Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Graph Neural Network-Based Discovery Of Novel Materials For Solid-State Batteries

Sana Kashaf
College of Material, Chemistry & Chemical Engineering, Chengdu University of Technology Changhua District, Chengdu 610059, Sichuan, China
Email: alikashaf107@gmail.com

Shakir Ullah
College of Geophysics, Chengdu University of Technology, Changhua District, Chengdu, 610059, China Email: shakirhayankhan365@gmail.com

ABSTRACT

Rapid identification of solid electrolytes is a central challenge in the development of all-solid-state batteries because high ionic conductivity must coexist with thermodynamic stability, electrochemical compatibility, electronic insulation, mechanical integrity, and scalable processing. Although graph neural networks (GNNs) have transformed prediction on large computed materials databases, their reliability is less established for small and heterogeneous experimental conductivity datasets. This study presents a leakage-controlled and externally validated public-data framework for solid-electrolyte screening. A stoichiometry-aware composition GNN was trained on the public OBELiX dataset containing 599 synthesized lithium solid electrolytes and its official grouped split (478 training; 121 testing). Element types were represented as graph nodes, atomic and stoichiometric properties as node features, and chemically weighted relationships as message-passing edges. A five-seed GNN ensemble was benchmarked against ridge regression, random forest, extremely randomized trees, and histogram gradient boosting (HGB). On the official hold-out set, HGB was the strongest standalone model (MAE = 1.158 log10(S cm-1), 95% bootstrap CI [0.955, 1.385]; R2 = 0.565; Spearman rho = 0.692), whereas the GNN ensemble achieved MAE = 1.700, R2 = 0.174, and rho = 0.299. The paired bootstrap difference favored HGB by 0.542 log units (95% CI 0.287-0.804), demonstrating that architectural complexity does not guarantee improved generalization in sparse experimental regimes. Nevertheless, the GNN captured complementary transferable signal. On 204 non-overlapping room-temperature records from Liverpool Ionics, an unoptimized equal-weight GNN-HGB consensus achieved MAE = 1.081 (95% CI [0.921, 1.253]), R2 = 0.440, and rho = 0.709. Rule-based isovalent substitutions generated 392 compositions absent from both source datasets, and uncertainty-aware ranking identified Li10.35Ge1.08P1.65S12Sn0.27 as the leading hypothesis, with predicted conductivity of 1.39 x 10-2 S cm-1 and combined uncertainty of 0.258 log units. These formulas are prioritized hypotheses rather than experimentally discovered materials. The study establishes a rigorous pathway in which leakage-resistant benchmarking, simple-model comparison, external validation, uncertainty analysis, and staged DFT and experimental verification jointly support credible data-driven materials discovery.

Keywords: All-Solid-State Batteries; Solid Electrolyte; Graph Neural Network; Ionic Conductivity; Materials Informatics; Uncertainty Quantification; Public Data; Candidate Screening

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Introduction

The transition from conventional lithium-ion batteries to all-solid-state batteries (ASSBs) is motivated by the need for safer electrochemical storage and the possibility of pairing high-capacity electrodes with nonflammable ion-conducting solids. Replacing the liquid electrolyte may suppress leakage and volatility and can support compact cell architectures, but the replacement also creates new transport and interface constraints. The solid electrolyte must conduct Li ions rapidly while blocking electrons, tolerate contact with highly reducing lithium metal and oxidizing cathodes, remain mechanically coherent during cycling, and be manufacturable at practical thickness and pressure [1-3]. Consequently, solid-electrolyte discovery is not a single-property optimization problem. Ionic conductivity is an essential early-stage filter, but a high conductivity value alone does not establish electrochemical stability, interfacial compatibility, mechanical reliability, cost-effectiveness, or cell-level performance.

The size and complexity of the candidate space make exhaustive experimental exploration unrealistic. Li-ion motion depends on crystal topology, site connectivity, bottleneck dimensions, lattice polarizability, mobile-ion concentration, disorder, defect chemistry, and collective diffusion. The measured conductivity additionally depends on variables that are not intrinsic to nominal composition, including phase purity, grain-boundary resistance, pellet density, porosity, processing atmosphere, electrode contact, thermal history, frequency window, and equivalent-circuit interpretation. Landmark sulfide conductors such as Li10GeP2S12 demonstrated that solid electrolytes can approach liquid-electrolyte transport, yet their success did not yield a universal composition-conductivity rule transferable across sulfides, oxides, halides, phosphates, and mixed-anion families [2].

Computational materials discovery has therefore moved toward staged screening. High-throughput density functional theory, structural descriptors, molecular dynamics, and machine learning can remove implausible compounds before expensive synthesis and electrochemical testing. Screening of more than 12,000 Li-containing structures and subsequent machine-learning-assisted transport simulations demonstrated that computational triage can improve the probability of identifying promising conductors compared with unguided selection [3,4]. Public resources such as the Liverpool Ionics database and OBELiX now make it possible to evaluate models against experimentally reported conductivity rather than relying only on computed proxy properties [5,6]. These datasets are especially valuable because experimental measurements contain the variability that a deployable discovery model must ultimately confront.

Graph neural networks are a compelling representation-learning approach because materials can be encoded as interacting entities rather than fixed descriptor vectors. Crystal graph convolutional neural networks, MEGNet, M3GNet, and large-scale graph-learning systems have shown that local atomic environments and message passing can predict diverse materials properties and support broad search campaigns [7-10]. However, success on millions of computed structures does not automatically transfer to several hundred experimental conductivity records. Experimental solid-electrolyte datasets are sparse, chemically imbalanced, affected by censored or inconsistent measurements, and frequently incomplete with respect to crystal

Online ISSN 3007-3197
Print ISSN 3007-3189

structure. Multiple records may share a composition while differing in synthesis, density, phase assemblage, or test method. Under these conditions, a flexible graph model may fit chemically familiar neighborhoods without learning transport rules that generalize to new families.

A central methodological gap is the difference between predictive accuracy and credible discovery. Many materials-learning studies rely on random record splits that allow closely related compositions or repeated literature measurements to appear in both training and test sets. Such leakage can produce optimistic performance estimates. Studies also commonly evaluate a single dataset, compare deep learning with insufficiently tuned baselines, omit uncertainty or applicability-domain analysis, and describe high model scores as material discovery before stability, novelty, synthesis, and transport have been verified. For solid electrolytes, these weaknesses are consequential because an error of one log10 conductivity unit corresponds to a tenfold error in the physical property.

The present study addresses this gap by treating model evaluation as part of the scientific claim. OBELiX provides a grouped hold-out split intended to separate related records, while Liverpool Ionics permits an independent test of transportability after overlapping compositions are removed [5,6]. The study therefore asks not only whether a GNN can fit public data, but whether its predictions survive leakage-resistant testing, whether it adds information beyond strong conventional models, and whether uncertainty can prevent aggressive extrapolation during candidate prioritization.

This work addresses three research questions: (i) can a composition-level GNN learn useful conductivity signal from all OBELiX records when reliable crystal structures are unavailable for many entries; (ii) how does its generalization compare with regularized linear and tree-ensemble baselines; and (iii) can cross-model agreement and ensemble variability support conservative screening of dataset-novel compositions? The study makes five specific contributions:

- It establishes a reproducible benchmark on the official grouped OBELiX split, avoiding the inflated accuracy that may arise from random splitting of chemically related records.
- It introduces a stoichiometry-aware composition graph that can represent the complete experimental dataset, including records without a fully resolved crystal structure.
- It compares the GNN with multiple strong baselines and quantifies uncertainty through seed variability, bootstrap confidence intervals, and cross-model disagreement.
- It evaluates transferability on a non-overlapping, independently curated room-temperature subset of Liverpool Ionics rather than relying solely on internal resampling.
- It generates and ranks isovalent substitution hypotheses while explicitly distinguishing dataset novelty, literature novelty, predicted promise, and experimentally confirmed discovery.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Scientific and computational background

Requirements for a solid electrolyte

A practical SSE must satisfy a multi-objective design problem. Room-temperature conductivity near or above 10^{-3} S cm^{-1} is commonly treated as a useful fast-ion benchmark, although cell-level requirements depend on electrolyte thickness, area-specific resistance, pressure, and electrode architecture [3,6]. Electronic conductivity must remain low to avoid internal leakage and lithium penetration. Thermodynamic and kinetic stability against lithium metal and high-voltage cathodes must be considered together because interphases can be either passivating or continuously reactive. The electrolyte must also possess sufficient fracture resistance and manufacturability for dense, low-defect layers. Figure 1 visualizes these coupled constraints.

Conductivity measurements are especially heterogeneous. Total conductivity includes bulk and grain-boundary contributions; sample density, porosity, electrode contact, fitting choices, frequency range, and thermal history can alter the reported value. A composition-only model cannot recover all such variables. Its predictions should therefore be interpreted as conditional expectations over the measurement practices represented in the training literature, not intrinsic material constants.

Why graphs are useful and why small data remain difficult

A graph representation encodes entities as nodes and relationships as edges. In crystal GNNs, atoms are commonly nodes and local geometric neighborhoods are edges [7–9]. When complete structures are unavailable, a composition graph can instead treat element types as nodes and use stoichiometry and periodic-table similarity to parameterize message passing. This preserves permutation invariance and relational reasoning while avoiding the false assumption that a clean, experimentally matched CIF exists for every record.

The compromise is important. Composition graphs cannot represent site occupancy, coordination, long-range diffusion pathways, or polymorphism. Furthermore, GNN flexibility increases estimation variance when the number of observations is small relative to the number of learnable parameters. Tree ensembles often remain competitive because they exploit sparse elemental fractions and nonlinear thresholds with stronger implicit regularization. A scientifically robust study must therefore benchmark the GNN against simpler alternatives rather than comparing only neural architectures.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

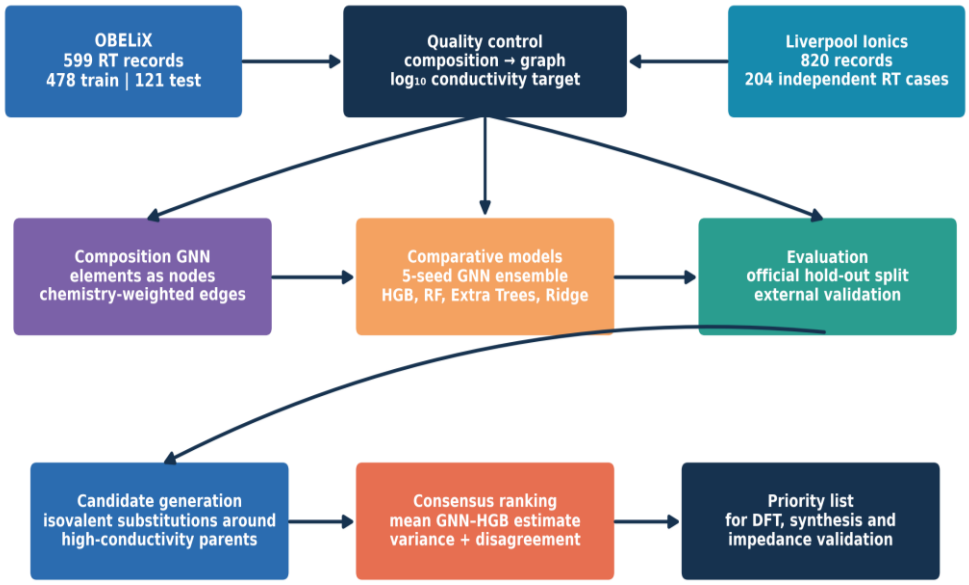

Figure 2: End-to-end study design:Public-data, leakage-controlled discovery workflow

Materials and methods

Public datasets and study design

OBELiX was downloaded from the authors' public repository on 31 July 2026 [6]. It contains 599 synthesized lithium SSE records with room-temperature ionic conductivity, chemical composition, space group, lattice parameters, family labels, and bibliographic provenance. The repository-provided train/test files were used without reassignment: 478 records for model development and 121 records for final evaluation. The split groups records from the same paper or with the same composition to reduce leakage. Thirty-seven conductivity entries were left-censored (reported using "<"). Their numeric upper bounds were retained as targets and a censoring flag was preserved for sensitivity analysis.

The Liverpool Ionics database contains 820 literature-derived conductivity entries measured between 5 and 873 °C [5]. For external validation, records measured from 20 to 35 °C were retained, chemical formulas had to be parsable, and normalized compositions appearing in OBELiX were removed. The resulting external set contained 204 records. It was never used for hyperparameter selection, early stopping, or candidate generation.

Table 1: Public data sources and analytical roles

Dataset	Role	Original size	Records analyzed	Key filtering / split
OBELiX	Development and official test	599	478 train; 121 test	Authors' grouped split; all records retained

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Dataset	Role	Original size	Records analyzed	Key filtering / split
Liverpool Ionics	Independent external validation	820	204	20–35 °C; parsable Li formulas; OBELiX overlaps removed
Virtual substitution pool	Candidate prioritization	Not applicable	392	Isovalent substitutions; absent from both datasets; ≤6 element types

Target definition and preprocessing

The regression target was $y = \log10(\sigma / S\ cm^{-1})$. Log transformation was required because the reported conductivities span more than fifteen orders of magnitude. Chemical formulas were converted to element–amount dictionaries, and each amount was divided by the total atom count to obtain atomic fractions. The parser supported decimal stoichiometries and parenthesized groups. Records failing the external-dataset parser were excluded before any prediction. Normalized composition tuples rounded to five decimal places were used for overlap detection.

No missing values were statistically imputed into the conductivity target. Space group and lattice parameters were not used in the principal models because the purpose was to provide a representation available for all 599 records and transferable to external formulas lacking matched structures. This deliberate choice makes the model broadly applicable but limits its ability to distinguish polymorphs and processing effects.

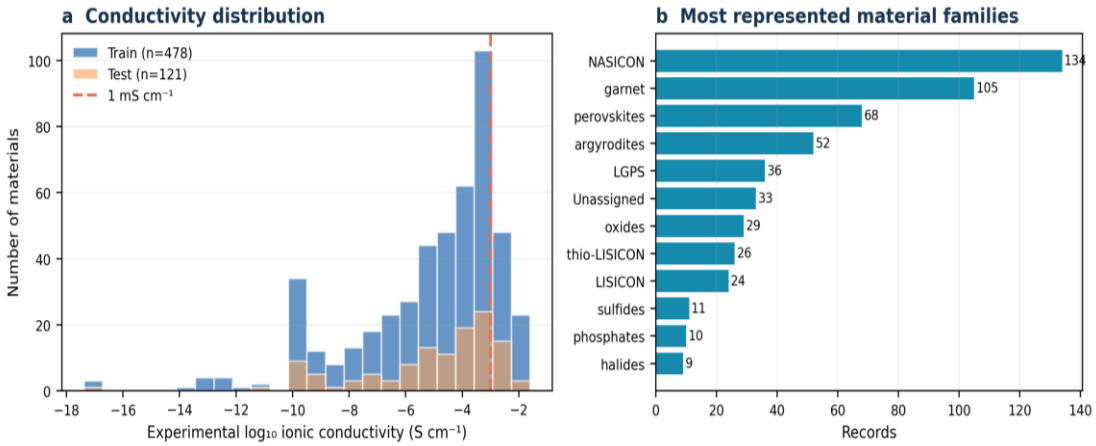

Figure 3: OBELiX dataset landscape. (a) Experimental log10 ionic-conductivity distributions for the official training and test sets. The dashed line marks 10^{-3} S cm^{-1}. (b) The twelve most represented family labels

Composition graph representation

Each material was represented as a complete graph G = (V, E). A node represented one element type in the composition; therefore, the graph contained between two and six nodes. The nine-dimensional node vector contained normalized

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

atomic number, atomic mass, covalent radius, van der Waals radius, periodic-table row, number of outer electrons, default valence, atomic fraction, and log atomic fraction. Elemental properties were obtained from the RDKit periodic table implementation.

The directed edge weight from node i to j was proportional to $f_j[0.15 + \exp(-|Z_i - Z_j|/35)]$, where f_j is the atomic fraction of the receiving element and Z is atomic number. Each row of the adjacency matrix was normalized to sum to one. This construction combines stoichiometric importance with a smooth chemical-similarity prior while ensuring that every element can exchange messages with every other element in the formula.

GNN architecture and training

The network contained an element-embedding multilayer perceptron, two residual message-passing blocks, masked graph pooling, and a regression head (Figure 4). Hidden width was 40. Each message update concatenated the current node state, the weighted neighbor aggregate, and their elementwise product. Layer normalization, SiLU activation, and dropout (0.12) were used. Weighted-mean, arithmetic-mean, and maximum graph summaries were concatenated with five global composition statistics before regression. The model contained 27,865 trainable parameters.

A stratified 16% validation subset was selected only from the official training split. Five models were trained with random seeds 11, 47, 101, 167, and 233. AdamW used learning rate 1.2×10^{-3} and weight decay 2×10^{-4}. The loss was smooth L1 with $\beta = 0.7$. Training was capped at 320 epochs, gradient norm was clipped at 3.0, and the learning rate was reduced by a factor of 0.6 after 25 validation epochs without improvement. The checkpoint with the lowest validation MAE was restored. The ensemble mean formed the GNN prediction and the across-seed standard deviation formed an epistemic-uncertainty proxy.

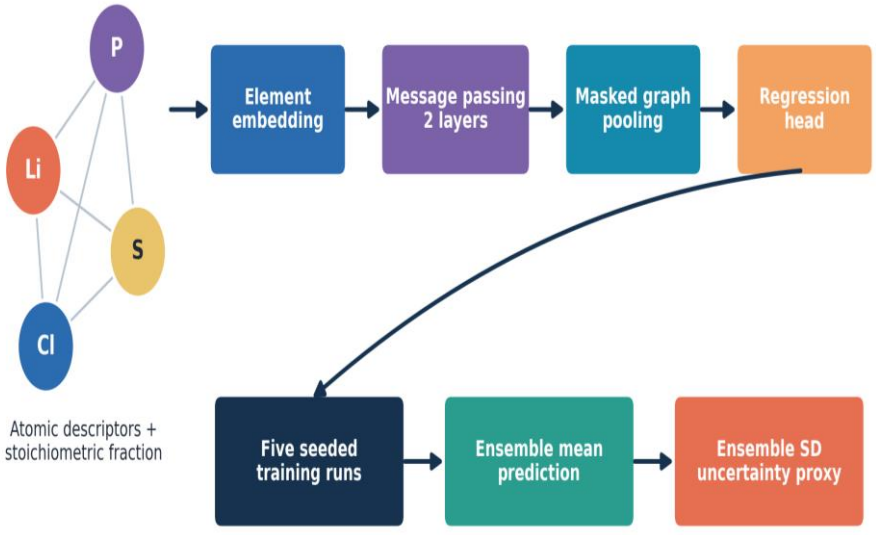

Figure 4: Composition GNN architecture. Element nodes receive atomic and stoichiometric

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

features, exchange weighted messages for two layers, and are pooled into a graph representation. Five independently seeded fits yield an ensemble mean and standard-deviation uncertainty proxy

Comparator models

Comparator models used a 140-dimensional tabular representation: 118 elemental fractions, weighted means and standard deviations of seven atomic properties, number of element types, composition entropy, Li/O/S/Se fractions, total halogen fraction, and total P/As/Sb fraction. Ridge regression used standardized inputs and internal alpha selection. Random forest used 600 trees, minimum leaf size two, and 70% feature sampling. Extremely randomized trees used 800 trees, minimum leaf size two, and 80% feature sampling. HGB used 500 boosting iterations, learning rate 0.04, maximum 15 leaves, and L2 regularization 1.0. All final metrics were computed once on the untouched official test set.

Table 2: Principal model and reproducibility settings

Component	Specification
Graph nodes	Element types; maximum six nodes per composition
Node features	9: normalized Z, mass, covalent radius, vdW radius, row, outer electrons, default valence, fraction, log fraction
Message passing	2 residual layers; hidden width 40; SiLU; layer normalization
Pooling	Stoichiometric weighted mean + arithmetic mean + max + 5 global statistics
Regularization	Dropout 0.12; AdamW weight decay 2×10^{-4}; gradient clipping 3.0
Training	5 seeds; maximum 320 epochs; smooth-L1 $\beta = 0.7$; validation checkpointing
Tabular representation	140 composition and elemental-statistics features
Software	Python; pandas; NumPy; scikit-learn; PyTorch; RDKit; SciPy; Matplotlib

Evaluation and uncertainty

Performance was summarized by MAE, RMSE, coefficient of determination (R^2), and Spearman rank correlation. MAE and RMSE are expressed in log10(S cm^{-1}). Nonparametric 95% confidence intervals were obtained from 3,000 bootstrap resamples of each evaluation set. A paired 5,000-resample bootstrap compared the HGB and GNN absolute errors on identical test records. The uncensored sensitivity analysis excluded left-censored test entries.

For external screening, an equal-weight consensus was defined as $p_c = (p_{GNN} + p_{HGB})/2$. It was not optimized on the external data. Model disagreement was $d = |p_{GNN} - p_{HGB}|$. Combined uncertainty was $u = \sqrt{s_{GNN}^2 + (d/2)^2}$, where s_{GNN} is the across-seed GNN standard deviation. This heuristic deliberately

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

penalizes candidates supported by only one model.

Virtual candidate generation

Parents were selected from the top 20% of OBELiX conductivity values, with a maximum of 90 parent records. Chemically plausible isovalent substitution rules were applied to halogens, chalcogens, group-14 and group-15 elements, selected group-13 elements, and selected transition-metal pairs. Examples included substitution among Cl, Br, and I; between S and Se or O; among Si, Ge, and Sn; among P, As, and Sb; between Zr and Hf; and between Ta and Nb. Any generated composition already present in OBELiX or Liverpool Ionics was removed, as were duplicates and formulas with more than six element types. This yielded 392 dataset-novel hypotheses.

Candidates were ranked by $U = p_c - 0.75u$, which rewards predicted conductivity while penalizing uncertainty. This is a triage score, not a thermodynamic objective. No formation energy, energy above hull, electronic band gap, electrochemical window, toxicity, cost, or synthesis model was applied. Consequently, a high-ranked formula indicates only that its composition resembles high-conductivity regions learned from the datasets.

Results

Dataset characteristics

The full OBELiX target ranged from -17.377 to -1.602 log10(S cm^{-1}), with median -4.290. The compositions contained 55 unique elements and at most six element types. Thirty-seven records (6.2%) were left-censored. The official test set represented 20.2% of all records, matching the source dataset design. The family distribution was concentrated in NASICON, garnet, perovskite, argyrodite, and LGPS-related records (Figure 3), creating a realistic but non-uniform chemical domain.

Table 3: Descriptive characteristics of the modelling dataset

Statistic	Value
Total OBELiX records	599
Official training records	478
Official test records	121
Left-censored records	37
Target minimum	−17.377 log10(S cm^{-1})
Target median	−4.290 log10(S cm^{-1})
Target maximum	−1.602 log10(S cm^{-1})
Unique elements	55
Maximum element types per formula	6

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Official test-set comparison

HGB obtained the lowest official test error, with MAE 1.158 (95% CI [0.955, 1.385]), RMSE 1.672, R^2 0.565, and Spearman ρ 0.692. Random forest and Extra Trees were close, with MAEs 1.173 and 1.188. The GNN–HGB equal-weight consensus obtained MAE 1.299, intermediate between HGB and the GNN. The five-seed GNN ensemble produced MAE 1.700 (95% CI [1.425, 1.990]), RMSE 2.305, R^2 0.174, and ρ 0.299. Ridge regression and the mean baseline were weaker.

The paired bootstrap difference, calculated as MAE(HGB) − MAE(GNN), was −0.542 log units (95% CI [−0.804, −0.287]); all 5,000 bootstrap draws favored HGB. Excluding censored test observations did not improve the GNN materially: MAE was 1.701, RMSE 2.325, R^2 −0.065, and ρ 0.162. The result therefore cannot be attributed solely to the upper-bound treatment of censored targets.

Table 4: Performance on the untouched official OBELiX test set (n = 121). Lower MAE/RMSE and higher R^2/ρ indicate better performance

Model	MAE	RMSE	R^2	Spearman ρ
Histogram gradient boosting	1.158	1.672	0.565	0.692
Random forest	1.173	1.737	0.531	0.686
Extra Trees	1.188	1.833	0.478	0.673
GNN–HGB equal-weight consensus	1.299	1.780	0.508	0.558
Composition GNN ensemble	1.700	2.305	0.174	0.299
Mean baseline	1.947	2.537	-0.000	—
Ridge composition	2.005	2.914	-0.320	0.340

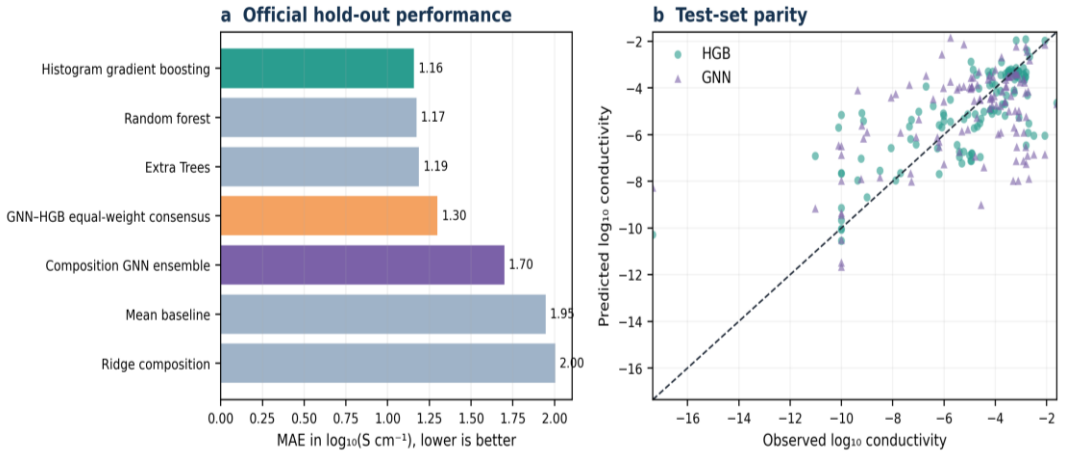

Figure 5:(a) MAE ranking of all models. (b) Observed-versus-predicted values for HGB and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

the composition GNN ensemble. The tree ensemble is substantially more accurate, while both models compress the extreme low-conductivity tail

GNN seed stability

Individual GNN fits exhibited non-trivial variation. Across five seeds, test MAE was 1.741 ± 0.077, R^2 was 0.131 ± 0.054, and Spearman ρ was 0.288 ± 0.037. All runs reached the 320-epoch cap, but the best validation checkpoint was retained. The spread confirms that point estimates from a single neural-network seed would overstate precision in this data regime.

Table 5: Stability of independently seeded GNN fits

Seed	Validation MAE	Epochs	Test MAE	Test RMSE	Test R^2	Test ρ
11	1.368	320	1.697	2.375	0.123	0.323
47	1.183	320	1.798	2.421	0.089	0.283
101	1.142	320	1.639	2.244	0.217	0.315
167	1.281	320	1.740	2.349	0.142	0.290
233	1.433	320	1.831	2.428	0.084	0.228
Mean ± SD	1.281 ± 0.122	320	1.741 ± 0.077	2.364 ± 0.074	0.131 ± 0.054	0.288 ± 0.037

Independent external validation

The independent Liverpool set provided a different picture of model complementarity. The GNN achieved MAE 1.220, R^2 0.341, and ρ 0.621. HGB achieved slightly lower MAE (1.181) and higher ρ (0.721), but its RMSE and R^2 were marginally weaker than those of the GNN. The equal-weight consensus improved the overall external metrics to MAE 1.081 (95% CI [0.921, 1.253]), RMSE 1.650, R^2 0.440, and ρ 0.709. Thus, although the consensus did not beat HGB internally, it reduced external absolute error by combining partially distinct model biases.

Table 6: Independent validation on 204 non-overlapping Liverpool room-temperature records

Model	MAE	95% CI for MAE	RMSE	R^2	Spearman ρ
Composition GNN	1.220	[1.051, 1.406]	1.791	0.341	0.621
HGB	1.181	[0.994, 1.378]	1.819	0.320	0.721
Equal-weight consensus	1.081	[0.921, 1.253]	1.650	0.440	0.709

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Candidate prioritization

Virtual substitution generated 392 unique compositions absent from both analytical datasets. The consensus predicted 118 candidates above 10^{-3} S cm^{-1} and 5 above 10^{-2} S cm^{-1}. However, the median combined uncertainty was 0.636 log units and median GNN–HGB disagreement was 1.105, indicating substantial extrapolation risk across much of the pool.

The highest-ranked hypotheses were dominated by LGPS-like sulfide compositions and Si/Ge/Sn substitutions. The top candidate, Li10.35Ge1.08P1.65S12Sn0.27, had GNN prediction −1.963, HGB prediction −1.750, consensus −1.856, corresponding to 13.9 mS cm^{-1}, with combined uncertainty 0.258 log units. The leading S→Se candidate, Li10Ge0.66P2Se12Sn0.33, had a lower consensus estimate of 7.81 mS cm^{-1} and greater combined uncertainty (0.308) because the two models disagreed by 0.524 log units.

Table 7: Highest-priority dataset-novel substitution hypotheses. Rankings are model outputs, not evidence of stability, synthesizability, or experimental conductivity

Rank	Candidate formula	Substitution	Consensus log σ	Predicted σ (mS cm^{-1})	Combined uncertainty
1	Li10.35Ge1.08P1.65S12Sn0.27	Si->Ge	-1.856	13.92	0.258
2	Li10.2Ge0.96P1.8S12Sn0.24	Si->Ge	-1.945	11.34	0.167
3	Li10Ge0.95P2S12Sn0.05	Si->Sn	-1.972	10.66	0.183
4	Li10.5Ge1.2P1.5S12Sn0.3	Si->Ge	-1.898	12.65	0.311
5	Li3.425Ge0.425P0.575S4	Si->Ge	-2.026	9.43	0.339
6	Li10P2S12Si0.416Sn0.584	Ge->Si	-2.161	6.90	0.187
7	Li10P2S12Si0.05Sn0.95	Ge->Sn	-2.247	5.66	0.118
8	Li10Ge0.66P2Se12Sn0.33	S->Se	-2.107	7.81	0.308
9	Li10P2S12Si0.33Sn0.66	Ge->Si	-2.236	5.81	0.138
10	Li9.54Cl0.3Ge1.74P1.44S11.7	Si->Ge	-1.942	11.43	0.534
11	Li3.475Ge0.475P0.525S4	Si->Ge	-2.080	8.32	0.361
12	Li3.335Ge0.33P0.67S4	Sn->Ge	-2.246	5.67	0.153

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

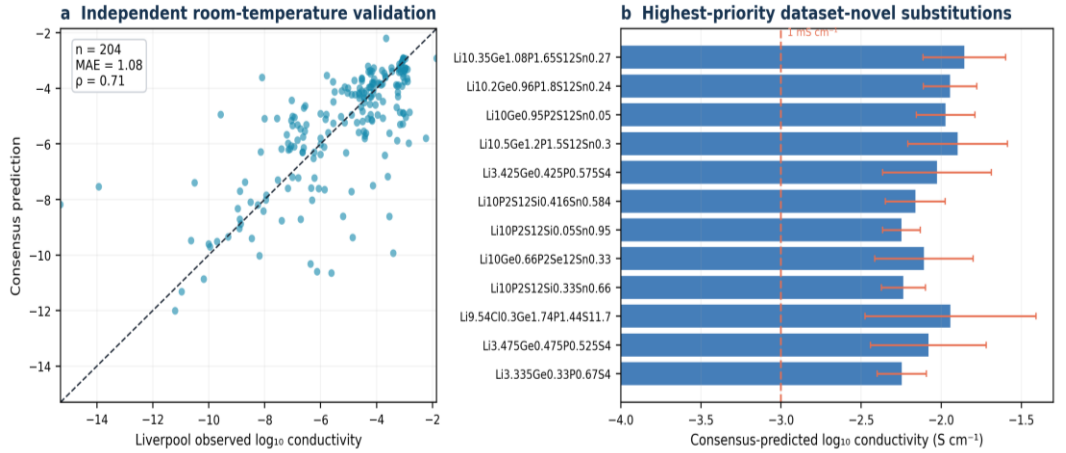

Figure 6: External validation and screening output. (a) Consensus predictions against 204 independent Liverpool room-temperature measurements. (b) Twelve highest-priority dataset-novel substitutions; horizontal error bars show combined uncertainty. The dashed line is 10^{-3} S cm^{-1}

Discussion

The principal finding: graph complexity did not guarantee accuracy

The strongest internal result is the superiority of HGB over the standalone GNN. This should not be treated as a failed study. It is an important empirical boundary condition for materials informatics: when 599 experimental measurements span many families and orders of magnitude, a 27,865-parameter graph model can have higher variance than a well-regularized tree ensemble. The finding is consistent with the OBELiX source benchmark, which reported that random forest and a simple multilayer perceptron outperformed several three-dimensional geometric architectures on the grouped test set [6].

Three factors likely contribute. First, the graph is composition based and therefore cannot encode diffusion topology, polymorphism, or partial occupancy. Second, conductivity labels integrate uncontrolled experimental variation. A structure-sensitive model cannot recover sample density, grain boundaries, electrode contact, or fitting practice if those covariates are absent. Third, family and publication grouping makes the official test set deliberately more difficult than a random record split. A model that memorizes local compositional neighborhoods receives less benefit under this design.

Why retain the GNN?

Despite weaker internal accuracy, the GNN should not be discarded. Its external rank correlation of 0.621 demonstrates non-random transportability, and the consensus improved external MAE and R^2. The models appear to make different errors: HGB exploits thresholded elemental fractions and summary statistics, whereas the GNN learns smooth relational interactions among elements. Averaging can reduce dataset-specific bias when these errors are not perfectly correlated. This benefit is visible only because the study used a genuinely independent dataset rather than a second random split of OBELiX.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The result supports a pragmatic model-selection policy. A GNN should earn its place through held-out and external evidence, not through architectural prestige. For near-term screening, HGB is the preferred single predictor. For conservative candidate ranking, the GNN contributes disagreement and seed variance, allowing uncertain extrapolations to be downgraded. This is a more defensible use of deep learning in small experimental data than claiming state-of-the-art accuracy from one split.

Chemical interpretation of the candidate list

The leading candidates are concentrated around sulfide LGPS-like parents. This is chemically plausible because sulfides offer polarizable anion frameworks and have produced some of the highest reported room-temperature Li conductivities. The model's preference for Si→Ge and related substitutions also reflects known compositional neighborhoods in the training data. At the same time, this concentration exposes a major risk: candidate ranking may reproduce the density of the source literature rather than uncover remote chemical families.

Several candidates contain substantial Ge, which raises cost and resource concerns. Direct isovalent substitution can also destabilize a framework, change site ordering, or produce an unsynthesizable nominal composition. Selenium substitution may improve polarizability but can alter redox stability and materials cost. Therefore, Table 7 is best viewed as a computational queue. The most valuable candidates are not necessarily those with the highest predicted conductivity, but those that jointly show high consensus, low disagreement, plausible charge balance, and a tractable route to phase-pure synthesis.

Proposed DFT and experimental validation cascade

A credible follow-up should use staged elimination so that expensive methods are applied only to defensible candidates. The proposed sequence is:

- Novelty and identity: search exact, reduced, substituted, and prototype-related formulas; identify likely crystal prototypes and competing phases.
- Charge balance and chemical sanity: assign plausible oxidation states; remove obviously unstable or toxic combinations; assess critical-element burden.
- Thermodynamic screening: enumerate structures; relax with density functional theory; calculate formation energy and energy above the convex hull.
- Electronic and electrochemical screening: require a sufficiently large electronic band gap or low electronic conductivity; estimate reduction and oxidation limits and electrode reaction energies.
- Transport simulation: perform bond-valence or geometric pre-screening, followed by NEB and/or AIMD at multiple temperatures; quantify activation energy and extrapolated room-temperature conductivity with uncertainty.
- Synthesis and phase verification: document precursors, atmosphere, milling, temperature profile, pressure, moisture control, and yield; use XRD/Rietveld analysis and compositional characterization.
- Electrochemical validation: measure total and bulk conductivity by impedance spectroscopy across temperature; report cell geometry, density, equivalent

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

circuit, repeat specimens, activation energy, and Li-ion transference.

- Interface and cell validation: test Li|SSE|Li cycling, cathode compatibility, critical current density, and full-cell retention under reported pressure and temperature.

Table 8: Recommended validation cascade for converting a model-ranked formula into a defensible materials claim

Stage	Primary output	Go / no-go example
1. Novelty search	Known phases, patents, prototypes	No unresolved prior identical claim
2. Chemical checks	Oxidation states, hazards, cost	Plausible charge balance and manageable elements
3. DFT stability	Formation energy; energy above hull	Prefer low metastability; threshold justified by synthesis route
4. Interface / electronic	Band gap; reaction energies	Electronically insulating and kinetically manageable interphases
5. Ion transport	Barrier; AIMD conductivity	Transport remains promising after structure relaxation
6. Experimental phase	Phase purity and density	Reproducible target-dominant phase
7. Impedance / cell	σ, Ea, transference, cycling	Replicated performance with complete reporting

Scientific implications for data-driven ASSB discovery

The results have a broader implication for materials informatics: the effective complexity of a model should be matched to the information content of the data, not to the visual sophistication of the architecture. In the present dataset, conductivity varies because of hidden structural and experimental variables that composition alone cannot resolve. Increasing neural-network capacity cannot reconstruct absent covariates. Progress therefore depends as much on standardized measurement metadata, repeat experiments, explicit crystal and defect descriptions, and negative results as it does on improved message-passing functions.

A high-value future platform would be multi-fidelity rather than purely

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

predictive. Experimental conductivity could be linked to relaxed structures, energy-above-hull calculations, electronic band gaps, electrochemical decomposition energies, elastic properties, diffusion-path descriptors, and synthesis conditions. A graph model could then learn from atomic geometry, while calibrated tree or probabilistic models integrate processing metadata and uncertainty. Active learning should select candidates that are both promising and informative, thereby expanding underrepresented chemical families instead of repeatedly sampling well-studied LGPS-like regions.

For practitioners, the immediate decision rule is conservative: use the strongest validated predictor for numerical ranking, retain the GNN when it contributes independent signal, penalize candidates with high model disagreement, and require physical verification before increasing the strength of the claim. This hierarchy converts machine learning from a source of overconfident lists into a transparent prioritization instrument. It also creates a direct route for future experimental results to update the model and progressively reduce uncertainty.

Limitations and threats to validity

First, the model predicts only ionic conductivity. It does not evaluate thermodynamic stability, electrochemical window, electronic leakage, mechanical properties, moisture stability, toxicity, cost, or manufacturability. Second, composition graphs cannot distinguish polymorphs, site disorder, vacancy topology, or microstructure. Third, the training labels are literature measurements collected under heterogeneous conditions; a prediction is therefore not equivalent to an intrinsic single-crystal value.

Fourth, 37 upper-bound values were treated numerically, which is not a full censored-regression solution. The uncensored sensitivity result suggests that this choice does not explain the GNN's weak test performance, but dedicated Tobit, interval, or survival-style objectives could be explored. Fifth, the external set was filtered by formula parsing and temperature, which may introduce selection bias. Sixth, the consensus and uncertainty formulas are transparent heuristics rather than calibrated posterior distributions.

Conclusions

This study developed a reproducible public-data framework for graph neural network-guided screening of lithium solid electrolytes for all-solid-state batteries. The analysis used experimentally reported data rather than fabricated laboratory measurements and preserved the leakage-controlled OBELiX split. The principal internal result was unambiguous: histogram gradient boosting achieved MAE = 1.158 log10(S cm-1), whereas the five-seed composition GNN achieved MAE = 1.700. Thus, the deepest model was not the strongest standalone predictor. This outcome directly answers the first research question and shows that model complexity cannot compensate for small sample size, chemical imbalance, incomplete structural information, and unobserved processing variables.

The weaker internal GNN performance is scientifically informative rather than merely negative. A composition graph captures smooth interactions among constituent elements, but conductivity is also governed by polymorphism, defects,

Online ISSN 3007-3197
Print ISSN 3007-3189

diffusion topology, density, grain boundaries, and measurement protocol. Because these factors were unavailable for much of the public corpus, the graph representation faced an irreducible information limit. HGB was better able to exploit stable thresholds and nonlinear combinations in the engineered composition descriptors. The finding supports a central principle for small-data materials research: sophisticated architectures must demonstrate incremental value against strong, well-regularized baselines under leakage-resistant evaluation.

External validation nevertheless showed why the GNN remained useful. On 204 non-overlapping room-temperature Liverpool records, the equal-weight GNN-HGB consensus achieved MAE = 1.081, R2 = 0.440, and Spearman rho = 0.709. The improvement indicates that the two models encode partially different chemical information and make non-identical errors. In practical terms, HGB provides the more accurate primary estimate, while the GNN contributes complementary signal, seed-based variability, and a second perspective for identifying unstable extrapolations. This is a stronger and more credible role for graph learning than assuming that it must dominate every benchmark.

The virtual screening stage generated 392 compositions absent from OBELiX and Liverpool Ionics, including 118 hypotheses predicted above 1 mS cm-1 and five above 10 mS cm-1. These results narrow the search space, but they do not prove global novelty, phase stability, synthesizability, or electrochemical performance. Even the highest-ranked formula, Li10.35Ge1.08P1.65S12Sn0.27, remains a computational hypothesis. A defensible discovery claim requires database and patent searching, structure identification, charge-balance and toxicity assessment, DFT stability and electrochemical-window calculations, diffusion simulations, controlled synthesis, phase-purity analysis, temperature-dependent impedance spectroscopy, transference measurements, interface testing, and full-cell validation.

The broader contribution of the paper is therefore methodological and practical. Reliable data-driven discovery of solid electrolytes should combine curated experimental data, family-aware splitting, transparent baselines, independent validation, calibrated uncertainty, chemical constraints, and sequential physical verification. Future gains are likely to come from richer multi-fidelity datasets that unite crystal structures, defect chemistry, synthesis conditions, microstructure, and standardized conductivity measurements. Within that framework, GNNs can become powerful components of an evidence chain rather than substitutes for evidence. The present study provides that chain: it identifies promising candidates, quantifies where confidence is limited, and defines the experiments required to convert a model-ranked formula into a validated material for all-solid-state batteries.

References

[1] Janek, J.; Zeier, W. G. A solid future for battery development. Nature Energy 2016, 1, 16141. https://doi.org/10.1038/nenergy.2016.141.

[2] Kamaya, N.; Homma, K.; Yamakawa, Y.; et al. A lithium superionic conductor. Nature Materials 2011, 10, 682–686. https://doi.org/10.1038/nmat3066.

[3] Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K.-A. N.; Cui, Y.; Reed, E. J. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. Energy & Environmental Science 2017, 10, 306–320.

Online ISSN 3007-3197
Print ISSN 3007-3189

https://doi.org/10.1039/C6EE02697D.

[4] Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J. Machine learning-assisted discovery of solid Li-ion conducting materials. Chemistry of Materials 2019, 31, 342–352. https://doi.org/10.1021/acs.chemmater.8b03272.

[5] Hargreaves, C. J.; Gaultois, M. W.; Daniels, L. M.; et al. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. npj Computational Materials 2023, 9, 9. https://doi.org/10.1038/s41524-022-00951-z.

[6] Therrien, F.; Abou Haibeh, J.; Sharma, D.; et al. OBELiX: a curated dataset of crystal structures and experimentally measured ionic conductivities for lithium solid-state electrolytes. Digital Discovery 2026, 5, 910. https://doi.org/10.1039/D5DD00441A.

[7] Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters 2018, 120, 145301. https://doi.org/10.1103/PhysRevLett.120.145301.

[8] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials 2019, 31, 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294.

[9] Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science 2022, 2, 718–728. https://doi.org/10.1038/s43588-022-00349-3.

[10] Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. Nature 2023, 624, 80–85. https://doi.org/10.1038/s41586-023-06735-9.

[11] Jain, A.; Ong, S. P.; Hautier, G.; et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. APL Materials 2013, 1, 011002. https://doi.org/10.1063/1.4812323.

[12] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. Nature 2018, 559, 547–555. https://doi.org/10.1038/s41586-018-0337-2.

[13] Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials 2019, 5, 83. https://doi.org/10.1038/s41524-019-0221-0.

[14] Breiman, L. Random forests. Machine Learning 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

[15] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. Machine Learning 2006, 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

[16] Friedman, J. H. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001, 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

[17] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.

[18] Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, *14*(2), 1518-1527.

[19] Paszke, A.; Gross, S.; Massa, F.; et al. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 2019, 32, 8024–8035.

[20] Landrum, G. RDKit: Open-source cheminformatics software. https://www.rdkit.org (accessed 31 July 2026).

[21] Efron, B.; Tibshirani, R. J. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1993.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Appendix A. Substitution rules used for virtual screening

Table A1: Rule-based substitutions

Source element	Allowed substitutes
Cl	Br, I
Br	Cl, I
I	Br, Cl
S	Se, O
Se	S
O	S
Si	Ge, Sn
Ge	Si, Sn
Sn	Ge
P	As, Sb
As	P, Sb
Sb	P, As
Al	Ga, In
Ga	Al, In
In	Ga
Zr	Hf
Hf	Zr
Ti	Zr, Hf
Ta	Nb
Nb	Ta
Y	Sc, Lu
Sc	Y
La	Nd, Sm
Nd	La, Sm

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Appendix B. Detailed diagnostics for the top candidates

Table B1. Model diagnostics for the fifteen leading candidate hypotheses

#	Candidate	Parent	Change	GNN log σ	HGB log σ	Disagreement	Uncertainty
1	Li10.35Ge1.08P1.65S12Sn0.27	Li10.35(Sn0.27Si1.08)P1.65S12	Si->Ge	-1.963	-1.750	0.213	0.258
2	Li10.2Ge0.96P1.8S12Sn0.24	Li10.2(Sn0.2Si0.8)1.2P1.8S12	Si->Ge	-1.983	-1.908	0.075	0.167
3	Li10Ge0.95P2S12Sn0.05	Li10Ge0.95Si0.05P2S12	Si->Sn	-2.080	-1.865	0.215	0.183
4	Li10.5Ge1.2P1.5S12Sn0.3	Li10.5(Sn0.2Si0.8)1.5P1.5S12	Si->Ge	-1.941	-1.855	0.087	0.311
5	Li3.425Ge0.425P0.575S4	Li3.425Si0.425P0.575S4	Si->Ge	-2.241	-1.810	0.431	0.339
6	Li10P2S12Si0.416Sn0.584	Li10(Ge0.416Sn0.584)P2S12	Ge->Si	-2.285	-2.038	0.247	0.187
7	Li10P2S12Si0.05Sn0.95	Li10Ge0.95Si0.05P2S12	Ge->Sn	-2.300	-2.195	0.105	0.118
8	Li10Ge0.66P2Se12Sn0.33	Li10Ge0.66Sn0.33P2S12	S->Se	-1.845	-2.369	0.524	0.308
9	Li10P2S12Si0.33Sn0.66	Li10Ge0.33Sn0.66P2S12	Ge->Si	-2.280	-2.192	0.088	0.138
10	Li9.54Cl0.3Ge1.74P1.44S11.7	Li9.54Si1.74P1.44S11.7Cl0.3	Si->Ge	-2.348	-1.536	0.812	0.534
11	Li3.475Ge0.475P0.525S4	Li3.475Si0.475P0.525S4	Si->Ge	-2.294	-1.867	0.427	0.361
12	Li3.335Ge0.33P0.67S4	Li3.335Sn0.33P0.67S4	Sn->Ge	-2.292	-2.201	0.091	0.153
13	Li10P2S12Si0.776Sn0.224	Li10(Ge0.776Sn0.224)P2S12	Ge->Si	-2.275	-2.130	0.145	0.219
14	Li10.35Ge0.27P1.65S12Si1.08	Li10.35(Sn0.27Si1.08)P1.65S12	Sn->Ge	-2.127	-2.249	0.123	0.242
15	Li10Ge0.776P2Se12Sn0.224	Li10(Ge0.776Sn0.224)P2S12	S->Se	-1.850	-2.424	0.574	0.328