Turning Literature into Knowledge: AI-Driven Discovery of Solid-State Electrolytes

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study proposes an integrated discovery framework for next-generation battery applications, aimed at elucidating composition–structure–property relationships in sulfide solid electrolytes. The framework combines literature-derived data extraction, knowledge graph (KG) construction, large language model (LLM)–driven candidate material proposal, and validation through molecular dynamics (MD) simulations with machine-learning interatomic potentials. From 125 collected articles, 56 meeting quality criteria were converted into 146 normalized records that contained composition, space group, density, measurement temperature, and ionic conductivity, and 46 high-confidence entries were used to build the knowledge graph. Conditioning an LLM on the KG resulted in the identification of two novel candidate materials absent from the graph ($\mathrm{Li_{11}Si_2PS_{12}}$ and $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$). Diffusion simulations using GRACE and MACE over 600–1400 K, along with Arrhenius extrapolation, predicted activation energies of 0.21 and 0.20 eV and room-temperature ionic conductivities of 12.7 and 14.0 mS cm^{-1}, which are sufficiently high to be considered practically applicable. This end-to-end AI-driven framework is expected to establish a generalizable, data-centric paradigm for materials discovery that transcends conventional design limits, regardless of the targeted application domain.

1 Introduction

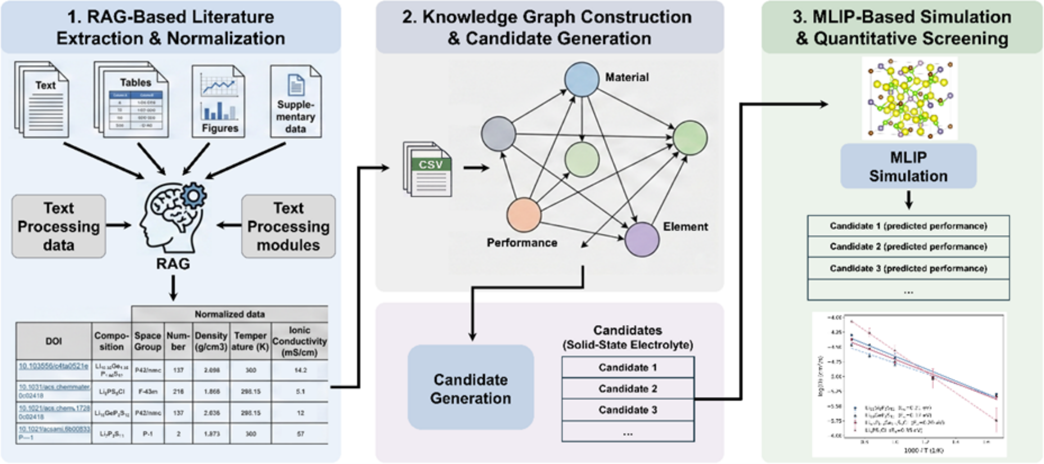
All-solid state batteries are emerging as a leading next-generation energy storage system capable of achieving both high energy density and safety [1]. Realizing their commercial potential requires the discovery of solid electrolyte materials that simultaneously satisfy conflicting requirements such as ionic conductivity, electrochemical stability, mechanical properties, and processability [2]. The primary challenge lies in the virtually infinite combinations of chemical compositions and crystal structures. Given this vast search space, traditional experimental trial-and-error methods face clear limitations due to excessive cost and time constraints.

To address these challenges, the materials science community has adopted high-throughput screening and artificial intelligence based on large-scale computational databases like the Materials Project and OQMD [3, 4, 5]. While these databases offer extensive data under unified protocols, they possess a critical limitation as they predominantly assume ideal crystal structures at absolute zero temperature without defects. In contrast, the actual performance of solid electrolytes is governed by realistic factors including operating temperatures, synthesis conditions, and grain boundary characteristics [6]. Consequently, this reality gap between idealized calculations and physical experiments becomes a major cause of failure during validation and undermines the reliability of data-driven exploration.

Scientific literature accumulated over decades represents the richest source of data capable of bridging this gap [7]. These documents contain actual ionic conductivity data across various synthesis conditions and temperatures. However, this information remains fragmented across unstructured text, tables, and figures using inconsistent unit systems [8]. Although human researchers can interpret these nuances, the structural limitations prevent artificial intelligence models from systematically learning or analyzing this knowledge. As a result, vast amounts of literature remain unutilized as data assets.

This study proposes an integrated framework that overcomes the unstructured nature of literature and combines it with computational science to present a new pathway for materials discovery (Fig. 1). Our primary contributions are summarized as follows:

- **Literature-Based Data Assetization:** We implemented Retrieval-Augmented Generation (RAG) to extract composition, temperature, and ionic conductivity from text and tables into a standardized schema with normalized units, converting literature into a machine-readable dataset and easing the data-availability bottleneck.
- **Structuring Condition-Performance Relations:** Beyond property lists, we organized the extracted data as a Knowledge Graph (KG) to capture structure–property links and preserve the context of processing conditions, enabling exploration under realistic conditions.
- **Knowledge Graph-Based Candidate Design:** We conditioned Large Language Models (LLMs) on KG-grounded literature evidence to propose new materials, reducing chemically invalid hallucinations and prioritizing feasible candidates consistent with established trends.
- **MLIP-Based Extended Validation:** To reduce DFT cost and move beyond 0 K assumptions, we performed large-scale simulations using Machine Learning Interatomic Potentials (MLIPs), enabling efficient evaluation of stability and ion diffusion at operating temperatures to shortlist promising candidates.

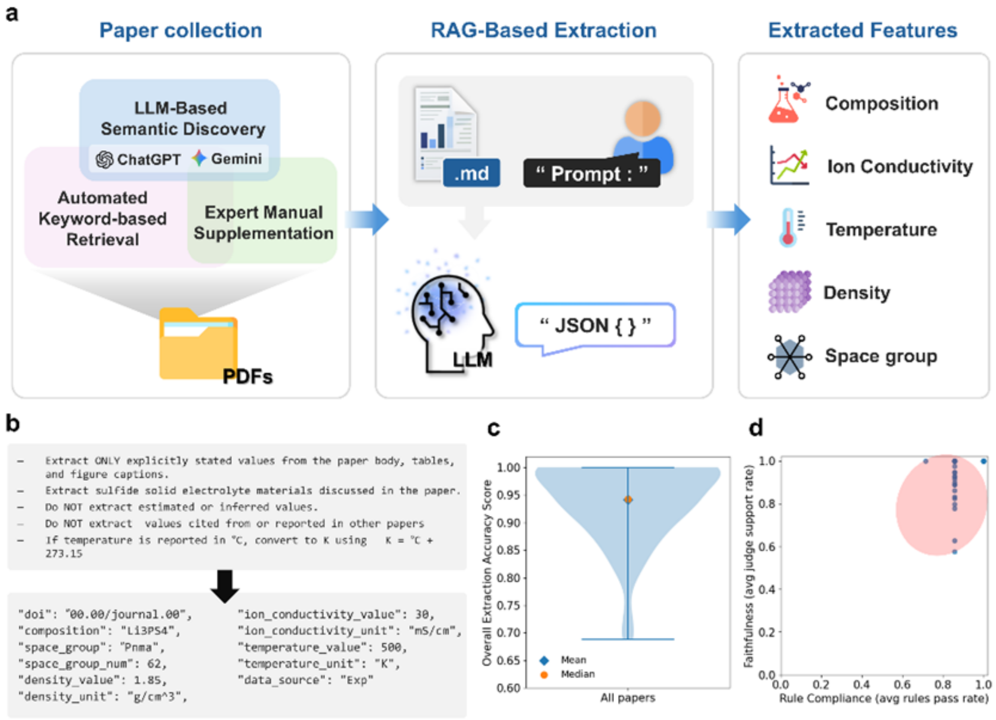

1. RAG-Based Literature Extraction & Normalization
Text
Tables
Figures
Supple-mentary data
Text Processing data
Text Processing modules
RAG
2. Knowledge Graph Construction & Candidate Generation
CSV
Material
Performance
Element
Candidate Generation
Candidates (Solid-State Electrolyte)
Candidate 1
Candidate 2
Candidate 3
...
3. MLIP-Based Simulation & Quantitative Screening
MLIP Simulation
Candidate 1 (predicted performance)
Candidate 2 (predicted performance)
Candidate 3 (predicted performance)
...
1000 / T (1/K)

		Normalized data				
DOI	Composition	Space Group	Number	Density (g/cm3)	Temperature (K)	Ionic Conductivity (mS/cm)
10.103556/c4ta0521e	$Li_{10.35}Ge_{1.35}P_{1.65}S_{12}$	P42/nmc	137	2.098	300	14.2
10.1031/acs.chemmater.0c02418	Li_6PS_5Cl	F-43m	216	1.868	298.15	5.1
10.1021/acs.chem.1728 0c02418	$Li_{10}GeP_2S_{12}$	P42/nmc	137	2.036	298.15	12
10.1021/acsami.6b00833 P—1	$Li_7P_3S_{11}$	P-1	2	1.873	300	57

Figure 1: Schematic of an integrated RAG-KG-MLIP framework for solid-state electrolyte discovery.

2 RAG-based Data Extraction

As shown in Fig. 2a, the overall workflow begins with a literature acquisition stage that integrates multiple search strategies and then expands into RAG-based automated information extraction and quantitative evaluation. We collected more than 100 papers on sulfide solid-state electrolytes via a multi-path acquisition protocol and selected 56 papers that satisfy pre-defined quality criteria as the final analysis set. Papers not directly relevant to sulfide solid electrolytes or lacking sufficient quantitative data were excluded. The same curated corpus was used consistently across all subsequent automated extraction and evaluation steps. Additional details of the acquisition strategy are provided in the Appendix. For knowledge graph construction and downstream new-material prediction, we define the target feature set to capture properties and structural information most directly linked to ionic conductivity in sulfide electrolytes. Concretely, we use LLM-guided feature selection step using ChatGPT-5.2 and Gemini-3, and finalize the extracted fields such as: DOI, composition, space group, space group number, density, ionic conductivity, temperature, and data source (calculation vs. experiment).



```
–   Extract ONLY explicitly stated values from the paper body, tables,
    and figure captions.
–   Extract sulfide solid electrolyte materials discussed in the paper.
–   Do NOT extract estimated or inferred values.
–   Do NOT extract  values cited from or reported in other papers
–   If temperature is reported in °C, convert to K using   K = °C +
    273.15
```

```
"doi": "00.00/journal.00",            "ion_conductivity_value": 30,
"composition": "Li3PS4",              "ion_conductivity_unit": "mS/cm",
"space_group": "Pnma",                "temperature_value": 500,
"space_group_num": 62,                "temperature_unit": "K",
"density_value": 1.85,                "data_source": "Exp"
"density_unit": "g/cm^3",
```

Figure 2: (a) LLM-RAG–based data collection pipeline for sulfide solid electrolytes. (b) Examples of the document-grounded extraction setup and structured outputs. (c) Distribution of extraction accuracy across all analyzed papers. (d) Relationship between rule compliance and faithfulness for individual documents.

Feature extraction from each paper is performed using a document-grounded extraction RAG setup, where a single document is mapped to a structured JSON record. Specifically, we convert each PDF into markdown text and provide it to the model together with a pre-defined prompt. The model is instructed to generate structured outputs only from information explicitly stated in the provided text, enforcing document-grounding and reducing unsupported inference. We implement this pipeline using LangChain [9] and Gemini-2.5-Flash and the average processing cost is approximately 24.3 KRW per PDF, and the average processing time is approximately 1 minute per PDF. Using this pipeline, we extract a total of 146 independent data records from the curated literature set. Fig. 2b presents the prompt template used in our extraction pipeline and a corresponding example of the structured output.

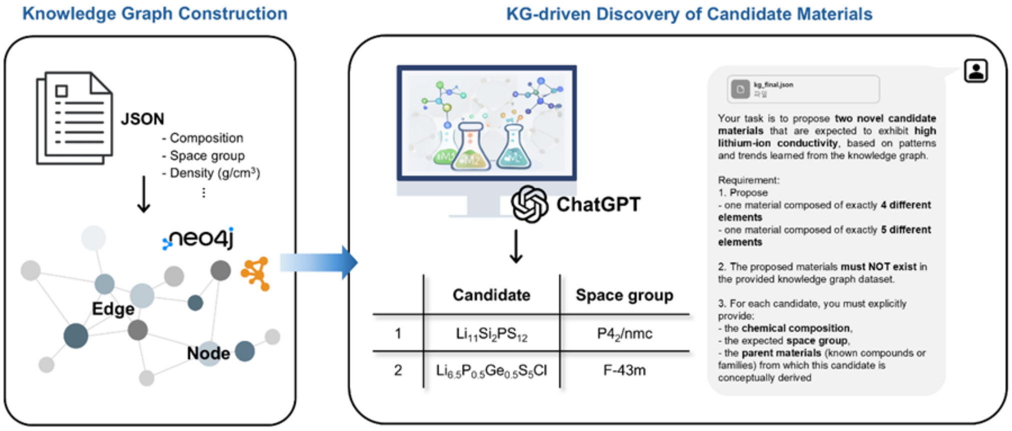
To assess the quality of the extracted records, we apply (i) rule-based validation that checks whether each output satisfies the required structural constraints, and (ii) an LLM-as-judge protocol to evaluate faithfulness, i.e., whether each extracted value is supported by evidence in the source document. This evaluation procedure is designed following a recently proposed reliability evaluation framework for RAG systems [10]. Fig. 2c summarizes the distribution of extraction accuracy computed over the full set of papers. We compute the accuracy as follows.

$$A_{\text{overall}} = 0.4\,A_{\text{rule}} + 0.6\,A_{\text{faith}}$$

The mean and median of the overall extraction accuracy are 0.942 and 0.943, respectively, indicating a highly concentrated score distribution in the high-accuracy regime. This suggests that the extraction performance exhibits neither strong systematic skews nor a large number of catastrophic failure cases. Fig. 2d shows a scatter plot of rule compliance versus faithfulness at the paper level. The average rule compliance and faithfulness are 0.918 and 0.957, respectively, and all analyzed papers lie in the upper-right region. This indicates that the outputs not only satisfy the structural constraints, but that the extracted numerical values are also explicitly supported by the source text.

After extraction, we apply an additional preprocessing filter to improve downstream reliability by removing records that do not explicitly report both ionic conductivity and density. This criterion is designed to minimize error propagation caused by incomplete inputs in subsequent stages, including knowledge graph construction, generative structure proposal, and MLIP-based physical validation. From 146 initially extracted records, after filtering, we obtained a final dataset of 46 sulfide solid-state electrolyte property records with clear provenance and improved numerical reliability.

3 Knowledge Graph Construction and KG-guided Materials Proposal



	Candidate	Space group
1	$Li_{11}Si_2PS_{12}$	$P4_2/nmc$
2	$Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl$	F-43m

Your task is to propose **two novel candidate materials** that are expected to exhibit **high lithium-ion conductivity**, based on patterns and trends learned from the knowledge graph.

Requirement:
1. Propose
- one material composed of exactly **4 different elements**
- one material composed of exactly **5 different elements**

2. The proposed materials **must NOT exist** in the provided knowledge graph dataset.

3. For each candidate, you must explicitly provide:
- the **chemical composition**,
- the expected **space group**,
- the **parent materials** (known compounds or families) from which this candidate is conceptually derived

Figure 3: Construct a Neo4j knowledge graph from RAG-extracted JSON records, then use the KG as context to propose novel electrolyte compositions with space-group hypotheses, excluding compositions already present in the KG.

In this section, we briefly describe how we use the constructed knowledge graph to propose new sulfide solid-state electrolyte candidates. As illustrated in **Fig. 3**, converting literature-extracted information into a KG provides a structured context that captures relationships among composition, crystallographic descriptors, properties, and measurement conditions, which in turn helps define a principled search space for candidate exploration. We leverage this KG context to guide an LLM to propose novel compositions that are not present in the curated dataset, and to provide a plausible space-group hypothesis for each candidate together with a brief rationale grounded in related compositional families observed in the KG. We emphasize that this section focuses on candidate proposal (composition and structural hypothesis), while the concrete procedure for crystal structure generation and subsequent physical validation is described in the following sections.

3.1 Element Frequency Analysis

We analyze the sulfide solid-state electrolyte dataset extracted via our RAG pipeline. We first compute the frequency of constituent elements and observe that Li, P, and S dominate the corpus (Fig. 4a). This reflects the fact that the most fundamental compositional family in sulfide electrolytes is the Li_2S–P_2S_5 (LPS) system, and that a substantial fraction of prior work has been developed around LPS-derived formulations [11, 12].

Among the next most frequent elements, Ge appears with relatively high prevalence, which we attribute to the influence of studies on $Li_{10}GeP_2S_{12}$ (LGPS)-type electrolytes [13, 14]. In addition, among halogens, Cl and I show meaningful occurrence, consistent with the prominence of argyrodite-structured sulfide electrolytes in the [15].

This elemental distribution indicates that prior studies on sulfide solid-state electrolytes have predominantly focused on a limited compositional domain. The elements identified are mainly drawn from the p-block, whereas transition metals are only rarely observed. This trend aligns with prior observations that transition metals can introduce electronic conduction pathways or trigger undesirable redox activity, thereby limiting their suitability as constituent elements in solid electrolytes [16]. Overall, our element-frequency analysis not only quantitatively confirms that sulfide solid-state electrolyte research largely progressed within a relatively constrained compositional subspace centered on LPS-derived chemistries.

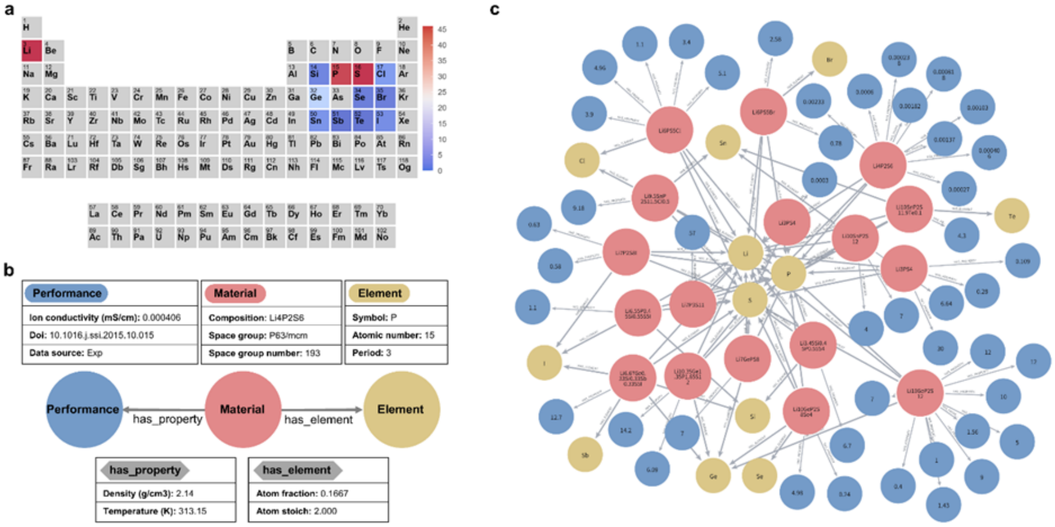


Figure 4: (a) Element frequency distribution of the constituent elements in the sulfide solid-state electrolyte dataset. (b) Knowledge graph schema, including definitions of node types and edge relations. (c) Overall structure of the knowledge graph constructed from the curated sulfide solid-state electrolyte dataset.

3.2 Knowledge Graph Construction

While element frequency analysis is useful for characterizing how individual elements are used, it is limited in capturing relationships across elements, compositions, and properties. To visualize and represent the complex interactions embedded in the sulfide solid-state electrolyte dataset in an inference-friendly form, we construct a knowledge graph (Fig. 4b, c). Methodologically, our knowledge graph differs from many prior studies that predominantly focus on bibliographic linking or simple element-material associations. By introducing sample-level material nodes, explicit measurement-condition nodes, and distinct performance nodes, our graph structurally decouples intrinsic compositional and structural effects from extrinsic measurement factors. This topology transcends flat correlation networks, enabling subgraph-level reasoning and message-passing inference along composition-structure-condition-property pathways. Consequently, it unveils latent design patterns that remain obscured in frequency analyses or simple pairwise statistics.

In the resulting graph, the element nodes Li, P, and S act as central hubs with the highest degrees, and most material nodes form a radially connected structure around them. This observation is consistent with the element-frequency analysis and suggests that performance improvements in sulfide solid electrolytes have largely been pursued through compositional tuning and structural variations within a constrained backbone chemistry, rather than by exploring fundamentally different elemental families.

The Ge node exhibits relatively high connectivity, and we also observe limited connections for Si and Sn, which are group-14 congeners of Ge. This pattern reflects emerging attempts to substitute within the same group to reduce cost and/or improve stability [17]. However, the substantially lower connectivity of Si and Sn compared to Ge suggests that such substitution strategies remain only partially explored in the current literature. For the halogens Cl, Br, and I, we find that these nodes connect selectively to material nodes belonging to specific structural families, such as argyrodites.

Among them, Cl shows the highest connectivity, which we attribute to its ionic-radius compatibility with S, enabling effective substitution on S sites with minimal lattice distortion [18].

Additional elements such as Sb, Se, and Te are also observed in the dataset; however, their corresponding nodes do not exhibit strong connectivity to the major compositional families. This suggests that, although these elements have been discussed as potential levers for performance improvement, they have not yet emerged as broadly adopted design variables in sulfide solid electrolytes.

Overall, the constructed knowledge graph provides an integrated representation that links compositional information with property signals such as ionic conductivity. By reassembling fragmented literature knowledge into a connected relational form, the KG offers a machine-interpretable substrate for capturing and reasoning over element-composition-property relationships in sulfide solid-state electrolytes.

3.3 Knowledge Graph-guided Candidate Proposal

Knowledge graphs serve as an effective tool for visualizing compositional biases reported in sulfide-based solid electrolyte research as well as the relational structures among materials, crystal structures, and ionic transport properties. However, because such representations are inherently limited to existing data, they face fundamental constraints in directly discovering novel compositional combinations that remain insufficiently explored. In contrast, LLMs possess the capability to perform combinatorial exploration and generation based on extensive literature-derived knowledge and relational reasoning.

In this study, we propose a framework that leverages a knowledge graph as structured input to an LLM in order to systematically identify new candidate sulfide solid electrolytes located in relatively unexplored regions of compositional space while remaining consistent with existing research trends. Specifically, relationships among material compositions, crystal structures, and ionic conductivities encoded in the knowledge graph were provided to the LLM, enabling the prioritized generation of new compositions that have not been reported in existing databases but are expected to exhibit high lithium-ion conductivity.

As a result, the LLM proposed two novel candidate compositions consisting of four and five elements, respectively. The first candidate, $\mathrm{Li_{11}Si_2PS_{12}}$, adopts an LGPS-type structure with a tetragonal $P4_2/nmc$ space group. Knowledge graph analysis revealed that LGPS-type structures constitute a representative structural family that repeatedly exhibits high ionic conductivity across the literature, and that LGPS compositions containing Si or Ge tend to show statistically superior transport performance. Moreover, multiple Li-Si-P-S compositions have already been reported within the same structural family. Based on these compositional distributions and property trends, the LLM derived the present candidate as an extension of the Si-based LGPS compositional space.

The second candidate, $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$, crystallizes in the cubic $F\bar{4}3m$ argyrodite structure. The knowledge graph indicates that numerous halide-containing argyrodite materials, including $\mathrm{Li_6PS_5Cl}$, as well as compositions featuring partial substitution of P by Ge, Si, or Sb, have been reported, with many of these exhibiting relatively high room-temperature ionic conductivity. After learning these composition-structure-property correlations, the LLM proposed a new candidate composition that preserves the Cl-based argyrodite framework while partially substituting Ge for P and adjusting the lithium content, thereby extending existing research trends into a previously unreported compositional regime.

4 Structural Construction and Ionic Transport Properties of KG-LLM Proposed Materials

4.1 Structural Construction of Candidate Materials

Based on the compositions and space groups proposed by the LLM, we constructed crystal structure models for each candidate material. In this process, the crystal structure files (POSCAR format) of the corresponding parent materials, which explicitly contain lattice parameters and atomic coordinates, were obtained from the Materials Project database and provided as input structures. The LLM then performed atomic substitutions and compositional adjustments directly on these frameworks to generate the structures of the proposed candidates. Specifically, experimentally reported sulfide solid electrolytes with well-established structural stability were adopted as reference frameworks, and their

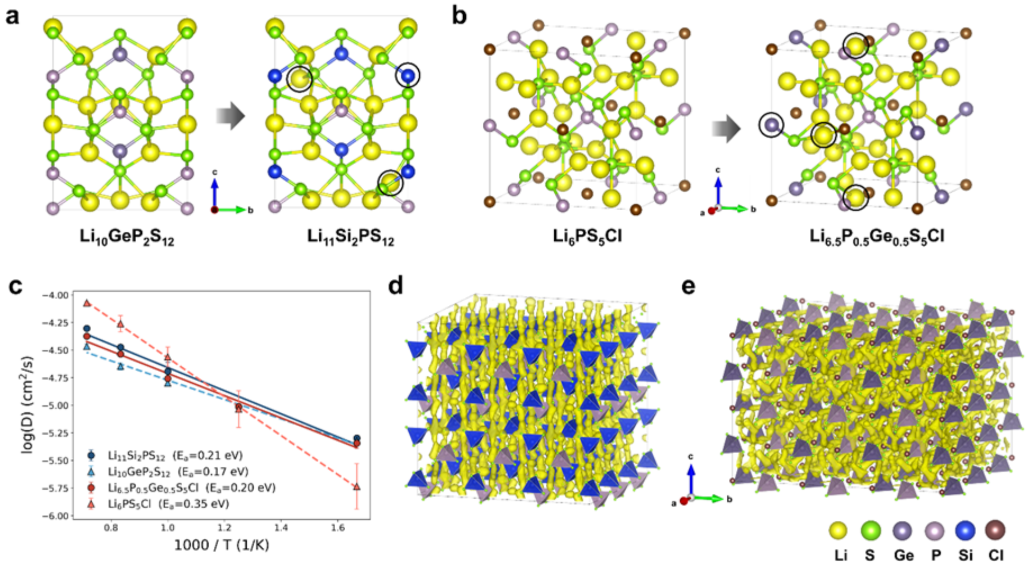


Figure 5: (a) Structural modification from $\mathrm{Li_{10}GeP_2S_{12}}$ to $\mathrm{Li_{11}Si_2PS_{12}}$ (LGPS-type). (b) Structural modification from $\mathrm{Li_6PS_5Cl}$ to $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ (argyrodite-type). (c) Arrhenius plots of ionic conductivity from MLIP-based simulations. (d) Li-ion diffusion pathways in $\mathrm{Li_{11}Si_2PS_{12}}$ at 300 K. (e) Li-ion diffusion pathways in $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ at 300 K.

structures were automatically modified by the LLM to satisfy the target compositions, resulting in the generation of POSCAR files for the candidate materials.

For the four-element candidate $\mathrm{Li_{11}Si_2PS_{12}}$, the POSCAR file of the LGPS-type structure [19] with the $P4_2/nmc$ space group was used as the parent framework. The LLM replaced the Ge atoms occupying the tetrahedral sites with Si atoms while preserving the lattice symmetry and the fundamental structural skeleton, and adjusted the Li occupancy to match the target stoichiometry, thereby generating a crystal structure model corresponding to the new composition (Fig. 5a).

For the five-element candidate $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$, the POSCAR file of the argyrodite-type $\mathrm{Li_6PS_5Cl}$ structure [20] with the $F\bar{4}3m$ space group was adopted as the reference. In this case, a partial substitution of Ge for P at the tetrahedral sites was introduced, and the Li occupancy was re-adjusted accordingly to construct a structure consistent with the proposed composition using the LLM (Fig. 5b). Throughout this procedure, the characteristic anion framework of the argyrodite structure and the connectivity of the Li sublattice were preserved.

4.2 MLIP-Based Analysis of Ionic Transport Properties

The optimized structures of the two candidate materials obtained from density functional theory (DFT) calculations were expanded into $3 \times 3 \times 2$ supercells and subjected to molecular dynamics simulations using two machine-learning interatomic potential models: Graph Representation and Atomic Simulation Engine (GRACE[21]) and Materials-Aware Atomic Cluster Expansion (MACE[22]). Under identical simulation conditions, using an NVIDIA RTX 4090 GPU, the computational cost for 100 ps of simulation was approximately 1.93 h for MACE(MPA-0) and 1.17 h for GRACE(2L-OAM-L), indicating that GRACE provides higher computational efficiency. For each material, ionic diffusion behavior was evaluated over a temperature range of 600–1400 K, and the room-temperature ionic conductivity was estimated through Arrhenius analysis. To benchmark the performance of the candidate materials, identical simulations were also performed for the parent compounds, LGPS and $\mathrm{Li_6PS_5Cl}$, under the same conditions. The mean squared displacement (MSD) curves of lithium ions obtained from both MLIP models for all materials are provided in the Appendix (Fig. 8), where stable diffusion behavior and consistent trends between the two models are observed across the entire temperature range. Fig. 5c presents the Arrhenius plots for both the candidate and parent materials. The reported values correspond to averages over the two MLIP models, with the standard deviations shown as error bars to represent model-to-model variability. All materials exhibit clear Arrhenius-type temperature dependence, and similar conductivity trends are reproduced by both MLIP models, confirming the consistency of the predictions.

Quantitatively, the four-element candidate $\mathrm{Li_{11}Si_2PS_{12}}$ exhibits an activation energy of 0.21 eV and a room-temperature ionic conductivity of 12.7 mS cm^{-1}, which is slightly higher in activation energy than its parent material $\mathrm{Li_{10}GeP_2S_{12}}$ (0.17 eV, 19.0 mS cm^{-1}), yet still maintains a high conductivity exceeding 10 mS cm^{-1}. In contrast, the five-element candidate $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ shows an activation energy of 0.20 eV and a room-temperature ionic conductivity of 14.0 mS cm^{-1}, representing a substantial improvement over the parent compound $\mathrm{Li_6PS_5Cl}$ (0.35 eV, 0.34 mS cm^{-1}), with a pronounced reduction in activation energy and an enhancement in conductivity by nearly two orders of magnitude. These results indicate that, particularly for the argyrodite family, partial cation substitution combined with Li stoichiometry tuning can effectively lower migration barriers and significantly enhance room-temperature ionic conductivity. Furthermore, the LGPS-type candidate demonstrates that Ge-free compositions can retain conductivity levels comparable to those of conventional LGPS, suggesting a promising alternative from both performance and resource-efficiency perspectives. In particular, replacing Ge with Si avoids the use of a relatively scarce and costly element, which is expected to reduce raw material costs and supply-chain constraints in large-scale electrolyte production. Overall, these findings quantitatively demonstrate that the integration of knowledge-graph-based compositional trend analysis with LLM-driven combinatorial design can successfully identify new sulfide solid electrolyte candidates exhibiting high ionic conductivity.

To elucidate the room-temperature ionic transport mechanisms of the candidate materials, additional MLMD simulations were performed for each candidate supercell at 300 K for 2 ns, and the spatial occupation and migration pathways of lithium ions were analyzed. As a result, long-range interconnected lithium-ion conduction networks were observed in both candidate materials, and representative visualizations of the diffusion pathways are shown in Fig. 5d and 5e. For the LGPS-type structure $\mathrm{Li_{11}Si_2PS_{12}}$, lithium ions preferentially migrate along well-defined channels oriented primarily along the crystallographic c-direction, followed by inter-channel hopping that enables continuous long-range diffusion (Fig. 5d and Fig. 9a). This behavior indicates that the channel-based diffusion mechanism commonly reported for LGPS-type electrolytes is preserved in the newly proposed composition. In contrast, for the argyrodite-type structure $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$, lithium ions form more isotropic three-dimensional interconnected pathways throughout the structure, consistent with a typical cage-to-cage hopping mechanism (Fig. 5e and Fig. 9b). Notably, in both candidates, the lithium occupation density forms continuous networks without fragmentation, demonstrating that meaningful long-range diffusion can occur even at room temperature. These pathway-based observations are qualitatively consistent with the low activation energies and high room-temperature ionic conductivities obtained from the Arrhenius analysis.

5 Conclusion

5.1 Conclusion

In this study, we proposed an integrated computational materials discovery framework that combines literature-based information extraction via retrieval-augmented generation (RAG), knowledge graphs (KG), large language models (LLMs), and machine-learning interatomic potentials (MLIPs) to systematically exploit composition–structure–property relationships in sulfide solid electrolytes. Unlike conventional computational screening approaches that rely primarily on high-cost density functional theory (DFT) calculations or random compositional sampling, the present strategy enables efficient identification of physically plausible and explainable candidate materials by structuring accumulated literature knowledge and incorporating it into generative models.

Specifically, key information such as chemical compositions, space groups, densities, measurement temperatures, and ionic conductivities was automatically extracted from the literature using an RAG-based pipeline and organized into a knowledge graph, enabling quantitative analysis of correlations among composition, structure, and performance. By providing this structured KG information as input to the LLM, two promising sulfide solid electrolyte compositions—$\mathrm{Li_{11}Si_2PS_{12}}$ (LGPS-type) and $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ (argyrodite-type)—were generated as candidates with high expected lithium-ion conductivity.

Although $\mathrm{Li_{11}Si_2PS_{12}}$ has been reported in previous experimental studies [23, 24], it was not included in the literature corpus used to construct our knowledge graph and was therefore treated as an out-of-graph composition by the model. This result indicates that the framework can systematically recover high-performance compositions that are consistent with learned composition–structure–property trends, even when such materials are absent from the training knowledge graph, rather than relying on direct memorization. In contrast, to the best of our knowledge, the argyrodite-type candidate $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ has not been previously reported, demonstrating that the proposed approach can also generate genuinely unexplored compositions within known structural classes.

The LLM-generated structural models were subsequently optimized using DFT and quantitatively validated through molecular dynamics simulations based on two independent MLIP models, GRACE and MACE. The results demonstrate that both candidates exhibit high room-temperature ionic conductivities of 12.7 mS $\mathrm{cm^{-1}}$ and 14.0 mS $\mathrm{cm^{-1}}$, respectively. Notably, the argyrodite candidate shows an improvement of nearly two orders of magnitude compared to its parent compound, while the LGPS-type candidate maintains conductivity comparable to conventional LGPS despite being free of Ge, highlighting its potential advantage in terms of resource efficiency.

Furthermore, long-timescale MLMD simulations at 300 K revealed that the channel-based diffusion mechanism along the crystallographic c-axis is preserved in the LGPS-type candidate, whereas the argyrodite candidate retains a three-dimensional cage-to-cage diffusion network. These diffusion characteristics are consistent with the low activation energies and high ionic conductivities obtained from Arrhenius analysis, providing a coherent physical interpretation of the observed transport behavior.

Overall, this work demonstrates that integrating structured literature knowledge (KG), generative reasoning (LLM), and large-scale property validation (MLIP) enables an explainable, reproducible, and computationally efficient pipeline for discovering high-performance solid electrolytes. The proposed framework is not limited to sulfide systems and can be readily extended to oxide, halide, and other functional materials, offering a general strategy to accelerate next-generation materials discovery when combined with targeted model refinement and experimental validation.

5.2 Limitation

Despite the advantages of the proposed framework, several limitations should be acknowledged. First, parts of the literature acquisition process still require manual intervention, as downloading and organizing full-text PDFs cannot be fully automated due to heterogeneous publisher access policies. Second, the constructed dataset and knowledge graph are inherently shaped by existing literature, resulting in denser coverage of well-studied material systems and literature-induced bias. This bias can limit the discovery of structurally or chemically underrepresented materials, including metastable or unconventional compositions. The dataset and knowledge graph are also not directly reusable across material domains, requiring reconstruction and prompt redesign for each new material class.

Methodologically, candidate discovery is constrained to substitutional exploration around known prototype structures. LLM-based structure generation relies on provided prototype POSCAR files and produces modified structures through compositional substitution or limited structural variation. While the current dataset size is sufficient for substitution-based exploration, it is not large or diverse enough to enable reliable prediction of entirely novel crystal structures beyond known motifs. Future work will address this limitation by expanding the dataset with experimental and computational structures, enabling exploration beyond local prototype substitutions. Moreover, the framework assumes that the extracted literature information is accurate, and any inaccuracies may propagate through the knowledge graph and affect downstream screening.

Finally, although MLIP-based molecular dynamics enable efficient large-scale screening, the use of pre-trained MACE and GRACE models introduces domain bias and uncertainty, particularly when extrapolating to unexplored compositions. Moreover, the simulations are limited to ideal crystalline structures and neglect microstructural effects, thermodynamic stability, and experimental synthesizability, collectively leading to a realism gap between predicted and experimentally observable behavior. Future work will focus on reducing this gap through targeted domain adaptation using system-specific training data and by incorporating stability- and disorder-aware modeling strategies, thereby improving transferability and practical relevance.

References

[1] Arumugam Manthiram, Xingwen Yu, and Shaofei Wang. Lithium battery chemistries enabled by solid-state electrolytes. *Nature Reviews Materials*, 2(4), 2017.

[2] Jürgen Janek and Wolfgang G. Zeier. A solid future for battery development. *Nature Energy*, 1(9), 2016.

[3] Anubhav Jain et al. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 2013.

[4] Scott Kirklin et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1, 2015.

[5] Stefano Curtarolo et al. The high-throughput highway to computational materials design. *Nature Materials*, 12(3):191–201, 2013.

[6] Theodosios Famprikis et al. Fundamentals of inorganic solid-state electrolytes for batteries. *Nature Materials*, 18(12):1278–1291, 2019.

[7] Vahe Tshitoyan et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98, 2019.

[8] Lauri Himanen et al. Data-driven materials science: Status, challenges, and perspectives. *Advanced Science*, 6(21), 2019.

[9] Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing LangChain. In *International Conference on Applied Engineering and Natural Sciences*, 2023.

[10] Yunfan Gao et al. Retrieval-augmented generation for large language models: A survey, 2023. arXiv:2312.10997.

[11] Fuminori Mizuno, Akitoshi Hayashi, Kiyoharu Tadanaga, and Masahiro Tatsumisago. High lithium ion conducting glass-ceramics in the system Li_2S–P_2S_5. *Solid State Ionics*, 177:2721–2725, 2006.

[12] Riku Maniwa, Marcela Calpa, Nataly Carolina Rosero-Navarro, Akira Miura, and Kiyoharu Tadanaga. Synthesis of sulfide solid electrolytes from Li_2S and P_2S_5 in anisole. *Journal of Materials Chemistry A*, 9(1):400–405, 2021.

[13] N. Kamaya, K. Homma, Y. Yamakawa, et al. A lithium superionic conductor. *Nature Materials*, 10:682–686, 2011.

[14] Chang Xu, Liquan Chen, and Fan Wu. Unveiling the power of sulfide solid electrolytes for next-generation all-solid-state lithium batteries. *Next Materials*, 6, 2025.

[15] Chuang Yu, Feipeng Zhao, Jing Luo, Long Zhang, and Xueliang Sun. Recent development of lithium argyrodite solid-state electrolytes for solid-state batteries. *Nano Energy*, 83, 2021.

[16] T. K. Schwietert, V. A. Arszelewska, C. Wang, et al. Clarifying the relationship between redox activity and electrochemical stability in solid electrolytes. *Nature Materials*, 19:428–435, 2020.

[17] Shyue Ping Ong, Yifei Mo, William Davidson Richards, Lincoln Miara, Yo-Sug Lee, and Gerbrand Ceder. Phase stability, electrochemical stability and ionic conductivity of the $Li_{10\pm1}MP_2X_{12}$ family of superionic conductors. *Energy & Environmental Science*, 6:148–156, 2013.

[18] Zhen Wang, Luyao Ding, Shenglong Yu, Huan Xu, Xiaohui Hao, Yi Sun, and Tianmin He. Effect of two different ZnO addition strategies on the sinterability and conductivity of $BaZr_{0.4}Ce_{0.4}Y_{0.2}O_{3-\delta}$. *Chemistry of Materials*, 31(21):8673–8678, 2019.

[19] Boran Tao et al. Thio-/LISICON and LGPS-type solid electrolytes for all-solid-state lithium-ion batteries. *Advanced Functional Materials*, 32(34), 2022.

[20] Chuang Yu et al. Synthesis, structure and electrochemical performance of the argyrodite Li_6PS_5Cl solid electrolyte. *Electrochimica Acta*, 215:93–99, 2016.

[21] Anton Bochkarev, Yury Lysogorskiy, and Ralf Drautz. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing. *Physical Review X*, 14(2):021036, 2024.

[22] Ilyes Batatia et al. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

[23] Alexander Kuhn et al. A new ultrafast superionic Li-conductor: Ion dynamics in $Li_{11}Si_2PS_{12}$. *Physical Chemistry Chemical Physics*, 16(28):14669–14674, 2014.

[24] Alexander Kuhn et al. Ultrafast Li electrolytes based on abundant elements: $Li_{10}SnP_2S_{12}$ and $Li_{11}Si_2PS_{12}$, 2014. arXiv:1402.4586.

Appendix / supplemental material

A LLM-Based Problem Formulation and Research Question Derivation

A.1 Multi-LLM Workflow: Setup, Integration, and Cross-Checking

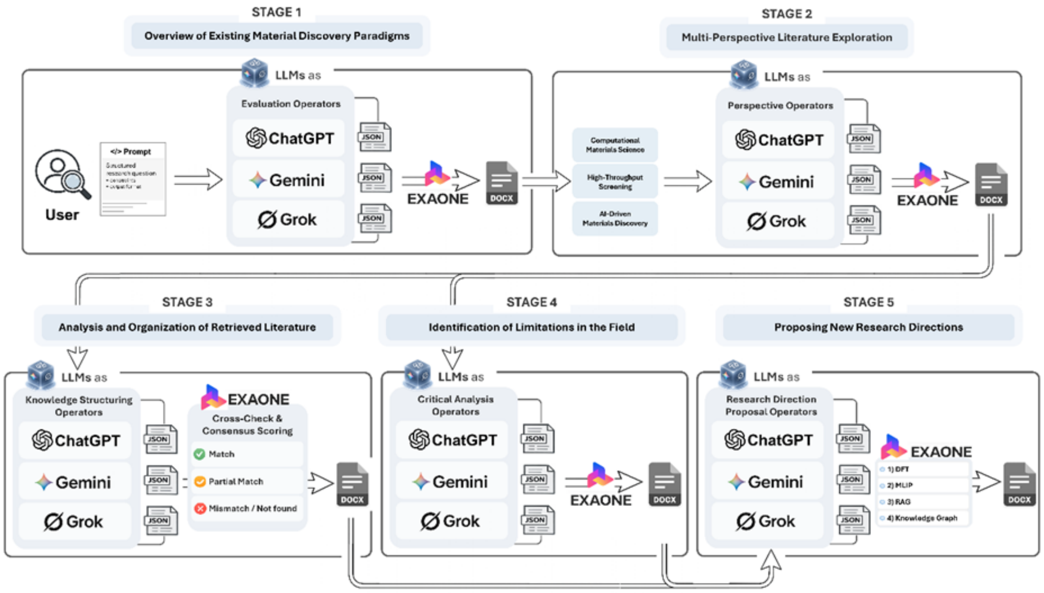

STAGE 1
Overview of Existing Material Discovery Paradigms
LLMs as
Evaluation Operators
ChatGPT
Gemini
Grok
JSON
User
</> Prompt
EXAONE
DOCX
STAGE 2
Multi-Perspective Literature Exploration
LLMs as
Perspective Operators
Computational Materials Science
High-Throughput Screening
AI-Driven Materials Discovery
STAGE 3
Analysis and Organization of Retrieved Literature
LLMs as
Knowledge Structuring Operators
EXAONE
Cross-Check & Consensus Scoring
Match
Partial Match
Mismatch / Not found
STAGE 4
Identification of Limitations in the Field
LLMs as
Critical Analysis Operators
STAGE 5
Proposing New Research Directions
LLMs as
Research Direction Proposal Operators
1) DFT
2) MLIP
3) RAG
4) Knowledge Graph

Figure 6: Multi-LLM workflow for Stages 1–3 (topic selection to literature analysis). ChatGPT, Gemini, and Grok generate independent, structured JSON outputs; EXAONE integrates them into a single Korean DOCX deliverable while enabling cross-checking and standardization.

In this appendix, we describe the detailed setup and integration procedure of the multi-LLM workflow used in Stages 1–3 (**Fig. 6**). In §A.1.1, we specify the prompting strategy and the structured output (JSON) schema. In §A.1.2, we describe the stage-wise execution procedure and the EXAONE-based integration process. In §A.1.3, we present the cross-check protocol and consensus labeling rules used in Stage 3.

A.1.1 Prompting and Structured Output Schema

Across all stages, we use role-specific instructions (e.g., evaluation, perspective exploration, and knowledge structuring). To ensure comparability across models, we constrain outputs to a pre-defined JSON schema. Each JSON record contains: (i) stage and role identifiers, (ii) core claims and supporting rationale, and (iii) provenance fields that indicate which model produced each item. This structured interface enables downstream integration and facilitates auditing when needed.

A.1.2 Stage-wise Procedure and Integration (Stages 1-5)

Stage 1 (Paradigm overview). We run ChatGPT, Gemini, and Grok as evaluation operators under the same prompt conditions. Each model independently summarizes existing materials discovery paradigms and produces a structured JSON output. We then consolidate these outputs using EXAONE to generate a single DOCX artifact that standardizes terminology and merges overlapping points while preserving provenance.

Stage 2 (Multi-perspective literature exploration). Using the domain axes obtained in Stage 1, we perform multi-perspective literature exploration by running the same three LLMs as perspective operators. Each model proposes candidate papers and returns structured entries (e.g., title/venue/year/DOI when available) in JSON. EXAONE merges the union of candidate lists, removes duplicates, and produces a consolidated literature list in DOCX form.

Stage 3 (Analysis and organization of retrieved literature). For each retrieved paper, we request independent analyses from multiple LLMs and enforce JSON outputs. EXAONE performs cross-checking by aligning paper identity (preferably via DOI; otherwise via normalized title matching) and aggregates model-level summaries into a unified record per paper. The final output is a DOCX report that includes both the merged analysis and model agreement signals.

Stage 4 (Identification of limitations in the field). Based on Stage 1–3 outputs, we run ChatGPT, Gemini, and Grok as critical analysis operators to extract recurring bottlenecks and failure modes in the target research area. Each model produces a structured JSON that separates (i) methodological limitations (e.g., assumptions, missing physics, evaluation gaps), (ii) data limitations (e.g., bias, coverage, label noise, reproducibility), and (iii) deployment limitations (e.g., scalability, robustness, integration barriers). EXAONE consolidates these limitation inventories into a single DOCX report, while retaining per-item provenance to enable traceable auditing of where each limitation originated.

Stage 5 (Proposing new research directions). Using the consolidated limitation report (Stage 4) as constraints, we run the three LLMs as research direction proposal operators to generate actionable research directions and testable research questions. Outputs are again serialized in JSON with fields such as: (i) proposed direction, (ii) hypothesis and expected mechanism, (iii) required resources/data, (iv) evaluation protocol and success criteria, and (v) risks/ablation plans. EXAONE integrates proposals into a single DOCX and additionally maps each direction onto an execution pathway (e.g., DFT, MLIP, RAG, and Knowledge Graph) to explicitly connect ideation with implementation modules and verification routes.

A.1.3 Cross-check Protocol and Consensus Labels

To reduce hallucination risk and ensure traceability, we assign a consensus label to each paper-level record. We use four categories—Match, Partial match, Mismatch, and Not found—based on (i) paper identity resolution (DOI/title) and (ii) agreement on the extracted core technical claims. The final DOCX report stores the consensus label together with per-model evidence fields, enabling manual verification when needed.

A.2 Multi-Agent Deliberation Protocol for Pipeline Synthesis

This appendix describes the design-time mechanism we used to synthesize the end-to-end discovery pipeline shown in Fig. 7. The core idea is to treat pipeline design as a structured deliberation problem: multiple role-specialized LLM agents iteratively propose, critique, and revise a candidate workflow until a verifiable and resource-aware specification is produced. We implement the interaction using an AutoGen-style multi-agent framework, with explicit turn-taking and grounding tools to reduce unsupported reasoning.

A.2.1 Agent Committee and Grounding Tools

We instantiate a small committee of LLM agents, each constrained to a distinct responsibility: (i) coordination and decision arbitration, (ii) target-property specification (e.g., ionic conductivity and stability), (iii) unstructured-data processing and representation, (iv) AI-method design, and (v) physics/validation constraints. The committee is equipped with web retrieval, enabling agents to fetch recent references during discussion rather than relying solely on parametric memory. Retrieved evidence is used to justify critiques, resolve disagreements, and update the evolving pipeline specification.

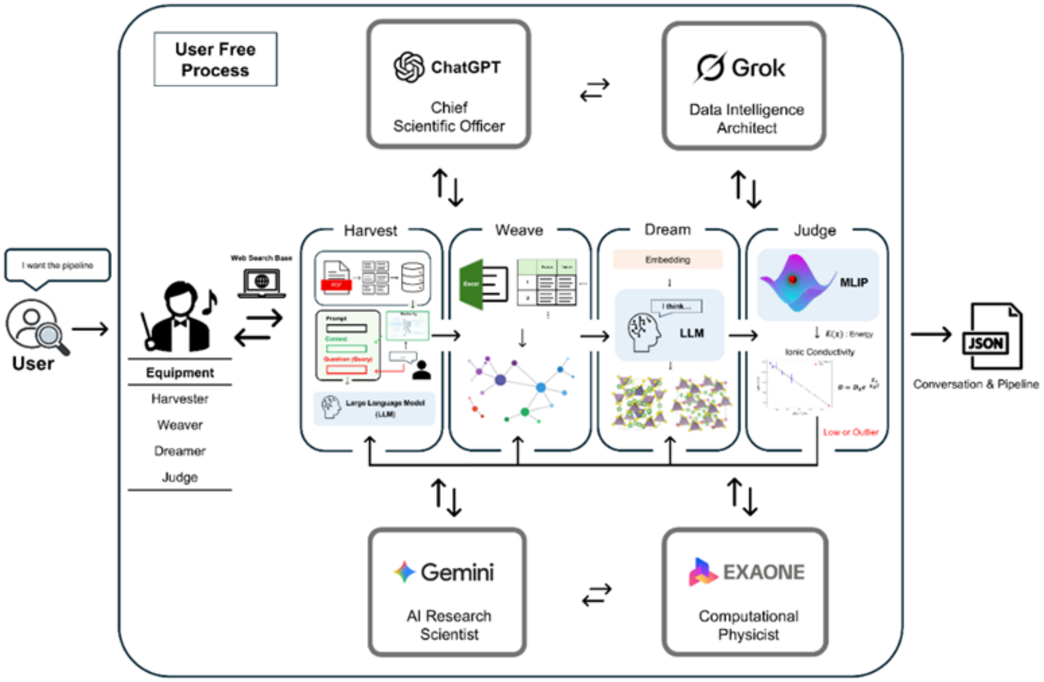


Figure 7: Multi-agent, web-grounded deliberation used to synthesize a closed-loop discovery pipeline (Harvest-Weave-Dream-Judge), with role-specialized agents interacting via round-robin critique and exporting a structured JSON specification.

A.2.2 Round-Robin Adversarial Refinement

The committee follows a round-robin deliberation protocol. Each proposal must be followed by a critique from the next agent, enforcing systematic stress-testing of assumptions (data availability, evaluation validity, physical consistency, and feasibility under compute constraints). We run multiple critique–revision rounds (≥ 5) and terminate when the pipeline description becomes (i) internally consistent, (ii) grounded with supporting evidence where applicable, and (iii) executable within the project's practical constraints. Importantly, the goal of this process is not to maximize creativity, but to converge to a pipeline with explicit checks against hallucinated steps and unverifiable claims.

A.2.3 Output Artifact and Reproducibility

Upon termination, the system exports the deliberation trace and the finalized pipeline as a structured JSON artifact. The JSON includes role-tagged messages, intermediate revisions, and the final stage definitions, enabling auditability of "why" each stage exists and what constraints motivated it. This makes the pipeline derivation reproducible as a controlled protocol, rather than a one-off design decision.

A.2.4 Resulting Four-Stage Loop

The deliberation produces a four-stage cyclic workflow (Harvest–Weave–Dream–Judge) that connects: (1) literature-scale extraction via retrieval and structured parsing, (2) relational consolidation into a knowledge graph, (3) KG-conditioned candidate suggestion, and (4) fast physics-oriented filtering via MLIP-based evaluation. Structural construction details for candidates are intentionally deferred to later sections; here we only document how the pipeline itself is synthesized and serialized.

B Method: Document-grounded Literature Collection and Extraction

We acquired sulfide solid-state electrolyte papers using a multi-route collection strategy that combines three complementary pathways. First, we used an LLM for semantic recommendation to gather relevant paper DOIs. Second, we performed keyword-based automated DOI collection using OpenAlex. Third, we manually supplemented the resulting list to recover papers that may be missed by specific keywords or by the automated retrieval process. All collected papers were standardized to PDF format and organized under a single directory to enable consistent use in the downstream automated extraction pipeline.

We extract information from each paper using a document-grounded extraction RAG setup that maps a single document to a structured JSON record. During extraction, we restrict evidence strictly to numerical values explicitly stated in the paper as primary results. In particular, we exclude values that appear only in narrative background or review text, values cited from other papers, and values obtained via inference, interpolation, or extrapolation. Only values that are directly measured by the authors or explicitly computed are permitted. To ensure format consistency, we normalize temperature to K and ionic conductivity to mS/cm. When a single composition is reported under multiple temperature conditions or with multiple conductivity values, we do not aggregate them (e.g., by averaging); instead, we store them as separate records to preserve condition-specific context.

C Machine Learning Interatomic Potentials

C.1 Simulation Settings

DFT was performed using the Vienna *Ab initio* Simulation Package (VASP). The core–valence interactions were treated using the projector-augmented wave (PAW) method. Spin-polarized DFT calculations employed the generalized gradient approximation (GGA) with the Perdew–Burke–Ernzerhof (PBE) functional. A kinetic energy cutoff of 500 eV was applied, and the Brillouin zone was sampled using a Γ-centered $3 \times 3 \times 3$ k-point grid. The electronic energy was converged to within 10^{-5} eV, and ionic relaxations were performed until the forces on all atoms were below 0.02 eV Å^{-1}.

Following the DFT structural optimization, each primitive unit cell was expanded into a $3 \times 3 \times 2$ supercell to construct large-scale atomic models for subsequent molecular dynamics simulations. As a result, the simulation cell for $\mathrm{Li_{11}Si_2P_2S_{12}}$ contained 936 atoms, while that for $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ comprised 972 atoms.

To evaluate lithium-ion diffusion behavior with high computational efficiency while maintaining reliable accuracy, molecular dynamics simulations based on machine-learning interatomic potentials (MLMD) were performed using two state-of-the-art models, GRACE and MACE, as implemented in the LAMMPS simulation package. For each candidate material, simulations were conducted at five different temperatures ranging from 600 K to 1400 K in increments of 200 K, and MD trajectories of 100 ps were generated at each temperature for Arrhenius analysis. In addition, to investigate lithium-ion transport mechanisms under operating conditions, extended MLMD simulations were carried out at 300 K for 2 ns for each candidate material. All MLMD simulations were performed using a time step of 2 fs under the NVT ensemble. The resulting trajectories were analyzed and visualized using VESTA to examine the structural and dynamic evolution of the systems.

C.2 Transport Property Analysis

The lithium-ion diffusion coefficients D were calculated from MLMD trajectories using the mean squared displacement (MSD) method. The MSD is defined as

$$\mathrm{MSD}(t) = \left\langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle, \tag{1}$$

where $\mathbf{r}_i(t)$ denotes the position of the i-th lithium ion at time t, and $\langle \cdots \rangle$ represents an average of all mobile lithium ions and different time origins.

The diffusion coefficient D was obtained from the slope of the MSD curve according to the Einstein relation,

$$D = \frac{1}{2d}\frac{d}{dt}\mathrm{MSD}(t), \tag{2}$$

where $d = 3$ is the dimensionality of the diffusion space. In practice, D was extracted by linear fitting of the MSD curve within the time window exhibiting diffusive behavior, where the MSD increases linearly with time.

For each material, diffusion coefficients were evaluated at five temperatures ranging from 600 to 1400 K using two independent machine-learning interatomic potential models, GRACE and MACE. Arrhenius plots were constructed by fitting the temperature-dependent diffusion coefficients to

$$D(T) = D_0 \exp\left(-\frac{E_a}{k_{\mathrm{B}}T}\right), \tag{3}$$

where D_0 is the pre-exponential factor, E_a is the activation energy, k_{B} is the Boltzmann constant, and T is the absolute temperature. For each material, the final Arrhenius parameters were obtained by averaging the results from the two MLIP models.

The ionic conductivity σ_T was subsequently estimated using the Nernst–Einstein relation,

$$\sigma_T = \frac{nz^2F^2D}{RT}, \tag{4}$$

where n is the lithium-ion concentration, z is the ionic charge number ($z = 1$ for Li^+), F is the Faraday constant, and R is the ideal gas constant.

The same analysis procedure was applied to both the proposed candidate materials and their corresponding parent compounds to enable direct comparison of transport properties under identical simulation conditions.

C.3 Lithium-Ion Probability Density and Diffusion Pathway Analysis

To visualize lithium-ion migration pathways in the proposed candidate solid electrolytes, lithium-ion probability density distributions were constructed from MLMD trajectories at 300 K for $\mathrm{Li_{11}Si_2PS_{12}}$ and $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$.

During the simulations, the supercell was discretized into a three-dimensional grid, and the number of occurrences of lithium ions within each grid voxel was accumulated over the entire trajectory. The local probability density P was computed as

$$P = \frac{N}{V},$$

where N is the number of lithium-ion occurrences within a voxel and V is the voxel volume. The average probability density over the simulation cell is denoted as P_0.

Three-dimensional isosurfaces of lithium-ion probability density were generated at multiple threshold values ranging from $2P_0$ to $6P_0$ to identify dominant diffusion channels as well as secondary migration pathways. Regions with higher probability density correspond to frequently occupied lithium sites, whereas lower isovalues reveal extended connections between neighboring sites, representing long-range diffusion networks.

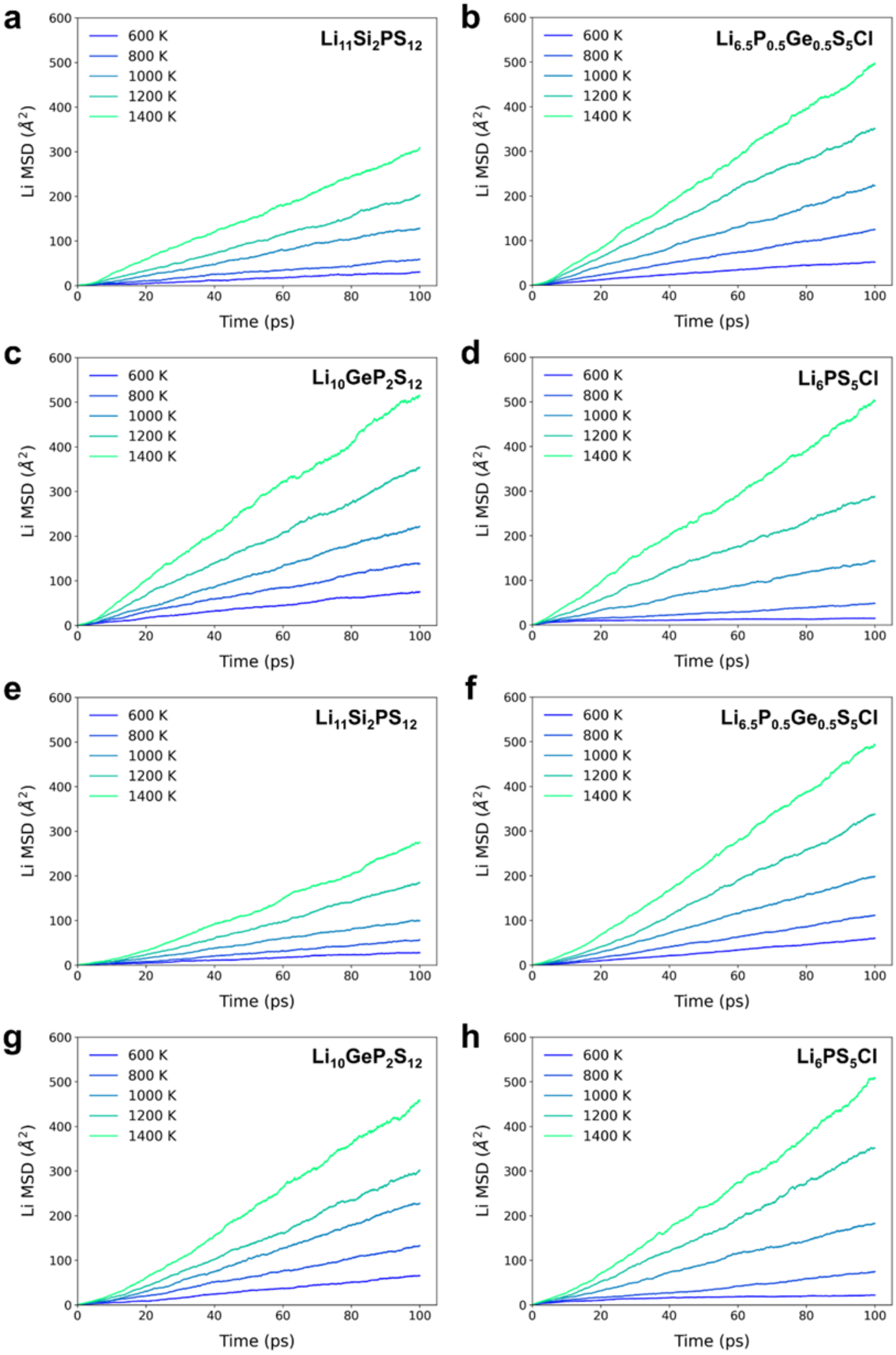


Figure 8: Lithium-ion MSD plots at 600-1400 K calculated using the MACE (a-d) and GRACE (e-h) models for the candidate materials and their parent compounds.

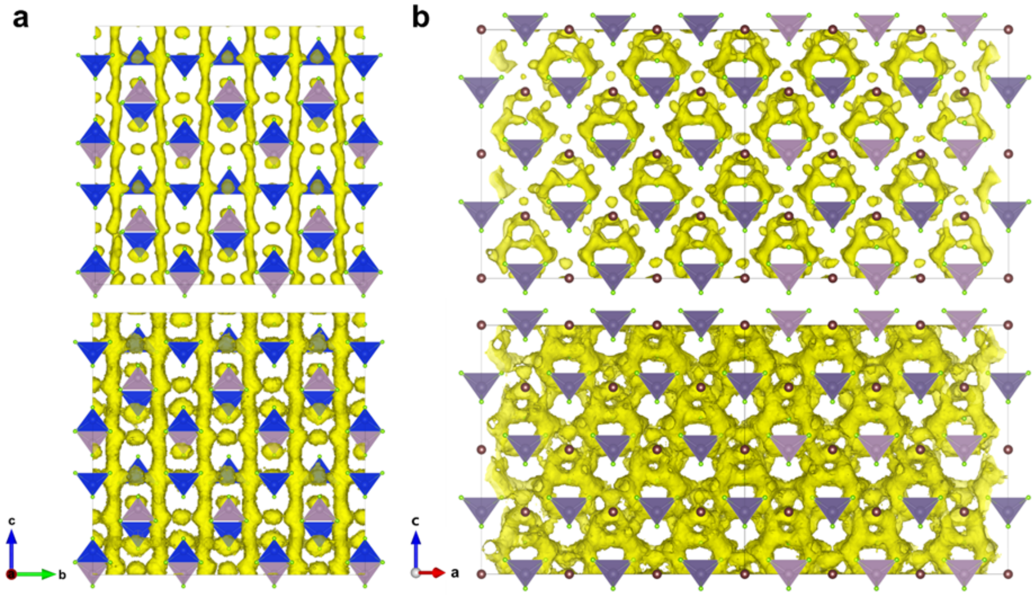


Figure 9: Lithium-ion probability density isosurfaces at 300 K for (a) $\mathrm{Li_{11}Si_2PS_{12}}$ and (b) $\mathrm{Li_{6.5}P_{0.5}Ge_{0.5}S_5Cl}$ obtained from MLMD simulations. The upper panels correspond to an isovalue of $6P_0$, highlighting the most frequently occupied lithium sites, while the lower panels show an isovalue of $2P_0$, revealing extended diffusion pathways. Yellow isosurfaces indicate regions of high lithium occupation probability, and tetrahedral units are shown as polyhedra for structural reference.

AI Co-Scientist Challenge Korea Paper Checklist

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

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

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

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

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

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “AI Co-Scientist Challenge Korea paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims are explicitly stated in the Abstract and Section 1 (Introduction) and are consistent with corresponding sections of the paper: RAG-based literature-to-dataset construction in Section 2, evidence-preserving knowledge graph construction and KG-grounded candidate proposal in Sections 3.2 and 3.3, and MLIP-based screening of ionic transport properties in Section 4.2.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the proposed framework, including literature-induced bias, constraints of substitution-based structure exploration due to dataset scale, assumptions on data accuracy, and the realism gap of pre-trained MLIP models, are explicitly discussed in Section 5.2 (Limitation).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

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

   Answer: [N/A]

   Justification: The paper does not present theoretical results or formal proofs.

   Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: The paper discloses all information required to reproduce the main experimental results that support the paper's claims, particularly the MLIP-based molecular dynamics simulations and ionic transport property evaluations. This includes the knowledge graph schema and candidate generation logic (Sections 3.2 and 3.3), the MLIP models used (MACE and GRACE), molecular dynamics configurations, system sizes, temperature ranges, and analysis protocols (Sections 4.1 and 4.2 and Appendix). While the detailed prompt configurations of the LLM-based literature extraction pipeline are not fully specified, the core simulation-based results and conclusions are reproducible based on the information provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The paper does not provide open access to the full curated literature corpus or the complete end-to-end pipeline code. Therefore, open access to data and code is not provided. This limitation is primarily due to publisher and licensing constraints on the source literature. Nevertheless, the paper provides detailed descriptions of the dataset structure, extracted features, and knowledge graph schema (Section 3.2 and Fig. 4), as well as final candidate structures and simulation settings, enabling methodological reuse and partial reproduction of the reported results.

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [No]

Justification: The paper provides detailed and reproducible descriptions of the molecular dynamics simulation and analysis settings, including the MLIP models (MACE and GRACE), system sizes, temperature ranges, and transport-property evaluation procedures (Section 4.2 and Appendix C). However, the detailed configurations of the LLM-based literature processing pipeline, such as prompt templates, inference parameters, and model-specific settings, are not fully specified. As a result, not all experimental settings necessary to fully reproduce the entire pipeline are disclosed.

Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: Statistical variability is explicitly reported using error bars to quantify model-to-model uncertainty. For each material, ionic transport properties were evaluated using two independent MLIP models (MACE and GRACE) under identical simulation conditions. The reported values correspond to averages over the two models, with error bars representing the standard deviation between model predictions (Fig. 5). Consistent lithium diffusion behavior is observed across the full temperature range, as confirmed by the MSD curves provided in the Appendix (Fig. 8). The agreement in Arrhenius trends between the two models supports the robustness and consistency of the reported results.

   Guidelines:

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: The computational resources and simulation settings required for reproduction are explicitly described in Section 4.2, including the MLIP models used (MACE and

GRACE), supercell sizes, simulation time scales, temperature ranges, and molecular dynamics parameters. Additional simulation outputs and analysis details are provided in the Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: No human subjects, no sensitive personal data, no deployment-related risks.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [N/A]

    Justification: Direct societal deployment is not studied; potential impacts are indirect (e.g., accelerating materials discovery).

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The study does not release data or models that pose a high risk of misuse or dual use and therefore does not require specific safeguards for responsible release.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All existing assets used in this work, including publicly available MLIP models (MACE and GRACE) and referenced literature sources, are properly cited in the paper in accordance with their respective licenses and terms of use.

    Guidelines:

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

13. **New Assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, models, or code assets as standalone public resources.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing, human participants, or any form of human subject experimentation.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [N/A]

    Justification: The study does not involve human subjects or activities that require Institutional Review Board approval or an equivalent ethical review.

    Guidelines:

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