International Journal of Engineering Innovation and Computational Science

Endpoint Fidelity and Chemical Transfer in Machine Learning for Lithium-Ion Cathode Voltage Prediction

Chao Gao [1], **Anniken Karlsen** [*1]
[1] Associate Professor, Department of ICT and Natural Sciences, Norwegian University of Science and Technology (NTU), Ålesund, Norway
*** Corresponding Author:** aniken.t.karlsen@ntu.no

Citation: C. Gao and A. Karlson, "Endpoint Fidelity and Chemical Transfer in Machine Learning for Lithium-Ion Cathode Voltage Prediction," *International Journal of Engineering Innovation and Computational Science*, vol. 1, no. 1, pp. 50–61, 2026.

Received: 18 Dec 2025
Accepted: 10 Mar 2026

Abstract

Composition-based voltage models offer inexpensive screening of lithium-ion electrode materials, but a completely Li-removed formula may misrepresent the charged endpoint associated with the target voltage. This study tests the value and deployment limits of endpoint information using 2,636 Materials Project electrode records spanning 1,763 discharged compositions and 601 chemical systems. Pointwise composition descriptors were regenerated, and discharged-only, Li-removed-proxy, and actual-endpoint representations were compared using five outer chemical-system folds with three-fold inner parameter selection. All fitted preprocessing remained within training partitions. The recorded charged endpoint retained Li in 882 records. Under chemical-system holdout, actual endpoints yielded mean absolute error 0.455 V, root mean squared error 0.619 V, and R^2 0.654, compared with 0.470 V, 0.642 V, and 0.628 for the proxy representation. The paired RMSE difference was -0.0234 V, with a conditional chemical-system bootstrap interval of [-0.0391, -0.0094] V. A secondary equal-dimensional endpoint control yielded RMSE 0.621 V. Composition-grouped evaluation produced actual-endpoint RMSE 0.573 V. Split-conformal intervals reached 96.2% empirical coverage at a nominal 95% level, with mean full width 2.861 V. Separately, canonical identity auditing reduced an archived 1,056-record filtered candidate set to 702 training-excluded records. These findings support a modest endpoint-information benefit while distinguishing reaction-informed prediction, formula-only screening, and independent materials verification.

Keywords: Lithium-Ion Cathodes, Voltage Prediction, Endpoint Composition, Nested Validation, Chemical Transfer, Conformal Prediction.

INTRODUCTION

Average electrode voltage connects redox chemistry to the energy stored in a rechargeable battery [1]. Its thermodynamic description involves the free-energy difference between charged and discharged states and the amount of Li transferred [2]. First-principles methods can address voltage, phase stability, and transport, although these questions require different calculations and assumptions [3]. A machine-learning voltage model therefore approximates a specified reaction-level property rather than establishing the complete suitability of a material as an electrode.

Public repositories such as the Materials Project and JARVIS support large-scale materials analysis [4], [5]. Composition descriptors offer a practical input representation when structures or reaction energies are unavailable. Magpie elemental statistics [6] and matminer featurizers [7] convert chemical formulas into numerical attributes at low computational cost. Their convenience makes it especially important to identify which chemical states the input formulas represent and which information would be available during deployment.

Machine-learning prediction of battery voltage is established. Joshi et al. studied metal-ion electrode voltages [8]. Li et al. developed a cathode-voltage model from charged and discharged formulas [9], and Nagappan et al. investigated endpoint composition descriptors and their differences using LightGBM [10]. Structure-sensitive attention models provide an alternative representation [11]. Consequently, the use of paired formulas, difference descriptors, or gradient boosting alone does not establish a new contribution. A

Copyright © 2026 by Author/s. IJEICS is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

useful empirical question is how much recorded endpoint information contributes under a specified chemical transfer task, relative to a formula-only approximation.

Evaluation design determines which transfer task is measured. MatFold provides chemically motivated splitting protocols [12], and materials out-of-distribution studies caution that an unfamiliar chemical label need not imply extrapolation in every representation [13]. Standardized benchmarks further demonstrate the value of fixed tasks and comparable evaluation procedures [14]. Hyperparameter selection and preprocessing must also respect the evaluation boundary. Otherwise, the reported test error can reflect information used during model development rather than a separate assessment of the fitted procedure [15], [16], [17].

Uncertainty and candidate identity introduce additional requirements. Ensemble spread can help rank errors, but calibrated uncertainty requires an explicit connection to prediction residuals [18]. Conformal prediction supplies a residual-based interval construction with coverage statements that depend on exchangeability assumptions [19]. Candidate screening also requires chemical identity to be defined consistently: a formula absent as a literal string can represent a composition already used in training. Training exclusion remains distinct from discovery or electrochemical validation. The BatteryMat preprint illustrates a staged workflow in which surrogate screening is followed by force-field and density functional theory assessments [20].

This study combines a new controlled benchmark with an operational audit of the supplied screening archive. The primary comparison replaces a completely Li-removed proxy with the recorded charged endpoint, using a common model family and nested chemical-system evaluation. A secondary composition-grouped protocol examines transfer within related chemistry. Additional analyses quantify finite-data composition ambiguity, evaluate matched split-conformal predictors, and correct candidate overlap. The contribution is an empirical assessment of endpoint information and screening reliability on a fixed dataset, with explicit separation of the information required by reaction-informed prediction and formula-only deployment

MATERIALS AND METHODS

Dataset and Provenance

The supervised dataset was the supplied cleaned Materials Project Li-electrode export [4]. It contained 2,636 records, 1,763 canonical discharged compositions, and 601 complete chemical systems. Chemical systems were the sorted sets of constituent elements in discharged formulas. The original notebook retained Li-containing entries with voltage strictly between 0.5 and 6.0 V, positive gravimetric capacity below 500 mAh g^{-1}, and positive gravimetric energy below 2,500 Wh kg^{-1}. It removed repeated combinations of formula, voltage, and gravimetric capacity. These upstream rules define the analyzed snapshot. The unfiltered database response and an upstream release identifier were unavailable.

The target was the exported average electrode voltage, representing a calculated insertion-electrode property. Charged and discharged formulas were retained, together with endpoint identifiers used only as metadata. Capacity, gravimetric energy, stability, endpoint identifiers, and target-derived fields were excluded from predictive inputs. In this dataset, gravimetric energy equaled voltage multiplied by gravimetric capacity to numerical precision, making their exclusion particularly important. The analysis used the supplied files without a new database query, electronic-structure calculation, or electrochemical experiment.

The prior notebook, its archived predictions, and candidate tables were inspected as separate evidence. The topic-definition document and previous audit supplied context rather than executable instructions. Two acoustic-transducer archives were unrelated to cathode materials and were excluded. New models reported below were fitted independently of archived serialized estimators and the archived selected-feature list.

Chemical Identity and Endpoint Information

Formulas were parsed into exact rational element amounts, including grouped formulas and decimal stoichiometry. Element amounts were normalized to total atom amount, giving a composition key invariant to element order and overall formula scale. Identical keys denote identical composition, not identical crystal structure. A second identity key combined the normalized discharged and charged compositions. Both keys

were used to describe repeated records and label spans. Computational composition handling used pymatgen [21]; exact identity checks were independently implemented.

Every recorded endpoint pair preserved the same normalized non-Li composition. In 882 records, the charged formula still contained Li and therefore differed from a proxy obtained by removing all Li from the discharged formula. Li contents were normalized by the total non-Li atom amount in each endpoint. If these normalized contents arex_d and x_c the extracted fraction is (Eq.1):

$$f_{\mathrm{Li}} = \frac{x_d - x_c}{x_d}. \quad (1)$$

Within each identity group, assignment of its observed mean minimizes finite-sample squared error, while assignment of its observed median minimizes absolute error. These empirical assignment errors were calculated for discharged composition and actual endpoint-pair identity. They describe ambiguity in the supplied labels under a deterministic composition mapping. They are in-sample summaries and do not estimate prospective performance or irreducible physical noise.

Regenerated Input Representations

Pointwise descriptors were regenerated using stoichiometry, Magpie elemental statistics, average valence-orbital attributes, and ionic-property featurizers [6], [7]. Each composition state contributed 145 descriptors. Compositions were reduced before featurization, and equivalent normalized compositions shared a cache entry. The numerical cache keys were checked for one-to-one correspondence with exact rational identities across all endpoint formulas. There were 3,808 distinct cached compositions. All featurizers completed successfully; the resulting matrices contained no missing values.

Three representations were specified before observing the new model results. The discharged-only representation contained its 145 descriptors plus five normalized composition features: element count, Li atomic fraction, O atomic fraction, transition-metal atomic fraction, and compositional entropy. The proxy representation concatenated discharged descriptors D, completely Li-removed descriptors P, their differences D-P, and the same five composition features. The actual-endpoint representation substituted recorded charged descriptors C and D-C, then added x_d, x_c, x_d-x_c, extracted Li fraction, and a non-Li composition-match indicator. Their raw dimensions were 150, 440, and 445, respectively.

The proxy transformation is deterministic from discharged composition. It adds a nonlinear descriptor representation but no independently observed endpoint. Actual charged composition supplies additional reaction information in the partially delithiated subset. The primary comparison concerns the complete representation bundle, including extraction features. To examine this distinction, an equal-dimensional control retained D, C, D-C and the original five composition features, giving 440 columns. This control was specified after preliminary fold inspection and is reported as a secondary exploratory analysis, separately from the frozen primary design.

Nested Evaluation and Model Selection

The primary protocol used five outer GroupKFold partitions by complete chemical system. All records in a system remained in one outer fold. Three inner GroupKFold partitions within each outer training set selected the model configuration by pooled inner RMSE. The secondary protocol used five outer and three inner partitions grouped by exact discharged composition. It held out compositions while allowing related chemical systems in training. Shared outer assignments and the same fixed configuration grid were used for all three primary representations. Explicit manifests accompany the results.

LightGBM [22] was selected as the common model family to compare input information under a controlled training budget. The fixed grid crossed 15 or 31 leaves, minimum child size 10 or 30, and L2 regularization 1 or 5, giving eight configurations. Every model used 300 boosting iterations, learning rate 0.035, feature and row sampling fractions 0.9, row-sampling frequency 1, and random seed 42. The lowest pooled inner RMSE selected the configuration, with the first minimum resolving ties. No outer-test early stopping was used.

Median imputation with missingness indicators and a zero-variance filter were fitted separately inside each inner training partition and refitted on each outer training partition. No full-dataset feature selection, correlation pruning, or scaling was performed. Although no missing values occurred in this regeneration, the imputation boundary remained explicit. Scikit-learn supplied the grouped split and preprocessing implementations [23]. A training-mean predictor was evaluated on each outer fold as a baseline. Nested selection separates parameter fitting from outer evaluation, as motivated by previous work on model-selection bias [16], [17]. The analysis plan and model grid were saved before new fitting; the plan was fixed after the earlier archive audit and was not a blinded prospective preregistration.

Error Metrics and Paired Resampling

Pooled mean absolute error, root mean squared error, and coefficient of determination were calculated from one outer-test prediction per record (Eq.2-4):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - p_i|. \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - p_i)^2}. \tag{3}$$

$$R^2 = 1 - \frac{\sum_i (y_i - p_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{4}$$

Chemical-system macro MAE gave equal weight to each of the 601 systems. Residuals were defined as observed minus predicted voltage. Diagnostics used true-voltage ranges below 3.0 V, 3.0 to below 4.2 V, and at least 4.2 V, together with charged-Li retention subgroups.

Conditional intervals used 2,000 bootstrap resamples of chemical systems with replacement, generated with seed 42. All records from a sampled system were retained at every selection. Paired representation differences used identical resampled systems, with difference defined as actual-endpoint error minus proxy error. Percentile bounds at 2.5% and 97.5% defined the reported 95% interval. These intervals condition on the dataset, chosen outer folds, and fitted predictions. They do not include retraining variation, alternative fold assignments, shared-training dependence between outer predictors, or upstream selection uncertainty. Subgroup and secondary analyses are descriptive extensions rather than multiplicity-adjusted confirmatory tests.

Matched Split-Conformal Analysis

Within each primary outer training set, 20% of its chemical-system groups were reserved for calibration using seed 1042 plus the outer-fold index. The remaining proper-training records underwent the same three-fold inner selection. One model was then fitted to proper training and used without refitting for both calibration and outer-test prediction. This procedure was repeated for proxy and actual-endpoint representations. Calibration, proper training, and outer test had disjoint chemical systems.

For calibration absolute residuals r_j, the empirical quantile used rank k in their sorted values (Eq.5):

$$k = \mathrm{ceil}((n_{cal} + 1)(1 - \alpha)), \qquad q_\alpha = r_{(k)}. \tag{5}$$

The interval for an outer-test prediction was (Eq.6):

$$I_\alpha(x) = [\hat{y}(x) - q_\alpha, \hat{y}(x) + q_\alpha]. \tag{6}$$

Nominal levels were 80%, 90%, and 95%. Intervals were not clipped to the observed target range. Coverage and mean full width were calculated on outer tests. This is a pointwise residual-quantile procedure

with group-separated calibration, not a group-coverage theorem. The within-system dependence and deliberate chemistry shift prevent treating the standard exchangeable-observation coverage guarantee as established here [19]. Its value is an empirical assessment with a matched prediction object. These proper-training models differ from the primary point predictors fitted to all outer-training records.

Candidate Audit and Reproducibility

The archive contained 4,326 unique JARVIS candidate records [5] and 1,056 passing the original heuristic filters. Canonical composition matching assessed overlap with the supplied training snapshot. A separately supplied 817-record list had excluded literal formula matches. Records matching a training composition were removed, and the remainder was classified by Li-free composition and chemical-system membership. Archived voltage predictions, hull energies, and ranks were treated as historical screening outputs. The new actual-endpoint model was not applied to these candidates because their recorded charged endpoints were unavailable.

Python scripts, raw regenerated matrices, out-of-fold predictions, split manifests, parameter scores, calibration intervals, and figure data are supplied. Software versions include matminer 0.10.1, pymatgen 2026.10.24, scikit-learn 1.9.1, and LightGBM 4.7.0. Independent checks recomputed metrics, verified group separation and complete row coverage, inspected feature inputs, and reproduced a selected outer model with maximum prediction difference below 10^-10 V. Input and output checksums support file identity. Full software details and commands are given in the supplementary information.

RESULTS

Endpoint Information And Composition Ambiguity

The charged endpoint retained Li in 882 of 2,636 records (33.5%), while 1,754 recorded endpoints were Li-free. All pairs preserved their normalized non-Li composition. Figure 1 shows the endpoint groups across the voltage distribution and the finite-data identity summaries. Discharged-composition grouping produced 1,763 identities, compared with 1,879 actual endpoint pairs. Adding the charged endpoint therefore distinguished some records that were compositionally identical in the discharged state.

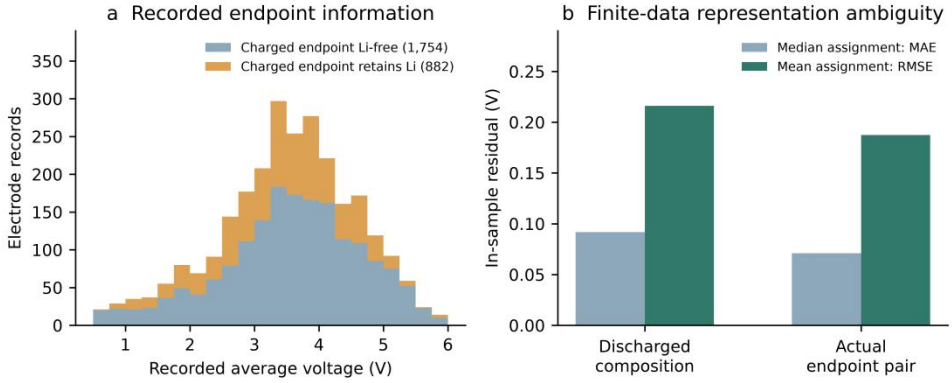


Figure 1. (a) Voltage distributions for records with Li-free and Li-retaining charged endpoints. (b) In-sample errors obtained by assigning each identity group its observed median or mean. Panel b describes finite-data ambiguity and is not an estimate of model test error

The empirical mean-assignment RMSE decreased from 0.216 V for discharged composition to 0.187 V for endpoint-pair identity. Corresponding median-assignment MAE decreased from 0.092 to 0.071 V. Repeated discharged compositions occurred in 336 groups containing 1,209 records, and their largest observed voltage span was 3.648 V. Endpoint-pair identity retained residual label variation, consistent with omitted structural or reaction detail, differences in database records, or both. The dataset alone does not resolve these possibilities.

Nested Chemical-System Evaluation

Table 1 reports the primary outer-test metrics. The training-mean baseline had RMSE 1.059 V, while the discharged-only model yielded 0.651 V. The proxy model yielded MAE 0.470 V, RMSE 0.642 V, and R2 0.628. The actual-endpoint model reduced these errors to 0.455 V and 0.619 V, with R2 0.654. Its macro-system MAE was 0.502 V, compared with 0.526 V for the proxy. Thus, the gain was also present when each chemical system received equal weight.

Table 1. Nested Outer-Test Performance on Held-out Chemical Systems. Brackets are Conditional 95% Chemical-System Bootstrap Intervals for Fixed Out-of-Fold Predictions. Units for MAE and RMSE are Volts

Representation	MAE [95% interval]	RMSE [95% interval]	R^2	Macro MAE
Training mean	0.834 [0.770, 0.903]	1.059 [0.996, 1.126]	-0.012	1.011
Discharged only	0.477 [0.444, 0.508]	0.651 [0.612, 0.688]	0.618	0.526
Li-removed proxy	0.470 [0.438, 0.501]	0.642 [0.603, 0.682]	0.628	0.526
Actual endpoints	0.455 [0.424, 0.483]	0.619 [0.582, 0.655]	0.654	0.502

The paired actual-minus-proxy MAE difference was -0.0149 V (95% interval [-0.0286, -0.0020] V), and the RMSE difference was -0.0234 V ([-0.0391, -0.0094] V). Both intervals excluded zero under this conditional resampling scheme. Relative to the proxy, the reduction was 3.16% in MAE and 3.65% in RMSE. These modest improvements support the value of the recorded endpoint information on this snapshot without establishing an architecture advance or a universal effect across databases.

The 440-feature endpoint control yielded MAE 0.456 V and RMSE 0.621 V. Its RMSE difference from the 440-feature proxy was -0.0210 V ([-0.0393, -0.0042] V). Its difference from the 445-feature actual-endpoint model was 0.0024 V ([-0.0050, 0.0088] V). The control supports a benefit from replacing the endpoint descriptors; it does not establish an additional RMSE benefit from the five extraction-related scalars, whose paired interval included zero. Because this control was added after preliminary inspection, it strengthens interpretation as a secondary check rather than an independent confirmatory experiment.

Composition Transfer and Error Structure

The secondary composition-grouped protocol gave actual-endpoint MAE 0.407 V, RMSE 0.573 V, and R2 0.704, compared with proxy RMSE 0.592 V (Table 2 and Figure 2). In this protocol, 83.3% of outer-test records belonged to chemical systems represented in training. The endpoint diagnostics also found 178 test discharged compositions matching a charged composition somewhere in that fold's training records. These are different transfer conditions from the primary system holdout. Composition separation alone should therefore not be described as assessment of unfamiliar chemistry or complete reaction-network separation.

Table 2. Secondary Nested Evaluation on Held-out Discharged Compositions. Related Chemical Systems Can Occur in Training and Test; The Protocol Measures a Different Transfer Task from Table 1**.**

Representation	MAE (V)	RMSE (V)	R^2	Macro MAE (V)
Training mean	0.831	1.055	-0.003	1.011
Discharged only	0.432	0.604	0.671	0.520
Li-removed proxy	0.420	0.592	0.683	0.514
Actual endpoints	0.407	0.573	0.704	0.491

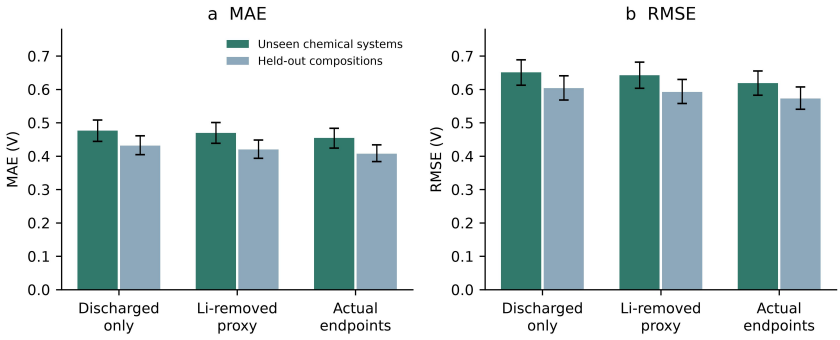


Figure 2. MAE and RMSE for the three prespecified representations under chemical-system and composition grouping. Error bars are conditional chemical-system bootstrap intervals of the fixed outer predictions. The secondary equal-dimensional control is reported separately in Table S1

Figure 3 compares primary outer predictions and their residual structure. The actual-endpoint model showed mean residuals -0.405 V below 3.0 V and 0.446 V at or above 4.2 V, indicating overprediction at the low end and underprediction at the high end. Its MAE was 0.564 V, 0.354 V, and 0.516 V in the three voltage ranges, respectively. The high-voltage MAE was slightly greater than the proxy's 0.511 V. The pooled benefit was therefore not a uniform improvement across voltage ranges and did not eliminate compression of extreme predictions toward the center of the target distribution.

In the Li-retaining charged-endpoint subgroup, the actual-minus-proxy MAE difference was -0.0195 V ([-0.0475, 0.0063] V). This interval included zero. In the Li-free subgroup it was -0.0125 V ([-0.0255, -0.0016] V). These subgroup contrasts describe fitted-model behavior rather than a causal isolation of Li retention: all model parameters were learned from mixed training data, and the representation bundles can alter the fitted function across both groups.

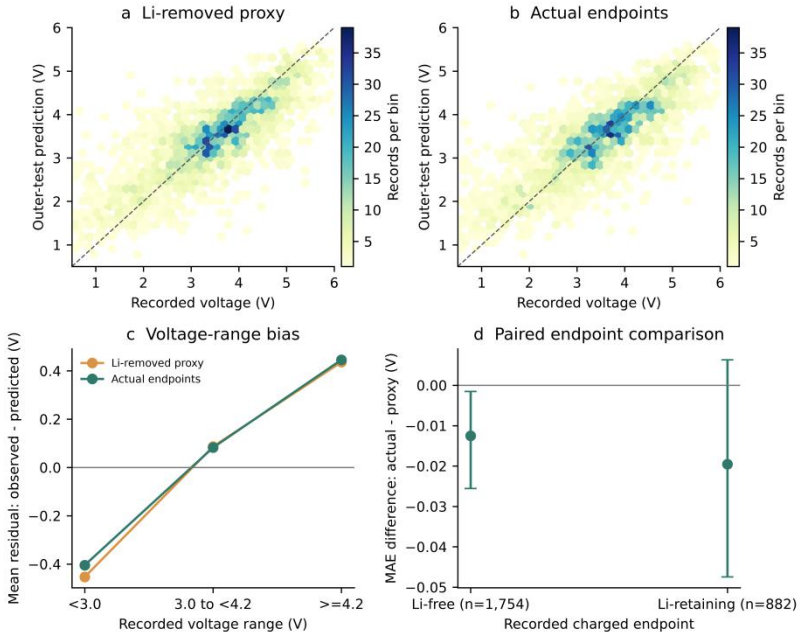


Figure 3. Primary chemical-system outer predictions for (a) the proxy and (b) actual endpoints; color density summarizes record counts. (c) Mean residual by recorded-voltage range. (d) Actual-minus-proxy MAE by charged-Li retention, with paired conditional chemical-system bootstrap intervals. Negative MAE differences favor actual endpoints

Empirical Interval Coverage and Resolution

The matched actual-endpoint split-conformal procedure attained 83.0%, 92.1%, and 96.2% pooled coverage at nominal levels of 80%, 90%, and 95%, respectively. Mean full widths were 1.591, 2.230, and 2.861 V (Table 3). At 95%, the proxy procedure attained 95.8% coverage with mean width 2.966 V. Figure 4 displays the coverage-width relationship.

Table 3. Outer-test Coverage and Mean Full Interval Width from Matched Proper-Training Predictors. These Pointwise Empirical Results Do Not Establish Coverage Guarantees Under Chemical Shift

Representation	Nominal level	Empirical coverage	Mean width (V)
Li-removed proxy	80%	82.7%	1.657
Li-removed proxy	90%	91.5%	2.291
Li-removed proxy	95%	95.8%	2.966
Actual endpoints	80%	83.0%	1.591
Actual endpoints	90%	92.1%	2.230
Actual endpoints	95%	96.2%	2.861

The observed coverage is substantially broader in voltage resolution than a narrow ranking claim would require. A mean width of 2.861 V spans a large part of the recorded voltage range. Useful pooled coverage therefore coexists with limited ability to distinguish candidates separated by a few tenths of a volt. Fold-specific coverage, quantiles, and proper-predictor errors are supplied in Tables S4 and S5; their variation is obscured by pooled summaries.

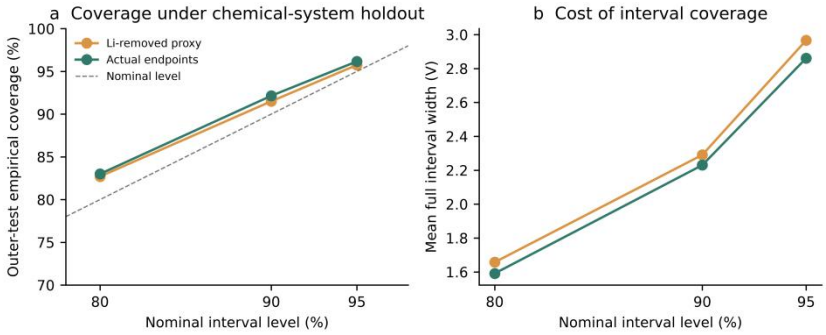


Figure 4. (a) Pooled empirical coverage and (b) mean full width at three nominal levels for group-separated split-conformal intervals. Each calibration and outer-test prediction came from the same proper-training estimator. The diagonal denotes equality of nominal and observed coverage, not a guarantee under the study's grouped distribution shift

Operational Consequences for the Archived Screen

Canonical matching identified 637 training-composition matches among 4,326 archived candidates and 354 among the 1,056 passing the original filters. The latter count included all ten highest-ranked records. Literal-string exclusion had missed 115 matches in the supplied 817-record list, leaving 702 records after canonical exclusion (Table 4). Within this corrected pool, 244 had a Li-free composition already represented in training, 240 had a new Li-free composition in a represented chemical system, and 218 belonged to systems absent from training.

Table 4. Training Overlap in the Archived Screening Outputs. Candidate Voltages Were Archived Predictions, Not External Voltage Labels

Candidate collection	Records	Canonical training matches	Remaining after exclusion
Full archived candidate pool	4,326	637	3,689
Archived filter-passing pool	1,056	354	702
Supplied string-excluded list	817	115	702

These corrections establish composition exclusion relative to one training snapshot. They do not establish literature novelty, a reversible insertion reaction, or measured electrode performance. Actual charged endpoints were unavailable for this candidate pool. Applying the actual-endpoint benchmark score to these formula-only inputs would therefore exceed the information available at deployment. The 702-record pool remains a starting point for reaction-level investigation; its historical ranks have not been relabeled as predictions from the new models.

DISCUSSION

The central result is a small, reproducible improvement from recorded endpoint information under an explicit chemical transfer task. The actual-endpoint representation reduced primary RMSE by 3.65% relative to the Li-removed proxy, with a paired conditional interval excluding zero. The equal-dimensional control produced a similar error level. This supports examining the reaction endpoints represented by the target rather than assuming that complete Li removal reconstructs the relevant charged state. It does not make endpoint descriptors a new invention, given prior voltage studies using paired compositions [9], [10]. The present contribution lies in the matched nested comparison, uncertainty analysis, and connection to deployable screening inputs.

The information distinction has a physical basis. Average insertion voltage depends on the free-energy change between specified endpoint states and the amount of transferred Li [2], [3]. Replacing an endpoint that retains Li with a completely Li-free composition changes the represented reaction interval. The empirical benefit observed here is consistent with that distinction. Nevertheless, formulas omit crystal geometry, electronic structure, and phase evolution. The remaining endpoint-pair label ambiguity and systematic voltage-range bias show why endpoint composition alone cannot define an exact reaction-energy model.

The proxy model retains practical value because its inputs are derived from a single discharged formula. Its lower information requirement makes it applicable earlier in a screening workflow. The actual-endpoint model answers a different question: voltage prediction conditional on a proposed or recorded charged composition. An unknown endpoint must be generated or independently established before that model can be used. If it is obtained through substantial reaction-level computation, the cost and selection consequences of that upstream stage belong in the assessment. The additional prediction accuracy should be considered alongside this information cost.

The primary and secondary validation scores refer to distinct populations. The composition-grouped protocol often retains related chemistry and can share endpoint compositions across different reaction records. The primary system holdout removes complete elemental sets from training, although this does not guarantee extrapolation in every numerical descriptor [13]. Neither protocol measures transfer to an independently collected experimental dataset. Explicitly naming the held-out unit and preserving split manifests prevents a single accuracy claim from obscuring these distinctions.

The interval experiment changes the interpretation of uncertainty. The earlier archive's diagnostic use of ensemble disagreement did not provide calibrated predictive scale; the present procedure connects residual calibration to the same proper-training estimator used on outer tests. Its pooled coverage is empirically close to or above the nominal levels, but the wide intervals limit fine candidate discrimination. Standard conformal guarantees cannot simply be transferred to correlated records and deliberately held-out chemical systems [19].

The measured coverage-width tradeoff is therefore a result to report, rather than proof that uncertainty has been solved.

Candidate identity adds a separate limitation that improved regression cannot repair. A high-scoring record can be compositionally identical to training material despite different formula text. Conversely, exclusion from training does not make a material new to the literature. A favorable hull energy also does not establish a viable charged state, Li mobility, electrolyte compatibility, or reversible capacity. The chemistry dependence of crystalline metastability further limits a universal hull-energy cutoff [24]. Staged frameworks illustrate the need to keep surrogate prioritization and reaction-level verification distinct [20].

Several limitations bound the present outcome. The dataset is one filtered computational snapshot, and the prior audit informed the choice of research question. Only one model family and a fixed eight-configuration budget were used; no state-of-the-art performance claim follows. The outer partition was not repeatedly randomized, and bootstrap intervals omit retraining and model-selection variability. Repeated records receive record-level weight in pooled errors, with macro-system MAE supplying a complementary view. The equal-dimensional control was added after preliminary inspection. Structure-aware controls, independent voltage labels, and candidate-level reaction calculations remain absent. These limitations support a computational reliability study of the supplied data rather than a claim of newly discovered cathodes.

CONCLUSION

A new nested benchmark shows that recorded charged-endpoint information modestly improves composition-based voltage prediction on the supplied 2,636-record dataset. Holding out complete chemical systems, the actual-endpoint model yielded MAE 0.455 V, RMSE 0.619 V, and R2 0.654. The paired RMSE improvement over the Li-removed proxy was 3.65%, and a secondary equal-dimensional control showed a similar error level. These results support the relevance of endpoint fidelity while preserving the distinction between reaction-informed prediction and formula-only screening.

Matched residual-calibrated intervals achieved 96.2% empirical coverage at a nominal 95% level, with mean width 2.861 V. The coverage required broad intervals and did not establish guaranteed performance under chemical shift. Canonical identity auditing separately reduced the archived filtered candidate pool to 702 training-excluded records. These records are computational leads without independent voltage or electrochemical validation. The resulting contribution concerns computational prediction and screening reliability, with independent reaction-level validation required for materials discovery.

REFERENCES

[1] J. B. Goodenough and K.-S. Park, “The Li-Ion Rechargeable Battery: A Perspective,” *Journal of the American Chemical Society*, vol. 135, no. 4, pp. 1167-1176, 2013, doi: 10.1021/ja3091438. https://doi.org/10.1021/ja3091438

[2] M. K. Aydinol, A. F. Kohan, G. Ceder, K. Cho, and J. Joannopoulos, “Ab initio study of lithium intercalation in metal oxides and metal dichalcogenides,” *Physical Review B*, vol. 56, no. 3, pp. 1354-1365, 1997, doi: 10.1103/PhysRevB.56.1354. https://doi.org/10.1103/PhysRevB.56.1354

[3] A. Urban, D.-H. Seo, and G. Ceder, “Computational understanding of Li-ion batteries,” *npj Computational Materials*, vol. 2, Art. no. 16002, 2016, doi: 10.1038/npjcompumats.2016.2. https://doi.org/10.1038/npjcompumats.2016.2

[4] A. Jain et al., “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation,” *APL Materials*, vol. 1, no. 1, Art. no. 011002, 2013, doi: 10.1063/1.4812323. https://doi.org/10.1063/1.4812323

[5] K. Choudhary et al., “The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design,” *npj Computational Materials*, vol. 6, Art. no. 173, 2020, doi: 10.1038/s41524-020-00440-1. https://doi.org/10.1038/s41524-020-00440-1

[6] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, “A general-purpose machine learning framework for predicting properties of inorganic materials,” *npj Computational Materials*, vol. 2, Art. no. 16028, 2016, doi: 10.1038/npjcompumats.2016.28. https://doi.org/10.1038/npjcompumats.2016.28

[7] L. Ward et al., “Matminer: An open source toolkit for materials data mining,” *Computational Materials Science*, vol. 152, pp. 60-69, 2018, doi: 10.1016/j.commatsci.2018.05.018. https://doi.org/10.1016/j.commatsci.2018.05.018

[8] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, and J. E. Peralta, “Machine Learning the Voltage of Electrode Materials in Metal-Ion Batteries,” *ACS Applied Materials & Interfaces*, vol. 11, no. 20, pp. 18494-18503, 2019, doi: 10.1021/acsami.9b04933. https://doi.org/10.1021/acsami.9b04933

[9] H. H. Li, Q. Chen, G. Ceder, and K. A. Persson, “Voltage Mining for (De)lithiation-Stabilized Cathodes and a Machine Learning Model for Li-Ion Cathode Voltage,” *ACS Applied Materials & Interfaces*, vol. 16, pp. 69379-69387, 2024, doi: 10.1021/acsami.4c15742. https://doi.org/10.1021/acsami.4c15742

[10] N. Nagappan, G. S. Priyanga, and T. Thomas, “Trends and insights into alloying elements impact on predicted battery voltage in metal-ion batteries,” *Journal of Energy Storage,* vol. 103, Art. no. 114412, 2024, doi: 10.1016/j.est.2024.114412. https://doi.org/10.1016/j.est.2024.114412

[11] S.-Y. Louis, E. M. D. Siriwardane, R. P. Joshi, S. S. Omee, N. Kumar, and J. Hu, “Accurate Prediction of Voltage of Battery Electrode Materials Using Attention-Based Graph Neural Networks,” *ACS Applied Materials & Interfaces*, vol. 14, no. 23, pp. 26587-26594, 2022, doi: 10.1021/acsami.2c00029. https://doi.org/10.1021/acsami.2c00029

[12] M. D. Witman and P. Schindler, “MatFold: systematic insights into materials discovery models’ performance through standardized cross-validation protocols,” *Digital Discovery*, vol. 4, pp. 625-635, 2025, doi: 10.1039/D4DD00250D. https://doi.org/10.1039/D4DD00250D

[13] K. Li, A. N. Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A. B. Dieng, and J. Hattrick-Simpers, “Probing out-of-distribution generalization in machine learning for materials,” *Communications Materials*, vol. 6, Art. no. 9, 2025, doi: 10.1038/s43246-024-00731-w. https://doi.org/10.1038/s43246-024-00731-w

[14] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, “Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm,” *npj Computational Materials*, vol. 6, Art. no. 138, 2020, doi: 10.1038/s41524-020-00406-3. https://doi.org/10.1038/s41524-020-00406-3

[15] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804. https://doi.org/10.1016/j.patter.2023.100804

[16] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, Art. no. 91, 2006, doi: 10.1186/1471-2105-7-91. https://doi.org/10.1186/1471-2105-7-91

[17] G. C. Cawley and N. L. C. Talbot, “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,” *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010. [Online]. Available: https://jmlr.org/papers/v11/cawley10a.html

[18] G. Palmer et al., “Calibration after bootstrap for accurate uncertainty quantification in regression models,” *npj Computational Materials*, vol. 8, Art. no. 115, 2022, doi: 10.1038/s41524-022-00794-8. https://doi.org/10.1038/s41524-022-00794-8

[19] A. N. Angelopoulos and S. Bates, “Conformal Prediction: A Gentle Introduction,” *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494-591, 2023, doi: 10.1561/2200000101. https://doi.org/10.1561/2200000101

[20] J. Lee, C. R. Campbell, K. Zhang, and K. Choudhary, “BatteryMat: a hierarchical machine-learning and DFT framework for average-voltage screening of lithium-ion cathode materials,” *arXiv:2607.06645v1*, preprint, 2026, doi: 10.48550/arXiv.2607.06645. https://doi.org/10.48550/arXiv.2607.06645

[21] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, “Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis,” *Computational Materials Science*, vol. 68, pp. 314-319, 2013, doi: 10.1016/j.commatsci.2012.10.028. https://doi.org/10.1016/j.commatsci.2012.10.028

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[23] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825-2830, 2011. [Online]. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html

[24] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and G. Ceder, “The thermodynamic scale of inorganic crystalline metastability,” *Science Advances*, vol. 2, no. 11, Art. no. e1600225, 2016, doi: 10.1126/sciadv.1600225. https://doi.org/10.1126/sciadv.1600225

[25] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18. https://doi.org/10.1038/sdata.2016.18