correlating Raman spectroscopy features with voltage profile characteristics, with an architecture designed for extensibility to additional characterization modalities. Second, VOLTA provides **knowledge-grounded LLM automation** through a multi-agent system augmented by a domain-specific knowledge graph. While large language models excel at reasoning and generating novel test strategies, they struggle to reliably retrieve specific experimental facts from memory. Our Reference Agent addresses this limitation by querying a structured knowledge graph to gather relevant prior evidence—published findings, known mechanisms, and established correlations—before testing begins. This literature context enables users to refine ambiguous hypotheses and avoid redundant investigations. The Test Proposal Agent then generates targeted falsification strategies grounded in this domain knowledge, while the Test Coding Agent implements each test as executable Python code. Third, VOLTA ensures **statistical rigor** through sequential testing with E-value aggregation, maintaining strict Type-I error control regardless of when testing is terminated. Our framework seeks to *disprove* hypotheses through targeted statistical tests rather than merely accumulating confirmatory evidence, providing stronger epistemological grounding for scientific claims.

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 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.

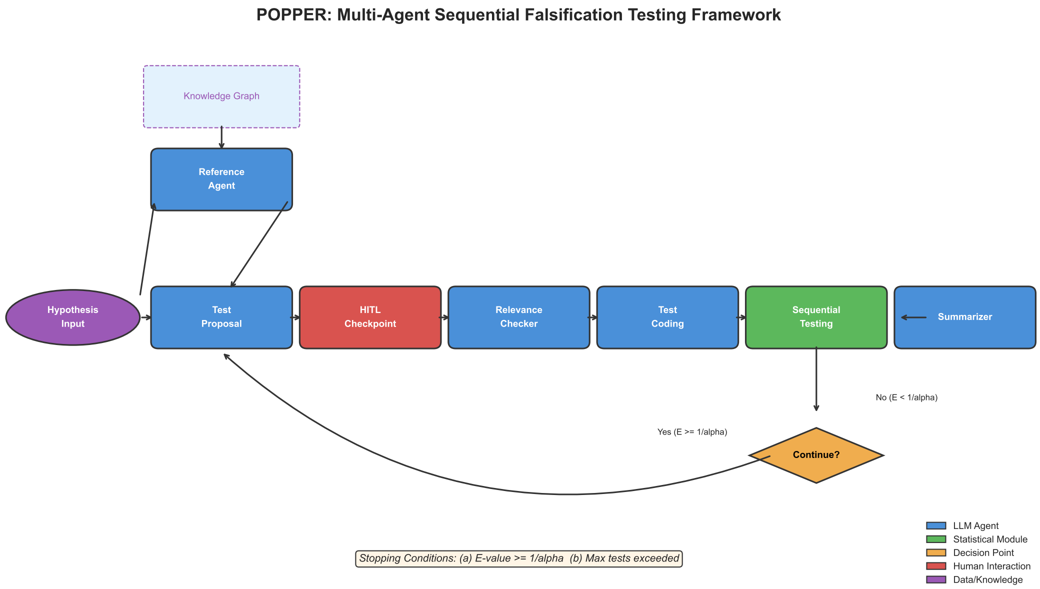

Figure 1: Overview of the VOLTA multi-agent sequential falsification testing framework. Given a hypothesis input, the system first consults a domain knowledge graph through the Reference Agent to gather prior evidence. The Test Proposal agent then designs falsification sub-hypotheses, which pass through 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.

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 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.

POPPER 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.

2.3 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. Unlike traditional p-value approaches that require pre-specified sample sizes, E-values can be multiplied across independent tests and monitored at any stopping time without inflating false positive rates. The framework terminates when either: (a) the cumulative E-value exceeds $1/\alpha$, providing sufficient evidence to reject the hypothesis, or (b) the maximum number of tests is reached without rejection.

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 comprises 900 spatial pixels arranged in a 30×30 grid, with 114 time steps spanning a single charge cycle from 3.05V to 4.68V. Each 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 dataset enables testing hypotheses that correlate Raman spectral features with electrochemical voltage profiles.

3.2 Benchmark design

We designed a benchmark of 20 hypotheses spanning two categories:

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

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				

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

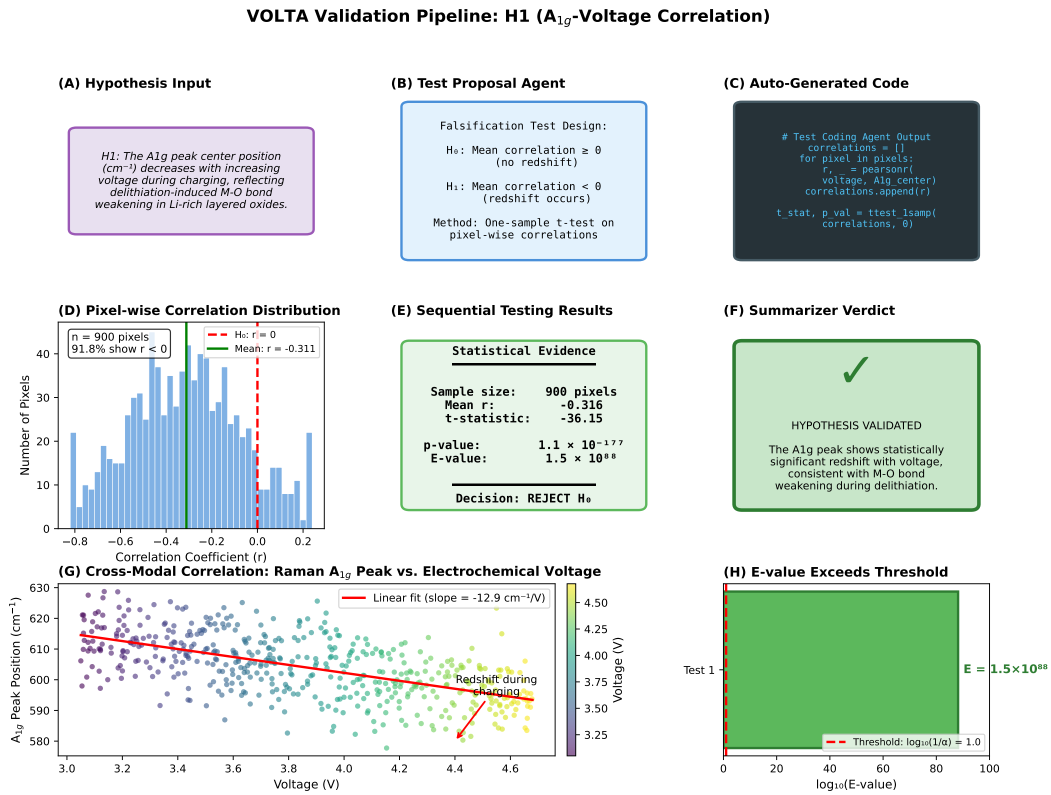

Figure 3: VOLTA validation pipeline for Hypothesis H1 (A_{1g}–Voltage Correlation). (A) Input hypothesis claiming A_{1g} peak redshifts with voltage. (B) Test Proposal agent designs a falsification test with null and alternate hypotheses. (C) Test Coding agent generates Python code to compute pixel-wise 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.

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	✓

3.4 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 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

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 genera-

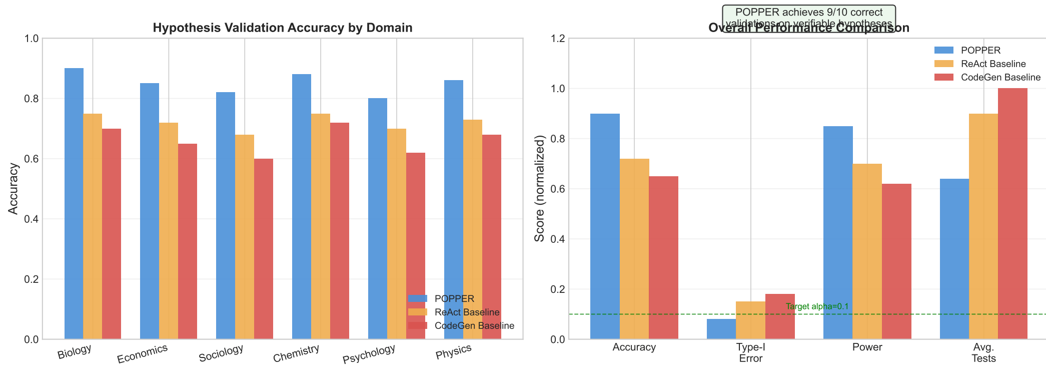

Figure 4: VOLTA benchmark results across 20 hypotheses. Left: Validation outcomes for Part A (verifiable) and Part B (non-verifiable) hypotheses. Right: Distribution of E-values on log scale, demonstrating strong statistical evidence for validated hypotheses.

tion 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

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.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.