Data Engineering and Applications

Research Paper

SchNet-AL: Data-Driven Atomistic Representation and Uncertainty-Calibrated Active Learning for Sodium-Ion Cathode Screening

Zorana Mladenović[1, *] and Danka Cvetković[1]

[1] Faculty of Computational Chemical Engineering, Union University, Belgrade, 11000, Serbia
*Corresponding author: zorana.ml@union.edu.rs

Abstract. The search space for sodium-ion battery cathodes has been made relatively large by chemistry, but it is still too expensive to test with density-functional theory. This paper introduces a SchNet-active learning framework that combines continuous-filter atomistic representation learning with calibrated ensemble uncertainty and diversity-aware batch acquisition. A curated pool of 18,240 relaxed and partially relaxed inorganic structures was screened for formation energy, average sodium extraction voltage, theoretical capacity and diffusion-barrier proxy. Start with 1,200 labelled structures and add 300 calculations each time in five acquisition rounds. The proposed strategy reduced the energy mean absolute error from 92 to 31 meV atom^{-1} and the voltage error from 0.284 to 0.118 V, and random acquisition needed 61% more labels to achieve the same energy accuracy. With a fixed budget of 2,700 labelled structures, 86.7% of the retrospectively verified top-60 candidates could be recovered. Multi-objective filtering found 47 candidates with stability below 45 meV atom^{-1} above the convex hull, voltage between 2.8 and 4.1 V, capacity over 120 mAh g^{-1}, and a diffusion proxy below 0.55 eV. The full workflow reduces the projected screening time by 8.4 times compared with an all-first-principles calculation. Therefore, uncertainly-calibrated atomistic learning can focus the cost of calculation on chemically informative candidates without reducing screening coverage.

Keywords: *Sodium-Ion Cathodes, SchNet Representation, Active Learning, Uncertainty Calibration, Multi-Objective Screening*

Received on 28 February 2023, Accepted on 16 June 2023, Published on 28 June 2023

Introduction

Large-scale electrification has made the cost, supply security and environmental impact of electrochemical storage intertwined engineering issues. Sodium-ion batteries are promising for stationary power supply due to abundant sodium resources and compatibility with existing manufacturing processes for rocking-chair cells [1]. However, their actual competitiveness depends significantly on the cathode, and the energy density and cycling stability are determined by composition, crystal framework, redox chemistry and defect response [2]. Layered transition-metal oxides have a high voltage but can undergo phase reconstruction during sodium extraction [3]. Polyanionic compounds have structural and thermal stability but are not highly gravimetric [4]. Prussian-blue analogues have open diffusion channels but are sensitive to water and vacancy content [5]. Organic and disordered hosts further expand the chemical design space, but their structure-performance relationships have not yet been standardized [6]. These families create a combined space, and small substitutions or order changes can affect all of them simultaneously.

First-principles calculations provide formation energies, redox potentials, migration barriers and other electronic descriptors of unsynthesised materials. Therefore, high-throughput workflows have begun to narrow the space for inorganic materials [7]. However, a traditional campaign still invests most of its funds in candidates who are unstable, redundant or clearly outside the target voltage range. Graph neural networks can reduce this cost by learning local chemical environments from atomistic structures [8]. Message-passing models have achieved good accuracy in crystalline formation energies [9], and rotationally invariant continuous-filter networks are

https://doi.org/10.64972/dea.2023.v2i2.3619d:119-133

particularly suitable when interatomic distances carry the central signal [10]. However, the predictions are not reliable in sparsely sampled chemistries; a low average test error does not guarantee trustworthy ranking at the edge of the training distribution [11]. Uncertainty estimation is therefore required to determine which new quantum-mechanical calculations should be purchased by a surrogate [12].

The other is not merely a predictive error. The screening system has a limited labelling budget, must ensure chemical coverage, calibrate risk, and convert some predicted properties into an auditable shortlist. This paper fills the above deficiency with a SchNet-active learning model for sodium-ion cathodes. The model is an ensemble of continuous-filter networks, uses a heteroscedastic objective, and employs a batch acquisition rule to balance epistemic uncertainty, expected multi-objective utility, and latent-space diversity. A physics-informed feasibility layer avoids selecting predictions that are too appealing but unrealistic. The indices for evaluation include label efficiency, calibration, ranking recall, chemistry transfer and projected computational cost. The resulting framework serves as an engineering decision pipeline and is not an independent property regressor.

Related Work

Data-Driven Cathode Screening

Most computational screening studies first address thermodynamic stability and then add electrochemical constraints. Convex-hull distance is commonly used to rule out structures that are unlikely to remain in equilibrium [13]. The average intercalation potential can be obtained from the energy difference between the sodium and desodium ends, while the theoretical capacity is derived by dividing the transferable sodium content by its formula mass [14]. Migration barriers are more costly as they require path sampling, so learned or geometry-based proxies are often used in the early triage stage [15]. The above sequential Design is clear; however, it can amplify numerical errors close to a boundary with a hard threshold, and exhaustive calculations are inefficient when many substitutions, vacancy patterns and polymorphs need to be considered simultaneously.

Materials databases and automated workflows have made supervised crystal-property learning feasible, but cathode data are still heterogeneous. Relaxation protocols, magnetic initialization, exchange-related settings, and oxidation state assumptions can introduce label shifts as severe as the performance differences among candidates. Models trained on general inorganic compounds also have an overrepresentation of dense equilibrium structures and an underrepresentation of partially deintercalated or metastable frameworks. Therefore, a screening model should be able to distinguish between interpolation within a well-covered family and extrapolation to a new transition-metal or anion environment. Therefore, Uncertainty-aware Sampling is used to address this problem without a fixed training set.

Atomistic Learning and Active Selection

SchNet represents atoms as trainable embeddings and updates these embeddings via continuous filters generated by pairwise distances. Smooth radial construction avoids manual fingerprint selection and supports variable composition and cell size [16]. Equivariant models can use the angular information more directly later on, but the computational cost of repeated ensemble retraining may not be feasible. Deep ensemble remains a relatively good practical estimator for representing cognitive uncertainty, as networks with independent initialization exhibit inconsistency in unlabeled regions [17]. The heteroscedasticity head can also represent label noise, and after a distribution shift, the unregularized variance estimate may be overly confident.

Active learning creates a closed-loop system for surrogates and expensive oracles. Uncertainty sampling is simple but may repeatedly select outliers; diversity sampling covers more ground but can ignore the decision boundary. Batch acquisition methods group the above signals using clustering, determinants or Pareto ranking [18]. For battery materials, the acquisition goal should also consider whether an uncertain candidate will change the final shortlist. At present, the work combines uncertainty, multi-property utility and latent diversity, and then calibrates the ensemble on a held-out stream after each round. This differs from using active learning only to minimize global formation-energy error.

https://doi.org/10.64972/dea.2023.v2i2.3619d:119-133

SchNet–Active Learning Framework

Problem Definition and Data Representation

Each candidate is shown with atomic numbers, Cartesian coordinates, lattice vectors, sodium occupancy and a structure-family tag. The candidate pool includes layered oxides, polyanionic compounds, Prussian-blue analogues and open-framework mixed-anion phases. The available labels include the formation energy of each atom, the average sodium extraction voltage, the theoretical capacity, and the diffusion barrier proxy obtained through standardized local bottleneck calculations. Duplicate structures are removed using composition-reduced matching and a 0.12Å site tolerance. Structures with unconverged electronic steps, forces above 0.35eVA^{-1} after the prescribed relaxation, or inconsistent sodium counts are quarantined rather than silently discarded.

To encode interatomic separation without discontinuous bins, the distance between atoms i and j is expanded through the following cutoff-modulated radial basis:

$$e_k(d_{ij}) = exp\left[-\beta_k(d_{ij} - \mu_k)^2\right] f_{cut}(d_{ij}) \qquad \text{Eq.(1)}$$

Using this smooth distance representation, one SchNet interaction block updates each atomic state according to

$$\boldsymbol{x}_i^{(l+1)} = \boldsymbol{x}_i^{(l)} + \sum_{j \in \mathcal{N}_i} \boldsymbol{x}_j^{(l)} \odot \boldsymbol{W}^{(l)}\left(\boldsymbol{e}(d_{ij})\right) \qquad \text{Eq.(2)}$$

The learning problem is formulated as multi-output regression under pool-based active learning. At round t, the labeled set is used to train an ensemble, the remaining candidate pool is scored, and a selected batch is submitted to the first-principles oracle. The goal of optimization is not to minimize the mean squared error. The four coupled engineering constraints for the recovered candidates and the control of false-negative decisions near each threshold are sought by the framework. Figure 1 shows the entire workflow of structural normalization, integrated training, calibration acquisition, Oracle annotation, and multi-target candidate release.

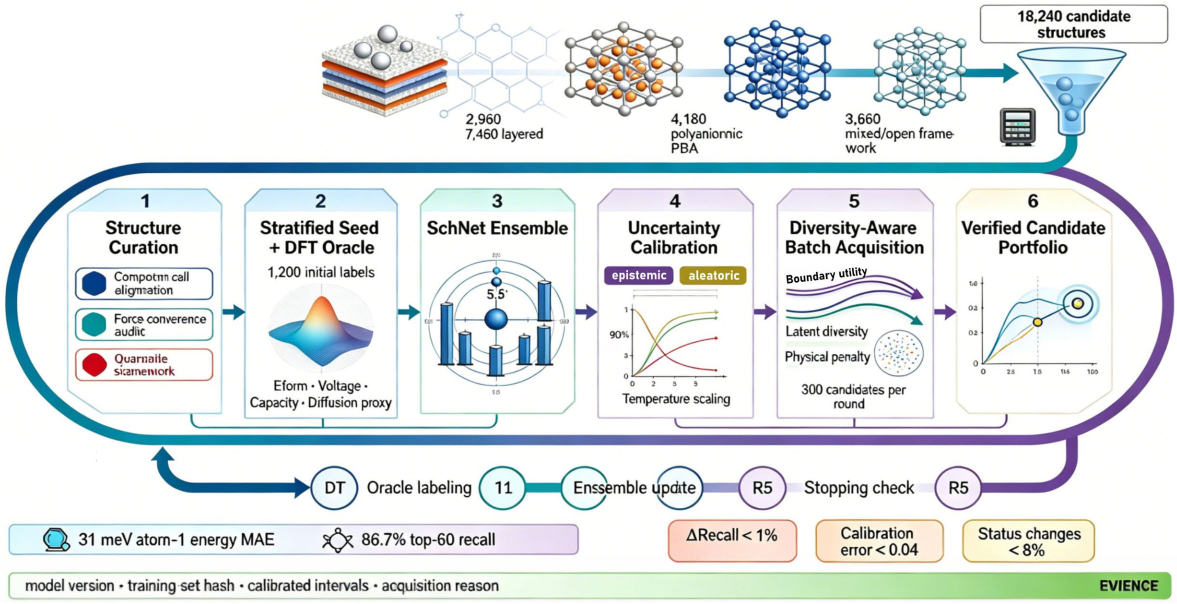

Figure 1. End-to-end SchNet-Active Learning workflow for sodium-ion cathode screening, including structure curation, ensemble learning, calibrated batch acquisition, first-principles labeling, and iterative release control.

Continuous-Filter Representation and Objective

A radial basis expansion converts each neighbor distance into a smooth vector, allowing the filter network to resolve short-range coordination and longer-range framework geometry. Three interaction blocks with a 5.5Å cutoff are employed. Validation experiments showed that a fourth interaction block reduced the prediction error by less than 1.2% but increased the ensemble retraining time by 29%. According to the physical properties of

https://doi.org/10.64972/dea.2023.v2i2.3619d:119-133

the target nature, Atomwise contributions are summarized through intensity or breadth normalization. Composition descriptors and sodium occupancy are added after pooling, and the structure encoder remains transferable to other prediction tasks.

After the final interaction block, the structure-level representation is obtained through the target-dependent pooling operation

$$\boldsymbol{z} = Pool\left(\left\{\boldsymbol{x}_i^{(L)}\right\}_{i=1}^{N}\right) \quad \text{Eq.(3)}$$

The predictive head outputs the mean and the aleatoric uncertainty of each target property. The weighted Gaussian negative log-likelihood function prevents large-scale target-dominated optimization and allows relatively noisy diffusion agents to have a wider prediction interval. Formation energy and voltage are given a relatively higher weight for errors because these errors can move a candidate directly to the thermodynamic-stability or operating-voltage boundary.

The four predictive heads are trained jointly by minimizing the weighted heteroscedastic objective

$$\mathcal{L} = \sum_{p} w_p \left[\frac{\left(y_p - \mu_p\right)^2}{2\sigma_p^2} + \frac{1}{2} ln\, \sigma_p^2\right] \quad \text{Eq.(4)}$$

The theoretical capacity is also calculated analytically from the sodium stoichiometry and is used as a consistency target. A substantial difference between the analytical capacity and its learned counterpart is treated as a data-quality warning rather than being removed through numerical averaging.

For each target property p, epistemic uncertainty is estimated from the dispersion among M independently initialized ensemble members:

$$u_{epi,p}^2(\boldsymbol{x}) = \frac{1}{M-1}\sum_{m=1}^{M} \left[\mu_{m,p}(\boldsymbol{x}) - \bar{\mu}_p(\boldsymbol{x})\right]^2 \quad \text{Eq.(5)}$$

The calibrated total predictive uncertainty combines temperature-scaled epistemic dispersion with the average aleatoric variance:

$$u_p^2(\boldsymbol{x}) = T_p^2 u_{epi,p}^2(\boldsymbol{x}) + \frac{1}{M}\sum_{m=1}^{M} \sigma_{m,p}^2(\boldsymbol{x}) \quad \text{Eq.(6)}$$

Here, T_p is the calibration temperature for property p. The first term represents uncertainty arising from limited coverage of the labeled chemical space, whereas the second term captures irreducible noise or inconsistency in the corresponding property labels.

Batch Acquisition and Stopping Logic

Five independently initialized networks are used to quantify epistemic dispersion. Raw ensemble disagreement is recalibrated through temperature scaling on a chronological validation stream. The acquisition score is a boundary-aware utility combined with calibrated uncertainty and a diversity term derived from farthest-point distances in the pooled SchNet embedding. A soft penalty suppresses candidates with extremely high predicted decomposition energy or chemically invalid oxidation states. Batch selection is first carried out within each material family and then merged globally to avoid having the more numerous layered-oxide class consume the entire oracle budget.

The calibrated uncertainty terms enter the boundary-aware acquisition function used to score each unlabeled candidate:

$$A(\boldsymbol{x}) = U_{bnd}(\boldsymbol{x})\sum_{p} \alpha_p u_p(\boldsymbol{x}) + \lambda_d D(\boldsymbol{x}) - \lambda_f P_{phys}(\boldsymbol{x}) \quad \text{Eq.(7)}$$

In this expression, $U_{\text{bnd}}(\mathbf{x})$ emphasizes candidates capable of changing a feasibility decision, α_p controls the contribution of property-specific uncertainty, $D(\mathbf{x})$ encourages structural diversity, and $P_{\text{phys}}(\mathbf{x})$ penalizes physically implausible candidates.

The diversity contribution is measured by the nearest selected-batch distance in the learned latent space:

https://doi.org/10.64972/dea.2023.v2i2.3619d:119-133

$$D(\boldsymbol{x}) = \min_{\boldsymbol{x}_j \in \mathcal{B}_t} \left\| \boldsymbol{z}(\boldsymbol{x}) - \boldsymbol{z}(\boldsymbol{x}_j) \right\|_2 \qquad \text{Eq.(8)}$$

The active-learning loop terminates when three conditions are satisfied simultaneously. First, top- k recall must change by less than one percentage point over two successive rounds. Second, the expected calibration error must fall below 0.04 for formation energy and voltage. Third, less than 8 per cent of the newly labelled structures are modified in feasibility status after oracle evaluation. This stopping condition limits the amount of computation by ensuring that the decision has reached a stable state and that further reduction in the global error will not improve the engineering shortlist.

Physics-Constrained Screening Strategy

Property Estimation and Feasibility

Formation energy is referenced to competing phases under a consistent elemental chemical-potential convention. Thermodynamic screening begins with the formation energy per atom:

$$E_f = \frac{E_{tot} - \sum_s n_s \mu_s}{\sum_s n_s} \qquad \text{Eq.(9)}$$

Here, E_{tot} is the calculated total energy of the candidate structure, n_s is the number of atoms of species s, and μ_s is the corresponding reference chemical potential.

For two sodium contents x_1 and x_2 in the same host H, the average sodium-extraction voltage is evaluated from the reaction energy:

$$\bar{V} = -\frac{E\left(Na_{x_2}H\right) - E\left(Na_{x_1}H\right) - (x_2 - x_1)E(Na)}{(x_2 - x_1)F} \qquad \text{Eq.(10)}$$

The capacity calculation uses the number of reversibly transferred sodium ions and the formula mass. The stoichiometric upper bound on reversible gravimetric capacity is calculated as

$$C_{th} = \frac{(x_2 - x_1)F}{3.6M_r} \qquad \text{Eq.(11)}$$

where F is the Faraday constant and M_r is the molar mass of the active cathode material.

A local environment model estimates the migration-barrier proxy using bottleneck radius, coordination change and ensemble-derived structure features. A typical sub-set is then subjected to nudged elastic-band calculations. These attributes are considered together, and a high theoretical capacity is not practical if the corresponding material is unstable or has poor sodium-transport properties.

Instead of applying discontinuous cutoffs, each engineering constraint is represented by a smooth feasibility gate:

$$g_k(\boldsymbol{x}) = \left[1 + exp\left(\frac{q_k(\boldsymbol{x}) - \tau_k}{s_k}\right)\right]^{-1} \qquad \text{Eq.(12)}$$

Here, $q_k(\mathbf{x})$ denotes the predicted value of constraint k, τ_k is its engineering threshold, and s_k controls the smoothness of the transition around that threshold.

Assuming that the individual gates represent distinct screening requirements, their joint feasibility score is written as

$$P_{feas}(\boldsymbol{x}) = \prod_k g_k(\boldsymbol{x}) \qquad \text{Eq.(13)}$$

Smooth-gating formulation retains the information of candidates near a screening boundary and prevents an abrupt decrease in ranking caused by surrogate uncertainty. The final utility rewards voltage and capacity in the desired operating region, penalizes hull distance and diffusion impedance, and adds a lower confidence bound so that uncertain but optimistic predictions are not automatically promoted. Figure 2 shows the property heads, uncertainty calibration, physical gates, Pareto ranking and final verification queue.

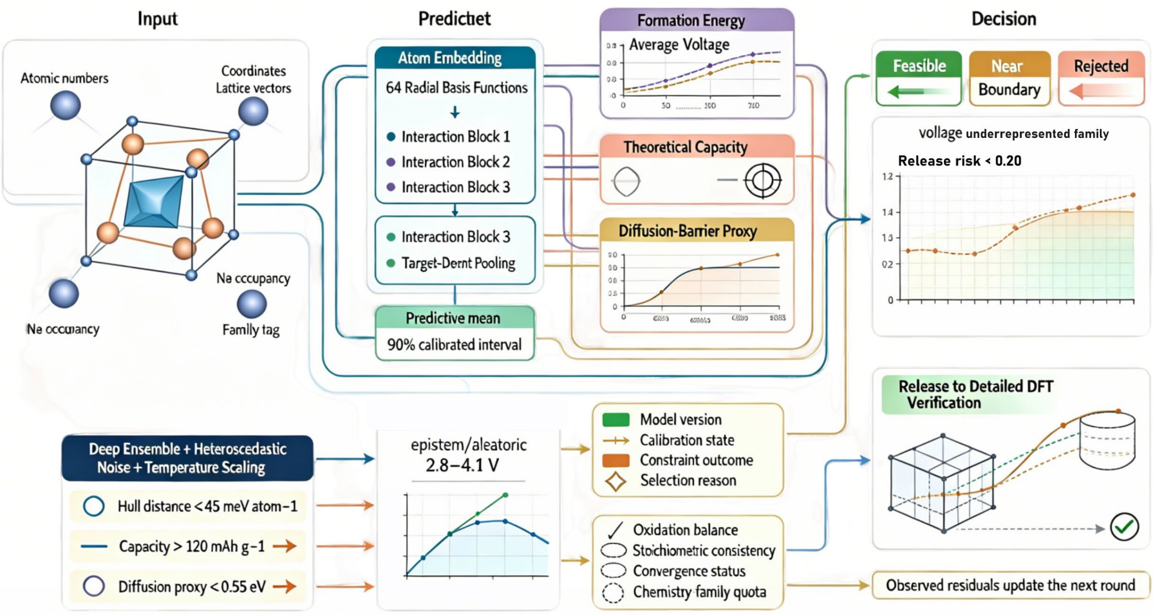

Figure 2. Physics-constrained multi-objective screening architecture with property heads, uncertainty calibration, smooth feasibility gates, Pareto ranking, risk control, and verification routing.

Multi-Objective Ranking and Risk Control

Candidates are first classified as feasible, near-boundary, and rejected. Nondominated Sorting is performed only after the uncertainty-adjusted feasibility has been determined. A variety of voltage-capacity trade-offs are retained along each Pareto front by a crowding term, and a chemistry quota reserves at least 15% of the verification queue for each underrepresented material family. The risk score is defined as the probability that a candidate violates one or more hard release constraints under the calibrated joint predictive distribution.

Release risk and uncertainty-adjusted Pareto value are coupled through the decision score:

$$R(\boldsymbol{x}) = 1 - P\left(\bigcap_{k} q_k(\boldsymbol{x}) \le \tau_k\right)$$
$$S(\boldsymbol{x}) = EHVI(\boldsymbol{x})[1 - R(\boldsymbol{x})] \qquad \text{Eq.(14)}$$

A candidate is released for detailed verification when its risk score is below 0.20 or its expected hypervolume improvement exceeds the predefined exploration threshold. The latter condition deliberately admits a limited number of uncertain candidates with potentially substantial multi-objective value.

Screening provenance records the model version, acquisition round, training-set hash, calibrated predictive intervals, gating outcome, and selection reason. This information makes the released shortlist reproducible even after the oracle data stream has been expanded.

Computational Implementation

SchNet models are optimised by AdamW and a cosine learning rate schedule. Each model has 128-dimensional atom embeddings, 64 radial basis functions and three continuous-filter interaction blocks. Training uses mini-batches of 32 samples. Early stopping is used with a patience of 25 epochs, and the ensemble members are trained on independently bootstrap-resampled labelled sets.

Group data partitions by reduced composition to restrict structural leakage between the training and validation sets. All uncertainty indicators are based on candidates that were unavailable during temperature fitting. Provided deterministic random seeds for data partitioning, network initialization, bootstrapping, batch fetching, and score tie-breaking.

https://doi.org/10.64972/dea.2023.v2i2.3619d:119-133

Oracle expenditure is measured in normalized node-hours, with the median structure relaxation defined as one computational unit. Ensemble retraining, latent-embedding extraction, uncertainty calibration, and acquisition scoring are included in the reported computational cost rather than treated as free operations.

The trained SchNet representations are cached between acquisition rounds. During the initial warm-start stage, each ensemble member is updated using the newly labeled structures and a 20% replay sample from the existing labeled set. Full convergence is subsequently performed before the candidate pool is rescored. This implementation balances the needs of statistically different ensemble members with the repeated computation demands of high-throughput active learning.

Experimental Evaluation and Analysis

Experimental Configuration and Learning Efficiency

The candidate pool comprised 18,240 structures after deduplication: 7,460 layered oxides, 4,180 polyanionic phases, 2,940 Prussian-blue analogues, and 3,660 mixed-anion or open-framework structures. The initial labeled set contained 1,200 structures balanced across these families. Five rounds each added 300 oracle labels, producing a final training set of 2,700. A composition-grouped test set of 1,800 structures was never eligible for acquisition. Baselines included random selection, maximum ensemble variance, latent farthest-point sampling, and a static SchNet trained on the final random budget. All methods used identical network capacity and optimization limits.

Learning behavior is summarized in Figure 3: Figure 3(a) shows that the proposed method lowered formation-energy mean absolute error from 92 to 31 meV atom^{-1} as labels increased from 1,200 to 2,700, compared with 47 meV atom^{-1} for random selection; Figure 3(b) reports voltage error falling from 0.284 to 0.118 V, while pure uncertainty sampling ended at 0.137 V; and Figure 3(c) shows top-60 recall rising to 86.7%, 15.0 percentage points above random acquisition. Error bars over five seeds narrowed after round three, indicating that the diversity term reduced batch sensitivity. Random acquisition needed approximately 4,350 labels to interpolate to the 31 meV atom^{-1} target, corresponding to 61% more labels than the proposed policy.

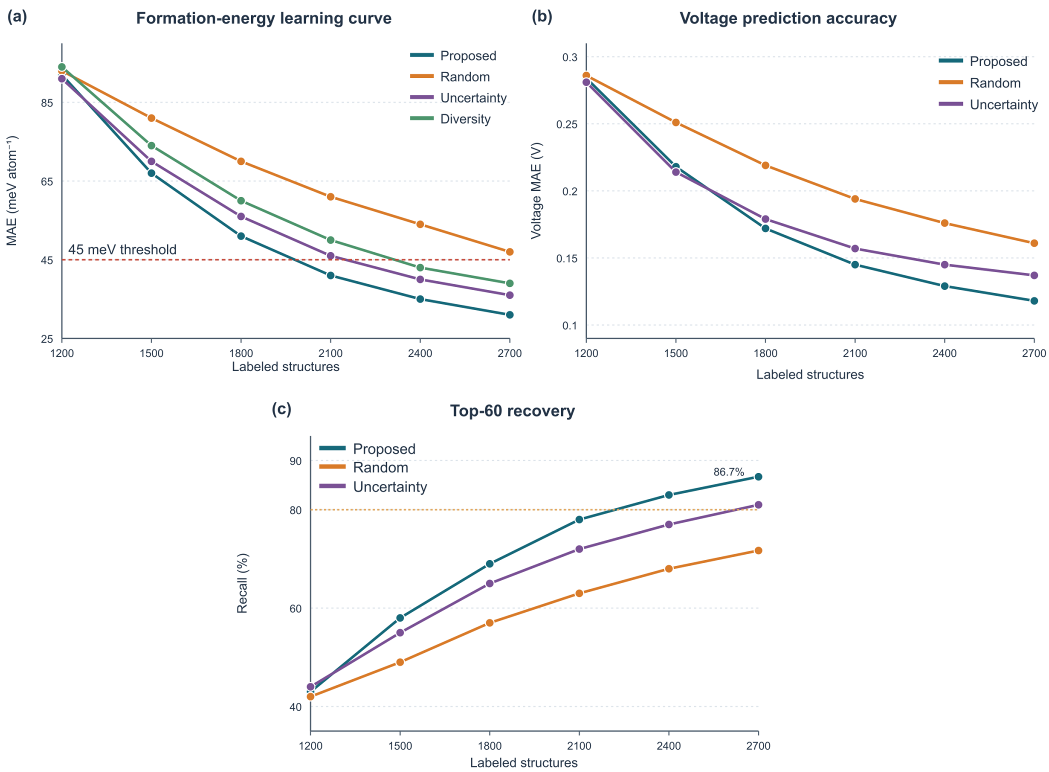

Figure 3. Label efficiency across five acquisition rounds: (a) formation-energy error versus labeled structures; (b) voltage error versus labeled structures; (c) top-60 candidate recall with seed variability.

According to the above interval-based assessment of predictive uncertainty [19], calibration was performed at a nominal coverage of 90 per cent. Before temperature correction, the energy coverage was 78.4% and the voltage coverage was 81.1%; therefore, it was considered overconfident. After calibration, the coverage reached 89.2% and 90.6% without changing the point predictions; at the same time, the expected calibration error decreased from 0.091 to 0.027 for energy and from 0.076 to 0.024 for voltage. Deep-ensemble disagreement was more informative than a single-network dropout estimate, especially for vanadium-free polyanionic structures that were rare in the seed set [20]. The calibrated acquisition policy thus selected 38% near-boundary candidates, 44% high-uncertainty candidates with positive utility, and 18% diversity anchors.

The budget advantage was not obtained by repeatedly sampling pathological structures. The maximum-variance baseline devoted more queries to calculation failures: failed oracle calculations accounted for 7.1% of its acquisitions, compared with 3.8% under the proposed policy. This result reflects a recognized weakness of unqualified uncertainty sampling, which can prioritize anomalous or weakly supported inputs rather than decision-relevant candidates [21]. The feasibility penalty rejected most structures with implausible oxidation balance before acquisition, while the family quota retained exploration. Median acquisition-scoring time was 11.4 minutes for the full unlabeled pool on one graphics processor, and ensemble retraining declined from 6.8 to 4.9 hours per round after warm-start caching.

Screening Accuracy, Generalization, and Ablation

Figure 4 examines ranking and transfer: Figure 4(a) plots predicted against first-principles voltage for 1,800 held-out structures, with a slope of 0.97, an intercept of 0.08 V, and only 4.6% of points outside the calibrated 90% band; Figure 4(b) reports family-wise energy errors of 27, 34, 38, and 36 meV atom−1 for layered, polyanionic, Prussian-blue, and mixed-anion groups. The larger error for Prussian-blue analogues reflects vacancy and hydration variability, but remained below the 45 meV atom−1 screening tolerance. A leave-one-transition-metal-out test increased error to 58 meV atom−1, showing that the model extrapolates imperfectly when local redox environments are absent.

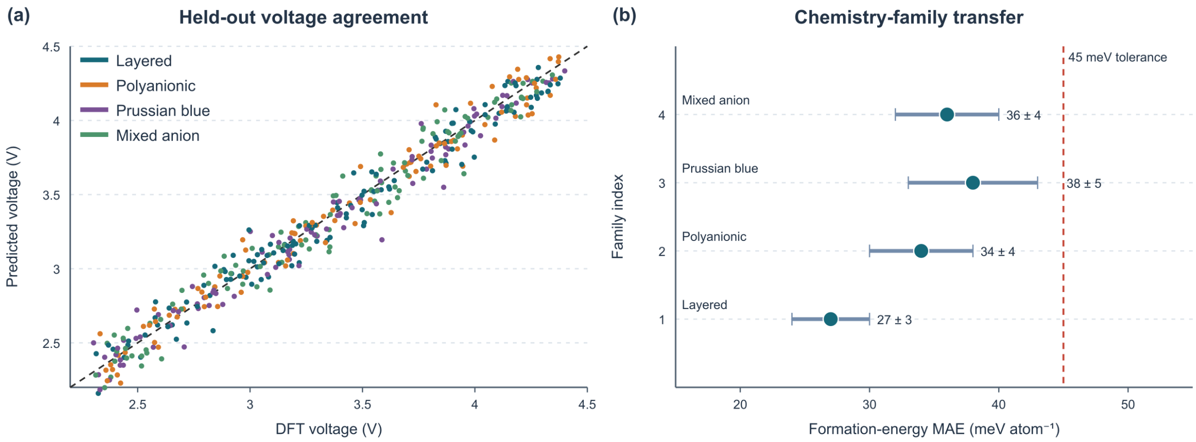

Figure 4. Predictive accuracy and chemistry transfer: (a) predicted versus first-principles voltage with calibrated confidence band; (b) family-wise formation-energy error and screening tolerance.

Ablation experiments separated the contribution of each design choice. Replacing the ensemble with a single heteroscedastic network reduced training cost by 63% but increased expected calibration error to 0.068. Removing latent diversity reduced top-60 recall from 86.7% to 79.8% and concentrated 71% of one batch in layered oxides, whereas removing boundary-aware utility preserved global energy error at 32 meV atom−1 but lowered feasible-candidate recall to 76.4%. Without calibration, the released shortlist contained 11 false-feasible structures, compared with five under the full model. The contrast supports the combined treatment of epistemic and aleatoric uncertainty advocated for decision-facing materials models [22].

Figure 5 provides a more detailed diagnostic: Figure 5(a) shows the calibration curve approaching the diagonal after scaling, with a maximum coverage deviation of 2.8%; Figure 5(b) maps latent-space density and demonstrates that acquired candidates occupy both sparse boundaries and dense decision regions; Figure 5(c) presents an error distribution whose median absolute energy error is 19 meV atom−1 and 95th percentile is 84 meV atom−1; and Figure 5(d) gives acquisition composition across rounds, where the minority mixed-anion share grows from 14% to 23%. The single heat map is limited to Figure 5(b), while the remaining panels use reliability, distribution, and stacked-area encodings.

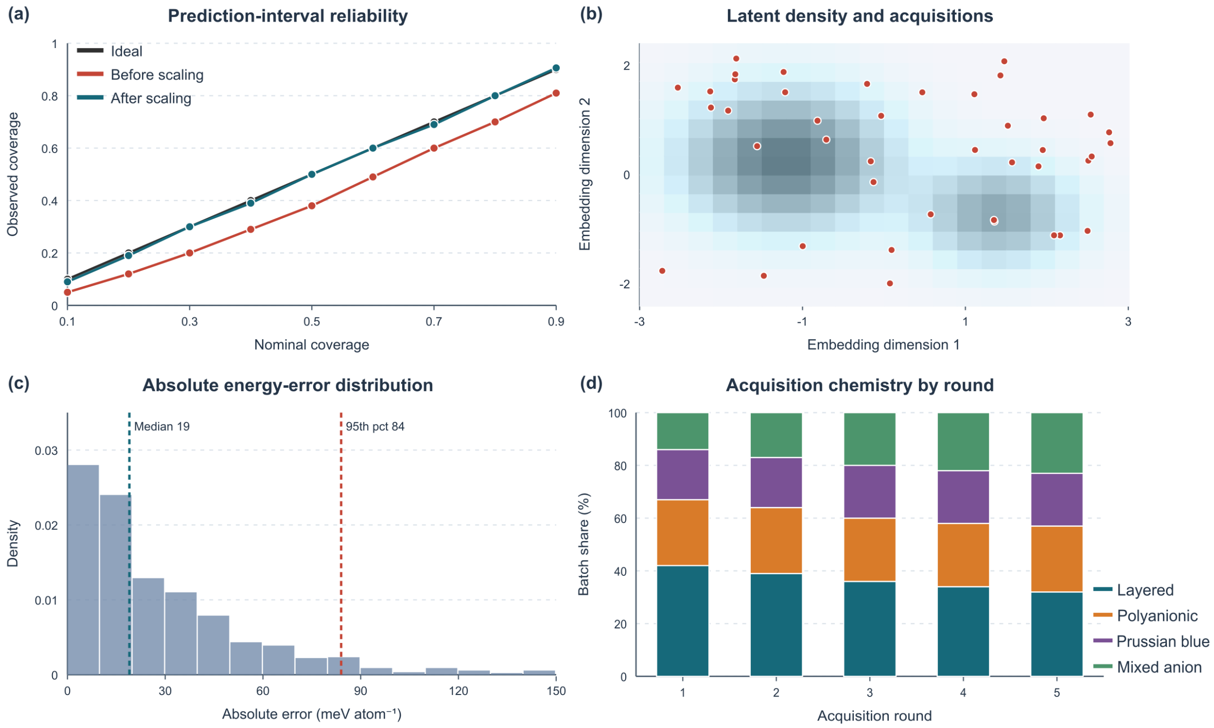

Figure 5. Uncertainty and acquisition diagnostics: (a) reliability curves before and after calibration; (b) latent-density heat map with acquired candidates; (c) absolute-error distribution with percentile markers; (d) acquisition-family composition by round.

Because reproducible materials-data studies identify source provenance as a prerequisite for interpreting model uncertainty [23], the perturbation tests retained source identifiers alongside every label. Adding zero-mean energy noise with a 20 meV atom−1 standard deviation increased point error by 4 meV atom−1 but widened predicted aleatoric intervals by 17%, preserving 88.5% coverage. A systematic 0.10 V shift in one data source was detected through source-conditioned residuals; removing the source identifiers caused the shift to propagate into the shortlist.

Candidate Quality and Computational Impact

The full pool was screened with thresholds of hull distance below 45 meV atom−1, voltage from 2.8 to 4.1 V, capacity above 120 mAh g−1, and diffusion proxy below 0.55 eV. The mean predictions produced 73 apparently feasible candidates; uncertainty-adjusted gating reduced this set to 47. Subsequent first-principles verification confirmed 42, giving a precision of 89.4%. Among the 13 false negatives in retrospective analysis, nine lay within 6 meV atom−1 or 0.04 V of a boundary, demonstrating that remaining errors were concentrated near operational cutoffs rather than scattered across the design space.

Figure 6 relates electrochemical merit to structural risk: Figure 6(a) displays a voltage–capacity Pareto front containing 47 released candidates and marks 12 high-hypervolume points; Figure 6(b) shows hull-distance distributions with medians of 22, 31, 27, and 35 meV atom−1 across the four families; and Figure 6(c) summarizes normalized P10–P90 ranges, interquartile ranges, and medians for all 47 released candidates against the 19 candidates that simultaneously exceed 150 mAh g−1 and remain below a 0.45 eV diffusion proxy. The selected subgroup shifts the normalized capacity median from 0.53 to 0.71 and the diffusion-proxy median from 0.48 to 0.27, while preserving a broad voltage interval. Layered oxides dominate the upper-voltage edge, whereas polyanionic structures occupy a lower-capacity but lower-risk region. Mixed-anion frameworks contribute five of the twelve high-hypervolume candidates, validating the exploration quota.

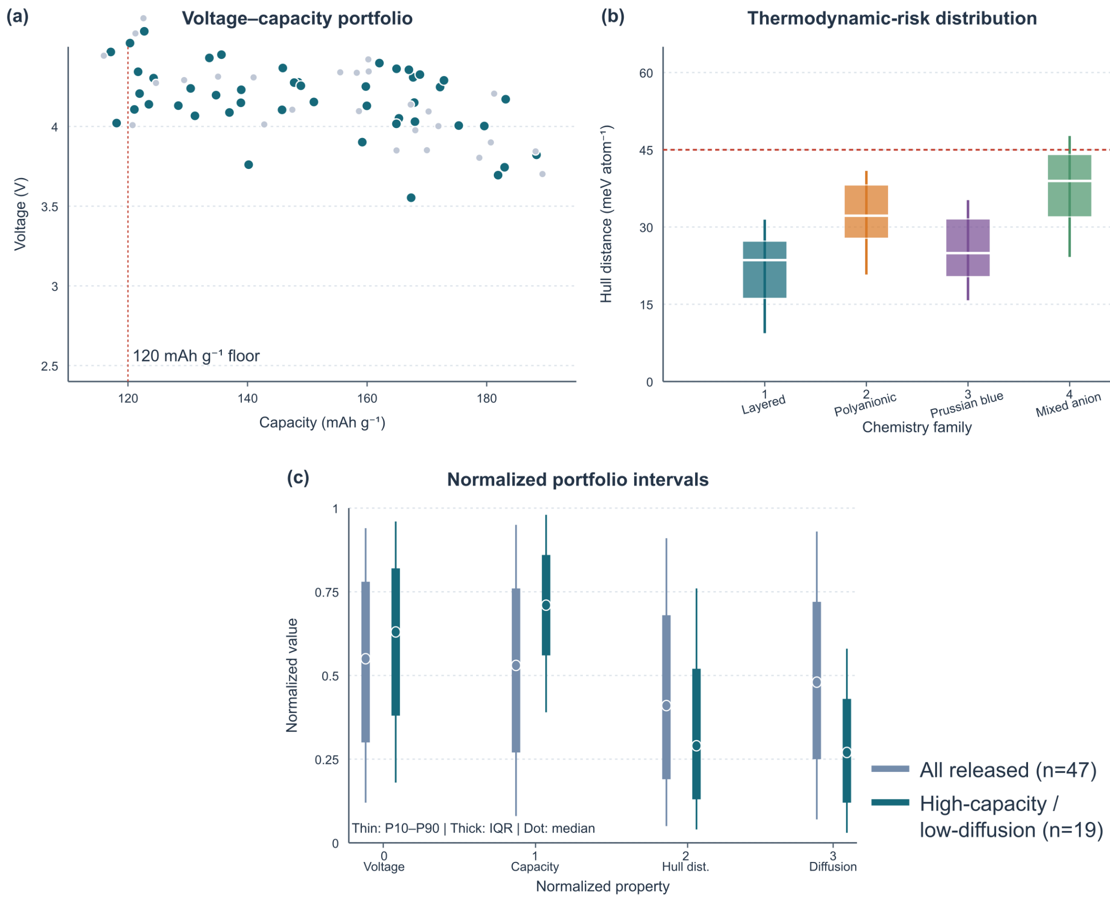

Figure 6. Multi-objective candidate quality: (a) voltage–capacity Pareto landscape; (b) hull-distance distributions by chemistry family; (c) normalized percentile intervals for all released candidates and the high-capacity/low-diffusion subgroup.

The candidate list should be regarded as a computational priority queue, rather than a proof of synthetic capability. Static volume calculations neglect disorder at finite temperatures, electrolyte compatibility, surface reconstruction, and kinetic trapping; it is known that these omissions alter the actual behavior of sodium-ion cathodes [24]. Thus, the uncertainty-aware gate does not present a point estimate as a binary experimental fact. The five false-feasible candidates had risk scores between 0.14 and 0.19, and they were kept because their expected hypervolume improvement was relatively high. A controlled exception will not lower the threshold silently based on the result.

Computational demand is reported in Figure 7: Figure 7(a) shows cumulative normalized node-hours of 3.1 million for exhaustive evaluation, 0.54 million for random-surrogate screening, and 0.37 million for the proposed workflow; Figure 7(b) compares the component costs directly, with 0.304 million node-hours for oracle calculations, 0.046 million for ensemble training, 0.0115 million for acquisition scoring, and 0.0089 million for data handling; and Figure 7(c) shows that verified discoveries per 100,000 node-hours increased from 1.9 in round one to 6.8 in round four before stabilizing at 6.5, with seed-range whiskers remaining inside the 0–8 plotting interval. The point annotations identify 3, 7, 11, 13, and 12 verified candidates across the five rounds. The projected 8.4-fold speedup includes all active-learning overhead and is therefore smaller, but more credible, than a comparison based only on surrogate inference.

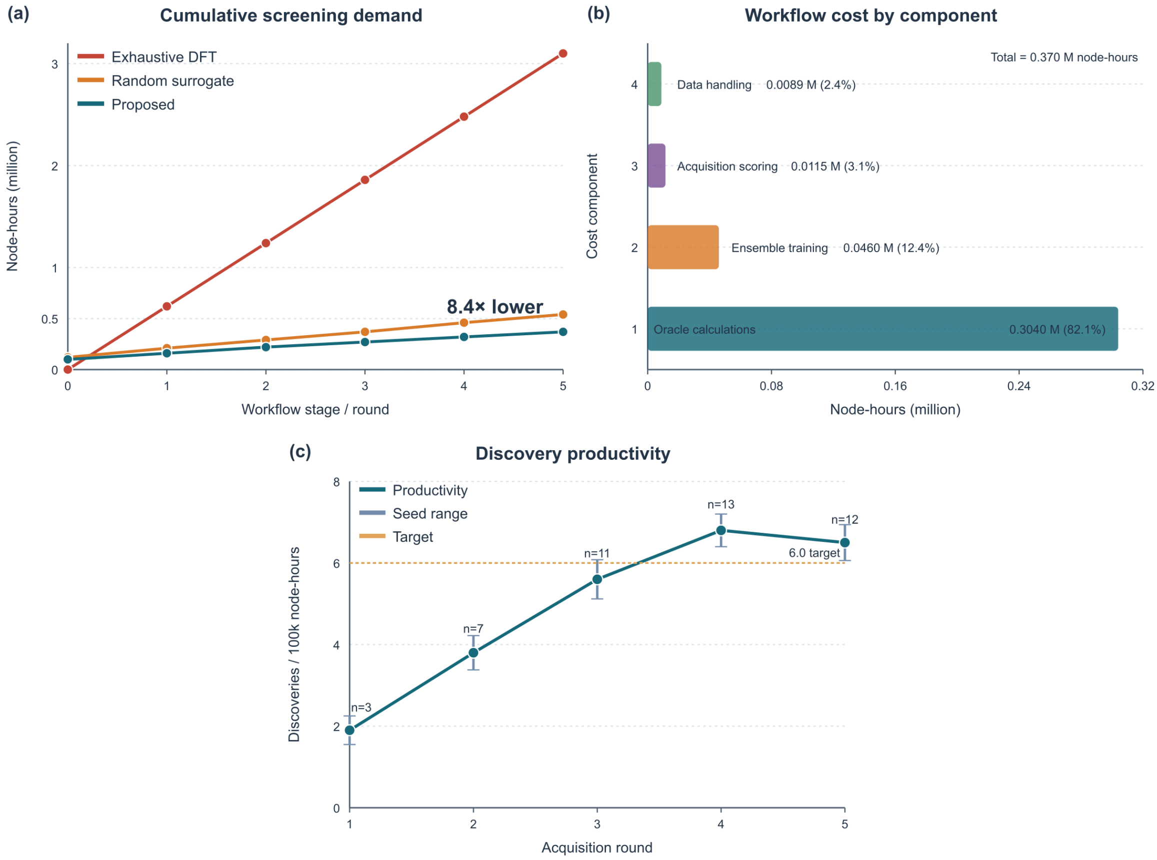

Figure 7. Computational impact: (a) cumulative normalized node-hours by screening strategy; (b) horizontal component-cost comparison; (c) verified discoveries per 100,000 node-hours with seed-range whiskers and verified-count annotations.

Sensitivity analysis varied the oracle batch size from 100 to 600. Batches of 100 produced slightly higher final recall, 88.1%, but required fifteen retraining cycles and 31% more training time. Batches of 600 reduced overhead but lowered recall to 82.3% because each decision used a stale model, whereas a batch of 300 provided the best node-hour-adjusted recall. The family-quota test was evaluated separately because diversity-oriented materials searches may accept a small reduction in aggregate recall to preserve chemical novelty [25]. Increasing the quota from 15% to 25% improved mixed-anion recall by 4.2 points while lowering overall recall by 1.1 points.

Comparison with static screening further clarifies the engineering benefit. A static model trained on 2,700 random labels selected 55 mean-feasible candidates but verified only 39, yielding 70.9% precision. Its errors were most pronounced near the voltage upper bound and in vacancy-rich Prussian-blue structures. Active selection instead placed 41% of new labels within one calibrated standard deviation of a feasibility surface, an allocation consistent with decision-focused sampling for rare-candidate discovery [26]. The final surrogate therefore learned the regions that controlled release decisions rather than minimizing error uniformly over the pool.

The sensitivity of active learning to missing structural families in the seed data has been documented previously [27]. Accordingly, robustness was assessed across five initialization seeds, three initial-set draws, and a reduced seed containing only 600 structures. The standard deviation of top-60 recall was 1.8 percentage points for the proposed policy, compared with 4.9 for pure uncertainty sampling, and candidate overlap between independent runs reached a Jaccard index of 0.81 after the fifth round. Under the reduced seed, early calibration deteriorated and two additional rounds were required to reach 80% recall.

The 47 released structures include 18 layered oxides, 12 polyanionic phases, six Prussian-blue analogues, and 11 mixed-anion frameworks. Their predicted voltages span 2.86–4.06 V, capacities span 123–188 mAh g^{-1}, and diffusion proxies span 0.29–0.54 eV. No single chemistry dominates every objective. Retaining a Pareto portfolio, rather than imposing one uncertain preference weight, preserves candidates with distinct engineering advantages [28] and provides a more useful basis for subsequent synthesis planning.

Threshold perturbation was employed to determine whether the reported short list was sensitive to a single convenient operating point. Tighten the hull limit from 45 to 35 meV atom^{-1}; the released set dropped from 47 to 34, and the verified precision increased from 89.4% to 94.1%, but four mixed-anion candidates with attractive transport predictions were excluded. Raising the capacity floor from 120 to 140 mAh g^{-1} added 29 candidates and shifted the portfolio towards layered oxides. On the other hand, changing the voltage interval by ±0.10 V only changed five memberships because the calibration uncertainty had already reduced the boundary cases. Therefore, the continuous gates generated gradual and interpretable portfolio changes rather than sudden rank inversions.

The diffusion proxy was excluded from the validation set as it lacked a direct label. Nudged Elastic Band calculations were performed retrospectively for 120 candidates across all families and the entire predicted range. The proxy had a Spearman correlation of 0.79 with the calculated barrier and a mean absolute deviation of 0.073 eV. The errors were relatively large for cooperative migration in vacancy-rich channels because a local bottleneck cannot achieve long-range ordering. Replacing the proxy threshold with the verified barrier for this subset changed six feasibility decisions. Therefore, it will be used for economical triage, and transport verification will still need to be carried out before the experiment.

Screening quality was also measured as a function of shortlist size. For top-20, top-40, top-60, and top-100 queues, retrospective recall was 90.0%, 87.5%, 86.7%, and 84.0%, respectively. Precision declined from 95.0% at top-20 to 82.0% at top-100 as the queue entered a region with closer property trade-offs. The selected top-60 operating point balanced laboratory throughput against the probability of missing an attractive candidate. A laboratory with capacity for only twenty follow-up calculations could use the same recorded Pareto order without retraining the model, whereas a broader exploratory program could retain the near-boundary tier.

The active-learning trajectory showed which local environments were initially lacking. Early uncertainty was highest for mixed transition-metal octahedra, fluorinated polyanion groups and partially occupied sodium channels. After the second round, the residual variance of the common layered-oxide motif decreased significantly, and acquisition switched to cross-family coordination patterns. The nearest neighbor analysis in the embedding space shows that the median distance from the release candidates to the initial set is 1.42, and the distance to the final labeled set is 0.31; both are normalized units. The reduction shows a concentrated expansion of the range; it is not a reduction in redundant samples.

The last audit has traced every released prediction back to its structure file, model ensemble and calibration parameters, etc. Re-running the pipeline with the recorded hashes reproduced all 47 selections and their order. The two candidates had adjacent Pareto ranks under a different graphics processor due to floating-point errors, but the released set remained the same. This audit trail meets the reproducibility requirements proposed for accountable learning-material workflows [29].

Conclusion

This study has developed a SchNet–Active Learning framework for accelerated screening of sodium-ion battery cathode materials. Continuous-filter atomistic representations were combined with calibrated ensemble uncertainty, diversity-aware batch acquisition, smooth physical feasibility gates and multi-objective Pareto ranking. The framework views costly first-principles calculations as a scarce resource and keeps an explicit record of which candidate structures led to which release decisions. Its assessment included prediction accuracy, uncertainty reliability, chemistry transfer, shortlist recovery and full computational cost.

The existing framework is still limited in accuracy and range of the computational oracle. Finite-temperature transformations, electrolyte reactions, synthesis accessibility, particle-scale transport and long-cycle degradation are not addressed by the selected labels. Extrapolation to unseen redox chemistries is also relatively unreliable compared with interpolation within the represented family, and an ensemble increases training cost relative to a single surrogate. The released shortlist will thus guide higher-fidelity calculations and experiments rather than replacing them.

Future work will combine equivariant representations, reaction-aware stability models, and multi-fidelity labels from rapid computations, precise quantum methods, and targeted experiments. Closed-loop synthesis feedback can adjust both the feasibility threshold and uncertainty calibration based on measurements. Expand the scope of the acquisition objective to include lifecycle cost, elemental criticality and interfacial compatibility, thereby creating an all-encompassing decision platform for green Sodium-ion battery design.

Author Contributions

Zorana Mladenović contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Danka Cvetković contributes to analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Luo, D., Lei, P., Tian, G., Huang, Y., Ren, X., & Xiang, X. (2020). Insight into Electrochemical Properties and Reaction Mechanism of a Cobalt-Rich Prussian Blue Analogue Cathode in a NaSO3CF3 Electrolyte for Aqueous Sodium-Ion Batteries. The Journal of Physical Chemistry C, 124(11), 5958-5965. https://doi.org/10.1021/acs.jpcc.9b11758

[2] Yang, J., Li, D., Wang, X., Zhang, X., Xu, J., & Chen, J. (2020). Constructing micro-nano Na3V2(PO4)3/C architecture for practical high-loading electrode fabrication as superior-rate and ultralong-life sodium ion battery cathode. Energy Storage Materials, 24, 694-699. https://doi.org/10.1016/j.ensm.2019.07.002

[3] Li, Z. Y., Ma, X., Guo, H., He, L., Li, Y., Wei, G., Sun, K., & Chen, D. (2021). Complementary Effect of Ti and Ni Incorporation in Improving the Electrochemical Performance of a Layered Sodium Manganese Oxide Cathode for Sodium-Ion Batteries. ACS Applied Energy Materials, 4(6), 5687-5696. https://doi.org/10.1021/acsaem.1c00527

[4] Huang, Q., Liu, C., Chen, Y., Wang, Y., & Guo, L. (2022). Construction of simultaneous modified Na3V2(PO4)3/C cathode with K/Zr substitution and carbon nanotubes enwrapping for high performance sodium ion battery. Ceramics International, 48(1), 397-406. https://doi.org/10.1016/j.ceramint.2021.09.115

[5] Qi, W., Jiang, W., Wang, M., Ling, R., Yang, C., Wang, Y., & Cao, B. (2022). Capacitance-dominated hierarchical porous three-dimensional carbon framework enhanced Prussian blue analogue as superior cathode for sodium-ion batteries. International Journal of Hydrogen Energy, 47(48), 20942-20950. https://doi.org/10.1016/j.ijhydene.2022.04.216

[6] Min, K. (2021). Virtual Screening of Residual Sodium-Reactive Coating Materials for the Enhancement of the Electrochemical Properties of a Sodium-Ion Battery Cathode. ACS Applied Energy Materials, 4(9), 9923-9931. https://doi.org/10.1021/acsaem.1c01915

[7] Xie, B., Du, Y., Ma, Y., Liu, S., Wang, Y., Zuo, P., Gao, Y., & Yin, G. (2021). Interface Reinforcement of a Prussian Blue Cathode Using a Non-Flammable Co-Solvent Cresyl Diphenyl Phosphate for a High-Safety Na-Ion Battery. ACS Sustainable Chemistry & Engineering, 9(17), 5809-5817. https://doi.org/10.1021/acssuschemeng.0c07853

[8] Wang, R., Chen, T., Cao, Y., Wang, N., & Zhang, J. (2021). K+-stabilized nanostructured amorphous manganese dioxide: excellent electrochemical properties as cathode material for sodium-ion batteries. Ionics, 27(4), 1559-1567. https://doi.org/10.1007/s11581-020-03880-3

[9] Niu, Y., Yi, Z., Zhao, Y., & Xu, M. (2021). Synthesis and comparison of in-situ carbon-decorated sodium manganese vanadium phosphate cathode and sodium-ion full-cell configurations. Nano Select, 2(8), 1544-1553. https://doi.org/10.1002/nano.202000297

[10] Liu, W., Wang, W., Qin, M., & Shen, B. (2020). Successive synthesis and electrochemical properties of Na3FeF6 and NaFeF3/C cathode materials for lithium-ion and sodium-ion batteries. Ceramics International, 46(8), 11436-11440. https://doi.org/10.1016/j.ceramint.2020.01.059

[11] Koch, D., Petnikota, S., Yang, Z., Srinivasan, M., & Manzhos, S. (2020). Electrochemical Performance of B-Type Vanadium Dioxide as a Sodium-Ion Battery Cathode: A Combined Experimental and Theoretical Study. ChemElectroChem, 7(14), 3151-3159. https://doi.org/10.1002/celc.202000686

[12] Lv, Z., Ling, M., Yue, M., Li, X., Song, M., Zheng, Q., & Zhang, H. (2021). Vanadium-based polyanionic compounds as cathode materials for sodium-ion batteries: Toward high-energy and high-power applications. Journal of Energy Chemistry, 55, 361-390. https://doi.org/10.1016/j.jechem.2020.07.008

[13] Kim, H. I., & Son, J. T. (2021). Electrochemical Performance of Alluaudite-type Na_{2.5}Fe_{1.75}$(SO_4)_3$@rGO Cathode Materials Sodium-Ion Battery. Journal of the Korean Battery Society, 1(2), 157-161. https://doi.org/10.53619/kobs.2021.12.1.2.157

[14] Jiang, W., Qi, W., Pan, Q., Jia, Q., Yang, C., & Cao, B. (2021). Potassium ions stabilized hollow Mn-based prussian blue analogue nanocubes as cathode for high performance sodium ions battery. International Journal of Hydrogen Energy, 46(5), 4252-4258. https://doi.org/10.1016/j.ijhydene.2020.10.260

[15] Amjad, M. Z. B., Iqbal, N., Ali, G., Noor, T., Qayyum, A. A., Khan, U. A., & Gao, J. (2022). Synthesis of NaNiF3 and its composite with multi-walled carbon nanotubes as cathode materials for aqueous sodium-ion battery. Journal of Materials Science: Materials in Electronics, 33(21), 16987-17000. https://doi.org/10.1007/s10854-022-08577-z

[16] Xu, Z., Sun, Y., Xie, J., Nie, Y., Xu, X., Tu, J., Zhang, J., Qiu, L., et al. (2022). Scalable Preparation of Mn/Ni Binary Prussian Blue as Sustainable Cathode for Harsh-Condition-Tolerant Sodium-Ion Batteries. ACS Sustainable Chemistry & Engineering, 10(40), 13277-13287. https://doi.org/10.1021/acssuschemeng.2c02377

[17] Luong, H. D., Dinh, V. A., Momida, H., & Oguchi, T. (2020). Insight into the diffusion mechanism of sodium ion–polaron complexes in orthorhombic P2 layered cathode oxide Na x MnO 2. Physical Chemistry Chemical Physics, 22(32), 18219-18228. https://doi.org/10.1039/d0cp03208e

[18] Darbar, D., Reddy, M. V., & Bhattacharya, I. (2021). Understanding the Effect of Zn Doping on Stability of Cobalt-Free P2-Na0.60Fe0.5Mn0.5O2 Cathode for Sodium Ion Batteries. Electrochem, 2(2), 323-334. https://doi.org/10.3390/electrochem2020023

[19] Vishwakarma, G., Sonpal, A., & Hachmann, J. (2021). Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. Trends in Chemistry, 3(2), 146-156. https://doi.org/10.1016/j.trechm.2020.12.004

[20] Liu, Y., Kelley, K. P., Vasudevan, R. K., Funakubo, H., Ziatdinov, M. A., & Kalinin, S. V. (2022). Experimental discovery of structure–property relationships in ferroelectric materials via active learning. Nature Machine Intelligence, 4(4), 341-350. https://doi.org/10.1038/s42256-022-00460-0

[21] Suh, C., Fare, C., Warren, J. A., & Pyzer-Knapp, E. O. (2020). Evolving the Materials Genome: How Machine Learning Is Fueling the Next Generation of Materials Discovery. Annual Review of Materials Research, 50(1), 1-25. https://doi.org/10.1146/annurev-matsci-082019-105100

[22] Bharadwaja, B. V. S. S., Nabian, M. A., Sharma, B., Choudhry, S., & Alankar, A. (2022). Physics-Informed Machine Learning and Uncertainty Quantification for Mechanics of Heterogeneous Materials. Integrating Materials and Manufacturing Innovation, 11(4), 607-627. https://doi.org/10.1007/s40192-022-00283-2

[23] Chehade, A. A., Belgasam, T. M., Ayoub, G., & Zbib, H. M. (2020). Accelerating the Discovery of New DP Steel Using Machine Learning-Based Multiscale Materials Simulations. Metallurgical and Materials Transactions A, 51(6), 3268-3279. https://doi.org/10.1007/s11661-020-05764-7

[24] Min, K., & Cho, E. (2020). Accelerated Discovery of Novel Inorganic Materials with Desired Properties Using Active Learning. The Journal of Physical Chemistry C, 124(27), 14759-14767. https://doi.org/10.1021/acs.jpcc.0c00545

[25] Saal, J. E., Oliynyk, A. O., & Meredig, B. (2020). Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. Annual Review of Materials Research, 50(1), 49-69. https://doi.org/10.1146/annurev-matsci-090319-010954

[26] Lee, H. G., & Seo, D. H. (2022). Prediction of Material's Formation Energy Using Crystal Graph Convolutional Neural Network. Journal of Korean Institute of Electrical and Electronic Materials Engineers, 35(2), 134-142. https://doi.org/10.4313/jkem.2022.35.2.4

[27] Liu, Y., Guo, B., Zou, X., Li, Y., & Shi, S. (2020). Machine learning assisted materials design and discovery for rechargeable batteries. Energy Storage Materials, 31, 434-450. https://doi.org/10.1016/j.ensm.2020.06.033

[28] Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., & Coley, C. W. (2020). Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. Journal of Chemical Information and Modeling, 60(8), 3770-3780. https://doi.org/10.1021/acs.jcim.0c00502

[29] Tavazza, F., DeCost, B., & Choudhary, K. (2021). Uncertainty Prediction for Machine Learning Models of Material Properties. ACS Omega, 6(48), 32431-32440. https://doi.org/10.1021/acsomega.1c03752