Open Electrolyte Databank: Unlocking Molecular-Mixture Intelligence for Battery Electrolyte Discovery through a Standardized, Multimodal Foundation

Hance Su [1] **Luca Mondonico** [1] **Yiyang Shen** [2]

Abstract

Meeting AI's data-center energy demands will require higher-energy-density storage, and electrolyte chemistries govern performance, stability, and safety. Yet the data foundation for electrolyte discovery remains misaligned with the real problem: benchmarks like MoleculeNet standardized AI on single molecules, while electrolyte performance is governed by molecular mixtures, compositions, and experimental context. Current battery datasets remain centered on full-cell aging and performance with limited chemical diversity. We therefore propose **Open Electrolyte Databank**, a standardized, multimodal resource centered on molecular-mixture chemistry for battery electrolytes, initially focused on metal-anode systems where electrolyte design is especially critical for achieving very high energy density. The databank will transform fragmented literature into machine-readable, formulation-native records linking standardized compositions to protocols, outcomes, and multimodal data, including spectroscopy, molecular dynamics, and DFT, while growing through contributions from the broader electrolyte community, including labs developing automated battery workflows. Our early work already establishes proof of concept, with a Coulombic Efficiency (CE) dataset about an order of magnitude larger than previous public efforts; we treat CE and capacity retention as protocol-conditioned cell outcomes rather than composition-only electrolyte properties. By moving battery AI beyond single-molecule prediction and toward formulation-aware learning, **Open Electrolyte Databank** would provide shared infrastructure for mixture-aware benchmarking, retrieval, and closed-loop electrolyte discovery.

[1]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA [2]University of Southern California, Los Angeles, CA 90089, USA. Correspondence to: Hance Su <hancesu@stanford.edu>.

Accepted by ICML 2026 AI for Science Workshop.

1. Introduction

Datasets often determine which scientific machine learning problems become tractable. In practical lithium-metal batteries, the liquid electrolyte is the nexus connecting energy density, stability, rate capability, and safety (Wang et al., 2022). Electrolytes therefore represent a particularly timely target for AI-driven discovery, especially as high-energy-density storage becomes increasingly important for electrified transportation, grid-scale buffering, and the rapidly growing infrastructure surrounding AI compute. Yet electrolyte science is poorly matched to the dominant benchmark paradigm: much of chemistry AI still centers on single-molecule property prediction, whereas electrolyte behavior is governed by multi-component formulations, concentration, protocol, and device context. Recent work on chemical-mixture learning makes the same point, but existing mixture benchmarks like CheMixHub remain narrow, focusing on a single modality—ionic conductivity—rather than the multimodal, protocol-aware evidence required for practical battery development (Rajaonson et al., 2025).

The battery literature reflects this gap. Existing public battery datasets are still weighted toward aging and device-level performance histories rather than formulation-centric electrolyte chemistry (Lappemic, 2026; Mayemba et al., 2024), while experimental electrolyte development remains slow and iterative despite the growing promise of AI-assisted search (Ma et al., 2025). Several recent model-driven datasets, including BAMBOO, DiffMix, and GeoMix, show that partially standardized electrolyte-mixture data are already emerging and should be incorporated rather than rebuilt from scratch (Gong et al., 2025; Zhu et al., 2024; Li et al., 2026). We therefore introduce the Open Electrolyte Databank, a standardized, continuously updated, multimodal resource organized around electrolyte formulations and their associated experimental context. As an open framework, it can lower the barrier to entry for computational groups with limited experimental infrastructure, strengthen synergy between dry-lab and wet-lab research, provide algorithm developers with a realistic benchmark for mixture-aware prediction, retrieval, and agentic reasoning, and help experimental researchers navigate formulation landscapes more efficiently using AI-assisted workflows.

By linking formulations to protocols, electrochemical outcomes, and multimodal evidence, the dataset enables models to reason not only about what works, but also why specific formulations succeed under particular experimental conditions.

2. Dataset Design

Each example in the dataset is a formulation-level record that jointly captures standardized composition, experimental protocol, measured outcomes, supporting evidence, and provenance. The atomic unit is one formulation evaluated under one protocol and outcome context, not an entire paper or a molecule detached from its mixture. This joint structure is a central design choice rather than a matter of convenience: electrolyte behavior emerges from interactions across all five components, so they must be preserved together rather than reduced to paper-level metadata. In practice, each record includes canonical identities for solvents, salts, and additives; normalized composition expressed in standardized ratio and concentration units; experimental context such as current density, temperature, and cell configuration; reported measurements; and confidence metadata for every extracted field. Cell-level measurements such as CE and capacity retention are retained only with the electrode, separator, voltage-window, pressure, temperature, and cycling-protocol fields needed to interpret them as conditional outcomes, while electrolyte-intrinsic labels such as conductivity or viscosity can be modeled with narrower protocol context. Appendix A details the full schema, example records, and the standardization pipeline, including conversion rules for weight-, volume-, and molar-ratio reports, together with automated extraction, rule-based normalization, and targeted expert adjudication.

Multimodality is not a cosmetic feature of electrolyte research; it is intrinsic to how the field generates and interprets evidence. Modern electrolyte studies rarely rely on a single reported metric alone, instead linking electrochemical performance to spectroscopic, microscopic, and computational evidence such as NMR, Raman, XPS, SEM, and simulation-derived structural descriptors (Cadiou et al., 2026; Shi et al., 2025; Ma et al., 2025). A useful benchmark must therefore preserve these cross-modal connections, enabling not only scalar prediction but also mechanistic reasoning about why a formulation behaves as it does. This design also connects naturally to adjacent data efforts. NMRNet illustrates the value of standardized NMR resources for chemical-shift modeling (Xu et al., 2025), while Electrolyte Genome provides complementary computed molecular records aligned with the broader goal of battery electrolyte discovery (Spotte-Smith et al., 2023; Qu et al., 2015).

3. Data Construction Roadmap

We begin with metal-anode systems because electrolyte choice is especially consequential in shaping the overall performance (Xu et al., 2014). This provides a sharp initial milestone while keeping the schema extensible to conventional lithium-ion, sodium-based, and zinc systems. The first release is a literature-derived core dataset assembled from metal-anode electrolyte papers and supplementary information, including standardized versions of existing electrolyte-mixture datasets when licenses permit. This phase converts the literature into standardized records with source-level traceability and release-level coverage statistics for papers, formulations, chemical families, protocols, labels, and extraction confidence. Our early CE pilot already suggests sufficient data density for a meaningful first benchmark release. Later phases expand the core resource rather than delay it: computational enrichment adds MD- and DFT-derived descriptors, continuous ingestion broadens coverage, and robotic platforms can provide gold-standard validation subsets (Lee et al., 2026; Svaluto-Ferro et al., 2025). Appendix B details the phase-by-phase roadmap and milestones.

4. Benchmark Tasks

The dataset supports three especially useful benchmark families with public splits and baselines. First, formulation-to-property prediction uses composition together with protocol and device context to predict conductivity, viscosity, CE, capacity retention, or oxidation stability. Conductivity and viscosity are evaluated as mixture-property regression tasks, whereas CE and capacity retention are evaluated as protocol-conditioned cell-outcome tasks to avoid attributing electrode and test-protocol effects solely to electrolyte composition. Second, out-of-distribution evaluation holds out unseen solvents, salts, additive families, ratio regimes, or protocol templates to test whether models generalize in chemically meaningful ways. Third, retrieval and evidence synthesis asks systems to identify relevant precedent formulations and summarize mechanistic support rather than return only a scalar score; evaluation combines source-retrieval recall with expert- or rubric-scored evidence summaries. Appendix C extends this list with spectroscopy-conditioned prediction and closed-loop optimization tasks.

5. Feasibility, Openness, and Governance

The project is feasible because its first milestone relies on data already available at scale in the literature. An initial release can be built directly from published electrolyte papers and supplementary information, focusing on high-availability labels. This makes the first benchmark scientifically meaningful without requiring new large-scale

experiments, while leaving later computational or robotic enrichment as optional extensions.

We support openness through a shared data format, open-source preprocessing tools, versioned releases, public splits, and clear provenance for each field. Where publisher licensing restricts redistribution of raw text or figures, the release can still provide structured facts, normalized metadata, and provenance pointers. We also distinguish directly reported measurements, expert-normalized metadata, machine-extracted fields, and computationally derived descriptors so downstream models can reason about confidence rather than treat all fields alike. Each release will include a changelog, data card, schema version, extraction model version, benchmark split hash, and known-limitations file so that improvements in standardization do not silently change the benchmark. Appendix D provides a fuller discussion of openness, governance, and impact.

6. Conclusion

Open Electrolyte Databank addresses a key gap in AI4S data: most existing chemistry resources are built around single molecules, while battery electrolyte behavior is governed by multi-component formulations and experimental context. By organizing the field around standardized, multimodal records, we hope to better align AI infrastructure with the real problem of electrolyte discovery and support more reliable formulation-aware prediction and reasoning.

References

Cadiou, F. et al. A large scale multi-modal workflow for battery characterization: From concept to implementation, 2026. arXiv:2602.09771.

Gong, S. et al. A predictive machine learning force-field framework for liquid electrolyte development. *Nature Machine Intelligence*, 7(4):543–552, 2025.

Lappemic. Open Source Battery Data. https://github.com/lappemic/open-source-battery-data, 2026. GitHub repository. Accessed 2 July 2026.

Lee, H.-G. et al. Albatross: a robotised system for high-throughput electrolyte screening via automated electrolyte formulation, coin-cell fabrication, and electrochemical evaluation. *Digital Discovery*, 2026.

Li, A. et al. Geometric mixture models for electrolyte conductivity prediction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 13236–13264, 2026.

Ma, P., Kumar, R., Wang, K.-H., et al. Active learning accelerates electrolyte solvent screening for anode-free lithium metal batteries. *Nature Communications*, 16: 8396, 2025.

Mayemba, Q. et al. Aging datasets of commercial lithium-ion batteries: A review. *Journal of Energy Storage*, 2024.

Qu, X. et al. The electrolyte genome project: a big data approach in battery materials discovery. *Computational Materials Science*, 103:56–67, 2015.

Rajaonson, E. M. et al. CheMixHub: Datasets and benchmarks for chemical mixture property prediction. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Shi, Z. et al. Liquid-state nmr spectroscopy for battery electrolyte design. *ACS Energy Letters*, 2025.

Spotte-Smith, E. W. C. et al. A database of molecular properties integrated in the materials project. *Digital Discovery*, 2023.

Svaluto-Ferro, E. et al. Toward an autonomous robotic battery materials research platform powered by automated workflow and ontologized findable, accessible, interoperable, and reusable data management. *Batteries & Supercaps*, 2025.

Wang, H. et al. Liquid electrolyte: The nexus of practical lithium metal batteries. *Joule*, 6:588–616, 2022.

Xu, W., Wang, J., Ding, F., Chen, X., Nasybulin, E., Zhang, Y., and Zhang, J.-G. Lithium metal anodes for rechargeable batteries. *Energy & Environmental Science*, 7:513–537, 2014.

Xu, Y. et al. Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts. *Nature Computational Science*, 2025.

Zhu, S. et al. Differentiable modeling and optimization of non-aqueous li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15(1): 8649, 2024.

Appendices

A. Extended Schema and Standardization

Each formulation record can be written schematically as

$$r = (c, p, y, m, s), \tag{1}$$

where c denotes standardized composition, p the experimental or computational protocol, y the reported outcomes, m supporting multimodal measurements, and s provenance plus confidence metadata.

Table 4 summarizes the standardization rules used to convert formulation data reported as weight ratios, volume ratios, or molar ratios into consistent molality and molarity representations.

Modern battery research increasingly combines electrochemistry with spectroscopy, imaging, and other heterogeneous measurements, and recent battery characterization workflows emphasize the importance of handling multimodal metadata consistently (Cadiou et al., 2026). For electrolyte systems in particular, liquid-state NMR is becoming increasingly important for design (Shi et al., 2025), and the active-learning lithium-metal electrolyte study of Ma et al. (2025) illustrates why linked electrochemistry, NMR, Raman, SEM, and XPS data are so useful.

Electrolyte standardization is not a purely mechanical ETL task. Names, abbreviations, ratios, and concentration units vary widely across papers and supplementary files. A shorthand such as "DEE" may refer to different compounds across contexts, and comparable formulations may be reported in molarity, molality, volume fraction, weight fraction, or informal mixture notation. We therefore advocate a schema-first, expert-in-the-loop pipeline with three layers: automated extraction from papers and supplementary material, rule-based normalization of units and identifiers, and targeted expert adjudication for chemically ambiguous or high-impact records.

Table 1 gives a compact example of the intended record granularity. The example is schematic, but it illustrates the key point that the label is attached to a formulation–protocol pair with explicit provenance and confidence, rather than to a solvent molecule alone.

B. Extended Data Construction Roadmap

The release plan is staged so that the community receives usable data early, while later versions improve coverage and modality depth. Each release reports the number of papers, unique formulations, solvent/salt/additive families, protocol templates, outcome labels, multimodal attachments, and human-adjudicated fields.

Phase I: literature-derived core dataset. The initial release is built from large-scale ingestion of metal-anode electrolyte papers and their supplementary information. We target Li-metal and anode-free Li-metal systems first, with emphasis on concentrated and localized high-concentration electrolytes and common solvent, salt, and additive families. The pipeline extracts formulations, molar ratios, device metadata, protocol fields, and reported outcomes, while preserving paragraph-, table-, or figure-level provenance. This phase is the most feasible near-term milestone and already produces a benchmark-ready resource.

Phase II: computational enrichment. The literature-derived core is augmented with computed descriptors tailored to formulation-aware modeling. Candidate additions include molecular dynamics statistics such as density, diffusion coefficients, coordination numbers, and solvation-shell descriptors, as well as density-functional-theory proxies for stability, binding, and decomposition behavior. These additions are useful because formulation-property relationships in electrolytes often emerge through local coordination structure and interfacial chemistry rather than through single-molecule descriptors alone.

Phase III: continuous updates and community contribution. The dataset is designed as a living resource rather than a one-time release. Scheduled literature ingestion, improved extraction models, ontology-based deduplication, public issue tracking, and versioned releases make it possible to improve coverage over time without sacrificing reproducibility.

Phase IV: robotic validation and closed-loop generation. Longer term, the dataset can incorporate data from automated battery research platforms. Recent robotic systems demonstrate realistic paths toward automated electrolyte formulation, coin-cell fabrication, and electrochemical testing under controlled conditions (Lee et al., 2026; Svaluto-Ferro et al., 2025). These systems are not required for the first milestone, but they provide a natural route to producing gold-standard, protocol-controlled subsets for validation and active learning.

C. Extended Benchmark Tasks

Table 3 makes the benchmark definitions explicit. Every benchmark is released with fixed splits, a baseline feature set, baseline models, and a metric card specifying which contextual fields are mandatory.

Beyond the three core benchmarks highlighted in the main text, the dataset also supports spectroscopy-conditioned prediction, where multimodal models combine composition with NMR, Raman, or surface-analysis evidence to infer performance or mechanistic labels, and closed-loop opti-

Table 1. Schematic example of a formulation-level record. Values are illustrative and show the intended structure rather than a released data point.

Field group	Example value	Why it is retained
Composition	LiFSI in DME/TTE with $LiNO_3$ additive; reported as salt:solvent:additive ratios and normalized to molality/mole fraction	Preserves mixture identity and allows ratio-aware modeling.
Protocol	Li‖Cu or anode-free cell; current density, areal capacity, voltage window, temperature, separator, pressure, and formation cycles when reported	Prevents cell-level outcomes from being interpreted as electrolyte-only properties.
Outcomes	CE, capacity retention, conductivity, viscosity, or stability window with units, cycle index, and uncertainty when available	Separates electrolyte-intrinsic labels from protocol-conditioned performance labels.
Evidence	Linked NMR, Raman, XPS, SEM, MD descriptors, or DFT descriptors when present	Supports mechanistic retrieval and multimodal prediction.
Provenance	Paper DOI, table/figure/paragraph locator, extraction method, expert-review flag, and field-level confidence	Makes noisy literature extraction auditable and correctable.

Table 2. Planned release milestones and artifacts.

Milestone	Primary contents	Released artifacts
Pilot	Metal-anode CE and capacity-retention records from high-yield literature sources, plus a small expert-adjudicated validation subset	Schema, normalized records, provenance pointers, confidence annotations, and baseline train/validation/test splits.
Core release	Broader metal-anode electrolyte records with conductivity, viscosity, stability-window, and device-context labels when available	Versioned dataset, coverage report, extraction scripts, normalization rules, baseline models, and benchmark cards.
Enriched release	MD/DFT descriptors and linked characterization evidence for records with sufficient composition and protocol detail	Descriptor tables, multimodal links, modality-specific missingness masks, and spectroscopy-conditioned benchmark splits.
Community release	Contributions, corrections, and robotic or protocol-controlled validation subsets	Contribution workflow, changelog, issue tracker, validation subset, and frozen benchmark hashes.

mization, where agents propose new formulations under practical constraints such as target conductivity, target CE, fluorine-free composition, or low-temperature operation.

Together, the full benchmark suite reflects the actual work of electrolyte design more closely than isolated single-property tasks, while also supporting retrieval, extrapolation, and mechanism-aware evaluation.

D. Openness, Governance, and Impact Details

The openness plan includes an open schema and ontology, open-source extraction and normalization tools, versioned dataset releases, public benchmark splits, baseline models, provenance tracking for every extracted field, explicit confidence annotations, and community contribution or correction workflows.

This project is intended to accelerate electrolyte discovery for higher-energy batteries and improve the quality of data available for AI-guided scientific reasoning. Positive impacts include more efficient reuse of published evidence, better reproducibility through standardized metadata, and more realistic benchmarks for mixture-aware machine learning. Potential risks include overconfidence in noisy or weakly standardized literature data, selective use of reported performance metrics, and the possibility that downstream models confuse machine-inferred fields with direct measurements. We mitigate these risks through record-level provenance, explicit confidence annotations, expert review of ambiguous cases, and clear separation between observed, normalized, and computed quantities.

Table 3. Core benchmark definitions.

Task	Input and target	Split strategy	Evaluation
Mixture-property regression	Composition, concentration, and temperature to conductivity, viscosity, or stability proxy	Random, scaffold/family, and ratio-regime holdouts	MAE/RMSE, calibration error, and uncertainty coverage.
Protocol-conditioned cell outcome	Composition plus electrode, separator, current density, areal capacity, voltage window, temperature, pressure, and formation protocol to CE or capacity retention	Held-out formulations, additive families, and protocol templates	MAE/RMSE, rank correlation for candidate prioritization, and calibration.
Out-of-distribution generalization	Same targets as above under withheld solvents, salts, additive families, concentrations, or temperatures	Chemically meaningful family and extrapolation splits	In-domain versus OOD performance gap and uncertainty calibration.
Retrieval and evidence synthesis	Query formulation or design goal to relevant precedent records and evidence summaries	Paper-level and family-level holdouts to avoid leakage	Source recall, citation precision, and rubric-scored mechanistic summary quality.

Table 4. Standardization rules for converting formulation data reported as weight ratios, volume ratios, or molar ratios into unified molality and molarity representations. Here w_i, ϕ_i, r_i, ρ_i, M_i, m_i, n_i, and V_i denote the reported weight fraction, volume fraction, molar ratio, density, molar mass, mass, moles, and volume of component i, respectively.

	Weight Ratios	**Volume Ratios**	**Molar Ratios**
Molality (mol kg^{-1})	$m_i = 1000 w_i\,\mathrm{g},$ $n_i = \frac{m_i}{M_i} = \frac{1000 w_i}{M_i}$	$V_i = 1000\phi_i\,\mathrm{mL},$ $m_i = \rho_i V_i = 1000\rho_i\phi_i,$ $n_i = \frac{m_i}{M_i} = \frac{1000\rho_i\phi_i}{M_i}$	$\sum_i m_i = \sum_i n_i M_i,$ $= \alpha \sum_i r_i M_i = 1000\,\mathrm{g},$ $\alpha = \frac{1000}{\sum_i r_i M_i},$ $n_i = \frac{1000 r_i}{\sum_j r_j M_j}$
Molarity (mol L^{-1})	$m_i = 1000 w_i,$ $n_i = \frac{1000 w_i}{M_i}$	$V_i = 1000\phi_i\,\mathrm{mL},$ $m_i = \rho_i V_i = 1000\rho_i\phi_i,$ $n_i = \frac{m_i}{M_i} = \frac{1000\rho_i\phi_i}{M_i}$	$\sum_i V_i = \sum_i \frac{m_i}{\rho_i},$ $= \alpha \sum_i \frac{r_i M_i}{\rho_i} = 1000\,\mathrm{mL},$ $\alpha = \frac{1000}{\sum_i \frac{r_i M_i}{\rho_i}},$ $n_i = \frac{1000 r_i}{\sum_j \frac{r_j M_j}{\rho_j}}$

Table 5. Representative fields in a formulation record.

Category	Representative fields
Composition	Solvent, salt, and additive identities; canonical molecular identifiers; stoichiometric roles; reported ratios; standardized molality, mole fraction, or salt-to-solvent ratios; water content.
Protocol	Cell type; electrode identities and loading; separator and current collector; formation procedure; current density or C-rate; voltage window; temperature; pressure when available.
Outcomes	CE, capacity retention, conductivity, viscosity, impedance, oxidative or reductive stability proxies, gas evolution markers, and interphase-relevant observations.
Evidence	NMR, Raman, FTIR, mass spectrometry, XPS, SEM, linked figures, and simulation-derived descriptors.
Provenance	Paragraph, table, or figure source; extraction route; expert review flag; confidence score; version metadata.