Computational and Data-Driven Discovery of Li Solid-State Electrolytes: From Representation to Experimental Realization

Thesis by
Daniel Brendan McHaffie

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended September 24, 2025

Daniel Brendan McHaffie
ORCID: 0000-0002-7265-7584

ACKNOWLEDGEMENTS

I first want to thank my PhD advisor, Prof. Kimberly See. One of your most admirable qualities as an advisor is your commitment to encouraging your students to follow the research questions that inspire them. In my case, this meant pursuing directions that were well outside what you may have anticipated, or even preferred. Despite this, your support for me was unwavering and allowed me to shape my PhD in the way that was aligned with my interests and goals. I admire your dedication to scientific rigor and your emphasis on clear communication. I aspire to uphold these standards in my future work. The positivity and collaborative spirit of the See Group serves as a testament to the kindness and compassion that you show your students.

I am grateful to the members of my committee, Professor Katherine Faber, Professor Harry Atwater, and Professor Marco Bernardi, for their continued support.

To all the members of the See Group, thank you for making every day in the lab enjoyable. I have always felt that I could turn to anyone in the group for support, and I would especially like to thank a few individuals who had an outsized influence on my time at Caltech.

Forrest, the positive impact you had on shaping my PhD journey could not be overstated. Your infectious enthusiasm for your research was one of the primary reasons for my desire to join the See Group. I am deeply grateful for your mentorship, both during your time in the group and afterwards. Your expertise in electrochemistry and data science, together with the generosity with which you shared your knowledge, was invaluable. I still consider your creation of the database that drove much of my research as nothing short of a Herculean feat. I feel incredibly fortunate to have had the opportunity to work with you, and only wish that our time had overlapped for longer.

Zac, though not official, I always considered you a mentor and am grateful for the guidance and support you shared throughout my PhD. The perspective that you brought to our frequent discussions was invaluable, and I was inspired by your ability to persevere through difficult research problems to tell an insightful story. Though it was sometimes met with protests in the moment, I truly appreciate your regular reminders to enjoy my time at Caltech, and I know that one of the reasons I was able to do so was having you in the lab and as a friend along the way.

Thank you, Jadon for your dedication and positivity, which allowed us to explore a problem more deeply than would have been possible without our combined efforts. Your insightful questions continually pushed me to refine my own ideas. Mentoring you was a truly rewarding experience, and I am excited to see where your future research takes you.

I have not, and doubt I ever will, meet someone who thinks about their scientific questions with the same depth and passion as you, Eshaan. This quality regularly inspired me to explore my own research more deeply. I could always count on you for an unlimited supply of thoughtful questions after talks, fascinating discussions about the global supply chain for battery raw materials, and your company during late nights in the lab, complimentary vocals included.

Thank you, Steve, for consistently demonstrating how to ask the right questions and identify the next measurements that provide the most value. Your critical thinking in research is something that I have tried to emulate in my own work. I am also grateful for your willingness to answer all of my questions and for your friendship throughout my PhD.

Michelle, your dedication to mentorship is inspiring. In addition to setting an example as a talented scientist, your generosity in supporting others has had an exceedingly positive impact on every member of the See Group. Thank you for being a wonderful officemate and for putting up with my standing-desk experiments.

Abhiroop, your exceptional command of electrochemistry fundamentals and your generosity in sharing this knowledge have been a tremendous asset to the group. I always enjoyed our many conversations about the broader impact of our research and your unique humor always made time in the lab much more entertaining.

I have had the privilege of being mentored by an incredible group of scientists during my internships in both my undergraduate and graduate studies. Each has been instrumental in shaping my research journey thus far.

Prof. Adam Wei Tsen, thank you for giving me my very first opportunity in academic research. You instilled the importance of making the most of experiments when they are working, in addition to a foundational work ethic that has persisted to this day.

Dr. Lia Kouchachvili, thank you for starting my path in electrochemistry. I continue to rely on the electrochemical techniques that I first learned in your lab today.

Dr. Andrew Weng, thank you for being an outstanding mentor and for demonstrating the value of clear scientific communication. I know that my passion for batteries is due in no small part to the inspiring introduction you gave me, and I am deeply grateful to have had the opportunity to work with you.

Dr. John W. F. To, thank you for your invaluable advice and mentorship. Your thoughtful insight into the merits of pursuing a PhD and navigating graduate school was instrumental in helping me through a pivotal stage of my education.

Dr. Joseph Montoya and Dr. Steven Torrisi, thank you for giving me the opportunity to work with you and for broadening my perspective on the scientific questions central to my PhD research. Dr. Torrisi, your approach to scientific communication has been inspiring and has set a standard that I strive to follow. Dr. Montoya, thank you for sharing your expertise and creative insight in solving difficult technical challenges, as well as for your continued support.

This may not be a common sentiment among graduate students, but I had an incredibly positive and rewarding experience in graduate school. A primary reason for this is that I had the privilege of sharing the journey with my two best friends, Jared and Phillippe. Your friendship and support during our undergraduate years motivated and enabled me to pursue a PhD in the first place. In graduate school, even though we no longer worked together directly, living together and seeing your dedication as exceptional scientists continually inspired me to do my best work. I have no doubt that my PhD would not have been nearly as fulfilling or anywhere near as fun without you. Ben, moving in with people that had already been roommates for five years could have been a difficult situation, but you quickly became part of our close-knit group, something not many people could have done so easily. Thank you for your friendship and for making our time together all the more enjoyable. Charles, I feel lucky to have met such a great friend at Caltech. I've always appreciated your humor, your friendship, and your willingness to share your expertise in computational chemistry.

I have my family to thank most of all for their unconditional love and support. To my siblings, Ryan and Mara, I hope you know how much I have always looked up to you both in different ways. Mom and Dad, through your careers, you set an example for me to find joy and fulfillment in my own work. Dad, you instilled in me the value of effort through your simple but effective saying, "hard work pays off." Your constant curiosity about the natural world was no doubt passed on to me and helped guide my path to science. Mom, the care and dedication that you showed to your

work and students has always inspired me and is a huge part of who I am. I love you all.

Methely, thank you for the joy you bring to every moment. Your constant support has meant so much to me. Being with you has been the most important part of my graduate school journey, and I am so excited for our future together. I love you so much.

ABSTRACT

Improvements in energy storage are required to facilitate the transition to renewable energy and the electrification of transport. Lithium-ion batteries (LIBs) are a promising solution, but the current leading chemistry, consisting of a layered oxide cathode and a graphite anode separated by a liquid electrolyte, has been optimized to near-theoretical limits. Replacing the graphitic carbon with Li metal would significantly improve energy density but the instability of the Li metal–electrolyte interface introduces performance and safety challenges. Using a solid-state electrolyte (SSE) to construct an all-solid-state battery (ASSB) could mitigate these issues. However, an ideal SSE material has yet to be identified.

Thousands of known Li-containing materials have not yet been evaluated as SSEs. Data-driven methods could prioritize materials for experimental study but have historically lacked sufficient data and optimal representations. Chapter 2 presents the largest structure–ionic conductivity database to date and uses semi-supervised learning to determine the highest-performing descriptors. From ~26,000 Li-containing materials, 212 candidates are identified and screened using semi-empirical and first-principles calculations. Li_3BS_3 exhibits ionic conductivity above 10^{-3} S cm^{-1} with defect engineering through substitution and mechanical milling.

Chapter 3 explores Cl, Al, and Si substitution in Li_3BS_3 to reveal mechanisms of ionic conductivity enhancement. At low substitution levels, conductivity improvements are driven by disordered environments from reduced crystallinity and microstructural effects. For Cl and Al, higher substitution generates fully amorphous phases with ionic conductivity above 10^{-4} S cm^{-1}. Sufficient Si substitution produces novel crystalline phases with conductivities exceeding 10^{-3} S cm^{-1}.

Previous approaches, such as that in Chapter 2, could not represent disordered compounds, excluding much of the training data and candidate materials. This is particularly significant given the importance of disorder highlighted in Chapters 2 and 3 and the prevalence of disorder in known superionic conductors. Chapter 4 implements a transfer-learned graph representation compatible with disordered structures. A larger database is curated and used to train models for screening all known Li-containing materials. Experimental validation of superionic conductivity in an identified candidate demonstrates the utility of this graph-based approach for discovering experimentally relevant, high-performance materials.

PUBLISHED CONTENT AND CONTRIBUTIONS

Laskowski, F. A. L.; McHaffie, D. B.; See, K. A. Identification of Potential Solid-State Li-Ion Conductors with Semi-Supervised Learning. *Energy Environ. Sci.* **16** (2023), 1264–1276.
Contributions: F. A. L. L. and D. B. M. contributed equally. D. B. M. performed semi-empirical and first-principles calculations. F. A. L. L. and D. B. M. both participated in experimental validation and contributed to manuscript preparation.

McHaffie, D. B.; Bienz, J. M.; Hwang, S.-J.; Laskowski, F. A. L.; See, K. A. Substitution of Li_3BS_3: Revealing New Superionic Conductor Phases and the Significance of Crystallinity. *In preparation.*
Contributions: D. B. M. and J. M. B. contributed equally. Together they conceived and designed the study, performed experiments, and wrote the manuscript.

McHaffie, D. B.; Iton, Z. W. B.; Bienz, J. M.; Laskowski, F. A. L.; See, K. A. Classification of (Dis)Ordered Structures as Superionic Lithium Conductors with an Experimental Structure–Conductivity Database. *Digital Discovery* **4** (2025), 1518–1533.
Contributions: D. B. M. conceived and designed the study, performed computational analysis, experimental validation, and wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation and Background

Urgent action is needed to reduce greenhouse gas emissions and mitigate the impacts of global warming [1]. Meanwhile, global energy demand continues to rise. Sustaining the prosperity historically enabled by fossil fuel consumption while decreasing emissions requires a transition to renewable sources, complemented by other low-carbon sources such as nuclear power [2]. However, the inherent intermittency of renewables like solar and wind energy demands a simultaneous increase in energy storage capacity to maintain electricity security. Highlighting this, the International Energy Agency estimates that global storage capacity must increase sixfold by 2030 to permit the tripling of global renewable energy capacity pledged at COP28 [2, 3]. Advanced portable energy storage is also required to electrify the transport sector, which accounted for 19% of net global greenhouse gas emissions in 2019. Batteries have emerged as a competitive solution for energy storage due to their high round-trip efficiency, modularity, and rapidly declining cost [2, 4, 5].

Li-ion batteries (LIBs) have transformed human interaction with technology by providing an unprecedented means of portable energy storage. Their commercialization enabled the widespread adoption of mobile electronics and more recently, has made the electrification of transportation possible. The prototypical LIB consists of a layered oxide cathode and a graphitic carbon anode, separated by an electrolyte composed of $LiPF_6$ dissolved in carbonate solvents. When charged, Li ions are intercalated between the layers of the graphite anode. The difference in electrochemical potential between the anode and cathode in the charged state drives spontaneous redox reactions. When the electrodes are connected through an external load, an oxidation reaction occurs at the anode and electrons move through the external circuit from the anode to the cathode. Simultaneously, Li^+ ions move from the anode to the cathode through the electrolyte. At the cathode, a reduction reaction occurs and the Li^+ ions are incorporated into the host structure [6].

Despite the profound impact of LIBs, higher performance is needed. The use of Li metal anodes, which have the highest possible theoretical capacity of 3860 mAh g^{-1}, in place of graphite anodes (372 mAh g^{-1}) is desired. However, the substitution of

Li metal for graphite presents significant challenges. Similarly to graphite, Li metal reacts with conventional electrolytes to form a solid-electrolyte interphase (SEI). Unlike the relatively stable SEI on graphite anodes, the Li metal SEI is susceptible to cracking due to volume changes during plating and stripping of Li. This repeatedly exposes the Li metal surface, leading to continuous side reactions which increase impedance and consume active Li, decreasing performance. Furthermore, nonuniform Li plating and stripping during cycling can lead to dendrite growth, which may contact the cathode and cause internal short circuits. The organic carbonate-based electrolytes used in commercial LIBs are flammable and thermally unstable, requiring operation within a relatively narrow temperature window. Localized heating from short circuits caused by dendrites or mechanical abuse can initiate exothermic reactions that lead to thermal runaway [7, 8].

All-solid-state batteries (ASSBs), wherein the liquid electrolyte is replaced with a solid-state electrolyte (SSE), could alleviate some of the challenges associated with the use of Li metal anodes. It was originally suggested that the higher shear moduli of SSEs would suppress dendrite penetration [7, 9–11]. However, dendrite growth through the grain boundaries or interfacial defects in SSEs has been observed [12–14]. Despite this, SSEs are less flammable than liquid carbonates, mitigating some of the safety risks in the event of an internal short. They also have greater thermal stability, enabling operation over a wider temperature window [15]. Moreover, SSEs exhibit high transference numbers for Li^+ because the anion framework is fixed. This reduces the formation of salt concentration gradients that limit current in liquid electrolytes and allows higher current densities in ASSBs [16, 17].

The development of ASSBs has been hindered by the absence of a suitable SSE material. An ideal SSE has high ionic conductivity, low electronic conductivity, electrochemical stability at both electrodes, and suitable mechanical properties. However, in the most commonly studied SSE families, trade-offs between these properties are often observed. Figure 1.1 summarizes trends in ionic conductivity and stability window for different SSEs, grouped by anion chemistry. Ionic conductivity decreases with increasing anion electronegativity within each group, whereas the opposite trend is observed for the stability window. More electronegative anions (O^{2-}, F^-) are less polarizable and harder to oxidize. This gives them wide stability windows but also leads to stronger Li-anion interactions and a rigid lattice, increasing the migration energy barrier for Li^+ transport. By contrast, anions such as S^{2-} or Br^- form more flexible lattices and have weaker Li-anion interactions,

facilitating high ionic conductivity. However, compounds with these anions are more easily oxidized, resulting in lower anodic stability limits. The discovery of novel SSEs outside of traditionally studied families is required to achieve a more favorable balance of these properties.

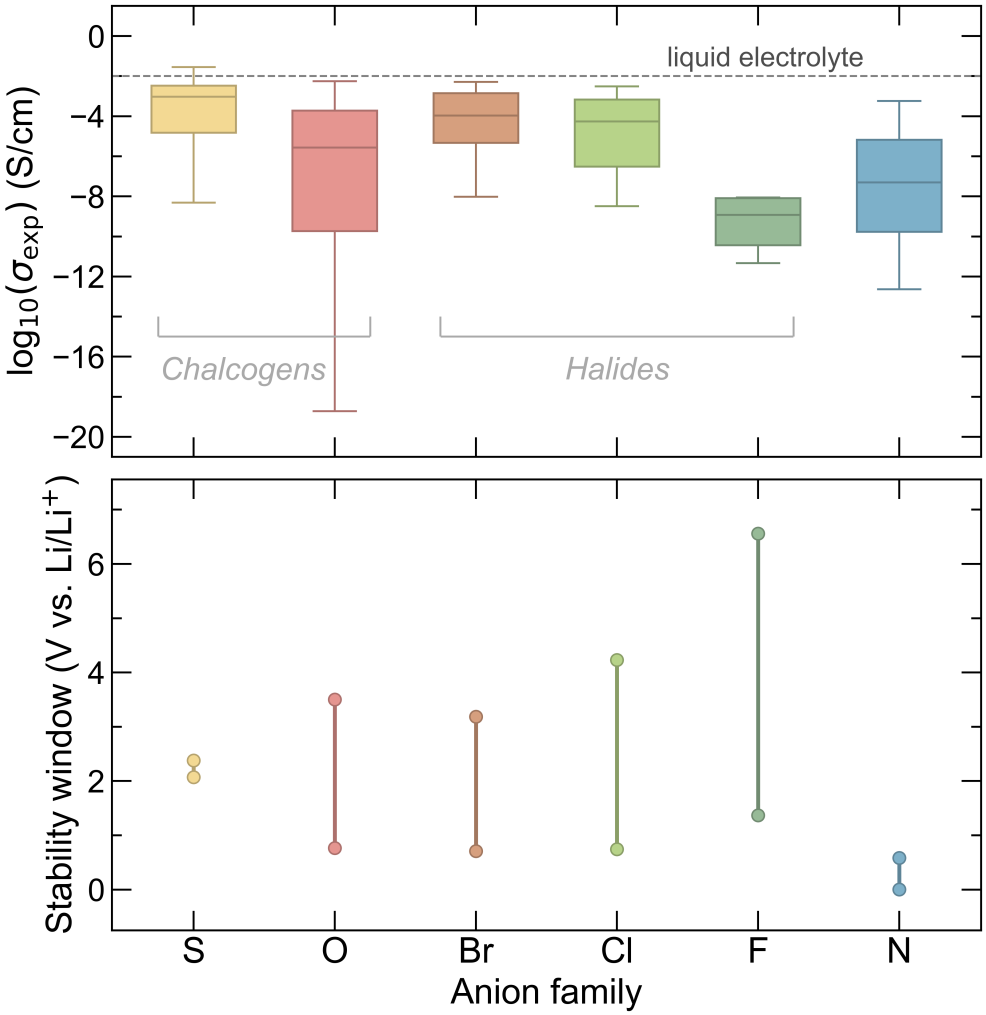

Figure 1.1: Stability and conductivity trade-offs for solid-state electrolytes (SSEs). (a) Average calculated stability window for each anion family. Data sourced from references [18] and [19]. (b) Ionic conductivity distributions for each anion family shown as box plots, based on an in-house curated database. (c) The highest-conductivity SSE from each anion family is represented by a circular marker, with the marker area proportional to the order of magnitude of its ionic conductivity. An ideal SSE is represented by the red circular marker, with ionic conductivity equal to 10^{-2} S cm^{-1}, a reductive limit of 0 V vs. Li/Li$^+$, and an oxidative limit greater than 5 V vs. Li/Li$^+$.

Only a small fraction of known Li-containing materials have been experimentally evaluated as potential SSEs. We have curated the largest dataset to date of crystal

structures from the Inorganic Crystal Structure Database (ICSD) and their experimentally measured ionic conductivities, as will be expanded upon in later chapters. Of the 11,295 Li-containing materials in the ICSD, only 571 have experimentally characterized ionic conductivities (Figure 1.2). A further 12,974 and 21,593 theoretical Li-containing materials are predicted to be stable or metastable in the Materials Project and GNoME computational databases, respectively. This vast chemical space is intractable for experimental study alone, highlighting the need for *in silico* approaches to guide the discovery of novel SSEs.

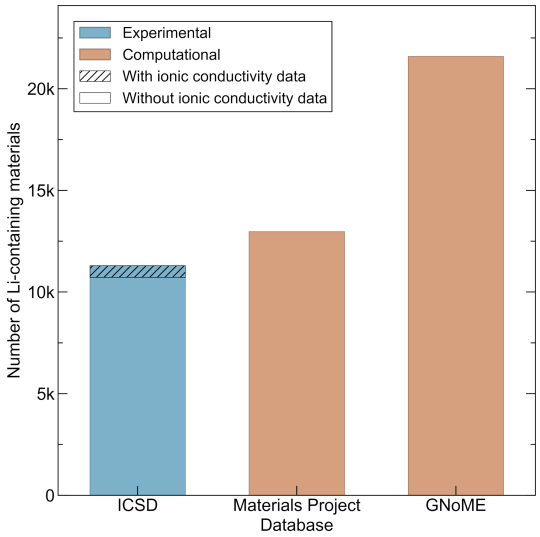

Figure 1.2: Number of known Li-containing materials in prominent databases. The Inorganic Crystal Structure Database (ICSD) contains experimentally reported materials, with the subset of measured ionic conductivities shown in hatched shading. The Materials Project and GNoME databases consist of computationally predicted structures, with GNoME representing the largest to date.

Computational methods are an increasingly important tool in the search for new high-conductivity SSEs. The conductivity of a material can be expressed as

$$\sigma = \sum_i n_i q_i \mu_i \tag{1.1}$$

where σ is the conductivity, n_i is the concentration of charge carriers of type i, q_i is the charge of those carriers, and μ_i is their mobility. The Nernst-Einstein equation relates the mobility of a charge carrier to its diffusivity by

$$\mu = \frac{qD_\sigma}{k_\mathrm{B}T} \tag{1.2}$$

where T is the temperature and k_B the Boltzmann constant. D_σ is the long-range macroscopic diffusion coefficient which can be related to the tracer diffusion coefficient, D^*, through the Haven ratio

$$H_R = \frac{D^*}{D_\sigma} \tag{1.3}$$

where $H_R \leq 1$ accounts for ion-ion correlation effects and $H_R \approx 1$ indicates weak or no correlations. D^* is the proportionality constant in Fick's First Law and refers to the diffusivity in a chemically uniform host. Arrhenius behavior is typically observed for the tracer diffusivity,

$$D^*(T) = D_0 \exp\left(-\frac{E_A}{k_B T}\right) \tag{1.4}$$

where E_A is the activation energy for diffusion. In the intrinsic regime, this is equal to $E_m + E_f/2$ where E_m and $E_f/2$ are the migration energy barrier and defect formation energy, respectively. When doping or substitution creates extrinsic defects that significantly outnumber intrinsic defects, the E_f term is no longer relevant and instead a trapping energy term, $E_t/2$, can be included if there are significant interactions between the substituted ions and mobile carriers.

The tracer diffusivity can be determined from molecular dynamics calculations. If the assumption $H_R = 1$ is made, the ionic conductivity can then be calculated through Equations 1.1–1.3. However, obtaining accurate D^* values from molecular dynamics requires the sampling of a sufficient number of diffusion events. This requires large simulation cells and long durations, especially if the diffusivity of the mobile species is low. *Ab initio* molecular dynamics (AIMD), wherein the most accurate description of the potential energy surface is obtained through first-principles methods, is prohibitively expensive for high-throughput screening. Molecular dynamics with machine-learned interatomic potentials trained on first-principles data (ML-MD) is a promising emerging technique, providing accuracy approaching AIMD at a significantly reduced cost. However, first-principles data is still required for fitting or fine-tuning parameters and the transferability of these potentials to non-equilibrium states remains uncertain. Density functional theory-based nudged-elastic band (NEB) calculations can be used to calculate the E_m but require identification of all possible ion transport pathways and the E_m itself can be affected by complex diffusion mechanisms, including the cooperative motion of

multiple ions. NEB is therefore typically reserved as a detailed evaluation method for a limited number of structures, rather than as a screening tool.

1.2 Thesis Overview

This thesis focuses on accelerating the discovery of novel high-performance SSEs. A strong emphasis is placed on addressing the entire discovery pipeline, from computational modeling to experimental demonstration and the development of mechanistic understanding.

Chapter 2 presents a semi-supervised learning approach for identifying fast Li-ion conductors. A database pairing experimental ionic conductivities with crystal structures is created, and semi-supervised learning is used to identify material representations that best correlate to observed ionic conductivity. The optimal representation is used to identify promising conductors. Subsequent semi-empirical and first-principles calculations are used to prioritize experimental candidates. One candidate, Li_3BS_3 is shown to have superionic conductivity through defect engineering.

Chapter 3 develops a mechanistic understanding of the factors influencing ionic conductivity in the Li_3BS_3 system. The effects of chemical substitution and processing conditions on both local and long-range structure are investigated using x-ray diffraction, Raman spectroscopy, solid-state nuclear magnetic resonance, and electrochemical impedance spectroscopy. Cl, Al, and Si substitution reduces crystallinity, and the microstructure is highly sensitive to grinding duration. Both means of introducing disordered environments significantly enhance the ionic conductivity in Li_3BS_3, while high levels of Si substitution drive the formation of a novel superionic phase.

The profound impact of disorder on ion transport is highlighted in Chapters 2 and 3. Computational studies of SSEs often inadequately address disorder due to the lack of suitable representations for imperfect crystals. Chapter 4 introduces a graph-based strategy for representing disorder that enables learning and prediction in disordered materials. The largest database of crystal structures with corresponding ionic conductivities assembled to date is used to train models capable of identifying fast ion conductors, aided by this representation and transfer-learning techniques. The value of this approach is demonstrated through experimental validation of superionic conductivity in the highly disordered phase $Li_9B_{19}S_{33}$.

1.3 Bibliography

[1] Calvin, K. et al. *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (Eds.)]. IPCC, Geneva, Switzerland.*; **2023**.

[2] IEA *Batteries and Secure Energy Transitions*; International Energy Agency: Paris, **2024**.

[3] COP28 UAE Global Renewables and Energy Efficiency Pledge. **2023**.

[4] Dunn, B.; Kamath, H.; Tarascon, J.-M. Electrical Energy Storage for the Grid: A Battery of Choices. *Science* **2011**, *334*, 928–935.

[5] Choi, D.; Shamim, N.; Crawford, A.; Huang, Q.; Vartanian, C. K.; Viswanathan, V. V.; Paiss, M. D.; Alam, M. J. E.; Reed, D. M.; Sprenkle, V. L. Li-Ion Battery Technology for Grid Application. *Journal of Power Sources* **2021**, *511*, 230419.

[6] Goodenough, J. B.; Park, K.-S. The Li-Ion Rechargeable Battery: A Perspective. *J. Am. Chem. Soc.* **2013**, *135*, 1167–1176.

[7] Goodenough, J. B. Rechargeable Batteries: Challenges Old and New. *J Solid State Electrochem* **2012**, *16*, 2019–2029.

[8] Kaliaperumal, M.; Dharanendrakumar, M. S.; Prasanna, S.; Abhishek, K. V.; Chidambaram, R. K.; Adams, S.; Zaghib, K.; Reddy, M. V. Cause and Mitigation of Lithium-Ion Battery Failure—A Review. *Materials* **2021**, *14*, 5676.

[9] Monroe, C.; Newman, J. The Impact of Elastic Deformation on Deposition Kinetics at Lithium/Polymer Interfaces. *J. Electrochem. Soc.* **2005**, *152*, A396.

[10] Ni, J. E.; Case, E. D.; Sakamoto, J. S.; Rangasamy, E.; Wolfenstine, J. B. Room Temperature Elastic Moduli and Vickers Hardness of Hot-Pressed LLZO Cubic Garnet. *J Mater Sci* **2012**, *47*, 7978–7985.

[11] Masias, A.; Felten, N.; Garcia-Mendez, R.; Wolfenstine, J.; Sakamoto, J. Elastic, Plastic, and Creep Mechanical Properties of Lithium Metal. *Journal of materials science* **2019**, *54*, 2585–2600.

[12] Nagao, M.; Hayashi, A.; Tatsumisago, M.; Kanetsuku, T.; Tsuda, T.; Kuwabata, S. In Situ SEM Study of a Lithium Deposition and Dissolution Mechanism in a Bulk-Type Solid-State Cell with a Li_2S–P_2S_5 Solid Electrolyte. *Physical Chemistry Chemical Physics* **2013**, *15*, 18600–18606.

[13] Ren, Y.; Shen, Y.; Lin, Y.; Nan, C.-W. Direct Observation of Lithium Dendrites inside Garnet-Type Lithium-Ion Solid Electrolyte. *Electrochemistry Communications* **2015**, *57*, 27–30.

[14] Liu, H.; Cheng, X.-B.; Huang, J.-Q.; Yuan, H.; Lu, Y.; Yan, C.; Zhu, G.-L.; Xu, R.; Zhao, C.-Z.; Hou, L.-P.; others Controlling Dendrite Growth in Solid-State Electrolytes. *ACS Energy Letters* **2020**, *5*, 833–843.

[15] Inoue, T.; Mukai, K. Are All-Solid-State Lithium-Ion Batteries Really Safe?– Verification by Differential Scanning Calorimetry with an All-Inclusive Microcell. *ACS Appl. Mater. Interfaces* **2017**, *9*, 1507–1515.

[16] Kato, Y.; Hori, S.; Saito, T.; Suzuki, K.; Hirayama, M.; Mitsui, A.; Yonemura, M.; Iba, H.; Kanno, R. High-Power All-Solid-State Batteries Using Sulfide Superionic Conductors. *Nat Energy* **2016**, *1*, 16030.

[17] Janek, J.; Zeier, W. G. A Solid Future for Battery Development. *Nat Energy* **2016**, *1*, 16141.

[18] Richards, W. D.; Miara, L. J.; Wang, Y.; Kim, J. C.; Ceder, G. Interface Stability in Solid-State Batteries. *Chem. Mater.* **2016**, *28*, 266–273.

[19] Wang, Z.-Y.; Zhao, C.-Z.; Sun, S.; Liu, Y.-K.; Wang, Z.-X.; Li, S.; Zhang, R.; Yuan, H.; Huang, J.-Q. Achieving High-Energy and High-Safety Lithium Metal Batteries with High-Voltage-Stable Solid Electrolytes. *Matter* **2023**, *6*, 1096–1124.

Chapter 2

IDENTIFICATION OF POTENTIAL SOLID-STATE LI-ION CONDUCTORS WITH SEMI-SUPERVISED LEARNING

Laskowski, F. A. L.; McHaffie, D. B.; See, K. A. Identification of Potential Solid-State Li-Ion Conductors with Semi-Supervised Learning. *Energy Environ. Sci.* **2023**, *16* (3), 1264–1276. **F.A.L.L. and D.B.M. contributed equally to this work.**

2.1 Abstract

Despite ongoing efforts to identify high-performance electrolytes for solid-state Li-ion batteries, thousands of prospective Li-containing structures remain unexplored. Here, we employ a semi-supervised learning approach to expedite identification of superionic conductors. We screen 180 unique descriptor representations and use agglomerative clustering to cluster ~26,000 Li-containing structures. The clusters are then labeled with experimental ionic conductivity data to assess the fitness of the descriptors. By inspecting clusters containing the highest conductivity labels, we identify 212 promising structures that are further screened using bond valence site energy and nudged elastic band calculations. Li_3BS_3 is identified as a potential high-conductivity material and selected for experimental characterization. With sufficient defect engineering, we show that Li_3BS_3 is a superionic conductor with room temperature ionic conductivity greater than 1 mS cm^{-1}. While the semi-supervised method shows promise for identification of superionic conductors, the results illustrate a continued need for descriptors that explicitly encode for defects.

2.2 Introduction

Identifying new materials that could improve solid-state ion battery prospects is an ongoing challenge. The search for an ideal solid-state Li electrolyte is a prime example. Research has focused on eight classes of materials: LISICON-type structures, argyrodites, garnets, NASICON-type structures, Li-nitrides, Li-hydrides, perovskites, and Li-halides [1]. However, only three compounds with near-liquid-electrolyte conductivity (10^{-2} S cm^{-1}) have been discovered: $Li_{10}GeP_2S_{12}$ (LGPS) [2], Li_6PS_5Br argyrodite [3], and $Li_7P_3S_{11}$ ceramic-glass [1, 4]. Although promising discoveries, all three high-conductivity structures are unstable against the Li

anode [5–10]. While investigations to limit instability are ongoing [11, 12], identification of additional superionic structures is desirable. Discovery of new structures that support superionic conductivity improves the odds of identifying or engineering a stable electrode|SSE interface. For example, engineering solutions that fail to stabilize the Li|argyrodite interface may prove more successful when applied to not-yet-discovered superionic conductors. Discovery of new superionic conductors may also enable stable architectures *via* multi-electrolyte approaches which have been proposed as more promising than single-electrolyte architectures for achieving stability against Li metal and cathode materials [13]. High-performing structures that enable new battery chemistries may exist outside of the eight classes. However, exploration under the traditional Edisonian approach prioritizes small perturbations to well-known variable spaces. Machine learning (ML) is a promising tool for expediting the discovery of useful solid-state materials. By describing prospective materials with physically meaningful descriptors, ML models can identify high-dimensional patterns in large datasets that are not readily apparent [14–20]. Ongoing descriptor engineering [21–26] has enabled discovery of battery components [27, 28], electrocatalysts [15, 29], photovoltaic components [16, 30], piezoelectrics [31], new metallic glasses [14] and new alloys [32]. However, application of ML for discovery of SSEs and other emerging technologies can be challenging. Supervised ML approaches require empirical data for use as "labels". For example, graph neural network (GNN) approaches have been successful in many domains but generally require thousands to tens of thousands of labels to avoid overfitting [33]. By contrast, relatively few SSEs have been experimentally characterized compared to the ~26,000 known Li-containing structures [19, 34–36]. Characterized materials often exhibit ill-defined properties owing to the variety of synthetic approaches and non-standardized testing methods [37]. Well-performing materials often contain charge-carrying defects that are not explicitly characterized or reported [38]. Negative examples, i.e. materials with undesirable properties, are useful for ML models but are seldom reported. Semi-supervised ML can guide synthetic prioritization of SSEs by overcoming the issues associated with label scarcity. Supervised ML requires labels because it infers correlation functions by mapping the input descriptors to the labels [39]. Semi-supervised ML prioritizes comparison of descriptors to identify relationships between the descriptors in a dataset [36, 39]. The input compositions are clustered (or grouped) by comparison of descriptors using a similarity metric. The clustering process does not consider labels, and thus circumvents the need for abundant labels. The resultant clusters can be labeled *ex post facto* to

examine correlation between the descriptor and a physical property of interest. For semi-supervised ML, ideal descriptors result in a set of clusters where each cluster has similar labels and thus the label variance is minimized. Promising synthetic targets may then be identified by their membership in clusters that contain desirable labels. A key insight of this work is that semi-supervised ML can be used to rank descriptors in terms of their correlation to physical properties of interest. Descriptors are representations of the input materials that encode the chemistry, composition, structure, and/or other system properties. An ideal descriptor should be a unique representation, a continuous function of the structure, exhibit rotational/translational invariance, and be readily comparable across all structures in the dataset [24–26]. Recently, Zhang et al. demonstrated that a modified X-Ray diffraction (mXRD) descriptor lead to favorable clustering for Li SSEs [34]. By labeling the resultant clusters with experimental room-temperature Li-ion conductivities, they identified 16 prospective fast-ion conductors. However, an ideal descriptor is not known *a priori*, and no comprehensive descriptor screening has yet been pursued for correlation with SSE properties. Descriptor screening is desirable for both experimentalists and computationalists. For experimentalists, ranking of descriptors affords insight into what aspects of materials are most correlated with target properties. For computationalists, descriptors rankings enable improved regression and supervised learning models by guiding the selection of input representation(s). Descriptor transformations for inorganic structures have been curated in a variety of software packages, including Matminer [24], Dscribe [25], SchNet [40], and Aenet [41].

Herein, we employ hierarchical agglomerative clustering to screen many descriptors, without assuming correlation to ionic conductivity. The performance of 20 descriptors is assessed for semi-supervised identification of Li SSEs. Each descriptor is paired with 9 structural simplification strategies, yielding a total of 180 unique representations per input structure. The approach is applied to a dataset of ~26,000 Li-containing phases, encompassing all Li-containing structures contained in the Inorganic Crystal Structure Database (ICSD - v.4.4.0) and the Materials Project (MP - v.2020.09.08) database (Figure 2.1). A set of 220 experimental room temperature ionic conductivities ($\sigma_{25^{\circ}C}$) are aggregated from literature reports and used as labels. Experimental labels are selected because they may bias models towards identifying structures that are synthetically tractable and processable. Descriptors that encode the spatial environment are found to be most correlated with the ionic conductivity labels, whereas descriptors that encode the electronic, compositional, or bonding environment have less predictive power. For the structural descriptors,

simplifications that neglect the mobile ion perform best. The descriptor screening results suggest that ionic conductivity is most sensitive to the spatial environment of the framework lattice.

Using the descriptors, the semi-supervised approach can identify potential fast solid-state Li-ion conductors. By selecting structures in clusters containing high conductivity labels, the ~26,000 input structures are down selected to just 212 promising structures. Practical considerations, a semi-empirical bond valence site energy (BVSE) method [42], and the Nudged Elastic Band (NEB) method are employed to rank the structures. From the ten highest ranking structures, Li_3BS_3 is selected for model validation. Synthesis of pure Li_3BS_3 yields a poor conductor. However, by employing defect engineering strategies we demonstrate that Li3BS3 is a superionic conductor with an ionic conductivity greater than 10^{-3} S cm^{-1}.

2.3 Results and Discussion

Screening simplification-descriptor combinations. A set of 20 descriptors is selected for screening the semi-supervised learning approach (Table 2.1). The descriptors generally encode four types of information: the spatial environment, the chemical bonding environment, the electronic environment, and composition. All descriptors are implemented in Python using the Matminer [24] or Dscribe [25] libraries. The code is published to a GitHub repository and is available for download (https://github.com/FALL-ML/materials-discovery). Zhang *et al.* illustrated that structure simplification prior to learning can produce lower variance outcomes [34]. Their mXRD descriptor was found to work best with removal of all cations, all the anions replaced by a single representative anion, and the structure volume scaled to 40 $Å^3$ per anion. Inspired by the previous success in using structure simplification, we screen eight structure simplifications in addition to the unperturbed structure. For simplifications the following categories of atoms are replaced with a representative specie: (1) Cations are represented as Al, (2) Anions are represented as S, (3) Mobile ions are represented as Li, and (4) Neutral atoms are represented as Mg. Categories of atom are removed as to yield the four simplifications: CAMN (all atoms retained), CAN (mobile ions removed), AM (cations and neutral atoms removed), and A (only anions retained). Four additional simplifications are formed by scaling each lattice volume to 40 $Å^3$ per anion: CAMN-40, CAN-40, AM-40, and A-40.

Agglomerative clustering is performed on all Li-containing structures from the ICSD and MP repositories. Agglomerative clustering is a "bottom-up" approach

Table 2.1: The descriptors used for agglomerative clustering. Descriptor vectors are attained by simplifying the input structures and then applying the descriptor transformation. In total, 180 unique descriptor vectors are screened for each structure.

Descriptor	**Descriptor description**	**Ref.**
Bond fraction	"Bag of bonds" approach described in Hansen *et al*. wherein pairwise nuclear charges and distances are encoded.	[43]
Band center	Estimation of band center from constituent atoms' electronegativity values described by Butler *et al.*	[44]
Crystal structure analysis by voronoi decomposition (CAVD)	Calculation of the largest sphere that can pass through the lattice-sans-mobile-ion using Voronoi decomposition of structures.	[45]
Chemical ordering	Warren-Cowley-like ordering method to determine how different the structure's ordering is from random.	[46]
Density features	Calculates density, volume per atom, and the packing fraction.	[47]
Electronegativity difference	Composition-weighted calculation of the electronegativity difference between cations and anions.	[48]
Ewald energy	Sum of coulomb interaction energies across all lattice sites described by Ewald *et al.*	[49]
Global instability index	Averaged square root of the sum of squared differences over the bond valence sums.	
Jarvis	Diverse set of descriptors from the Jarvis-ML library.	[50]
Maximum packing efficiency	A measure of the void space within the unit cell.	[46]
Meredig	Composite descriptor from Meredig *et al.*	[51]
Modified XRD (mXRD)	Powder diffraction pattern calculated using Bragg's law.	[47]
Orbital field matrix	Descriptor that encodes the distribution of valence shell electrons for each input structure.	[52]
Oxidation states	Concentration-weighted oxidation state statistics.	[48]
Radial distribution function	Radial distribution function for each structure.	[47]
Sine coulomb matrix	Coulomb matrix for periodic lattices, developed by Faber et al.	[53, 54]
Smooth overlap of atomic positions (SOAP)	Geometric encoder that is rotationally/transitionally invariant through use of spherical harmonics and radial basis functions. Atoms are represented by a smeared Gaussian.	[25]
Structural complexity	The Shannon information entropy for a given structure.	[55]
Structure variance	Bond length and atomic volume variance for each structure.	[46]
Valence orbital	Structure-averaged number of valence electrons in each orbital.	[48, 56]
Control	A control descriptor is not explicitly used. Instead, clustering outcomes are randomly assigned. For composite intracluster variance calculations, 100 control iterations are averaged.	

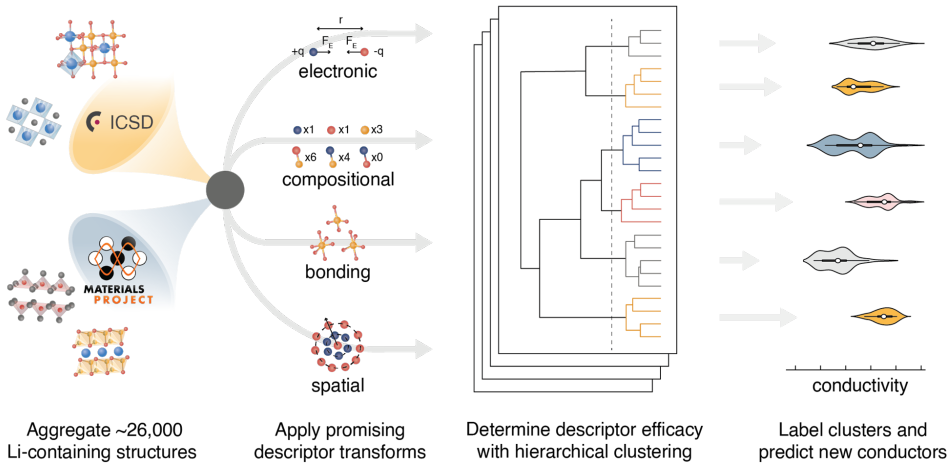

Figure 2.1: Schematic of the semi-supervised machine learning approach. Li-containing structures are aggregated from the ICSD and MP database. Each input structure is simplified and transformed to yield a unique descriptor representation. The descriptor representations are clustered with hierarchical agglomerative clustering. Each cluster is then labeled with experimental $\sigma_{25^\circ C}$ data and the intracluster conductivity variance is calculated. Comparison of the composite intracluster conductivity variance (intracluster conductivity variance summed across all clusters) enables identification of descriptors that are well correlated with ionic conductivity.

to clustering where each structure starts in its own cluster of one. Clusters are merged according to Ward's Minimum Variance criterion in Euclidean space, which minimizes the global descriptor variance [57]:

$$W = \sum_{k=1}^{n_C} \sum_{i \in C_k} \left[d_i - \bar{d}_k \right]^2$$

where n_C is the number of clusters in a set, C_k is cluster k, d_i is a descriptor representation for structure i, and $\bar{d}_k$ is the average descriptor representation in cluster k. Other common linkage criteria (average, complete, and single linkages) and metrics (l1, l2, Manhattan, cosine) were screened but are found to result in clustering outcomes with larger W). For each simplification-descriptor combination, all clustering sets from 2-300 are computed. Physically relevant labels are applied to the resultant clustering sets to assess how well each simplification-descriptor combination performs. To compare between the 180 different simplification-descriptions combinations, the data is labeled with 155 experimental room temperature conductivity (σ_{RT}) values aggregated from the literature reports. A secondary label set is also screened, comprising 6,845 activation energies (E_a) computationally generated using a bond valence energy approach (see Appendix A.2).

An ideal simplification-descriptor combination results in clustering where each cluster contains labels with similar σ_{RT} values. Ward's minimum variance method is applied to the conductivity labels as a measure of clustering efficacy [34]:

$$W_\sigma = \sum_{k=1}^{n_C} \sum_{i \in C_k} \left[\log(\sigma_{\mathrm{RT}})_i - \overline{\log(\sigma_{\mathrm{RT}})_k} \right]^2$$

where n_C is the number of clusters in a set, C_k is cluster k, and $\overline{\log(\sigma_{\mathrm{RT}})_k}$ denotes the mean for all labels in cluster k. Since clusters containing only one label effectively drop out of the W_σ calculation, a frozen-state strategy is employed when needed (see Appendix A.1). Each descriptor's W_σ results are shown in Figure 2.2 for the first 50 clustering outcomes (*i.e.* the W_σ is shown for each set of 2, 3, ..., 49, and 50 clusters). For simplicity, only the best-performing simplification-descriptor combination is shown for each descriptor.

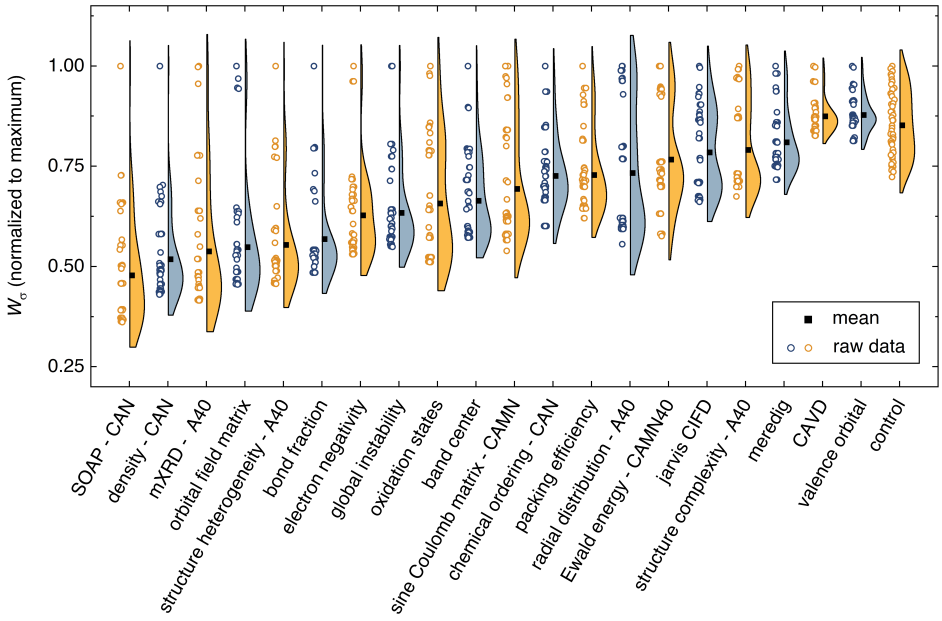

Figure 2.2: The composite intracluster conductivity variance (W_σ) for the first 50 clusters generated using each descriptor. Half-violin plots show the raw W_σ score for each depth of clustering as symbols next to the violin distribution. Simplification-descriptor combinations are sorted in order of ascending mean. The control is a random assignment of clusters, with W_σ values averaged over 100 randomly assigned sets. The smooth overlap of atomic positions (SOAP) descriptor outperforms all other descriptors. Although not shown here, SOAP continues to outperform for all depths of clustering through 300 clusters.

Using $\sigma_{25^\circ C}$ labels, the best semi-supervised ML performance is attained when using the SOAP descriptor. SOAP is a spatial descriptor that employs smeared Gaussians to represent atomic positions for each crystal structure [25]. Predictions using the SOAP descriptor have exhibited similar performance to state-of-the-art graph neural networks (GNNs) on a variety of materials science datasets [58]. Optimization of SOAP hyper-parameters (radial cutoff, number of radial basis functions, degree of spherical harmonics) is explored in Appendix A.3. SOAP is found to perform best when combined with the CAN structure simplification. That is, the simplification where the mobile Li atoms are removed, and the remaining atoms are simplified into three representative species: cations, anions, and neutral atoms. SOAP outperforms all other descriptors for all depths of clustering. The SOAP descriptor can be modestly improved (2-3% decrease in W_σ) by mixing with other descriptors to make a 2nd-order SOAP descriptor (see Appendix A.3).

Semi-supervised identification of prospective Li-ion conductors. Agglomerative clustering with the 2nd-order SOAP descriptor is used to identify prospective ionic conductors. W_σ minimization is prioritized over W_{E_a} minimization because E_a alone is not necessarily a good predictor of conductivity; $\sigma_{25^\circ C}$ may be affected by properties including the ionic carrier concentration, hopping attempt frequency, and the presence of concerted migration modes [59]. The agglomerative dendrogram for the 2nd-order SOAP clustering is shown in Figure 2.3. The agglomerative dendrogram is depicted to 241 clusters, after which the W_σ does not appreciably decrease. To facilitate discussion, an arbitrary cutoff is placed to yield 9 large clusters. The results show that although cluster #2 contains only 15% of the input structures, it accounts for over half of the high-conductivity ($\sigma_{25^\circ C} > 10^{-5}$ S cm^{-1}) labels. By the 17th clustering step, the densest cluster accounts for 6.2% of the structures while containing over half (52%) of the high-conductivity labels.

Candidates for next-generation SSEs can be identified by evaluating clusters that either contain or are near high conductivity labels. Clusters #2, #4, and #7 are promising because they account for 85% of the high $\sigma_{25^\circ C}$ labels. However, targeting these clusters would necessitate screening thousands of structures. Instead, we search from the 241st cluster depth, targeting all clusters that contain or are directly adjacent (*i.e.*, the nearest cluster in the Euclidean feature space) to high $\sigma_{25^\circ C}$ labels. The promising structures are further screened using calculated stability (E *vs.* E_{hull}) and band gap (E_g) properties from the Materials Project, and the BVSE E_a values. We select the structures that have (1) an E_{hull} of 70 meV or lower [60], (2) an E_g of

at least 1 eV, and (3) a BVSE-calculated E_a below a conservative 0.6 eV. We note that while a true E_g value of 1 eV would be problematic for an SSE, the bandgaps reported on Materials Project are typically underestimated by about 40% [61]. The approach identifies 212 structures as prospective ionic conductors. Climbing image nudged elastic band (CI-NEB) is employed to calculate the E_a for Li-ion hopping on the ten materials with the lowest BVSE-calculated E_a and an E_{hull} of 0 eV. The CI-NEB computational details can be found in Appendix A.4. The top 10 prospective structures are tabulated in Table 2.2.

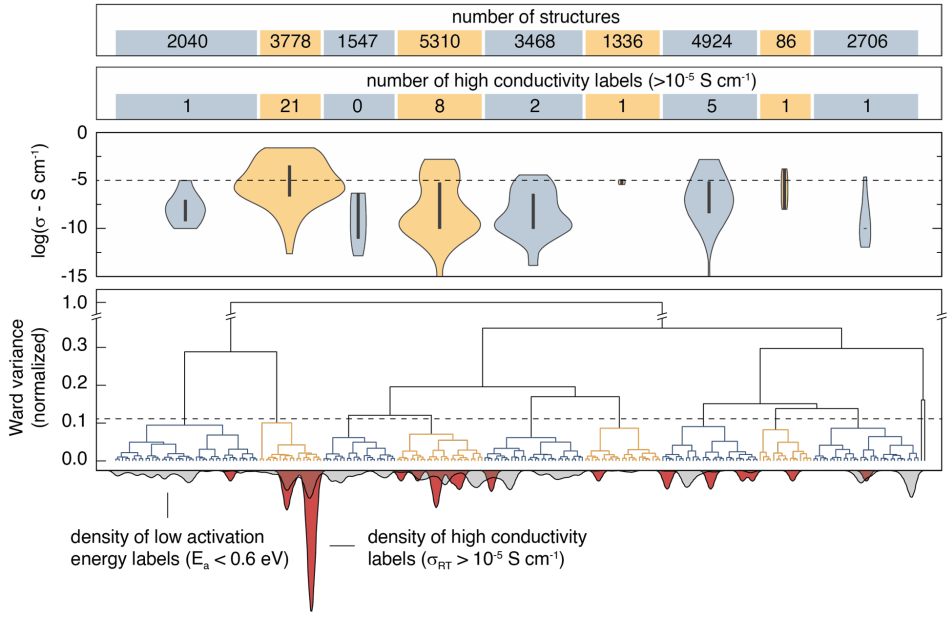

Figure 2.3: Agglomerative clustering dendrogram for the 2nd-order SOAP descriptor. The hierarchical clustering representation is shown for the first 241 clusters. An arbitrary variance cutoff is placed such that 9 large clusters are produced to facilitate analysis. The violin plots show the $\sigma_{25°C}$ distribution for the labels within the 9 large clusters. Three outlier clusters are grouped into two additional clusters and are hereafter ignored. The density (per 241 clusters) of low E_a (< 0.6 eV) and high conductivity ($\sigma_{25°C} > 10^{-5}$ S cm^{-1}) labels is shown underneath the agglomerative dendrogram. The results illustrate that agglomerative clustering on the 2nd-order SOAP descriptor results in favorable aggregation of most high-conductivity labels.

The CI-NEB calculations generally agree with the BVSE calculated E_a values, suggesting favorable activation energies (< 500 meV). Discrepancies between the two values may arise because BVSE does not allow framework ions to relax during Li^+ migration and does not account for repulsive interactions between atoms of the mobile ion species. BVSE also does not capture cooperative conduction mechanisms or those involving the so-called paddlewheel effect. Despite these limitations, we

Table 2.2: The top 10 prospective structures from the semi-supervised learning model as ranked by BVSE-calculated E_a. Structures in or directly adjacent to high-conductivity clusters were identified as promising. The list of promising structures was then further simplified by removing structures with Materials Project reported E_{hull} values greater than 0 V and E_g values less than 1 eV. To rank the remaining structures, the E_a was calculated using BVSE and CI-NEB approaches.

Compound	**Space group**	**MP_ID**	**ICSD_ID**	E **vs.** E_{hull} **(eV per atom)**	E_g **(eV)**	$E_{a,\text{calc}}$ **(meV)**	
						BVSE	**NEB**
Li_3VS_4	$P\bar{4}3m$ (#215)	mp-760375	NA	0	1.88	160	390
$Na_3Li_3Al_2F_{12}$	$Ia\bar{3}d$ (#230)	mp-6711	9923	0	7.85	230	340
Li_2Te	$Fm\bar{3}m$ (#225)	mp-2530	60434	0	2.49	260	320
$LiAlTe_2$	$I\bar{4}2d$ (#122)	mp-4586	280226	0	2.46	260	310
$LiInTe_2$	$I\bar{4}2d$ (#122)	mp-20782	658016	0	1.49	270	450
Li_6MnS_4	$P4_2/mmc$ (#137)	mp-756490	NA	0	1.55	270	470
$LiGaTe_2$	$I\bar{4}2d$ (#122)	mp-5048	162555	0	1.59	270	340
Li_3BS_3	$Pnma$ (#62)	mp-5614	380104	0	2.89	280	260
KLi_6TaO_6	$R\bar{3}m$ (#166)	mp-9059	73159	0	4.27	300	400
Li_3CuS_2	$Ibam$ (#72)	mp-1177695	NA	0	2.03	310	440

note that the model identifies numerous diverse structures beyond those routinely explored. Table 2.2 includes four tellurides, a vanadium sulfide, and multiple transition-metal-containing structures. Of the structures in Table 2.2, 70% avoid the space groups for the best-performing SSEs discovered to date: LPS (62), LGPS (137), the argyrodites (216), and LLZO (230).

Experimental validation of the semi-supervised learning model: Li_3BS_3. From the ten most promising candidates, Li_3BS_3 was selected for synthesis and characterization. Li_3BS_3 is noteworthy because it has been explored experimentally and computationally before. Experimentally, Vinatier *et al.* previously determined that Li_3BS_3 has a total DC conductivity of 2.5×10^{-7} S cm^{-1} with an activation energy of 700 meV [62]. The DC measurement was not included in our label set because DC measurements cannot differentiate between ionic and electronic conductivity, so they were categorically discounted from the label set (see Appendx A for more details on label selection). Although the conductivity and activation energy values reported by Vinatier *et al*. are underwhelming, there are promising theoretical reports. Density functional theory and molecular dynamics (DFT-MD) simulations from Sendek *et al*. [63] suggest that Li_3BS_3 should have a room temperature conductivity between 3.1×10^{-6} and 9.7×10^{-3} S cm^{-1}. Our NEB-calculated activation energy for Li_3BS_3 is 260 meV, corroborating a previous NEB result from Bianchini *et al*. [64]. Additionally, Li_3BS_3 is practically attractive because (1) Li_3BS_3

contains no redox-active metals, (2) band edge calculations have suggested stability against metallic Li [65], (3) DFT-MD calculations have suggested a kinetic barrier for decomposition against metallic Li [63], and (4) the synthesis is reported [66]. It is simpler to avoid redox active metals in the SSE as they may be reduced and oxidized at electrode interfaces. However, we note that $Li_{0.5}La_{0.5}Ti_3$ is a widely studied SSE that contains redox active Ti [67, 68] so the compounds we report here that contain Mn, V, and Cu should not be categorically discounted. It is important to note that while studying Li_3BS_3 as a candidate Li-ion conductor for model validation, Kimura *et al.* reported that a so-called "Li_3BS_3 glass" exhibits an ionic conductivity of 3.6×10^{-4} S cm^{-1} at 25°C [69].

Li_3BS_3 is prepared using solid-state synthesis from Li_2S, B, and S precursors. The diffraction and quantitative Rietveld refinement are shown in Figure 2.4 (a), indicating a phase pure material. Electrochemical impedance spectroscopy (EIS) is employed at various temperatures and the measured conductivity is plotted according to the Arrhenius-like relationship (Figure 2.4 (b):

$$\sigma = \frac{\sigma_0}{T} e^{-\frac{E_a}{k_B T}} \tag{2.1}$$

where T is the temperature, k_B is Boltzmann's constant, σ_0 is the conductivity prefactor, and E_a is the activation energy. The room temperature ionic conductivity ($\sigma_{25°C}$) is $7.2(\pm 3.0) \times 10^{-7}$ S cm^{-1} and the activation energy is 400 ± 47 meV. The low conductivity and high activation energy may be due to lack of charge-carrying defects in the Li_3BS_3 lattice [70, 71]. Although a sufficient carrier concentration is necessary for facile ionic conduction in most materials, the descriptors in the semi-supervised model not explicitly encode for charge-carrying defects. In the label set, conductivity is likely influenced by the presence of defects that are typically not reported. Still, the semi-supervised model may infer a structure's capacity to support conductive defects *via* correlation with the descriptors. To test the hypothesis, we use two strategies to engineer vacancies: aliovalent substitution and amorphization via extended ball milling. Aliovalent substitution has been shown to improve conductivity in Li-argyrodites, -sulfides, and -garnets by introducing vacancies [70, 71]. Similarly, amorphization can introduce defects and vacancies that enable Li^+ hopping [69, 71–73].

Aliovalent substitution of Li_3BS_3 is achieved by substituting Si for B. The XRD patterns and quantitative Rietveld refinements of $Li_{2.975}B_{0.975}Si_{0.025}S_3$ and $Li_{2.95}B_{0.95}$

$Si_{0.05}S_3$ are shown in Figure 2.4 (a). The lattice parameters from the refinements are plotted vs. stoichiometry with the Li_3BS_3 end-member in Figure 2.4 (e). The linear trend shows that the materials obey Vegard's law and confirms that Si incorporates into the lattice as a solid-solution. Substitution to 7.5% Si continues the Vegard trend until unidentified impurities are apparent in the XRD pattern. With 5% Si substitution, the ionic conductivity is improved to $1.82(\pm0.21) \times 10^{-5}$ S cm^{-1} and the activation energy is decreased to 333 ± 47 meV (Figure 2.4 (d)). All error bars reported for electrochemical measurements represent the standard deviation of three replicate cells. Kimura *et al.* demonstrated that extended ball milling of Li_3BS_3 causes amorphization and improves ionic conductivity, likely due to the introduction of defects [69]. Extended ball milling is attempted on the 5%-substituted Li_3BS_3 to assess whether both defect engineering strategies are compatible. Planetary ball milling of the 5%-substituted Li_3BS_3 for 100 h achieves amorphization (a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$), as verified by the lack of distinct peaks in the XRD pattern shown in Figure 2.4 (a).

We find that amorphization significantly improves Li-ion conductivity. EIS measurements of a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ are shown in Figure 2.4e. A high-frequency semicircle is partially resolved which may represent grain boundary or bulk ionic transport. A Warburg tail is evident at lower frequencies, indicating that electronic charge transfer is blocked. Although multiple high-frequency semicircles may exist (see Appendix A.5), a conservative estimate of the ionic conductivity is determined by linear fit of the Warburg tail and extrapolation to the x-intercept. The $\sigma_{25^\circ C}$ of a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ is $1.07(\pm0.08) \times 10^{-3}$ S cm^{-1} with an activation energy of 345 ± 2 meV (Figure 2.4 (d)). The electronic conductivity as measured by DC polarization is less than 4×10^{-10} S cm^{-1}. To determine if the local structure in the crystalline material is maintained after amorphization, we turn to ^{7}Li and ^{11}B NMR. If the local structure is not altered by amorphization, then it is likely that the ion diffusion pathways are similar. Comparing Li-ion diffusion pathways is important because the machine learning points to the structure of the crystalline Li_3BS_3 phase. The ^{7}Li NMR spectra of Li_3BS_3, $Li_{2.95}B_{0.95}Si_{0.05}S_3$, and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ are shown in Figure 2.4 (d). All materials show a similar resonance at the same chemical shift, suggesting that the Li local environment remains unchanged. The resonance width is reduced in the amorphous material due to the higher mobility. The ^{11}B NMR measurements are shown in Figure 2.4 (e). ^{11}B NMR for Li_3BS_3 and $Li_{2.95}B_{0.95}Si_{0.05}S_3$ show a single, quadrupolar environment that can be assigned to the $[BS_3]^{3-}$ moieties [69, 74]. The signal for the amorphous

phase a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ shows a similar signal to that of the crystalline phases but the shape changes, similarly to the previous measurement for amorphous Li_3BS_3 [69]. Li_3BS_3, $Li_{2.95}B_{0.95}Si_{0.05}S_3$, and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ all exhibit a major peak at $\sim$ 60 ppm and a relatively minor peak $\sim$ 0 ppm. The major peak is assigned to trigonal planar $[BS_3]^{3-}$ while the minor peak likely indicates a minor impurity with tetrahedrally coordinated B. [75–77] The change in the shape of the ^{11}B spectrum upon amorphization is likely due to an averaging of the quadrupolar couplings due to the fast Li dynamics. Thus, Li_3BS_3 and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ have similar local structures and we can attribute the faster Li dynamics to the introduction of charge-carrying defects. Although investigation of interfacial stability is beyond the scope of the model, we note that the Si-substituted Li_3BS_3 is a promising candidate for future investigations into interfacial stability. Work by Park *et al.* suggests that the (010) facet for Li_3BS_3 has a conduction band minimum 0.5 eV above the Li/Li^+ couple [65]. Since decomposition of Li_3BS_3 is likely to be mediated by electron injection from Li, their results suggest that thermodynamic stability can be engineered via orientation. From a kinetic perspective, high-temperature DFT-MD simulations show no mobility for B and S, suggesting large kinetic diffusion barriers [63]. Since decomposition of Li_3BS_3 would entail the diffusion of these species, the reaction may be sluggish or wholly precluded. Interfacial stability has been previously demonstrated for a glassy electrode in the Li–B–S–Si–O phase space [78]. This result may indicate that stability can be engineered into Si-substituted Li_3BS_3 by partial isovalent substitution of O for S. Finally, recently-synthesized Li–B–S–X (X = Cl, Br, I) quaternaries have exhibited promising conductivities [79]. With similar elemental composition, the Si-substituted Li_3BS_3 may be a good candidate for a multi-electrolyte architecture with the halide-containing quaternaries [13]. In addition to our experimental model validation, another of the predicted materials, KLi_6TaO_6, was recently synthesized with aliovalent Sn-substitution by Suzuki *et al*. [80]. With a reported ionic conductivity near 10^{-5} S cm^{-1}, KLi_6TaO_6 is better than 70% of the SSEs in the semi-supervised labels. Further improvement may be possible *via* extended amorphization to introduce structural defects, as is observed for Li_3BS_3.

2.4 Conclusions

Identification of functional materials is critical for improving technologies. Here, we show the utility of using semi-supervised learning as a method for guiding next-generation materials discovery in emerging fields. The method's focus on identifying

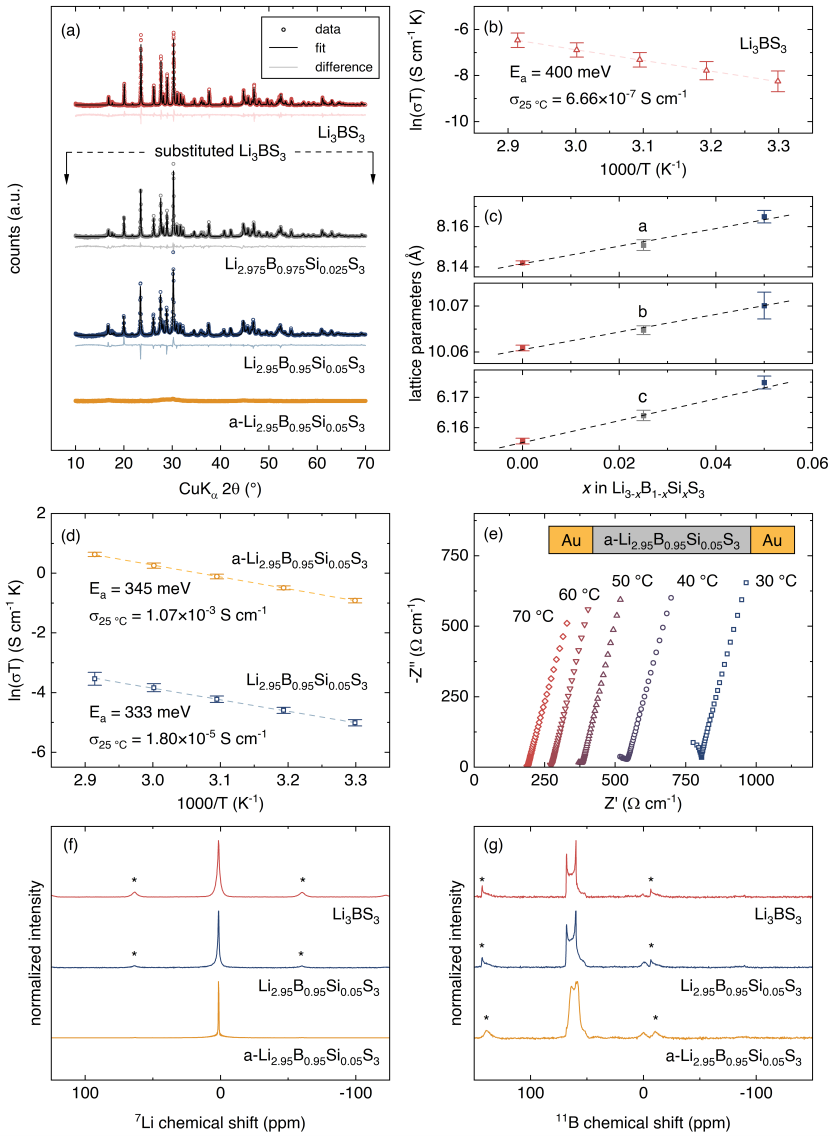

Figure 2.4: Characterization of Li_3BS_3 with vacancy engineering. (a) XRD patterns for Li_3BS_3, 2.5% Si substituted Li_3BS_3 ($Li_{2.975}B_{0.975}Si_{0.025}S_3$), 5% Si substituted Li_3BS_3 ($Li_{2.95}B_{0.95}Si_{0.05}S_3$), and amorphized 5% Si substituted Li_3BS_3 (a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$). No impurities are observed in any pattern. (b) Arrhenius fits for Li_3BS_3. (c) Lattice parameter comparison for Li_3BS_3, $Li_{2.975}B_{0.975}Si_{0.025}S_3$, and $Li_{2.95}B_{0.95}B_{0.05}S_3$. (d) Arrhenius fits for $Li_{2.95}B_{0.95}Si_{0.05}S_3$, and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$. (e) Electrochemical impedance spectroscopy for a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ at various temperatures. (f) ^{7}Li NMR and (g) ^{11}B NMR of Li_3BS_3, $Li_{2.95}B_{0.95}Si_{0.05}S_3$, and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$. Results show that combined aliovalent substitution and amorphization can improve the ionic conductivity of Li_3BS_3 by approximately four orders of magnitude.

the relationships between descriptors, prior to labeling, enables understanding of compositional spaces where most inputs are unlabeled. We demonstrate how semi-supervised learning can be used to identify descriptors correlated with superionic conductivity in Li SSEs. By analyzing all Li-containing structures from the ICSD and MP database, we identify 212 materials that show promise as SSEs. All 212 structures exhibit a BVSE-predicted E_a below 0.6 eV.

The results illustrate why careful screening of descriptors is useful when identifying new materials. While chemical intuition can be useful for descriptor selection, chemical intuition is often biased to favor previously investigated compositional spaces. For material discovery in emerging fields, the use of handpicked descriptors may miss complex phenomena that more generally describe the dataset. Descriptor screening reveals which material properties are correlated to a property of interest to help enhance chemical intuition. In the case of Li SSEs, spatial descriptors excel over compositional, bonding, and electronic descriptors: the Smooth Overlap of Atomic Positions (SOAP), modified X-ray diffraction (mXRD), and general density descriptors are within the top four models. For spatial descriptors, simplification of the input structure tends to improve clustering outcomes. Removing the mobile ions from the structure and simplifying the remaining atoms, i.e. the "CAN" simplification, is most effective. Thus, the placement of framework atoms, but not their precise identity, is most correlated with ionic conductivity. Specifying the mobile ion positions hurts the model performance, suggesting a low correlation of mobile ion positions with ionic conductivity.

Predictions from the semi-supervised method are promising starting points for the experimental identification of new superionic conductors but defects must be considered. The proposed materials are diverse, with the top thirty including halides, sulfides, tellurides, nitrides, oxides, and oxyhalides (see Appendix A.6). As a structure that falls outside of the eight routinely studied SSE classes, we demonstrate experimental characterization of Li3BS3 to confirm the utility of the approach. However, pure Li_3BS_3 exhibits poor ionic conductivity. Defects must be introduced into the material to achieve a superionic conductivity above 10^{-3} S cm^{-1}, a value that surpasses most reported SSEs. We note that the defects are introduced while maintaining the local structure of the crystalline material and thus the ionic conduction pathways are likely similar. The need to introduce defects highlights the paramount importance that defects play when measuring real materials. Many of the highest-performing SSEs contain charge-carrying defects that are not explicitly

encoded in their structure files. It is likely that some of the descriptors indirectly encode information about defects. By using experimental conductivity values as the evaluation metric, we may be prioritizing descriptors that encode information about a structure's ability to support charge-carrying defects. Although Li_3BS_3 is a poor conductor, it is clearly able to support charge-carrying defects. The large conductivity difference between pristine Li_3BS_3 and a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ highlights the importance of these defects. To improve predictive models and enhance chemical intuition, descriptors that explicitly encode defects are needed.

Now developed, the semi-supervised learning approach can serve as a template for material discovery beyond Li SSEs. The code is thoroughly documented following pythonic coding standards and made freely available on GitHub. Although the present effort focuses on Li SSEs, the approach is applicable to any material discovery space where labels are sparse. The discovery of new Li cathodes could be accomplished by using Li diffusivity, cathode capacity, and metal redox couple voltages as labels. The discovery of divalent SSEs (*e.g.*, Mg^{2+}, Ca^{2+}, Zn^{2+}) could foreseeably be accomplished in a similar manner. The semi-supervised learning strategy may accelerate identification of fast ionic conductors for ion exchange membranes, solid oxide fuel cells, and various sensor applications.

2.5 Methods

Data processing and semi-supervised learning

The ~26,000 input compositions are exported from the Inorganic Crystalline Structure Database (ICSD v.4.4.0) and Material's Project (MP – v.2020.09.08) as crystallographic information files (.cif). All structures containing Li are imported. Although transition metals could produce undesirable redox activity, transition metal-containing structures are not screened out. Some of the best-performing SSEs contain transition metals (*e.g.,* LLZO and LLTO). Entries that exist in both ICSD and MP are merged. Data manipulations and structure simplifications are performed using the Python libraries NumPy (v1.19.1), Pandas (v1.0.5), ASE (v3.19.1), and Pymatgen (v2020.8.3). Descriptor transformations are performed using the Python libraries Pymatgen (v2020.8.3), Matminer (v0.6.3), and Dscribe. Agglomerative hierarchical clustering is performed using the Python library scipy (v1.5.0). All code has been successfully executed on a custom-built CPU with an AMD Ryzen Threadripper 3990x Processor and 256 GB of RAM, in Ubuntu 20.04 running on Windows Subsystem for Linux 2. All code is made available on the GitHub

(https://github.com/FALL-ML/materials-discovery).

CI-NEB

Migration barriers for Li-ion hopping are evaluated with the Climbing Image – Nudged Elastic Band (CI-NEB) method as implemented in the QuantumESPRESSO PWneb software package [81–84]. Density-functional theory (DFT) calculations are performed using the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation functional and projector-augmented wave (PAW) sets [85, 86]. Convergence testing for the kinetic-energy cutoff of the plane-wave basis and the k-point sampling is performed for each structure to ensure an accuracy of 1 meV per atom. The lattice parameters and atomic positions of the as-retrieved structure are optimized. Supercells are created for each structure that are a minimum of 10 Å in each lattice direction to minimize interactions between periodic images of the mobile ion. To study the migration barrier in the dilute limit, a single Li vacancy is created in the boundary endpoint structures of each studied pathway. A uniform background charge is used to balance excess charge. Each boundary configuration is relaxed until the force on each atom is less than 3×10^{-4} eV Å^{-1}. Images are created by linearly interpolating framework atomic positions between the initial and final boundary configurations. The initial pathway for the mobile ion is generated from the BVSE output minimum energy pathway to promote faster convergence of the NEB calculation. An NEB force convergence threshold of 0.05 eV Å^{-1} is used. The calculation is first converged using the default NEB algorithm and then restarted with the CI scheme to allow for the maximum energy of the pathway to be determined.

Li_3BS_3 synthesis

Li_3BS_3 is synthesized by reaction of Li_2S (Alfa Aesar, 99.9%), S_8 (Acros Organics, >99.5%), and elemental B (SkySpring Nanomaterials, Inc. 99.99%). The reactants are first mixed stoichiometrically (300 rpm for 1 h) using a planetary ball mill (MSE PMV1-0.4L) in 50 mL ZrO_2 jars with ZrO_2 balls. Two grams of reactants are always combined with 2 large balls (10 mm diameter), 34 medium balls (5 mm diameter), and 8 grams of small balls (3 mm diameter). Loading of ball mill jars occurs in an Ar-filled glovebox (Mbraun) and the jars are sealed before removal. After the 1 h of milling, the precursor mixture is pumped back into the glovebox and

330–340 mg of the powder is loaded into carbon-coated vitreous silica ampoules (10 mm ID × 12 mm OD). The ampoules are evacuated (<10 mtorr) prior to sealing. Pure Li_3BS_3 is obtained *via* a four-step heating protocol in a Lindberg/Blue furnace: (1) ramp to 500 °C at 5 °C min^{-1}, (2) hold at 500 °C for 12 h, (3) ramp to 800 °C at 5 °C min^{-1}, and (4) hold at 800 °C for 6 h. The hot melt is then quenched from 800 °C into room-temperature water. Recovered ingots are typically covered in an amorphous shell. The shell is either sanded off or the ingot is ground into smaller pieces and the shell is manually removed.

Substituted Li_3BS_3

Aliovalent substitution is accomplished by adding elemental Si (Acros, 99+%) into the precursor mixture prior to the 1 h mix. Si-substitution stoichiometry assumed that each Si atom replaces one Li and B: $Li_{3-x}B_{1-x}Si_xS_3$. Aside from the addition of Si, all steps are the same as for the synthesis of Li_3BS_3. Amorphization is accomplished *via* extended planetary ball milling in Ar of the 5% Si-substituted Li_3BS_3 ($Li_{2.95}B_{0.95}Si_{0.05}S_3$). Approximately 1 g of $Li_{2.95}B_{0.95}Si_{0.05}S_3$ is combined in a ZrO_2 ball mill jar with 3 large balls (10 mm diameter), 51 medium balls (5 mm diameter), and 12 g of small balls (3 mm diameter). The powder is ground in a planetary ball mill (MSE PMV1-0.4L), under an Ar atmosphere, for 100 h.

Material characterization

Li_3BS_3 materials are characterized using powder X-ray diffraction (XRD) and electrochemical impedance spectroscopy (EIS). XRD patterns are attained on a Rigaku Smartlab by scanning from 10° to 70° 2θ at 2 degrees per minute. The Smartlab employs a Cu-Kα source with a 20 kV accelerating voltage. For EIS measurements, 50–100 mg of powder is first hot-pressed (100 °C, 5 min) into a 1/4" diameter pellet. The pellet faces are polished using diamond lapping powder (Allied High Tech Products Inc.) in sequentially finer grits: 60, 30, 6, 0.5, and 0.1 microns. Au contacts are sputtered (90 s at 40 mA) onto the polished surfaces using a 108 Auto Sputter Coater (Cressington). Pellets are then assembled into a Swagelok 1/4" cell with stainless steel current collectors. After applying pressure with a hand vise (~100 MPa), EIS data is collected on a VSP-300 with a Biologic low-current channel. All EIS data is collected to an upper frequency of 3 MHz. The lower frequency is case dependent, with a frequency cutoff selected such that the Warburg

polarization feature is visible. ^{7}Li and ^{11}B MAS MAS NMR spectra were acquired using a Bruker DSX-500 spectrometer with a 4 mm ZrO_2 rotor. The operating frequencies for ^{7}Li and ^{11}B are 190.5 and 160.5 MHz, respectively. The ^{7}Li and ^{11}B spectra were referenced to a 1 M LiCl aq. solution and BF_3-OEt_2, respectively. A spinning speed of 12 kHz was used, and the spectra were gathered after applying a single 0.5 μs to 15° pulse for both ^{7}Li and ^{11}B.

2.6 Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. The authors declare no competing financial interest.

2.7 Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

2.8 Acknowledgements

F. A. L. L. acknowledges the support of the Arnold and Mabel Beckman Foundation via a 2020 Arnold O. Beckman Postdoctoral Fellowship in Chemical Sciences. F. A. L. L. would also like to thank Andrew J. Martinolich for his guidance and insightful scientific input. The NEB computations presented here were conducted in the Resnick High Performance Computing Center, a facility supported by Resnick Sustainability Institute at the California Institute of Technology. K. A. S. acknowledges support from the David and Lucile Packard Foundation.

2.9 Bibliography

[1] Bachman, J. C.; Muy, S.; Grimaud, A.; Chang, H.-H.; Pour, N.; Lux, S. F.; Paschos, O.; Maglia, F.; Lupart, S.; Lamp, P.; Giordano, L.; Shao-Horn, Y. Inorganic Solid-State Electrolytes for Lithium Batteries: Mechanisms and Properties Governing Ion Conduction. *Chem. Rev.* **2016**, *116*, 140–162.

[2] Kamaya, N.; Homma, K.; Yamakawa, Y.; Hirayama, M.; Kanno, R.; Yonemura, M.; Kamiyama, T.; Kato, Y.; Hama, S.; Kawamoto, K.; Mitsui, A. A Lithium Superionic Conductor. *Nature Mater* **2011**, *10*, 682–686.

[3] Adeli, P.; Bazak, J. D.; Park, K. H.; Kochetkov, I.; Huq, A.; Goward, G. R.; Nazar, L. F. Boosting Solid-State Diffusivity and Conductivity in Lithium Superionic Argyrodites by Halide Substitution. *Angewandte Chemie International Edition* **2019**, *58*, 8681–8686.

[4] Seino, Y.; Ota, T.; Takada, K.; Hayashi, A.; Tatsumisago, M. A Sulphide Lithium Super Ion Conductor Is Superior to Liquid Ion Conductors for Use in Rechargeable Batteries. *Energy Environ. Sci.* **2014**, *7*, 627–631.

[5] Zhu, Y.; He, X.; Mo, Y. First Principles Study on Electrochemical and Chemical Stability of Solid Electrolyte–Electrode Interfaces in All-Solid-State Li-ion Batteries. *J. Mater. Chem. A* **2016**, *4*, 3253–3266.

[6] Richards, W. D.; Miara, L. J.; Wang, Y.; Kim, J. C.; Ceder, G. Interface Stability in Solid-State Batteries. *Chem. Mater.* **2016**, *28*, 266–273.

[7] Kerman, K.; Luntz, A.; Viswanathan, V.; Chiang, Y.-M.; Chen, Z. Review—Practical Challenges Hindering the Development of Solid State Li Ion Batteries. *J. Electrochem. Soc.* **2017**, *164*, A1731–A1744.

[8] Wenzel, S.; Randau, S.; Leichtweiß, T.; Weber, D. A.; Sann, J.; Zeier, W. G.; Janek, J. Direct Observation of the Interfacial Instability of the Fast Ionic Conductor Li $_{10}$ GeP $_{2}$ S $_{12}$ at the Lithium Metal Anode. *Chem. Mater.* **2016**, *28*, 2400–2407.

[9] Zhu, Y.; He, X.; Mo, Y. Origin of Outstanding Stability in the Lithium Solid Electrolyte Materials: Insights from Thermodynamic Analyses Based on First-Principles Calculations. *ACS applied materials & interfaces* **2015**, *7*, 23685–23693.

[10] Wenzel, S.; Sedlmaier, S. J.; Dietrich, C.; Zeier, W. G.; Janek, J. Interfacial Reactivity and Interphase Growth of Argyrodite Solid Electrolytes at Lithium Metal Electrodes. *Solid State Ionics* **2018**, *318*, 102–112.

[11] Ding, Z.; Li, J.; Li, J.; An, C. Interfaces: Key Issue to Be Solved for All Solid-State Lithium Battery Technologies. *Journal of The Electrochemical Society* **2020**, *167*, 070541.

[12] Wang, S.; Fang, R.; Li, Y.; Liu, Y.; Xin, C.; Richter, F. H.; Nan, C.-W. Interfacial Challenges for All-Solid-State Batteries Based on Sulfide Solid Electrolytes. *Journal of Materiomics* **2021**, *7*, 209–218.

[13] Sendek, A. D.; Cheon, G.; Pasta, M.; Reed, E. J. Quantifying the Search for Solid Li-Ion Electrolyte Materials by Anion: A Data-Driven Perspective. *J. Phys. Chem. C* **2020**, *124*, 8067–8079.

[14] Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-Throughput Experiments. *Science advances* **2018**, *4*, eaaq1566.

[15] Tran, K.; Ulissi, Z. W. Active Learning across Intermetallics to Guide Discovery of Electrocatalysts for CO2 Reduction and H2 Evolution. *Nature Catalysis* **2018**, *1*, 696–703.

[16] Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nature communications* **2018**, *9*, 3405.

[17] Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nature communications* **2016**, *7*, 1–9.

[18] 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 Environ. Sci.* **2017**, *10*, 306–320.

[19] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

[20] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *Journal of Materiomics* **2017**, *3*, 159–177.

[21] Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful Classification of Crystal Structures Using Deep Learning. *Nature communications* **2018**, *9*, 2775.

[22] Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature communications* **2017**, *8*, 15679.

[23] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nature communications* **2017**, *8*, 13890.

[24] Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; others Matminer: An Open Source Toolkit for Materials Data Mining. *Computational Materials Science* **2018**, *152*, 60–69.

[25] Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Computer Physics Communications* **2020**, *247*, 106949.

[26] Juan, Y.; Dai, Y.; Yang, Y.; Zhang, J. Accelerating Materials Discovery Using Machine Learning. *Journal of Materials Science & Technology* **2021**, *79*, 178–190.

[27] Suzuki, K.; Ohura, K.; Seko, A.; Iwamizu, Y.; Zhao, G.; Hirayama, M.; Tanaka, I.; Kanno, R. Fast Material Search of Lithium Ion Conducting Oxides Using a Recommender System. *Journal of Materials Chemistry A* **2020**, *8*, 11582–11588.

[28] Wang, Z.; Lin, X.; Han, Y.; Cai, J.; Wu, S.; Yu, X.; Li, J. Harnessing Artificial Intelligence to Holistic Design and Identification for Solid Electrolytes. *Nano Energy* **2021**, *89*, 106337.

[29] Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; others Accelerated Discovery of CO2 Electrocatalysts Using Active Machine Learning. *Nature* **2020**, *581*, 178–183.

[30] Wang, Z.; Zhang, H.; Li, J. Accelerated Discovery of Stable Spinels in Energy Systems via Machine Learning. *Nano Energy* **2021**, *81*, 105665.

[31] Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated Discovery of Large Electrostrains in BaTiO3-based Piezoelectrics Using Active Learning. *Advanced materials* **2018**, *30*, 1702884.

[32] Li, J.; Zhang, Y.; Cao, X.; Zeng, Q.; Zhuang, Y.; Qian, X.; Chen, H. Accelerated Discovery of High-Strength Aluminum Alloys by Machine Learning. *Communications Materials* **2020**, *1*, 73.

[33] Butler, K. T.; Oviedo, F.; Canepa, P. *Machine Learning in Materials Science*; American Chemical Society, 2022; Vol. 29.

[34] Zhang, Y.; He, X.; Chen, Z.; Bai, Q.; Nolan, A. M.; Roberts, C. A.; Banerjee, D.; Matsunaga, T.; Mo, Y.; Ling, C. Unsupervised Discovery of Solid-State Lithium Ion Conductors. *Nature communications* **2019**, *10*, 5260.

[35] Liu, Y.; Zhou, Q.; Cui, G. Machine Learning Boosting the Development of Advanced Lithium Batteries. *Small Methods* **2021**, *5*, 2100442.

[36] Forestier, G.; Wemmert, C. Semi-Supervised Learning Using Multiple Clusterings with Limited Labeled Data. *Information Sciences* **2016**, *361*, 48–65.

[37] Thangadurai, V.; Narayanan, S.; Pinzaru, D. Garnet-Type Solid-State Fast Li Ion Conductors for Li Batteries: Critical Review. *Chemical Society Reviews* **2014**, *43*, 4714–4727.

[38] Gorai, P.; Famprikis, T.; Singh, B.; Stevanovic, V.; Canepa, P. Devil Is in the Defects: Electronic Conductivity in Solid Electrolytes. *Chemistry of Materials* **2021**, *33*, 7484–7498.

[39] Van Engelen, J. E.; Hoos, H. H. A Survey on Semi-Supervised Learning. *Machine learning* **2020**, *109*, 373–440.

[40] Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Advances in neural information processing systems* **2017**, *30*.

[41] Artrith, N.; Urban, A. An Implementation of Artificial Neural-Network Potentials for Atomistic Materials Simulations: Performance for TiO2. *Computational Materials Science* **2016**, *114*, 135–150.

[42] Adams, S.; Rao, R. P. In *Bond Valences. Structure and Bonding*; Brown, I. D., Poeppelmeier, K. R., Eds.; Springer: Berlin, Heidelberg, 2014; Vol. 158; pp 129–159.

[43] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The journal of physical chemistry letters* **2015**, *6*, 2326–2331.

[44] Butler, MA.; Ginley, DS. Prediction of Flatband Potentials at Semiconductor-Electrolyte Interfaces from Atomic Electronegativities. *Journal of the Electrochemical Society* **1978**, *125*, 228.

[45] He, B.; Ye, A.; Chi, S.; Mi, P.; Ran, Y.; Zhang, L.; Zou, X.; Pu, B.; Zhao, Q.; Zou, Z.; others CAVD, towards Better Characterization of Void Space for Ionic Transport Analysis. *Scientific Data* **2020**, *7*, 153.

[46] Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including Crystal Structure Attributes in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Physical Review B* **2017**, *96*, 024104.

[47] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Computational Materials Science* **2013**, *68*, 314–319.

[48] Deml, A. M.; O'Hayre, R.; Wolverton, C.; Stevanović, V. Predicting Density Functional Theory Total Energies and Enthalpies of Formation of Metal-Nonmetal Compounds by Linear Regression. *Physical Review B* **2016**, *93*, 085142.

[49] Ewald, P. P. Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. *Annalen der physik* **1921**, *369*, 253–287.

[50] Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Hight Walker, A. R.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A.; others The Joint Automated Repository for Various Integrated Simulations (JARVIS) for Data-Driven Materials Design. *npj computational materials* **2020**, *6*, 173.

[51] Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Physical Review B* **2014**, *89*, 094104.

[52] Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine Learning Reveals Orbital Interaction in Materials. *Science and technology of advanced materials* **2017**, *18*, 756–765.

[53] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical review letters* **2012**, *108*, 058301.

[54] Faber, F.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *International Journal of Quantum Chemistry* **2015**, *115*, 1094–1101.

[55] Krivovichev, SV. Structural Complexity of Minerals: Information Storage and Processing in the Mineral World. *Mineralogical Magazine* **2013**, *77*, 275–326.

[56] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Computational Materials* **2016**, *2*, 1–7.

[57] Ward Jr, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American statistical association* **1963**, *58*, 236–244.

[58] Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Computational Materials* **2021**, *7*, 84.

[59] He, X.; Zhu, Y.; Mo, Y. Origin of Fast Ion Diffusion in Super-Ionic Conductors. *Nature communications* **2017**, *8*, 15893.

[60] Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Science advances* **2016**, *2*, e1600225.

[61] Jain, A.; Hautier, G.; Moore, C. J.; Ong, S. P.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations. *Computational Materials Science* **2011**, *50*, 2295–2310.

[62] Vinatier, P.; Ménétrier, M.; Levasseur, A. Structure and Ionic Conduction in Lithium Thioborate Glasses and Crystals. *Physics and Chemistry of Glasses* **2003**, *44*, 135–142.

[63] Sendek, A. D.; Antoniuk, E. R.; Cubuk, E. D.; Ransom, B.; Francisco, B. E.; Buettner-Garrett, J.; Cui, Y.; Reed, E. J. Combining Superionic Conduction and Favorable Decomposition Products in the Crystalline Lithium–Boron–Sulfur System: A New Mechanism for Stabilizing Solid Li-Ion Electrolytes. *ACS Applied Materials & Interfaces* **2020**, *12*, 37957–37966.

[64] Bianchini, F.; Fjellvåg, H.; Vajeeston, P. A First-Principle Investigation of the Li Diffusion Mechanism in the Super-Ionic Conductor Lithium Orthothioborate Li3BS3 Structure. *Materials Letters* **2018**, *219*, 186–189.

[65] Park, H.; Yu, S.; Siegel, D. J. Predicting Charge Transfer Stability between Sulfide Solid Electrolytes and Li Metal Anodes. *ACS Energy Letters* **2020**, *6*, 150–157.

[66] Vinatier, P.; Gravereau, P.; Ménétrier, M.; Trut, L.; Levasseur, A. Li3BS3. *Crystal Structure Communications* **1994**, *50*, 1180–1183.

[67] Zhang, L.; Zhang, X.; Tian, G.; Zhang, Q.; Knapp, M.; Ehrenberg, H.; Chen, G.; Shen, Z.; Yang, G.; Gu, L.; others Lithium Lanthanum Titanate Perovskite as an Anode for Lithium Ion Batteries. *Nature communications* **2020**, *11*, 3490.

[68] Belous, AG.; Novitskaya, GN.; Polyanetskaya, SV.; Gornikov, Y. I. Investigation into Complex Oxides of La 2/3-x Li 3x TiO 3 Composition. *Izv. Akad. Nauk SSSR, Neorg. Mater* **1987**, *23*, 470–472.

[69] Kimura, T.; Inoue, A.; Nagao, K.; Inaoka, T.; Kowada, H.; Sakuda, A.; Tatsumisago, M.; Hayashi, A. Characteristics of a Li3BS3 Thioborate Glass Electrolyte Obtained via a Mechanochemical Process. *ACS Applied Energy Materials* **2022**, *5*, 1421–1426.

[70] Zhou, L.; Minafra, N.; Zeier, W. G.; Nazar, L. F. Innovative Approaches to Liargyrodite Solid Electrolytes for All-Solid-State Lithium Batteries. *Accounts of chemical research* **2021**, *54*, 2717–2728.

[71] Zhao, W.; Yi, J.; He, P.; Zhou, H. Solid-State Electrolytes for Lithium-Ion Batteries: Fundamentals, Challenges and Perspectives. *Electrochemical energy reviews* **2019**, *2*, 574–605.

[72] Lacivita, V.; Artrith, N.; Ceder, G. Structural and Compositional Factors That Control the Li-ion Conductivity in LiPON Electrolytes. *Chemistry of Materials* **2018**, *30*, 7077–7090.

[73] Knauth, P. Inorganic Solid Li Ion Conductors: An Overview. *Solid State Ionics* **2009**, *180*, 911–916.

[74] Larink, D.; Eckert, H.; Martin, S. W. Structure and Ionic Conductivity in the Mixed-Network Former Chalcogenide Glass System [Na2S] 2/3 [(B2S3) x (P2S5) 1–x] 1/3. *The Journal of Physical Chemistry C* **2012**, *116*, 22698–22710.

[75] Hwang, S.-J.; Fernandez, C.; Amoureux, J. P.; Han, J.-W.; Cho, J.; Martin, SW.; Pruski, M. Structural Study of x Na2S+(1- x) B2S3 Glasses and Polycrystals by Multiple-Quantum MAS NMR of 11B and 23na. *Journal of the American Chemical Society* **1998**, *120*, 7337–7346.

[76] Kaup, K.; Bazak, J. D.; Vajargah, S. H.; Wu, X.; Kulisch, J.; Goward, G. R.; Nazar, L. F. A Lithium Oxythioborosilicate Solid Electrolyte Glass with Superionic Conductivity. *Advanced Energy Materials* **2020**, *10*, 1902783.

[77] Curtis, B.; Francis, C.; Kmiec, S.; Martin, S. W. Investigation of the Short Range Order Structures in Sodium Thioborosilicate Mixed Glass Former Glasses. *Journal of Non-Crystalline Solids* **2019**, *521*, 119456.

[78] Seino, Y.; Takada, K.; Kim, B.-C.; Zhang, L.; Ohta, N.; Wada, H.; Osada, M.; Sasaki, T. Synthesis and Electrochemical Properties of Li2S–B2S3–Li4SiO4. *Solid State Ionics* **2006**, *177*, 2601–2603.

[79] Kaup, K.; Assoud, A.; Liu, J.; Nazar, L. F. Fast Li-Ion Conductivity in Superadamantanoid Lithium Thioborate Halides. *Angewandte Chemie International Edition* **2021**, *60*, 6975–6980.

[80] Suzuki, N.; Lee, J.; Masuoka, Y.; Ohta, S.; Kobayashi, T.; Asahi, R. Theoretical and Experimental Studies of KLi6TaO6 as a Li-ion Solid Electrolyte. *Inorganic Chemistry* **2021**, *60*, 10371–10379.

[81] Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *The Journal of chemical physics* **2000**, *113*, 9901–9904.

[82] Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *The Journal of chemical physics* **2000**, *113*, 9978–9985.

[83] Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; others QUANTUM

ESPRESSO: A Modular and Open-Source Software Project for Quantum-simulations of Materials. *Journal of physics: Condensed matter* **2009**, *21*, 395502.

[84] Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M. B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; others Advanced Capabilities for Materials Modelling with Quantum ESPRESSO. *Journal of physics: Condensed matter* **2017**, *29*, 465901.

[85] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical review letters* **1996**, *77*, 3865.

[86] Dal Corso, A. Pseudopotentials Periodic Table: From H to Pu. *Computational Materials Science* **2014**, *95*, 337–350.

Chapter 3

SUBSTITUTION OF Li_3BS_3: REVEALING NEW SUPERIONIC CONDUCTOR PHASES AND THE SIGNIFICANCE OF CRYSTALLINITY

McHaffie, D. B.; Bienz, J. M.; Hwang, S.-J.; Laskowski, F. A. L.; See, K. A. Substitution of Li_3BS_3: Revealing New Superionic Conductor Phases and the Significance of Crystallinity. *In preparation.* **D.B.M. and J.M.B contributed equally to this work.**

[This chapter is temporarily embargoed.]

Chapter 4

CLASSIFICATION OF (DIS)ORDERED STRUCTURES AS SUPERIONIC LITHIUM CONDUCTORS

McHaffie, D. B.; Iton, Z. W. B.; Bienz, J. M.; Laskowski, F. A. L.; See, K. A. Classification of (Dis)Ordered Structures as Superionic Lithium Conductors with an Experimental Structure–Conductivity Database. *Digital Discovery* **2025**, *4* (6), 1518–1533.

4.1 Abstract

Solid-state electrolytes (SSEs) are critical for the development of high-performance all-solid-state batteries. Data-driven efforts to discover novel SSEs have been constrained by the absence of databases linking ionic conductivity with structure, as well as by challenges in encoding structural information for the disorder that is often found in superionic conductors. Here, we construct the largest database to date of experimentally measured ionic conductivity values paired with corresponding crystal structures, comprising 548 Li-containing compounds. Graph-based features, derived using a transfer learning framework, enable learning directly from disordered crystals, and AtomSets models leveraging these features outperform domain-specific features in a classification task. These models are employed to screen the Inorganic Crystal Structure Database (ICSD) and Materials Project for superionic Li-containing compounds. We identify 241 compounds with predicted superionic conductivity and band gaps greater than 1 eV. Experimental validation confirming superionic conductivity in of one of these candidates, $Li_9B_{19}S_{33}$, demonstrates the utility of this approach for the discovery and development of advanced SSEs for all-solid-state batteries.

4.2 Introduction

All-solid-state batteries represent a transformative frontier in energy storage technology, offering the potential for enhanced safety and performance compared to conventional lithium-ion batteries [1–4]. However, the realization of their full potential hinges critically upon the discovery and development of solid-state electrolytes (SSEs) exhibiting high ionic conductivity, low electronic conductivity, stability against both Li metal anodes and highly oxidative cathodes, and suitable mechan-

ical properties. The multi-objective search is further complicated by the observed trade-off between the conductivity and stability in commonly studied SSE material families [5–10]. Discovery of novel SSE materials is necessary to optimize these desired properties [8, 11].

To expedite the exploration for suitable SSEs, researchers have increasingly explored the integration of statistical and machine learning approaches [8, 12–31]. These methodologies often rely on databases consisting of compounds labelled with their experimental ionic conductivity (σ_{exp}) or related quantities such as migration energy (E_m), serving as the foundation for training predictive models. Features derived from these compounds serve as inputs to the models, encapsulating information about the material's composition and/or structure. Models trained solely on composition information have successfully predicted various material properties in other domains [32–36]. Such an approach has also been explored for predicting ionic conductivity, as demonstrated by Hargreaves et al., who achieved high performance using a composition-only model trained on 403 unique compositions [26]. However, since compounds are featurized by composition only, the model is unable to distinguish between polymorphs. Additionally, the ionic conductivity in solid-state materials is inherently linked to their crystal structure, as the arrangement of atoms and the pathways available for ion migration directly influence ion mobility. Structural features can capture information about coordination environments, atomic positions, and the potential for site disorder, all of which are critical in determining ion transport properties.

Incorporating structure-based information to identify fast ion conductors with data-driven methods has historically encountered two primary challenges: the lack of comprehensive datasets that provide both ionic conductivity values and corresponding crystal structures, and inadequate methods for representing the prevalent disorder in many fast ion conductors. In this context, disorder refers to the occurrence of atomic sites within a crystal structure that are not fully occupied by a single element. Instead, the sites are populated by a set of possible chemical species, with the partial occupancy describing the fraction of sites occupied by each species in the long-range average structure. Sendek et al. trained a logistic regression model capable of predicting if a material would exhibit superionic conductivity using interpretable structural features [15]. However, their training set contained only 40 entries, preventing evaluation with a holdout test set. Moreover, their probabilistic sampling method for feature construction for disordered compounds may become

computationally expensive when applied to large collections of known compounds such as the Inorganic Crystal Structure Database (ICSD), in which 6,860 of 11,925 Li-containing materials exhibit disorder (v5.2.0). Our own previous study implemented a semi-supervised learning strategy using a database of 219 ionic conductivity values and corresponding ICSD crystal structures. The structural descriptors used in the previous work were unable to represent disordered compounds, limiting the utilization to a subset that could be ordered through a costly supercell ordering procedure [27]. Excluding highly disordered compounds from consideration is particularly undesirable when searching for novel fast ion conductors, as site disorder is known to be a critical factor in realizing high conductivity in many systems [37–43]. The recently reported COSNet framework introduced by Wang et al., has shown promise in combining structural and compositional information using multimodal ensemble learning to predict material properties, including ionic conductivity [44]. However, the effectiveness of this approach for representing structural disorder was not explicitly examined in the study.

In the current work, we alleviate the data scarcity challenge by constructing the largest repository to date of 548 crystal structures and σ_{exp} values. To address the issue of disordered crystal structure representation, we use transfer-learned graph-based features. Graph-neural networks (GNNs) have emerged as powerful architectures for incorporating composition and crystal structure information to make property predictions [45–47]. Additionally, by representing the node features for disordered sites through combinations of elemental embeddings, GNNs have been used for learning from disordered compounds [48]. However, the flexibility of these models enabled by their vast number of trainable parameters necessitates thousands of labelled data points to be trained correctly [49–51]. To circumvent the issue, we represent our data with features derived from GNNs pre-trained on large datasets (e.g., Materials Project formation energies) and pass these graph-based features through comparatively simple multilayer perceptron (MLP) models using the AtomSets framework developed by Chen and Ong [50]. Such an approach has been demonstrated to achieve higher performance than GNNs for smaller datasets of similar size to our own [50, 52]. To further overcome the challenges of a small dataset, we implement transfer learning by pre-training AtomSets models on the MatBench metal classifier dataset containing 106,113 samples and use the trained weights to initialize the network weights for ionic conductivity prediction [53]. The AtomSets models are tasked with classifying input materials as superionic ($\sigma_{\text{exp}} > 10^{-4}$ S cm^{-1}) or not.

We examine the efficacy of representing disordered compounds using a linear combination of elemental embeddings and graph-based features. For comparison, ordered configurations are generated using a supercell approach. The performance of classification models trained using these two methods is found to be nearly equivalent. The optimal feature representation and model configurations for this task are explored through k-fold and leave-one-cluster-out (LOCO) cross-validation (CV). Transfer-learned features derived from early graph-convolutional layers of the parent model, which encode short-range structural information, achieve the highest performance for out-of-cluster predictions. Reducing the chemical diversity by replacing atoms with representative species improves extrapolation beyond the training set. A final ensemble of 100 AtomSets models is shown to achieve high test performance and is used to evaluate all Li-containing materials in the Inorganic Crystal Structure Database (v5.2.0) and Materials Project (v2023.11.1). An additional criterion requiring the electronic band gap (E_g) to be greater than 1 eV is used to prioritize compounds more likely to be electronically insulating, a critical property for SSEs. The screening identifies 241 compounds predicted to be superionic with $E_g > 1$ eV. To show the practical relevance of our approach, we experimentally validate superionic conductivity in one of these candidate phases, $Li_9B_{19}S_{33}$, achieving a σ_{exp} of 4.1×10^{-4} S cm^{-1}.

4.3 Results and Discussion

Structure-conductivity database. The database created for this study is comprised of experimental ionic conductivity values for 548 distinct Li-containing compounds and their corresponding crystal structures sourced from the Inorganic Crystal Structure Database (ICSD). All ionic conductivity measurements recorded are obtained from electrochemical impedance spectroscopy (EIS) data. The database includes ionic conductivity values that are both directly extracted from text and digitized from figures in reference sources. In solid-state ionics literature, particularly for low conductivity materials, measurements are frequently performed at elevated temperatures and presented in the form of Arrhenius-type plots where $\ln(\sigma T)$ or $\log_{10}(\sigma T)$ is plotted against T^{-1}. To capture these data, plots are digitized, and conductivity values are extrapolated to room temperature using an Arrhenius relationship. The resulting room-temperature conductivity values, along with the lowest measured temperature, are recorded in the database. To facilitate the inclusion of both structure and composition information for model training, conductivity values are paired with corresponding crystal structures. Wherever possible, crystallographic informa-

tion files (CIFs) associated with conductivity measurements are obtained from the ICSD using article DOIs. That is, reports are identified which included both conductivity measurements and sufficient structural characterization to generate an ICSD entry. Since the same nominal compound (i.e. $Li_{10}GeP_2S_{12}$) can have different lattice parameters, atomic positions, or defect concentrations depending on preparation conditions, direct matching of the CIF with the measured sample is prioritized. For articles containing conductivity measurements but lacking ICSD entries, associated crystal structures are identified by manual inspection. Only articles with sufficient structural characterization to enable matching of stoichiometry, space group, and lattice parameters to existing ICSD entries are included. Articles without structural characterization or containing conductivity values for non-crystalline compounds are excluded from the dataset. For a comprehensive list of ionic conductivity values corresponding to compounds without ICSD entries, readers are referred to the database compiled by Laskowski et al. [27]. Structures deemed identical within a specified tolerance are identified. In cases of multiple ionic conductivity values for identical structures, the entry corresponding to the median ionic conductivity is retained and duplicate entries are removed. Notably, this process preserves highly related structures, necessitating diverse forms of CV to assess model performance, as elaborated in subsequent sections. To ensure database accuracy, the database is constructed by a single author and is verified by the other authors. Any discrepancies found during the verification process are reviewed by a third author for validation. A summary of the database created in this work is presented in Figure 4.1. The compiled database contains a broad range of ionic conductivity values from crystal structures with 72 different space groups. However, certain space groups are more represented due to the bias in SSE material research which has primarily been confined to garnets, LISICON-type structures, argyrodites, NASICON-type structures, Li–nitrides, Li–hydrides, perovskites, and Li–halides [6]. Importantly, from the histogram in Figure 4.1 it is evident that most structures in the database, and especially those corresponding to materials with high conductivity, are disordered, further motivating the use of a compatible structural representation.

Experimental ionic conductivity measurements of the same compound with EIS can vary significantly across different laboratories [54]. Such variability has been attributed to inadequate control of sample temperature, sample geometry, the frequency range measured, choice of metal contact materials, and aging effects [54]. Extrapolating conductivity measurements performed at high-temperature to room-temperature is an additional source of error. The use of experimental conductivity

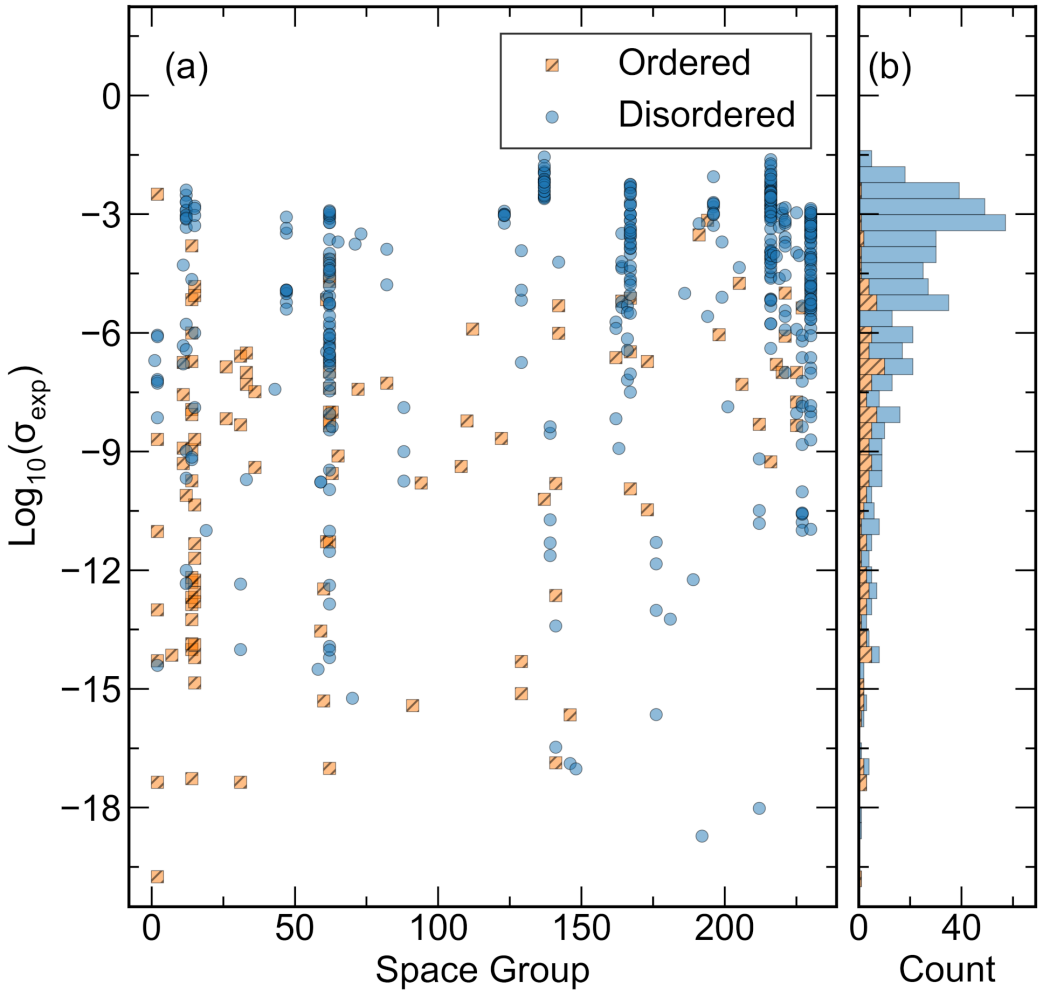

Figure 4.1: (a) The space group and corresponding Li-ion conductivity (σ) values plotted as $\log_{10}(\sigma_{\text{exp}})$ for each database entry. The database contains entries from 72 different space groups, with σ_{exp} values spanning over 10 orders of magnitude. (b) A histogram of the data in (a) showing the distribution of $\log_{10}(\sigma_{\text{exp}})$. Most superionic compounds contain site disorder, necessitating an appropriate featurization method. Note that seven compounds with $\sigma_{\text{exp}} < 10^{-20}$ S cm^{-1} are excluded from this figure for ease of visualization.

values from our database would thus introduce considerable noise into the training of a regression model. Thus, herein we do not endeavor to predict ionic conductivity but rather determine if a material is likely to be a good conductor or not. Framing a materials discovery problem as a classification task can enhance the prediction accuracy for identifying extraordinary compounds [32]. Classification models are designed to distinguish between distinct categories, allowing them to more effectively handle the binary nature of identifying extraordinary versus ordinary materials. In contrast, regression models predict continuous values, which can

introduce greater uncertainty and error, particularly when extrapolating beyond the training data.

The supervised learning performed in this study involves training a classifier neural network to determine if an input crystalline compound will exhibit superionic Li conductivity ($\sigma_{\text{exp}} > 10^{-4}$ S cm^{-1}). Table 4.1 provides summary statistics for the dataset used in this study. From the 548 labels, 10% are removed at the outset of this work and set aside as a final test set. The remainder of the data is used to determine optimal feature representations and hyperparameters using various CV techniques.

Table 4.1: A summary of the structure-conductivity database.

Description	Number
σ_{exp} values with crystal structure	571
Unique structures	548
Space groups	72
Ordered compounds	112
Disordered compounds	436
Positive class ($\sigma_{\text{exp}} \geq 10^{-4}$ S cm^{-1})	211
Negative class ($\sigma_{\text{exp}} < 10^{-4}$ S cm^{-1})	337

Training with disordered representations using AtomSets framework. Input compounds are transformed into a graphs following the MatErials Graph Network (MEGNet) formalism outlined by Chen et al. [45]. A graph is defined as $G = (\mathbf{u}, V, E)$ where $\mathbf{u}$, V, and E are the global state, atom (node), and bond (edge) attributes, respectively. A comprehensive description of the MEGNet architecture can be found in the original works [45, 48]. The graph representations are subjected to a specified number of graph-convolution (GC) layers within the pre-trained parent MEGNet model, after which atom features are extracted and provided as inputs for the AtomSets models. Within GC layers of the parent model, information is passed between atom, bond, and state vectors. Consequently, atom features following GC layers implicitly encapsulate both compositional and structural information, with a greater number of GC layers encoding longer-range interactions [50]. The AtomSets models accept the atom feature matrix $\mathbf{V}$ with dimensions $\mathbf{N}_a \times \mathbf{N}_f$ where $\mathbf{N}_a$ is the number of atoms in the structure and $\mathbf{N}_f$ is the number of features [50]. Consistent with the methodology implemented by Chen et al., the node feature for a disordered site is derived as a linear combination of elemental embeddings for the constituent elements, weighted by their reported occupancy. That is, $\mathbf{W}_{\text{disordered}} = \sum_i \mathrm{x}_i \cdot \mathbf{W}_{Z_i}$, where x_i is the reported site occupancy of element i and $\mathbf{W}_{Z_i}$ denotes the learned elemental embedding for the element with atomic number Z_i [48]. For the present

study, $\mathbf{W}_{Z_i}$ are learned embedding vectors of length 16 from a MEGNet model trained on 133,420 structures and their formation energies from the Materials Project database, downloaded on April 1, 2019. Importantly, this strategy for representing disorder does not consider possible occupancy correlations between disordered sites, instead treating each site independently. While the following analysis demonstrates that this approximation is sufficient for predicting superionic conductivity, we expect that other applications (e.g. force predictions between atoms) may require additional considerations to handle interactions between correlated sites.

The performance of models employing a linear combination of elemental embeddings is evaluated against those using ordered representations. To create the comparison set, ordered configurations without Li atoms are generated and ranked using the OrderDisorderedStructureTransformation in the Python Materials Genomics (Pymatgen) package, with the configuration exhibiting the lowest calculated Ewald energy selected for each structure [55]. Only disorder of the non-Li atoms is considered for this comparison because the extensive disorder in the mobile ion sublattice makes supercell generation computationally prohibitive for the entire dataset. An illustration of the two strategies to create graph representations from disordered crystals is shown in Figure 4.2 (a). AtomSets classification models, tasked with discerning whether an input structure is superionic, are trained using both ordered and linear combination of elemental embeddings representations. A comparative analysis is presented in Figure 4.2 (b) and (c) where the average area under the precision-recall curve (AUC-PR) and Matthew's correlation coefficient (MCC) assessed under k-fold CV for each model is shown over 500 training epochs. The AUC-PR is chosen as it provides a comprehensive evaluation of the model's precision and recall across different thresholds and is particularly well-suited for classification tasks with imbalanced datasets [56]. The AUC-PR score ranges from 0 to 1, with a perfect classifier obtaining a score of 1. MCC offers a balanced measure of classification performance, accounting for both true positives and true negatives, thereby providing a robust metric for our binary classification task [57–59]. The MCC score ranges from -1 to 1, where 1 indicates perfect agreement between predicted and actual labels and -1 indicates total disagreement between predicted and actual labels. As in the work by Hargreaves et al., these metrics are compared against those obtained from shuffled and mean controls, where predicted values are generated either by randomly shuffling the dataset labels or by using the training set mean as the prediction label [26]. Models trained with both ordered and disordered representations achieve AUC-PR and MCC scores significantly higher than those of

the controls and comparable performance levels by 500 training epochs. The results demonstrate that the linear combination of elemental embeddings representation enables similar efficacy to the ordered representation without necessitating the computationally intensive ordering transformation. Given the substantial computational costs associated with creating ordering configurations, which can scale combinatorially with the number of disordered sites and possible substitutions, the ability to use a disordered representation while maintaining performance parity offers expedited training [60]. Moreover, this capability facilitates efficient screening of experimental databases containing disordered compounds such as the ICSD, where over half of Li-containing compounds exhibit site disorder.

Feature and model evaluation. The present study explores two distinct feature engineering strategies: (1) the number of GC layers in the parent MEGNet model through which the graph is passed before the atom features are extracted and (2) input structure simplifications prior to graph generation. Models are trained using atom feature matrices $\mathbf{V_i}$ (i = 0, 1, 2, 3) where $\mathbf{V_0}$ is the atom feature matrix comprised solely of the learned elemental embeddings from the parent model and $\mathbf{V_i}$ (i = 1, 2, 3) denote the atom feature matrices after passing the graph through i GC layers. By nature of the message passing in each GC layer, higher i atom features encode longer-range interactions. The second feature engineering technique of pre-processing input structures before feature generation has been demonstrated to enhance learning outcomes for Li-ion conductor datasets [17, 27]. Laskowski et al. found that simplifying compounds by replacing categories of atoms with representative species and removing the position of the mobile ion improved clustering efficacy of known Li-ion conductors [27]. To evaluate this strategy within the model architecture under investigation, we explore structural modifications involving changes to the cations (**C**), anions (**A**), mobile Li ions (**M**), and neutral atoms (**N**) within the structures. Specifically, we investigate the following representations:

- CAMN: retaining all atom types
- CAN: removing the mobile Li ion
- CAMNS: retaining all atom types but simplifying the structure by substituting cations with Al, anions with S, and neutral species with Mg
- CANS: removing the mobile Li ion and performing the same substitutions as in CAMNS

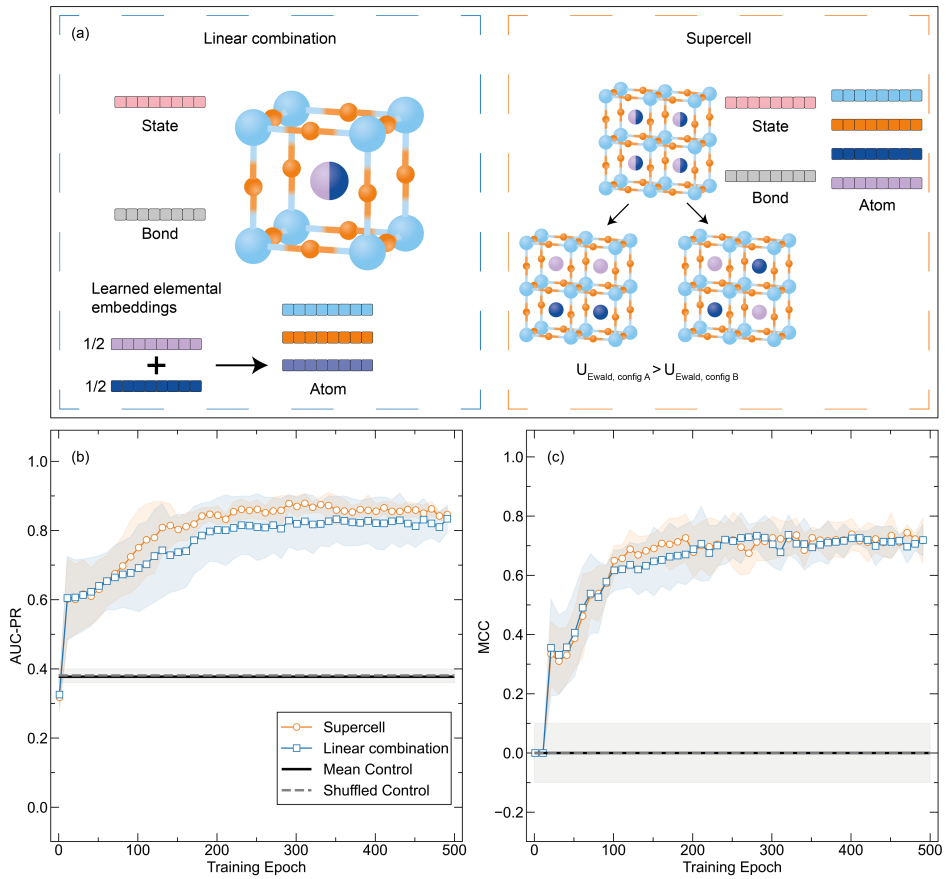

Figure 4.2: Different strategies to represent disordered structures. (a) On the left, the atom attributes are equal to a linear combination of elemental embeddings learned from a MEGNet model trained on a large database of Materials Project formation energies. On the right, ordered supercell configurations are generated. Configurations are compared using an Ewald summation and the lowest-energy configuration is used for graph creation. (b) The average area under the precision-recall curve (AUC-PR) and (c) Matthew's correlation coefficient (MCC) for AtomSets models trained with graph representations generated through the two approaches. Metrics are averaged over 5-fold random CV with the shaded regions indicating the standard deviation. Controls from randomly shuffling and using the mean of the training set as the predicted values are plotted as horizontal lines. Both methods for representing disordered structures offer comparable performance that exceeds the controls.

To compare the model performance for the different feature representations, we use both k-fold validation and LOCO CV. Experimental training data can exhibit a highly clustered distribution due to the inherent nature of scientific exploration — parent materials are systematically perturbed through various means (e.g. elemental substitution) to develop structure-property relationships, resulting in a large number of training data points confined to a relatively small number of parent structure

frameworks. The clustering of data can lead to the inclusion of highly related compounds in both training and validation sets when data is randomly segregated. Therefore, randomized k-fold validation provides insight into a model's interpolation ability but offers limited information regarding its capacity to predict in unseen chemical spaces. Predictive models intended for materials discovery also require an evaluation of their extrapolative capabilities. We assess this using LOCO CV, a clustering-based validation method for assessing a model's ability to predict on chemically distinct compounds not present in the training set [26, 61]. The dataset is clustered into n clusters using a chosen embedding presenting the chemical nature of the compounds and a clustering algorithm. Training is conducted on the compounds belonging to n-1 clusters and the model performance is evaluated on the compounds from the remaining cluster. In this work, we adhere to the procedure described by Hargreaves et al. [26]. The compounds in our labelled database are embedded using ElMD, a metric which captures the chemical similarity between compounds based on their chemical composition. Uniform Manifold Approximation and Projection (UMAP) is applied to obtain a low-dimensional representation that retains essential chemical relationships. Density-based spatial clustering of applications with noise (DBSCAN) is used to automate separation of the data into clusters for LOCO CV [26]. Detailed statistics for each cluster generated for this validation procedure are provided in the Table C.3. The effectiveness of data segregation from this clustering technique is examined by analyzing the compositional similarity between the entries in the test and validation sets and those in the training set. The results of this analysis displayed in Tables C.4, C.5, and C.6 show that except for one fold, the LOCO CV validation sets exhibit significantly lower compositional similarity with the training set compared to the k-fold validation or test sets. This motivates the use of LOCO-CV as our primary evaluation technique when comparing feature representations and performing hyperparameter optimization.

An additional benchmark for our AtomSets-based model is provided by comparing its performance with that of a logistic regression model trained using the database created in this work and a set of interpretable atomistic features defined by Sendek et al. [15]. This serves to validate our approach against a method that has been previously applied to the task of identifying SSE candidates using structure-based representations. The atomistic features used in the previous work include the average number of Li-Li bonds per Li atom in the crystal, the ionic character of bonds within the sublattice, the anion coordination environment, the shortest distance between Li ions and anions, and the shortest distance between Li ions [15]. These features were

chosen to encode information directly impacting Li mobility, potential pathways for ion conduction, and the ease of ion movement through the lattice. Detailed definitions for each feature are provided in the original work. Figure 4.3 depicts the AUC-PR and MCC for AtomSets models trained using $\mathbf{V_i}$ (i = 0, 1, 2, 3) atom features with different structural simplifications in addition to the logistic regression model trained using the atomistic feature set. All model variations are trained using the same dataset and k folds. Both the AtomSets and logistic regression models achieve higher AUC-PR and MCC than the controls, indicating significant predictive power. However, all variations of the AtomSets model outperform the logistic regression model that is based on atomistic features. For the CAN and CAMN representations, using $\mathbf{V_0}$ features achieves the highest performance, suggesting that composition-only information in the form of the learned elemental embeddings is sufficient for classifying ionic conductors in this dataset when assessed under k-fold CV. The CANS representation attains the lowest performance for all $\mathbf{V_i}$, but higher performance is enabled by i > 0 which incorporates longer-range structural information through additional graph convolutions. These highlights suggest that AtomSets models using transfer-learned features are able to better capture the complex relationships influencing ionic conductivity, leading to higher classification accuracy.

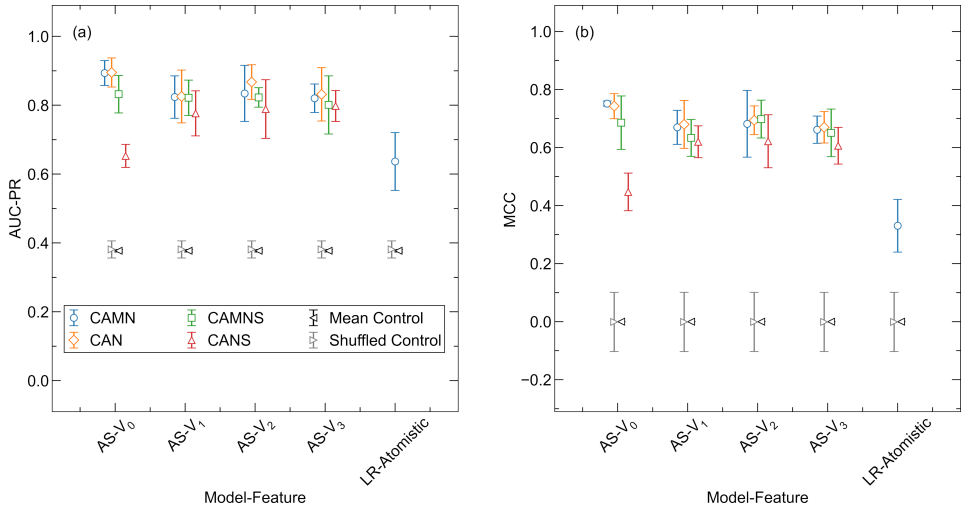

Figure 4.3: Classification performance of model-feature combinations assessed with k-fold cross-validation. (a) AUC-PR and (b) MCC of AtomSets (AS) models with graph-based atom features ($\mathbf{V_0}$ to $\mathbf{V_3}$) and a logistic regression model using atomistic features. Four different structural simplifications are shown (CAMN, CAN, CAMNS, and CANS) with mean and randomly shuffled controls. The symbol locations indicate the mean from random 5-fold cross validation and error bars represent the standard deviation.

LOCO CV is used as a complimentary method for evaluating model and feature representations in the context of materials discovery. Different from the case of k-fold CV, controls are calculated separately for each cluster due to the significant variation in the ratio of positive to negative labels across clusters (as shown in Table C.3). We perform hyperparameter optimization separately for each $\mathbf{V_i}$ and each validation cluster. For all subsequent results, figures, and discussion, we report and compare the metrics from the highest-performing hyperparameter configurations for each representation. This approach ensures that each representation and validation cluster is evaluated based on its optimal hyperparameter settings, allowing for a consistent comparison of performance. To capture the variance of the models, the metric value averaged over ten repeated runs is presented at the average best epoch across all folds. Reporting the variance in this way offers insight into the model performance under conditions akin to those encountered in materials discovery scenarios, where a final model is trained for a specified number of epochs before being used as a screening tool. For the logistic regression model, optimization of the regularization penalty term is performed in a similar manner and the results from the best value are shown for comparison to the AtomSets models. Figures 4.4 (a) and (c) depict the validation AUC-PR and MCC for AtomSets models trained with each $\mathbf{V_i}$, the logistic regression model, and controls across all clusters. The logistic regression model with atomistic features performs significantly worse than the AtomSets models, showing comparable MCC to the randomized and mean controls. No single $\mathbf{V_i}$ outperforms all others for every cluster, despite all surpassing the random and shuffled controls. The averaged AUC-PR and MCC scores across all clusters for each $\mathbf{V_i}$ are illustrated in Figures 4.4 (b) and (d). Descriptors capturing short range interactions ($\mathbf{V_0}$, $\mathbf{V_1}$) provide slightly higher classification performance than those derived from more GC layers. A similar finding was reported in the original AtomSets work where models trained using features from early GC layers exhibited higher accuracy across a variety of prediction tasks [50]. Additionally, it is observed that contrary to k-fold validation results, the averaged metrics for $\mathbf{V_1}$ are higher than those for $\mathbf{V_0}$, with the average AUC-PR and MCC being (0.86, 0.61) for $\mathbf{V_1}$ and (0.85, 0.58) $\mathbf{V_0}$, respectively. These findings suggest that incorporating some short-range structural information can enhance the model's ability to classify ion conductors with chemistry different from the training set beyond composition-only information. We note that while LOCO CV is designed to evaluate model extrapolation by grouping compounds based on chemistry, automated clustering does not always preserve chemically intuitive boundaries. For instance, clusters 6 and 7 both

include argyrodites with comparable compositions, potentially contributing to the higher observed performance.

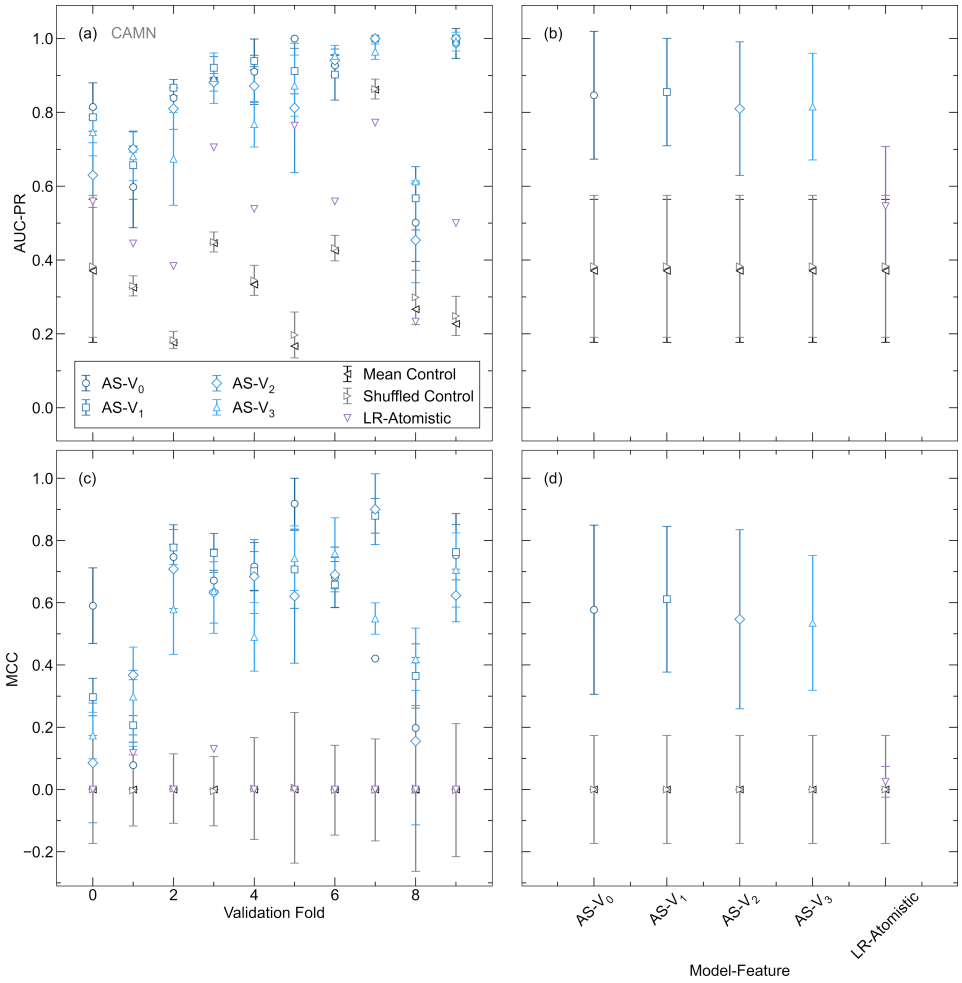

Figure 4.4: Classification performance comparison of different features assessed with leave-one-cluster-out cross validation. (a) AUC-PR and (c) MCC for each validation cluster of pre-trained AtomSets (AS) models with graph-based atom features ($\mathbf{V_0}$ to $\mathbf{V_4}$) and a logistic regression model using atomistic features. The average from ten repeated training runs with the optimal hyperparameters for each validation cluster are shown. Mean and shuffled controls are calculated for each validation cluster. (b) AUC-PR and (d) MCC from the optimal hyperparameter set for each model-feature combination averaged across all validation clusters. Error bars indicate the standard deviation. Metrics are from the best epoch across all runs and validation clusters.

Figure 4.5 illustrates the performance of AtomSets models trained using $\mathbf{V_1}$ atom features constructed from the CAMN, CAN, CAMNS, and CAN structure representations. The CAMN and CAMNS representation enable learning that surpasses the control tests for all validation clusters. Removing the mobile atom yields inferior

performance, with the CAN representation exhibiting slightly worse MCC compared to the controls for validation cluster 0, and the CANS representation showing lower MCC than controls for validation clusters 0 and 1. This emphasizes the value of the graph-based featurization in incorporating structural information while including the disordered mobile atom sites. The mobile ion sublattice typically constitutes the source of disorder in these compounds, and it is evident that neglecting these sites due to inadequate representation would overlook crucial information for prediction. Notably, the CAMNS representation, where the identity of the cation, anion, and neutral species remains constant for all compounds, achieves nearly the same performance as the model trained on the nominal structures (CAMN) while exhibiting lower variation between clusters. Most representations exhibit lower predictive performance on validation clusters 0 and 1, which primarily consist of garnets and other oxides. This may be due to the unique structural characteristics and ionic conduction mechanisms in these materials, which are more challenging for the models to capture compared to other clusters.

The best hyperparameter configuration is different between chosen validation clusters as shown in Table C.8. To introduce diversity in the final model parameters and reduce overfitting to one specific validation set, an ensemble comprised of AtomSets models with CAMNS-$\mathbf{V_1}$ features is trained with the most effective hyperparameters for each of the ten validation clusters. Variation for each model configuration is captured by training ten models for each parameter set, resulting in a total of 100 AtomSets models within the ensemble. The performance of the final ensemble is examined using the test partition, which is separate from the data used for the above k-fold and LOCO CV. It is noted that the test partition, while separate from training data, was partitioned randomly, similar to k-fold validation. This approach does not fully assess extrapolation to distinct chemistries, a limitation examined by LOCO CV in this study. Figure 4.6 shows the probability of a compound being superionic (P_{SI}) where superionic is defined as $\sigma_{exp} > 10^{-4}$ S cm^{-1} with the $\log_{10}(\sigma_{\mathrm{exp}})$ for the test set. The final model ensemble achieves a AUC-PR of 0.86 and an MCC of 0.60. By contrast, the logistic regression model only achieves an AUC-PR of 0.80 and an MCC of 0.26, highlighting the superior performance of the AtomSets ensemble approach. Test set compounds that are missclassified all have σ_{exp} values less than two orders of magnitude from the decision boundary. Overall, the pre-trained Atomsets CAMNS-$\mathbf{V_1}$ models display significantly higher predictive power than control metrics, as assessed through k-fold CV, LOCO CV, and a separate test set. The strong performance on out-of-cluster inputs suggests that this model architecture is

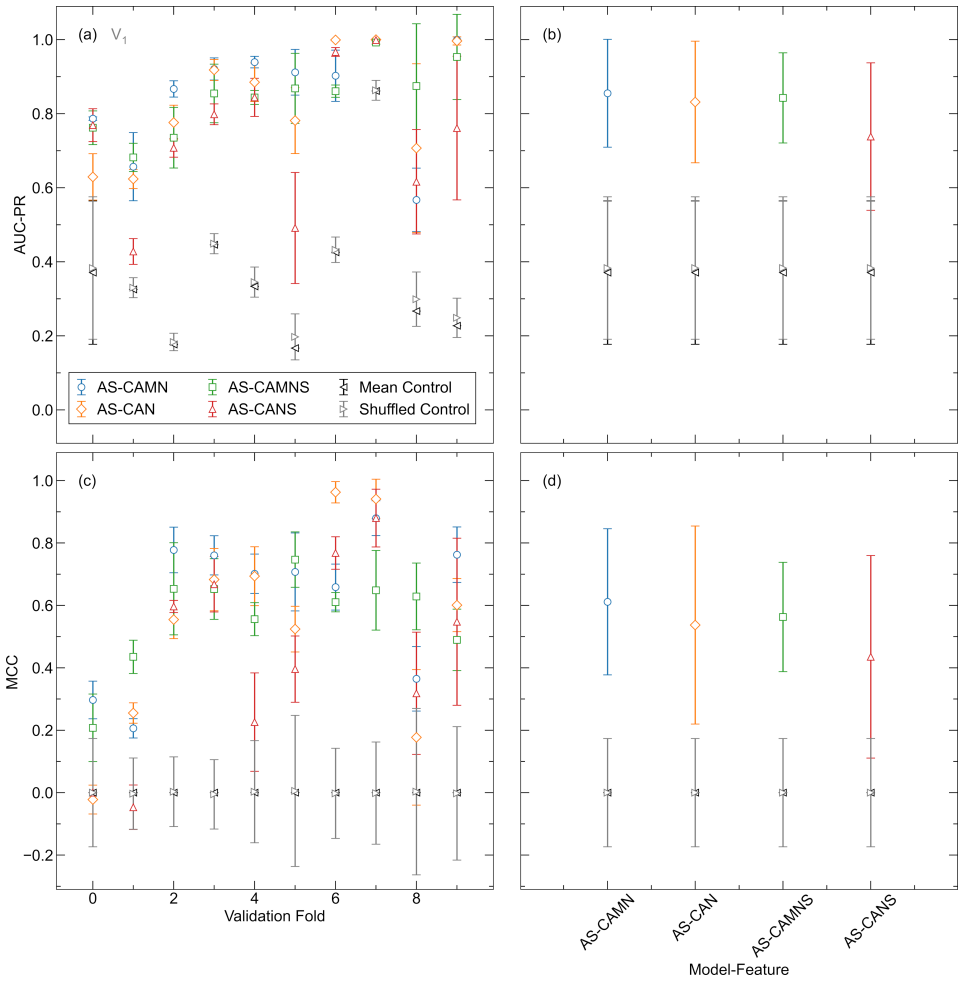

Figure 4.5: Classification performance of structural simplifications assessed with leave-one-cluster-out cross validation. (a) AUC-PR and (c) MCC for each validation cluster of pre-trained AtomSets (AS) models and $\mathbf{V_1}$ atom features for CAMN, CAN, CAMNS, and CANS structural simplifications. The average from ten repeated training runs with the optimal hyperparameters for each validation cluster are shown. Mean and shuffled controls are calculated for each validation cluster. (b) AUC-PR and (d) MCC from the optimal hyperparameter set for each model-feature combination averaged across all validation clusters. Error bars indicate the standard deviation. Metrics are from the best epoch across all runs and validation clusters.

well-suited for screening known Li-containing materials to discover novel fast ion conductors.

Screening of known Li-containing materials. All Li-containing materials present in the ICSD (v5.2.0) and Materials Project (v2023.11.1) are aggregated. Structures are featurized using the CAMNS structural simplification and $\mathbf{V_1}$ atom feature matrix. The P_{SI} is predicted for all compounds. To facilitate consideration of compounds as potential SSEs, the DFT-calculated E_g from the Materials Project

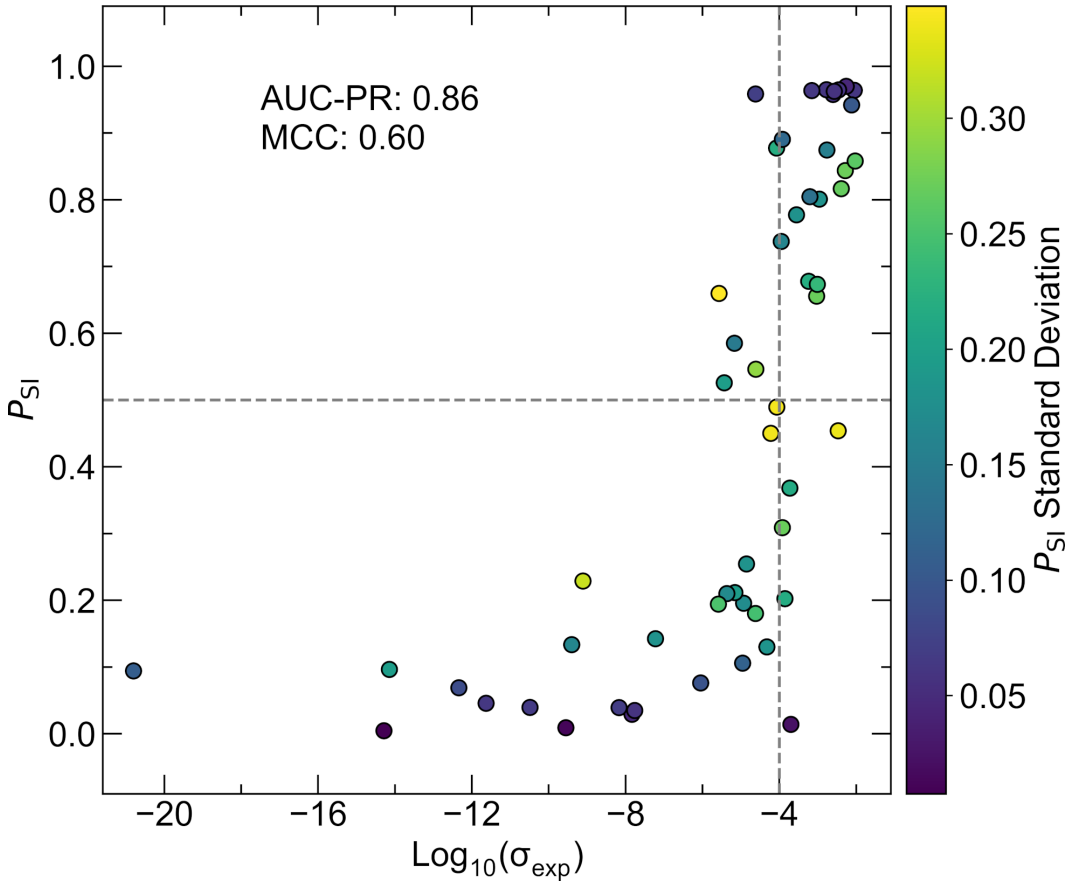

Figure 4.6: Test set evaluation of the AtomSets-V1 CAMNS model ensemble. The predicted likelihood of test set compounds exhibiting superionic conductivity (P_{SI}) is plotted against their reported $\log_{10}(\sigma_{exp})$. Dashed lines indicate boundaries for classification. The model ensemble achieves an AUC-PR of 0.86 and a MCC of 0.6. All incorrectly classified compounds have $\log_{10}(\sigma_{exp})$ values less than two orders of magnitude from the class boundary of 10^{-4} S cm^{-1}.

is retrieved if a corresponding ICSD entry can be identified. In cases where no matching entry exists in the Materials Project, the E_g is predicted using the MEGNet model developed by Chi et al. [45]. Compounds with E_g of less than 1 eV are excluded. The relatively low E_g for SSEs accounts for the systematic underestimation of experimental band gap values by approximately 40 percent in the Materials Project [62]. The MEGNet model is trained using Materials Project band gap data and so a similar systematic underestimation of experimental band gap values is expected. This value was chosen to balance the discovery of novel material families with practical considerations for electronic insulation.

A histogram of the P_{SI} for all 6,863 Li-containing materials with predicted $E_g > 1$ eV is shown in Figure 4.7. Most compounds are not predicted to be fast ion conductors

with 6,435 of 6,863 having P_{SI} less than 0.5. Of the 428 predicted to be superionic, 396 exhibit site disorder as highlighted in the inset of Figure 4.7 (a). This underscores the importance of choosing a compatible structural representation to ensure that disordered materials are retained in the screening process. The prediction confidence is quantified by the standard deviation of the ensemble P_{SI} and a calculated distance metric $d_{training}$. Lower standard deviations indicate greater agreement between ensemble models, increasing the confidence in the prediction. The distance metric is defined as the distance between the unlabelled compound and the nearest training sample in $\mathbf{N}_f$-dimensional space where $\mathbf{N}_f$ is the number of features in the atom feature matrix. Similar to previous work, we normalize the distances by the training data variance using principal component analysis (PCA) [15]. Figure C.2 shows the PCA embedding to two dimensions of the atom features for compounds in the training set, ICSD, and Materials Project. A smaller $d_{training}$ indicates that the prediction requires less extrapolation from the training data, increasing the confidence. Figure 4.7 (b) shows the P_{SI}, $d_{training}$ and P_{SI} standard deviation for each Li-containing compound in the ICSD with predicted E_g > 1 eV.

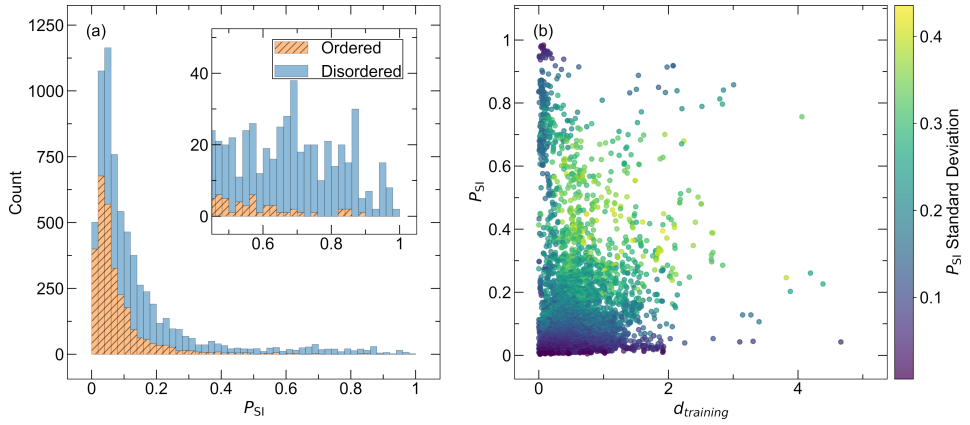

Figure 4.7: Results of screening Li-containing compounds in the ICSD using the AtomSets-V1 CAMNS model ensemble. (a) Histogram of the likelihood of superionic conductivity (P_{SI}) for ordered and disordered Li-containing compounds with predicted E_g > 1 eV. Inset shows region of high P_{SI} where most compounds are disordered. (b) P_{SI} vs. the distance from the nearest training sample $d_{training}$ for Li-containing materials with E_g > 1 eV.

To identify novel materials that could be interesting in battery applications, we filter out any compounds with chemical formula similar to those in our training set. Specifically, compounds whose normalized compositions have all constituent elements within five percent of any training sample composition are excluded. This screening results in 241 materials from the ICSD and Materials Project predicted to

be superionic with $E_g > 1$ eV. The ICSD compounds with the top 20 highest P_{SI} values are detailed in Table 4.2 for discussion. Intermetallic compounds with predicted $E_g >$ 1 eV are also omitted. The standard deviation of the ensemble predictions is provided in parentheses next to the P_{SI} in addition to the $d_{training}$. Among these candidates, conductivity measurements for five compounds were reported recently and were not captured during the database creation process. These values are included in Table 4.2. Although these compounds do not directly contribute to identifying new useful materials, they serve as additional validation of the model's effectiveness, as all were correctly classified based on the experimental measurements. While a measurement of $Li_{1.251}Cd_{1.671}In_{0.471}Cl_6$ could not be identified, its structure was described as resembling that of the high-temperature polymorph of $LiMnInCl_6$, which adopts a layered $CdCl_2$-type structure, with Li^+, Cd^{2+}, and In^{3+} ions randomly distributed across the octahedral sites [63]. $Li_2Zr_6MnCl_{15}$ is composed of abundant elements, has a straightforward reported synthesis method, and a high P_{SI} with low standard deviation, making it a strong candidate for experimental investigation [64]. In a recent computational study, LiP_5 was found to have the highest ionic conductivity of all known Li-P binaries, predicted to exceed 1 mS cm^{-1} at room temperature through molecular dynamics simulations [65]. The same study did not observe significant Li conduction in LiP_7. Nevertheless, the predictions from this study in addition to the work by Maltsev et al. suggests that these phases, particularly LiP_5, may warrant further investigation. The Li dynamics of B_xS_y compounds $Li_5B_7S_{13}$ and $Li_9B_{19}S_{33}$ studied via Li^7 nuclear magnetic resonance (NMR) have suggested high Li mobility and ab-initio molecular dynamics has also predicted high conductivity in these materials [66–68]. However, experimental measurements of the ionic conductivity are not reported in the literature. Another promising candidate, $LiBSi_2$, features an open tetrahedral framework with three-dimensional channels that may facilitate fast ion conduction [69]. Additional considerations such as the abundance or toxicity constituent elements could make candidates such as $Li_{6.55}Ga_{0.05}La_{2.91}Zr_2O_{12}$, $Li_7La_{1.8}Eu_{1.2}Zr_2O_{12}$, $Li_{6.43}Ga_{0.52}La_{2.67}Zr_2O_{12}$, LiCaAs, and $LiNdS_2$ less desirable. However these additional screening criteria are not applied for all compounds in the present work.

Experimental demonstration of $Li_9B_{19}S_{33}$. $Li_9B_{19}S_{33}$ is chosen for experimental characterization. Originally synthesized by Hiltmann et al., the crystal structure of $Li_9B_{19}S_{33}$ is composed of corner-sharing $B_{19}S_{36}$ units form large channels populated by highly disordered Li^+ cations, offering potential pathways for ion migration [74]. NMR studies by Bertermann et al. indicate anisotropic Li^+ diffusion within

Table 4.2: The top 20 candidate materials from the ICSD as ranked by the average P_{SI} from the AtomSets-V1 CAMNS model. Composition stoichiometries are rounded to two decimal places where appropriate.

Compound	ICSD Code	P_{SI} (SD)	E_{g} (eV)	$d_{training}$	σ_{exp} (mS cm^{-1})
$Li_{1.25}Cd_{1.67}In_{0.47}Cl_6$	98583	0.94 (0.09)	3.15	0.38	NA
$Li_2Zr_6MnCl_{15}$	71146	0.91 (0.13)	1.29	0.62	NA
$Li_{9.9}SnP_2S_{11.9}Cl_{0.1}$	48716	0.9 (0.13)	2.15	0.06	0.26[70]
LiP_5	23620	0.89 (0.15)	1.26*	1.54	NA
$Li_5B_7S_{13}$	143927	0.89 (0.16)	2.16	0.30	NA
$Li_{6.75}La_{2.75}Ca_{0.25}Zr_{1.5}Nb_{0.5}O_{12}$	63870	0.87 (0.12)	2.68	0.05	0.20[71]
$Li_{6.55}Ga_{0.05}La_{2.91}Zr_2O_{12}$	430602	0.86 (0.13)	2.47	0.05	NA
LiP_7	23621	0.84 (0.17)	1.65*	1.51	NA
LiCaAs	428102	0.84 (0.20)	1.1*	2.13	NA
$LiSrAlSb_2$	412654	0.83 (0.17)	1.01	1.95	NA
$LiBSi_2$	425643	0.83 (0.14)	1.17*	1.40	NA
$Li_{7.03}La_{2.87}Sr_{0.08}Zr_{1.39}Ta_{0.58}O_{12.22}$	45740	0.83 (0.20)	3.16	0.19	0.72[72]
$Li_9B_{19}S_{33}$	73151	0.82 (0.29)	2.27	0.29	NA
$Li_{6.41}La_{2.90}Sr_{0.10}Zr_{1.6}Mo_{0.4}O_{12}$	42738	0.81 (0.21)	2.74	0.14	0.33[73]
$Li_{0.5}ZrS_2$	642338	0.79 (0.26)	1.33	0.38	NA
$Li_{1.66}W_6I_{14}$	256678	0.79 (0.24)	1.13	2.52	NA
$Li_7La_{1.8}Eu_{1.2}Zr_2O_{12}$	27177	0.79 (0.24)	2.94	0.29	NA
$Li_{6.43}Ga_{0.52}La_{2.67}Zr_2O_{12}$	196425	0.79 (0.17)	2.27	0.09	NA
$LiNdS_2$	642202	0.78 (0.22)	1.5	2.21	NA
$Li_{7.10}La_{2.83}Sr_{0.16}Zr_{1.38}Ta_{0.61}O_{11.76}$	45741	0.78 (0.20)	3.16	0.19	0.85[72]

* Value retrieved from corresponding entries in the Materials Project. All other E_{g} values are predicted from the pre-trained MEGNet model.

these channels, associated with a low activation energy [67]. Computational work by Sendek et al. predicted that $Li_9B_{19}S_{33}$ possesses the widest electrochemical stability window and highest oxidative stability among the materials in the Li-B-S ternary phase space, including $Li_5B_7S_{13}$, Li_3BS_3, and $Li_2B_2S_5$ [68]. Experimental studies of materials in the Li-B-S ternary phase space are relatively rare in the context of fast ion conductors, partly due to synthesis challenges posed by the reactivity of their precursors with conventional reaction vessels and the difficulty in obtaining phase-pure products. In previous work, we developed a solid-state synthesis protocol for Li_3BS_3 using Li_2S, B, and S [27] that we find is readily adapted to the synthesis of $Li_9B_{19}S_{33}$. The powder X-ray diffraction (XRD) pattern and Rietveld refinement to the reported structure shown in Figure 4.8 (a) confirms phase-purity. Variable-temperature EIS is used to characterize the ionic conductivity of of $Li_9B_{19}S_{33}$. Although challenges with densification yield a pellet that is only 78% of the theoretical density, the material demonstrates a conductivity of 4.1×10^{-4} S cm^{-1}. The slope of the Arrhenius plot of $\ln(\sigma T)$ versus T^{-1} presented in Figure 4.8 (b) yields an activation energy E_a of 364 meV. Although improved pelletization is expected to increase conductivity, these findings nevertheless affirm the superionic conductivity of $Li_9B_{19}S_{33}$, a candidate identified by the model ensemble. True experimental validation of this approach's predictive capabilities would require the

synthesis and characterization of a significant number of the identified candidates. However, this task is beyond the scope of a single group and is not pursued in the present work.

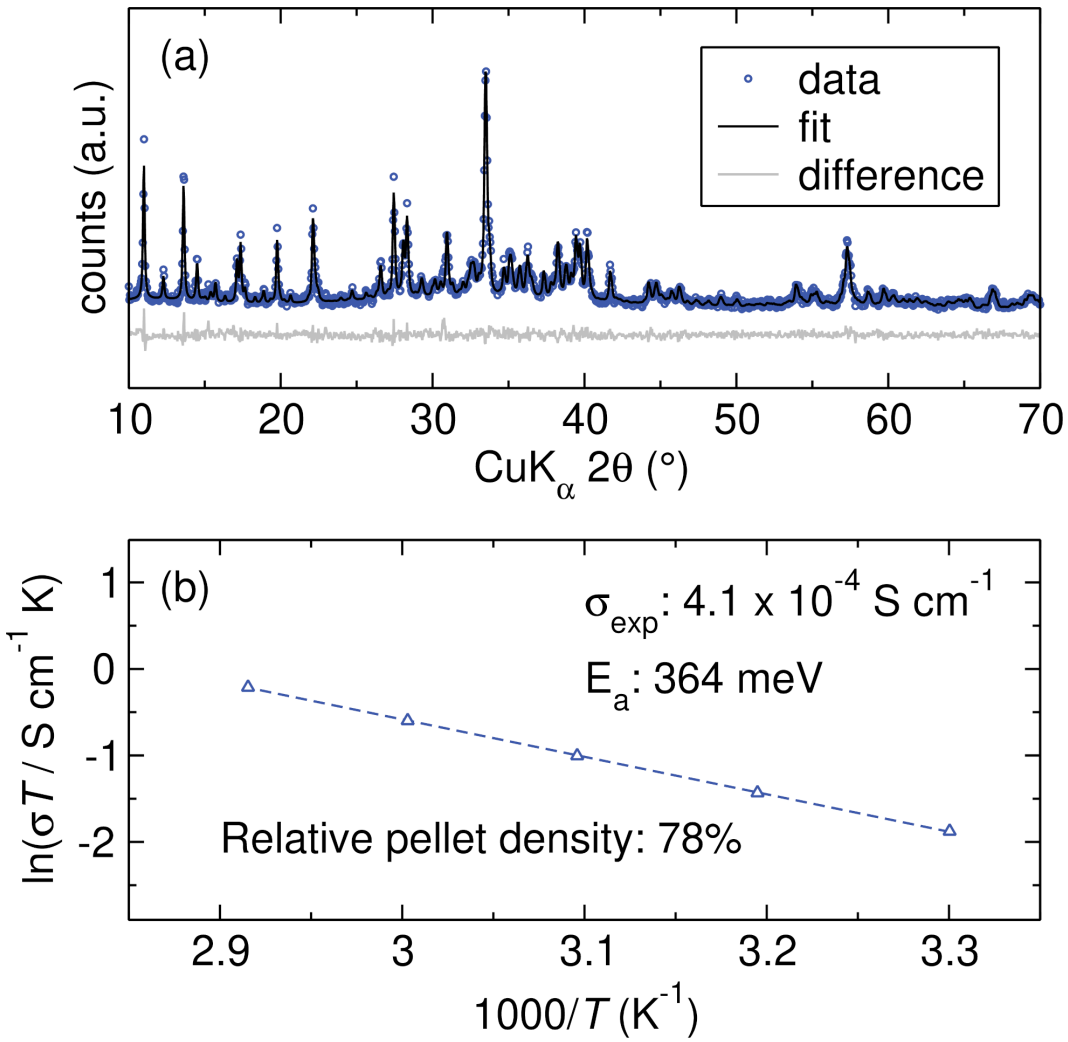

Figure 4.8: Experimental characterization of $Li_9B_{19}S_{33}$. (a) XRD pattern and Rietveld refinement for as-prepared $Li_9B_{19}S_{33}$. (b) Arrhenius-type fit for $Li_9B_{19}S_{33}$ with ionic conductivity values obtained from electrochemical impedance spectroscopy.

4.4 Conclusions

We have constructed the largest known database of experimental ionic conductivity and corresponding crystal structure information of 548 unique Li-containing compounds. By comparing with ordered configurations generated through a supercell sampling approach, we demonstrate that using linear combinations of elemental embeddings is an effective means of representing the prevalent site disorder in our

database with graph-based features, thereby enabling the training of structurally-aware predictive models to identify potential superionic conductors.

Using this representation and a transfer-learning approach, we train AtomSets models that display classification performance surpassing our controls under both k-fold and LOCO CV. As compared to a benchmark logistic regression model trained using domain-specific features, the AtomSets models employing transfer learning exhibit superior predictive power. We find that short-range interactions are most critical for accurate predictions, emphasizing the need to capture local structural environments. Properly including and representing Li atom positions significantly enhances predictive accuracy. Interestingly, the specific identity of anions is found to be less important, as models using simplified structural representations (e.g., CAMNS) showed high performance. This observation aligns with previous findings, suggesting that capturing the overall structural framework may be sufficient for effective identification of fast ion conductors within this database [17, 27, 75].

An ensemble of AtomSets models is used to screen all Li-containing materials in the ICSD and Materials Project repositories. Through this screening, we find 241 materials predicted to be superionic with $E_g > 1$ eV and compositions significantly different from those in our training database. The prediction confidence is quantified by reporting the standard deviation of the ensemble predictions and the distance from each screened compound to the nearest training sample. The predicted likelihood of superionic conductivity P_{SI} for all Li-containing materials in the ICSD and Materials Project are provided for consideration. To validate the effectiveness of the model ensemble for screening, we experimentally demonstrate superionic conductivity in a candidate phase, $Li_9B_{19}S_{33}$.

Importantly, while our approach facilitates screening of materials containing disorder in the Li framework, it does not account for changes in conductivity due to defect introduction. It is possible that compounds with $P_{\mathrm{SI}} < 0.5$ could be modified to be fast ion conductors through appropriate defect engineering strategies. Despite the strengths of the AtomSets architecture, it does not enable the direct determination of interpretable structural features to guide SSE design. While our results show that the logistic regression model using domain-specific features was less effective in this case, the identification of more refined or relevant features could potentially improve its performance. By making the structure-conductivity database used in this study public, we hope to enable future works to explore and develop better structure-property relationships for ion conduction, facilitating design-focused

methodologies.

4.5 Methods

Database processing. All 11,295 Li-containing compounds cataloged in the ICSD (v5.2.0) are compiled. The constructed database for this study encompasses the experimentally measured ionic conductivities of 571 compounds alongside their corresponding ICSD crystal structures. Consequently, there remain 10,724 Li-containing compounds in the ICSD without reported ionic conductivity measurements in the literature. To identify duplicate structures in the labelled database, the StructureMatcher tool within the Python Materials Genomics (Pymatgen) (v2023.11.12) library is employed. Briefly, all pairs of structures are converted to primitive cells, and checks are conducted to ensure that the number of sites, lattice parameters, unit cell angles, and atomic positions do not match within a default tolerance. Duplicate structures are consolidated by retaining the entry with the median ionic conductivity value. The resulting database, devoid of duplicate structures, is comprised of 548 entries.

Data partitioning and clustering. From the database, 10% of entries are randomly allocated to a test set, which is exclusively assessed with the final model ensemble after determination of the final structure representation and the completion of hyperparameter optimization. The remaining data is divided into training and validation sets using two distinct methods. Initially, the data undergoes random splitting for k-fold CV, with folds of equal size (80:20 training and validation). When assessing model performance using k-fold CV, the training and validation portion of the database is randomly partitioned into k different folds. The model is trained on k-1 of the folds and the predictive power of the model is assessed using the remaining fold. The process is repeated for all k folds to obtain the average and variation of the model performance. In the present study, five folds are used for CV. Additionally, the data is partitioned into non-random training and validation sets for LOCO-CV. In this scenario, the data is initially represented using the ElMD description, followed by the application of UMAP with a spread parameter of 5, which controls the scale of local neighborhood preservation, to acquire a low-dimensional representation that maintains essential chemical relationships. Subsequently, DBSCAN, using an epsilon of 4, which defines the maximum distance between points to be considered neighbors, is employed to automatically segregate the data into clusters for LOCO CV. Ten clusters of compounds are obtained, with a statistical summary of each cluster provided in the supplementary information. The ElMD description, the

spread parameter for UMAP, and the epsilon parameter for DBSCAN were selected to align with the leave-one-cluster-out procedure described in previous studies. Intuitive clustering of known families of ion conductors is observed, as detailed in previous works [26].

Descriptor generation and ML models. Crystallographic information files (CIFs) for each compound are parsed with Pymatgen (v2023.11.12). Simplified versions of each structure are generated by systematically removing or modifying groups of atoms. For the CAN representation the Li atoms in each structure are removed. The CAMNS representation is created by checking the oxidation state from the CIF file for each non-Li atom in the structure. Atoms with positive oxidation states are substituted with Al, negative oxidation states converted to S, and oxidation states of 0 converted to Mg. For CANS, this simplification is performed and the Li atoms are removed as well. Graph representations are created using a modified version of the MatErials Graph Network (MEGNet) library (https://github.com/materialsvirtuallab/megnet v1.3.2)[45] to accommodate disordered crystals. The MAterials Machine Learning (maml) libary (https://github.com/materialsvirtuallab/maml v2023.9.9) is then used to create the atom matrix features which are used as the intputs for the AtomSets models [50]. The AtomSets models pass the atoms features matrix through a series of fully connected layers before a set2set symmetry function is used to generate a readout vectors of a defined length with permutation invariance of the atom order [76]. The output of the symmetry function is subsequently passed through additional dense layers and a final sigmoid activation for classification. Atomistic features are generated using the definitions provided by Sendek et al. [15]. The scikit-learn library is used for training of logistic regression models with default parameters excluding the penalty for regularization [77].

Hyperparameter optimization. The default AtomSets architecture does not include conventional regularization techniques to avoid overfitting. Therefore, dropout layers and L2 kernel regularization is added. The optimal hyperparameters for each validation cluster within the LOCO-CV framework are determined using the Ray library (v2.9.3). Model weights are updated using the LAMB optimizer with lookahead mechanism and a triangular 2 cyclical learning rate schedule [78, 79]. A comprehensive listing of the hyperparameter ranges explored is provided in Table C.7. For each cluster, 250 configurations are tested. The top 10 performing configurations for each are then repeated 10 times to account for run-to-run variability.

Subsequently, the best performing configuration across these 10 runs is selected as the optimal configuration for that particular validation cluster. Hyperparameter trial runs are orchestrated using the Asynchronous Successive Halving Algorithm (ASHA) [80]. ASHA is an advanced optimization algorithm that efficiently allocates computational resources to hyperparameter configurations, enabling parallelization and faster optimization by iteratively promoting promising configurations while discarding under performing ones through successive halving. The search space is explored employing HyperOpt, which employs Bayesian optimization to find the optimal configuration [81].

$Li_9B_{19}S_{33}$ synthesis. $Li_9B_{19}S_{33}$ is prepared from lithium sulfide (Li_2S, 99.9%, Thermo Fisher Scientific), elemental boron (99.99%, SkySpring Nanomaterials, Inc.) and sulfur (S_8, > 99.5%, Acros Organics). In an Ar-filled glovebox (Mbraun), a two gram stoichiometric mixture of the precursor materials is combined in a 50 ml YSZ milling jar along with milling media (two 10 mm diameter balls, 34 5 mm diameter balls, and eight grams of 3 mm diameter balls). The jar is sealed before removing from the glovebox to minimize exposure to air. The precursors are milled in a planetary ball mill (MSE PMV1-0.4L) for 45 minutes at 300 rpm. After milling, the precursor mixture is extracted under Ar and 333 mg of the powder is transferred to a glassy carbon crucible (SPI Supplies). Two repeated heating steps are required to obtain pure $Li_9B_{19}S_{33}$. The crucible containing the powder is placed into a carbon-coated vitreous silica ampoule (inner diameter 14 mm, outer diameter 16 mm), which is evacuated to <10 mtorr and sealed. The sealed ampoule is heated to 700 °C at a rate of 1 °C/min, held at 700 °C for 16 h, and then cooled to room temperature at 1 °C/min. After the first annealing step, the material is removed under Ar, ground with a mortar and pestle, and reloaded into the crucible. The crucible is then sealed in a second carbon-coated vitreous silica ampoule, and the heating procedure is repeated to yield the desired phase.

Experimental characterization of $Li_9B_{19}S_{33}$. Powder X-ray diffraction is used to assess the phase purity of the prepared $Li_9B_{19}S_{33}$ material. The sample powder is loaded into a Rigaku air-free sample holder under Ar to prevent exposure to air during the measurement. Diffraction patterns are collected using a Rigaku Smartlab diffractometer with a Cu Kα X-ray source. The scan range is from 10° to 70° 2θ at a rate of 3° min^{-1} with a step size of 0.04°. Rietveld refinement of the diffraction patterns was performed using GSAS-II software [82]. To characterize the ionic conductivity of $Li_9B_{19}S_{33}$, 40 - 60 mg of the material is hot pressed

(Col-Int Tech Manual Hydraulic press) at 250°C under 2 tons of pressure for 5 minutes, forming pellets with 6 mm diameter. The pellet surfaces are polished with 1500-grit abrasive sheets before the pellet thickness is measured. Indium metal foil is placed on stainless steel current collector rods and the pellet is assembled into Swagelok cells under ~100 MPa of pressure using a manual vise. Electrochemical impedance spectroscopy (EIS) is performed with a Biologic VSP-300 potentiostat over a frequency range of 3 MHz to 1 Hz and an amplitude of 25 mV, across a temperature range of 25°C to 70°C.

4.6 Data and Code Availability

The database of σ_{exp} values and ICSD collection codes for corresponding crystal structures is made available as a supplementary comma-separated values file. The dataset is additionally available through CaltechDATA at `https://doi.org/10.22002/23mvv-6gk43`. The version of the codebase used to train models, perform screening, and analyze results is archived at `https://doi.org/10.22002/cgx0v-wqq34`.

4.7 Author Contributions

Conceptualization, D.B.M., Z.W.B.I, J.M.B, F.A.L.L.,; Data curation, F.A.L.L., D.B.M., Z.W.B.I, J.M.B,; Formal analysis, D.B.M,; Investigation, D.B.M.; Methodology, D.B.M.,; Software, D.B.M; Validation, D.B.M., Z.W.B.I., J.M.B.; Visualization, D.B.M.; Writing – original draft, D.B.M.; Writing – review & editing, D.B.M., Z.W.B.I, J.M.B, F.A.L.L., K.A.S.; Supervision, K.A.S.; Funding acquisition, K.A.S.

4.8 Acknowledgements

This research was supported by the Arnold and Mabel Beckman Foundation through the Beckman Young Investigator Award. The computations presented here were conducted in the Resnick High Performance Computing Center, a facility supported by Resnick Sustainability Institute at the California Institute of Technology. K.A.S. acknowledges support from the Packard Fellowship for Science and Engineering, the Alfred P. Sloan Foundation, and the Camille and Henry Dreyfus Foundation.

4.9 Bibliography

[1] Goodenough, J. B. Rechargeable Batteries: Challenges Old and New. *J Solid State Electrochem* **2012**, *16*, 2019–2029.

[2] Janek, J.; Zeier, W. G. A Solid Future for Battery Development. *Nat Energy* **2016**, *1*, 16141.

[3] Kato, Y.; Hori, S.; Saito, T.; Suzuki, K.; Hirayama, M.; Mitsui, A.; Yonemura, M.; Iba, H.; Kanno, R. High-Power All-Solid-State Batteries Using Sulfide Superionic Conductors. *Nat Energy* **2016**, *1*, 16030.

[4] Inoue, T.; Mukai, K. Are All-Solid-State Lithium-Ion Batteries Really Safe?– Verification by Differential Scanning Calorimetry with an All-Inclusive Microcell. *ACS Appl. Mater. Interfaces* **2017**, *9*, 1507–1515.

[5] Kamaya, N.; Homma, K.; Yamakawa, Y.; Hirayama, M.; Kanno, R.; Yonemura, M.; Kamiyama, T.; Kato, Y.; Hama, S.; Kawamoto, K.; Mitsui, A. A Lithium Superionic Conductor. *Nature Mater* **2011**, *10*, 682–686.

[6] Bachman, J. C.; Muy, S.; Grimaud, A.; Chang, H.-H.; Pour, N.; Lux, S. F.; Paschos, O.; Maglia, F.; Lupart, S.; Lamp, P.; Giordano, L.; Shao-Horn, Y. Inorganic Solid-State Electrolytes for Lithium Batteries: Mechanisms and Properties Governing Ion Conduction. *Chem. Rev.* **2016**, *116*, 140–162.

[7] Richards, W. D.; Miara, L. J.; Wang, Y.; Kim, J. C.; Ceder, G. Interface Stability in Solid-State Batteries. *Chem. Mater.* **2016**, *28*, 266–273.

[8] Sendek, A. D.; Cheon, G.; Pasta, M.; Reed, E. J. Quantifying the Search for Solid Li-Ion Electrolyte Materials by Anion: A Data-Driven Perspective. *J. Phys. Chem. C* **2020**, *124*, 8067–8079.

[9] Sheng, O.; Jin, C.; Ding, X.; Liu, T.; Wan, Y.; Liu, Y.; Nai, J.; Wang, Y.; Liu, C.; Tao, X. A Decade of Progress on Solid-state Electrolytes for Secondary Batteries: Advances and Contributions. *Advanced Functional Materials* **2021**, *31*, 2100891.

[10] Janek, J.; Zeier, W. G. Challenges in Speeding up Solid-State Battery Development. *Nat Energy* **2023**, *8*, 230–240.

[11] Hu, Y.; Li, W.; Zhu, J.; Hao, S.-M.; Qin, X.; Fan, L.-Z.; Zhang, L.; Zhou, W. Multi-Layered Electrolytes for Solid-State Lithium Batteries. *Next Energy* **2023**, *1*, 100042.

[12] Jalem, R.; Aoyama, T.; Nakayama, M.; Nogami, M. Multivariate Method-Assisted *Ab Initio* Study of Olivine-Type $LiMXO_4$ (Main Group M^{2+}–X^{5+} and M^{3+}–X^{4+}) Compositions as Potential Solid Electrolytes. *Chem. Mater.* **2012**, *24*, 1357–1364.

[13] Fujimura, K.; Seko, A.; Koyama, Y.; Kuwabara, A.; Kishida, I.; Shitara, K.; Fisher, C. A. J.; Moriwake, H.; Tanaka, I. Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms. *Advanced Energy Materials* **2013**, *3*, 980–985.

[14] Jalem, R.; Nakayama, M.; Kasuga, T. An Efficient Rule-Based Screening Approach for Discovering Fast Lithium Ion Conductors Using Density Functional Theory and Artificial Neural Networks. *J. Mater. Chem. A* **2014**, *2*, 720–734.

[15] 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 Environ. Sci.* **2017**, *10*, 306–320.

[16] Ahmad, Z.; Xie, T.; Maheshwari, C.; Grossman, J. C.; Viswanathan, V. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Cent. Sci.* **2018**, *4*, 996–1006.

[17] Zhang, Y.; He, X.; Chen, Z.; Bai, Q.; Nolan, A. M.; Roberts, C. A.; Banerjee, D.; Matsunaga, T.; Mo, Y.; Ling, C. Unsupervised Discovery of Solid-State Lithium Ion Conductors. *Nat Commun* **2019**, *10*, 5260.

[18] 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. *Chem. Mater.* **2019**, *31*, 342–352.

[19] Cubuk, E. D.; Sendek, A. D.; Reed, E. J. Screening Billions of Candidates for Solid Lithium-Ion Conductors: A Transfer Learning Approach for Small Data. *The Journal of Chemical Physics* **2019**, *150*, 214701.

[20] Choi, E.; Jo, J.; Kim, W.; Min, K. Searching for Mechanically Superior Solid-State Electrolytes in Li-Ion Batteries *via* Data-Driven Approaches. *ACS Appl. Mater. Interfaces* **2021**, *13*, 42590–42597.

[21] Zhao, Q.; Avdeev, M.; Chen, L.; Shi, S. Machine Learning Prediction of Activation Energy in Cubic Li-argyrodites with Hierarchically Encoding Crystal Structure-Based (HECS) Descriptors. *Science Bulletin* **2021**, *66*, 1401–1408.

[22] Guo, H.; Wang, Q.; Urban, A.; Artrith, N. Artificial Intelligence-Aided Mapping of the Structure–Composition–Conductivity Relationships of Glass–Ceramic Lithium Thiophosphate Electrolytes. *Chem. Mater.* **2022**, *34*, 6702–6712.

[23] Lu, Z.; Adeli, P.; Yim, C.-H.; Jiang, M.; Rempel, J.; Chen, Z. W.; Yadav, S.; Mercier, P.; Abu-Lebdeh, Y.; Singh, C. V. Automatically Capturing Key Features for Predicting Superionic Conductivity of Solid-State Electrolytes Using a Neural Network. *ACS Appl. Energy Mater.* **2022**, *5*, 8042–8048.

[24] Adhyatma, A.; Xu, Y.; Hawari, N. H.; Satria Palar, P.; Sumboja, A. Improving Ionic Conductivity of Doped $Li_7La_3Zr_2O_{12}$ Using Optimized Machine Learning with Simplistic Descriptors. *Materials Letters* **2022**, *308*, 131159.

[25] Kim, K.; Siegel, D. J. Machine Learning Reveals Factors That Control Ion Mobility in Anti-Perovskite Solid Electrolytes. *J. Mater. Chem. A* **2022**, *10*, 15169–15182.

[26] Hargreaves, C. J. et al. A Database of Experimentally Measured Lithium Solid Electrolyte Conductivities Evaluated with Machine Learning. *npj Comput Mater* **2023**, *9*, 9.

[27] Laskowski, F. A. L.; McHaffie, D. B.; See, K. A. Identification of Potential Solid-State Li-ion Conductors with Semi-Supervised Learning. *Energy Environ. Sci.* **2023**, *16*, 1264–1276.

[28] Lin, Y.-Y.; Qu, J.; Gustafson, W. J.; Kung, P.-C.; Shah, N.; Shrivastav, S.; Ertekin, E.; Krogstad, J. A.; Perry, N. H. Coordination Flexibility as a High-Throughput Descriptor for Identifying Solid Electrolytes with Li+ Sublattice Disorder: A Computational and Experimental Study. *Journal of Power Sources* **2023**, *553*, 232251.

[29] Sun, J.; Kang, S.; Kim, J.; Min, K. Accelerated Discovery of Novel Garnet-Type Solid-State Electrolyte Candidates via Machine Learning. *ACS Appl. Mater. Interfaces* **2023**, *15*, 5049–5057.

[30] Guo, X.; Wang, Z.; Yang, J.-H.; Gong, X.-G. Machine-Learning Assisted High-Throughput Discovery of Solid-State Electrolytes for Li-ion Batteries. *J. Mater. Chem. A* **2024**, *12*, 10124–10136.

[31] Kim, J.; Lee, D.; Lee, D.; Li, X.; Lee, Y.-L.; Kim, S. Machine Learning Prediction Models for Solid Electrolytes Based on Lattice Dynamics Properties. *J. Phys. Chem. Lett.* **2024**, *15*, 5914–5922.

[32] Kauwe, S. K.; Graser, J.; Murdock, R.; Sparks, T. D. Can Machine Learning Find Extraordinary Materials? *Computational Materials Science* **2020**, *174*, 109498.

[33] Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

[34] Kauwe, S. K.; Graser, J.; Vazquez, A.; Sparks, T. D. Machine Learning Prediction of Heat Capacity for Solid Inorganics. *Integr Mater Manuf Innov* **2018**, *7*, 43–51.

[35] Jha, D.; Ward, L.; Paul, A.; Liao, W.-k.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci Rep* **2018**, *8*, 17593.

[36] Wang, A. Y.-T.; Kauwe, S. K.; Murdock, R. J.; Sparks, T. D. Compositionally Restricted Attention-Based Network for Materials Property Predictions. *npj Comput Mater* **2021**, *7*, 77.

[37] Phani Dathar, G. K.; Balachandran, J.; Kent, P. R. C.; Rondinone, A. J.; Ganesh, P. Li-Ion Site Disorder Driven Superionic Conductivity in Solid Electrolytes: A First-Principles Investigation of β-Li_3PS_4. *J. Mater. Chem. A* **2017**, *5*, 1153–1159.

[38] Di Stefano, D.; Miglio, A.; Robeyns, K.; Filinchuk, Y.; Lechartier, M.; Senyshyn, A.; Ishida, H.; Spannenberger, S.; Prutsch, D.; Lunghammer, S.; Rettenwander, D.; Wilkening, M.; Roling, B.; Kato, Y.; Hautier, G. Superionic Diffusion through Frustrated Energy Landscape. *Chem* **2019**, *5*, 2450–2460.

[39] Hogrefe, K.; Minafra, N.; Hanghofer, I.; Banik, A.; Zeier, W. G.; Wilkening, H. M. R. Opening Diffusion Pathways through Site Disorder: The Interplay of Local Structure and Ion Dynamics in the Solid Electrolyte $Li_{6+x}P_{1-x}Ge_xS_5I$ as Probed by Neutron Diffraction and NMR. *J. Am. Chem. Soc.* **2022**, *144*, 1795–1812.

[40] Morgan, B. J. Mechanistic Origin of Superionic Lithium Diffusion in Anion-Disordered Li_6PS_5X Argyrodites. *Chem. Mater.* **2021**, *33*, 2004–2018.

[41] Botros, M.; Janek, J. Embracing Disorder in Solid-State Batteries. *Science* **2022**, *378*, 1273–1274.

[42] Zeng, Y.; Ouyang, B.; Liu, J.; Byeon, Y.-W.; Cai, Z.; Miara, L. J.; Wang, Y.; Ceder, G. High-Entropy Mechanism to Boost Ionic Conductivity. *Science* **2022**, *378*, 1320–1324.

[43] Gamon, J.; Dyer, M. S.; Duff, B. B.; Vasylenko, A.; Daniels, L. M.; Zanella, M.; Gaultois, M. W.; Blanc, F.; Claridge, J. B.; Rosseinsky, M. J. $Li_{4.3}AlS_{3.3}Cl_{0.7}$: A Sulfide–Chloride Lithium Ion Conductor with Highly Disordered Structure and Increased Conductivity. *Chem. Mater.* **2021**, *33*, 8733–8744.

[44] Wang, S.; Gong, S.; Böger, T.; Newnham, J. A.; Vivona, D.; Sokseiha, M.; Gordiz, K.; Aggarwal, A.; Zhu, T.; Zeier, W. G.; Grossman, J. C.; Shao-Horn, Y. Multimodal Machine Learning for Materials Science: Discovery of Novel Li-Ion Solid Electrolytes. *Chem. Mater.* **2024**, *36*, 11541–11550.

[45] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572.

[46] Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Comput Mater* **2021**, *7*, 84.

[47] Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; Van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Commun Mater* **2022**, *3*, 93.

[48] Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P. Learning Properties of Ordered and Disordered Materials from Multi-Fidelity Data. *Nat Comput Sci* **2021**, *1*, 46–53.

[49] Butler, K. T.; Oviedo, F.; Canepa, P. *Machine Learning in Materials Science*; American Chemical Society: Washington, DC, USA, 2022.

[50] Chen, C.; Ong, S. P. AtomSets as a Hierarchical Transfer Learning Framework for Small and Large Materials Datasets. *npj Comput Mater* **2021**, *7*, 173.

[51] De Breuck, P.-P.; Hautier, G.; Rignanese, G.-M. Materials Property Prediction for Limited Datasets Enabled by Feature Selection and Joint Learning with MODNet. *npj Comput Mater* **2021**, *7*, 83.

[52] Gupta, V.; Choudhary, K.; DeCost, B.; Tavazza, F.; Campbell, C.; Liao, W.-k.; Choudhary, A.; Agrawal, A. Structure-Aware Graph Neural Network Based Deep Transfer Learning Framework for Enhanced Predictive Analytics on Diverse Materials Datasets. *npj Comput Mater* **2024**, *10*, 1.

[53] Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput Mater* **2020**, *6*, 138.

[54] Müller, M.; Auer, H.; Bauer, A.; Uhlenbruck, S.; Finsterbusch, M.; Wätzig, K.; Nikolowski, K.; Dierickx, S.; Fattakhova-Rohlfing, D.; Guillon, O.; Weber, A. Guidelines to Correctly Measure the Lithium Ion Conductivity of Oxide Ceramic Electrolytes Based on a Harmonized Testing Procedure. *Journal of Power Sources* **2022**, *531*, 231323.

[55] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Computational Materials Science* **2013**, *68*, 314–319.

[56] Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432.

[57] Matthews, B. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1975**, *405*, 442–451.

[58] Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21*, 6.

[59] Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Mining* **2021**, *14*, 13.

[60] Hart, G. L. W.; Forcade, R. W. Algorithm for Generating Derivative Structures. *Phys. Rev. B* **2008**, *77*, 224115.

[61] Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.

[62] Jain, A.; Ong, S. P.; Hautier, G.; Moore, C.; Munro, J. Electronic Structure | Materials Project Documentation. https://docs.materialsproject.org/methodology/materials-methodology/electronic-structure, 2023.

[63] Nagel, R.; Wickel, Ch.; Lutz, H. Crystal Structure of the Quaternary Compounds $Li_{1.25}Cd_{1.67}In_{0.47}Cl_6$ and $Li_{0.21}Mn_{1.71}In_{0.79}Cl_6$. *Solid State Sciences* **2003**, *5*, 827–832.

[64] Zhang, J.; Corbett, J. D. Zirconium Chloride Cluster Phases Centered by Transition Metals Mn-Ni. Examples of the Nb_6F_{15} Structure. *Inorg. Chem.* **1991**, *30*, 431–435.

[65] Maltsev, A. P.; Chepkasov, I. V.; Kvashnin, A. G.; Oganov, A. R. Ionic Conductivity of Lithium Phosphides. *Crystals* **2023**, *13*, 756.

[66] Grüne, M. Complex Lithium Dynamics in the Novel Thioborate Li5B7S13 Revealed by NMR Relaxation and Lineshape Studies. *Solid State Ionics* **1995**, *78*, 305–313.

[67] Bertermann, R.; Muller-Warmuth, W.; Jansen, C.; Hiltmann, F.; Krebs, B. NMR Studies of the Lithium Dynamics in Two Thioborate Superionic Conductors: $Li_9B_{19}S_{33}$ and $Li_{4-2x}Sr_{2+x}B_{10}S_{19}$ ($x \approx 0.27$). *Solid State Ionics* **1999**,

[68] Sendek, A. D.; Antoniuk, E. R.; Cubuk, E. D.; Ransom, B.; Francisco, B. E.; Buettner-Garrett, J.; Cui, Y.; Reed, E. J. Combining Superionic Conduction and Favorable Decomposition Products in the Crystalline Lithium–Boron–Sulfur System: A New Mechanism for Stabilizing Solid Li-Ion Electrolytes. *ACS Appl. Mater. Interfaces* **2020**, *12*, 37957–37966.

[69] Zeilinger, M.; van Wüllen, L.; Benson, D.; Kranak, V. F.; Konar, S.; Fässler, T. F.; Häussermann, U. $LiBSi_2$: A Tetrahedral Semiconductor Framework from Boron and Silicon Atoms Bearing Lithium Atoms in the Channels. *Angew Chem Int Ed* **2013**, *52*, 5978–5982.

[70] Wang, Q.; Liu, D.; Ma, X.; Zhou, X.; Lei, Z. Cl-Doped $Li_{10}SnP_2S_{12}$ with Enhanced Ionic Conductivity and Lower Li-Ion Migration Barrier. *ACS Appl. Mater. Interfaces* **2022**, *14*, 22225–22232.

[71] Limpert, M. A.; Atwater, T. B.; Hamann, T.; Godbey, G. L.; Hitz, G. T.; McOwen, D. W.; Wachsman, E. D. Achieving Desired Lithium Concentration in Garnet Solid Electrolytes; Processing Impacts on Physical and Electrochemical Properties. *Chem. Mater.* **2022**, *34*, 9468–9478.

[72] Ning, T.; Zhang, Y.; Zhang, Q.; Shen, X.; Luo, Y.; Liu, T.; Liu, P.; Luo, Z.; Lu, A. The Effect of a Ta, Sr Co-Doping Strategy on Physical and Electrochemical Properties of $Li_7La_3Zr_2O_{12}$ Electrolytes. *Solid State Ionics* **2022**, *379*, 115917.

[73] Zhou, X.; Huang, L.; Elkedim, O.; Xie, Y.; Luo, Y.; Chen, Q.; Zhang, Y.; Chen, Y. Sr^{2+} and Mo^{6+} Co-Doped $Li_7La_3Zr_2O_{12}$ with Superior Ionic Conductivity. *Journal of Alloys and Compounds* **2022**, *891*, 161906.

[74] Hiltmann, F.; Zum Hebel, P.; Hammerschmidt, A.; Krebs, B. $Li_5B_7S_{13}$ und $Li_9B_{19}S_{33}$: Zwei Lithiumthioborate mit neuen hochpolymeren Anionengerüsten. *Zeitschrift anorg allge chemie* **1993**, *619*, 293–302.

[75] Wang, Y.; Richards, W. D.; Ong, S. P.; Miara, L. J.; Kim, J. C.; Mo, Y.; Ceder, G. Design Principles for Solid-State Lithium Superionic Conductors. *Nature Mater* **2015**, *14*, 1026–1031.

[76] Vinyals, O.; Bengio, S.; Kudlur, M. ORDER MATTERS: SEQUENCE TO SEQUENCE FOR SETS. **2016**,

[77] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **2011**, *12*, 2825–2830.

[78] You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; Hsieh, C.-J. Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes. *Proc. International Conference on Learning Representations (ICLR)* **2020**.

[79] Zhang, M.; Lucas, J.; Ba, J.; Hinton, G. E. Lookahead Optimizer: k Steps Forward, 1 Step Back. *Advances in neural information processing systems* **2019**, *32*.

[80] Li, L.; Jamieson, K.; Rostamizadeh, A.; Gonina, E.; Ben-Tzur, J.; Hardt, M.; Recht, B.; Talwalkar, A. A System for Massively Parallel Hyperparameter Tuning. *Proceedings of machine learning and systems* **2020**, *2*, 230–246.

[81] Bergstra, J.; Yamins, D.; Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *International Conference on Machine Learning*. **2013**; pp 115–123.

[82] Toby, B. H.; Von Dreele, R. B. GSAS-II: The Genesis of a Modern Open-Source All Purpose Crystallography Software Package. *Journal of Applied Crystallography* **2013**, *46*, 544–549.

Chapter 5

CONCLUSION

5.1 Summary

This thesis demonstrates the combined computational and experimental discovery of next-generation SSEs. A strong emphasis is placed on leveraging existing experimental data to search for promising materials outside of traditionally studied chemistries. Careful consideration of the complex nature of experimentally derived data is a central theme throughout this work.

We construct the largest known database containing the structures and ionic conductivities of experimentally characterized SSEs. In Chapter 2, we demonstrate a semi-supervised learning approach to determine the best material representation for the task of identifying fast ion conductors. A description of the host lattice's local environment performs best in agglomerative clustering, grouping compounds with similar conductivities. This representation is used to screen Li-containing materials in the ICSD and Materials Project. Candidates are prioritized with semi-empirical and first principles calculations, allocating the most computational resources to the most promising compounds. From this tiered workflow, Li_3BS_3 is selected for experimental demonstration and shown to exhibit ionic conductivity through defect engineering via chemical substitution and mechanical milling.

Chapter 3 explores the primary factors controlling ionic conductivity in the Li_3BS_3 system. Previous work on this material attributed conductivity improvements from chemical substitution to specific mechanisms without consideration of other possible structural changes. A combination of Raman spectroscopy, solid-state nuclear magnetic resonance, and x-ray diffraction indicates that sample crystallinity decreases upon substitution and is highly dependent on the substituting element. At high levels of Cl and Al substitution, the product phase becomes completely amorphous. By contrast, high Si substitution drives the formation of previously unidentified crystalline phases distinct from the parent Li_3BS_3 framework. The amorphous Cl- and Al-substituted phases exhibit ionic conductivities exceeding 10^{-4} S cm^{-1}, presenting a pathway to high conductivity through substitution rather than extended milling. The novel crystalline phase formed through Si substitution shows an order-of-magnitude higher ionic conductivity ($> 10^{-3}$ S cm^{-1}). Microstructure is also

shown to be highly sensitive to processing conditions, with small changes in manual grinding of powders causing significant variations in crystalline domain size and microstrain. Importantly, the conductivity improvements from Cl, Al, and Si substitution cannot be solely attributed to the creation of additional mobile carriers. Instead, enhanced ion transport is facilitated by increased disorder, either through reduced crystallinity or microstructural effects such as a greater volume fraction of grain boundaries associated with smaller domain sizes or microstrain.

Many known fast ion conductors contain disorder, and Chapters 2 and 3 illustrate the effects of defects and disorder on ion transport. However, digital representations of disorder in solid-state materials are limited. As a result, only a subset of the curated database could be used in Chapter 2 and screening was restricted to ordered Li-containing materials. As a step towards addressing this limitation, a graph-based representation of disorder is implemented in Chapter 4. This approach enables the utilization of the full training set and allows prediction of ionic conductivity for all known Li materials. We identify 241 compounds as potential fast ion conductors with estimated band gaps greater than 1 eV. Experimental validation confirms superionic conductivity in $Li_9B_{19}S_{33}$, a highly disordered compound that was excluded from consideration in Chapter 2. This demonstrates the value of disorder-compatible representations in the discovery of novel SSEs.

5.2 Outlook

Data-driven methods, like those presented in this thesis, can accelerate the discovery of new materials with high ionic conductivity. However, methods for high-throughput evaluation of other required properties are necessary to efficiently explore available chemical space to realize an ideal SSE material. Prediction of electrochemical stability, another important requirement, has been demonstrated through the use of thermodynamic calculations [1, 2]. This approach provides stability windows for SSEs, but does not include kinetic considerations. Additionally, SSEs that are thermodynamically unstable may still be suitable if their decomposition forms an interphase that is self-passivating (electronically insulating) and allows ion transport similar to the solid-electrolyte interphase formed in conventional Li-ion batteries. Methods to address this are being developed and will likely benefit from accelerated molecular dynamics using machine-learned interatomic potentials and experimental feedback [3].

In the introduction of this thesis, a clear trade-off between stability window and ionic

conductivity is illustrated. Within each group, anions with greater electronegativity provide higher anodic stability at the cost of reduced conductivity. This suggests that mixed-anion systems may provide a better compromise between these properties. It has also been proposed that disordered mixed-anion systems could permit even higher conductivities by flattening the energy for ion migration [4, 5]. The recently reported oxyhalide SSEs demonstrate the potential of this class of materials, achieving ionic conductivities comparable to best-in-class sulfide SSEs while providing significantly enhanced oxidative stability [6]. Development of multi-phase systems, composed of multiple crystalline or amorphous phases, is another strategy that shows promise for meeting the diverse property requirements of SSEs. Importantly, both paths represent a dramatic expansion of the design space, further emphasizing the need for computational approaches to guide experimentation.

The effectiveness of these data-driven approaches hinges on the availability of high-quality data. Rapid progress in machine-learning-accelerated simulations could provide a means for generating expansive datasets for training predictive models. However, these models must be validated against experimental measurements. The database of structures and experimental ionic conductivities presented in this thesis represents a substantial increase in size over previous efforts but it remains a relatively small and biased coverage of known Li-containing materials. Additionally, it was sourced from 285 different scientific papers and measurement differences across laboratories can introduce significant variability which lowers model performance [7]. Advancements in high-throughput, automated experiments will be critical to accelerating the discovery of high-performance SSEs. These approaches will enable large amounts of experimental data to be produced under highly repeatable conditions and can provide insight into systems not captured in computational screening, such as multi-phase mixtures, metastable phases, or novel crystalline phases. The material with the highest ionic conductivity reported in this thesis is a novel crystalline phase without a solved structure, underscoring the role of serendipitous experimental discovery.

Finally, a key advancement required for closing the gap between model predictions and experiment will be the development of representations better suited to capture the complexities observed in real materials. As illustrated in this thesis, defects, disorder, and microstructure can all have pronounced effects on the measured properties of SSEs. Chapter 4 describes a method that enables training models with input structures containing site disorder, but the assumption was still made that all were

crystalline with long-range order. Representations that can encode local atomic arrangements without presuming periodicity may be better suited to amorphous materials or systems where local ordering controls the properties [8]. Additionally, future models and representations will need to be capable of combining multimodal inputs to capture a more complete picture of the entire system, including structure, composition, and microstructure, to accelerate the discovery of next-generation SSEs.

5.3 Bibliography

[1] Richards, W. D.; Miara, L. J.; Wang, Y.; Kim, J. C.; Ceder, G. Interface Stability in Solid-State Batteries. *Chem. Mater.* **2016**, *28*, 266–273.

[2] Wang, Z.-Y.; Zhao, C.-Z.; Sun, S.; Liu, Y.-K.; Wang, Z.-X.; Li, S.; Zhang, R.; Yuan, H.; Huang, J.-Q. Achieving High-Energy and High-Safety Lithium Metal Batteries with High-Voltage-Stable Solid Electrolytes. *Matter* **2023**, *6*, 1096–1124.

[3] Lomeli, E. G.; Ransom, B.; Ramdas, A.; Jost, D.; Moritz, B.; Sendek, A. D.; Reed, E. J.; Devereaux, T. P. Predicting Reactivity and Passivation of Solid-State Battery Interfaces. *ACS Appl. Mater. Interfaces* **2024**, *16*, 51584–51594.

[4] Botros, M.; Janek, J. Embracing Disorder in Solid-State Batteries. *Science* **2022**, *378*, 1273–1274.

[5] Zeng, Y.; Ouyang, B.; Liu, J.; Byeon, Y.-W.; Cai, Z.; Miara, L. J.; Wang, Y.; Ceder, G. High-Entropy Mechanism to Boost Ionic Conductivity. *Science* **2022**, *378*, 1320–1324.

[6] Tanaka, Y.; Ueno, K.; Mizuno, K.; Takeuchi, K.; Asano, T.; Sakai, A. New Oxyhalide Solid Electrolytes with High Lithium Ionic Conductivity >10 mS Cm^{-1} for All-Solid-State Batteries. *Angew Chem Int Ed* **2023**, *62*, e202217581.

[7] Müller, M.; Auer, H.; Bauer, A.; Uhlenbruck, S.; Finsterbusch, M.; Wätzig, K.; Nikolowski, K.; Dierickx, S.; Fattakhova-Rohlfing, D.; Guillon, O.; Weber, A. Guidelines to Correctly Measure the Lithium Ion Conductivity of Oxide Ceramic Electrolytes Based on a Harmonized Testing Procedure. *Journal of Power Sources* **2022**, *531*, 231323.

[8] Billinge, S. J. Do Materials Have a Genome, and If They Do, What Can Be Done with It? *Matter* **2024**, *7*, 3714–3727.

Appendix A

SUPPORTING INFORMATION FOR CHAPTER 2: IDENTIFICATION OF POTENTIAL SOLID-STATE LI-ION CONDUCTORS WITH SEMI-SUPERVISED LEARNING

A.1 W_σ optimization

Ward's minimum variance method applied to the conductivity labels (W_σ) is used to assess the utility of each descriptor-simplification combination. The W_σ is calculated after agglomerative clustering, for each clustering set:

$$W_\sigma = \sum_{k=1}^{n_c} \sum_{i \in C_k} \left[\log(\sigma_{RT})_i - \overline{\log(\sigma_{RT})}_k \right]^2$$

where n_c is the number of clusters in a set, C_k is cluster k, and where $\log(\sigma_{RT})_k$ denotes the mean for all labels in cluster k. Lower W_σ values indicate that the descriptor-simplification combination results in clustering where structures with similar conductivity are grouped together, whereas a large W_σ indicates that the clusters have little correlation to the conductivity labels.

A frozen-state strategy is employed to prevent any label from dropping out of the W_σ calculation. The frozen-state strategy operates by calculating the partial variance (PV) for each label at each clustering depth:

$$PV_{x,C_k} = \left[\log(\sigma_{RT})_x - \overline{\log(\sigma_{RT})}_k \right]^2$$

where PV_{x,C_k} is the partial variance for label x, when label x is assigned to cluster k. The PV for each label is saved before summing all the partial variances to yield the W_σ. At each subsequent clustering depth, all new clusters are checked to determine whether any cluster contains a single label. If a label is the only label in a cluster, then that label's partial variance is frozen: its PV_{x,C_k} becomes equal to the saved state from the previous cluster depth:

$$PV_{x,C_j} = PV_{x,C_k}$$

where C_j denotes the cluster with only one label and C_k denotes the cluster that label x previously resided in. Without the frozen state strategy, poor models will reach desirable W_σ values at sufficient depths of clustering. The artificial depression of the W_σ value occurs because clusters that contain a single label evaluate to 0 (the label mean and cluster mean are the same). Whereas the frozen state strategy effectively "remembers" how well (or poorly) the label was clustered before it drops out.

Hyperparameter tuning was employed for some of the descriptors. At least one W_σ representation exists for each unique combination of structure simplification and descriptor. However, some of the descriptors can be altered by tuning associated hyperparameters, resulting in more W_σ representations. The descriptors with hyperparameter tuning are the global instability index, radial distribution function, smooth overlap of atomic positions (SOAP), and mXRD. A grid search was done over the hyperparameters, for each descriptor, with parameters shown in Table A.1.

Table A.1: Hyperparameters used in grid search.

Descriptor	Hyper-parameter	Description	Values attempted in grid search
Global instability index	r_{cut}	The distance, in angstroms, to search for neighbors when calculating bond valences.	[1.0, 1.1, . . ., 5.9, 6.0]
Radial distribution function	cutoff	The distance, in angstroms, over which the radial distribution function should be calculated.	[1, 2, . . ., 29, 30]
	bin_size	The radial distance, in angstroms, for each bin.	[0.01, 0.02, . . ., 0.09, 0.1, 0.2, . . ., 0.9, 1.0]
Smooth overlap of atomic positions (SOAP)	r_{cut}	The radial cutoff for the local region in angstroms.	[1, 2, . . ., 29, 30]
	n_{max}	The number of radial basis functions used.	[2, 3, . . ., 8, 9]
	l_{max}	The maximum degree of spherical harmonics used.	[1, 2, . . ., 8, 9]
	average	The averaging mode.	['outer', 'inner']
mXRD	pattern_length	The number of 2θ values calculated between 0° and 90°.	[101, 201, . . ., 901, 1001]

Ultimately, the SOAP-CAN descriptor-simplification outperforms all other descriptor-simplifications when the averaging hyperparameter is set to 'outer'. Setting the 'outer' hyperparameter results in averaging over the power spectrum of different sites. Whereas the 'inner' setting averages over the sites first, before summing up the magnetic quantum numbers. The other three hyperparameters (r_{cut}, n_{max}, and l_{max}) are less consequential, with most combinations tested outperforming all other non-SOAP descriptors. To illustrate the point, three different SOAP–CAN

outcomes are depicted in Figure A.1, plotted against the best-performing outcomes from density–CAN, mXRD–A40, orbital field matrix, and structure heterogeneity–A40. The three SOAP-CAN outcomes are those with the lowest W_σ mean for the depth of clustering ranges: 2–100, 101–200, and 201–300. The respective hyperparameters for the three SOAP-CAN descriptors are [r_{cut}=2, n_{max}=4, l_{max}=2], [r_{cut}=4, n_{max}=2, l_{max}=2], and [r_{cut}=3, n_{max}=5, l_{max}=3].

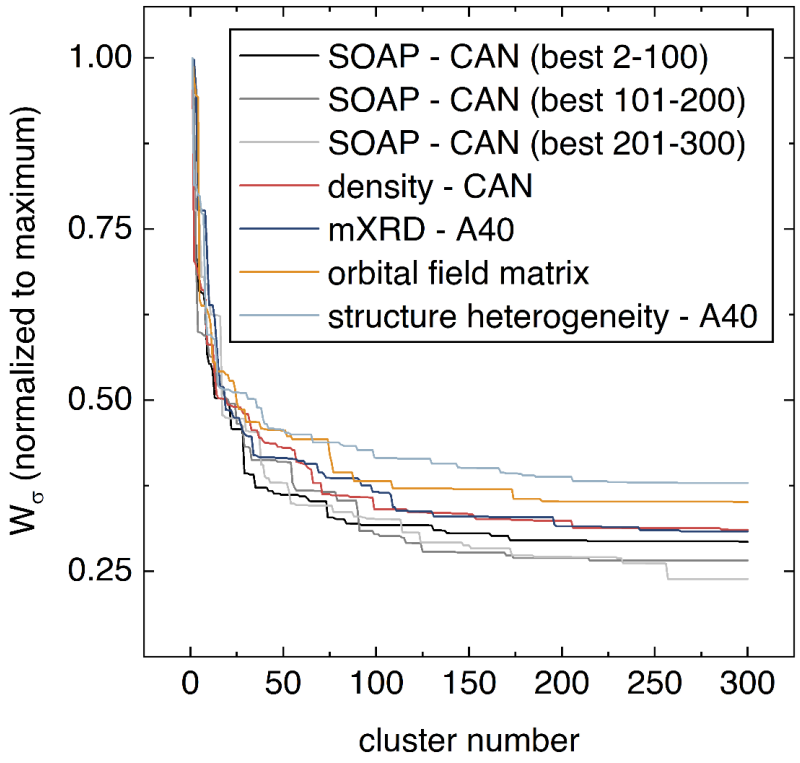

Figure A.1: W_σ vs. cluster number for three different SOAP–CAN models compared with the best-performing models for density–CAN, mXRD–A40, orbital field matrix, and structure heterogeneity–A40. The three SOAP–CAN models are those with the lowest W_σ mean for the clustering ranges: 2–100, 101–200, and 201–300. Almost all SOAP–CAN models outperformed the best non-SOAP models, irrespective of the specific combination of r_{cut}, n_{max}, and l_{max} hyperparameters.

A.2 W_{E_a} optimization

Each clustering outcome is also assessed by labeling with approximate activation energies for ion hopping. The activation energies are calculated using a bond valence site energy (BVSE) method developed by Adams and Rao [1, 2]. The strategy approximates the E_a as the sum of an attractive Morse-type potential term and a repulsive Coulombic interaction term. The Morse-type potential term represents mobile ion interactions with lattice anions. While the Coulombic interaction term represents mobile ion interactions with lattice cations. Relative to DFT-based methods, the BVSE method is a computationally lean approach that can be used to rapidly assess thousands of structures. However, the BVSE method tends to overestimate activation energies because it (1) does not allow for structural relaxation as the mobile ion moves and (2) does not consider repulsive interactions between mobile ions [1, 2]. The BVSE method has been implemented by He et al. and is available for use through their python API [3]. Using the BVSE method, we label 6,845 structures with activation energies (6,845 is the number of structures successfully solved given a computing time cutoff of 20-minutes for each structure). Ward's minimum variance method applied to the activation energy labels (W_{E_a}) is calculated in a similar manner to the W_σ:

$$W_{E_a} = \sum_{k=1}^{n_c} \sum_{i \in C_k} \left[(E_{a,BVSE})_i - \overline{(E_{a,BVSE})}_k \right]^2$$

where n_c is the number of clusters in a set, C_k is cluster k, and where $\overline{(E_{a,BVSE})}_k$ denotes the mean for all labels in cluster k. Each descriptor's W_{E_a} results are shown in Figure A.2 for the first 50 clustering sets. For simplicity, only the best-performing simplification-descriptor combination is shown for each descriptor.

For E_a labels, all descriptor-simplification pairings result in better semi-supervised ML performance than randomized clustering. The SOAP descriptor performs well relative to most, but five other descriptors outperform it: CAVD, orbital field matrix–CAN, density, mXRD–CAMN, and the packing efficiency descriptors. The favorable performance of CAVD is anticipated because the BVSE calculation directly uses the CAVD descriptor as a parameter. The favorable performance of the density and packing efficiency descriptors may be explained by their similarity to CAVD: the Voronoi decomposition to encode void space is dependent on the density and packing efficiency of the structure. Similarly, the orbital field matrix descriptor relies on calculation of Voronoi polyhedra to understand the coordination environment for

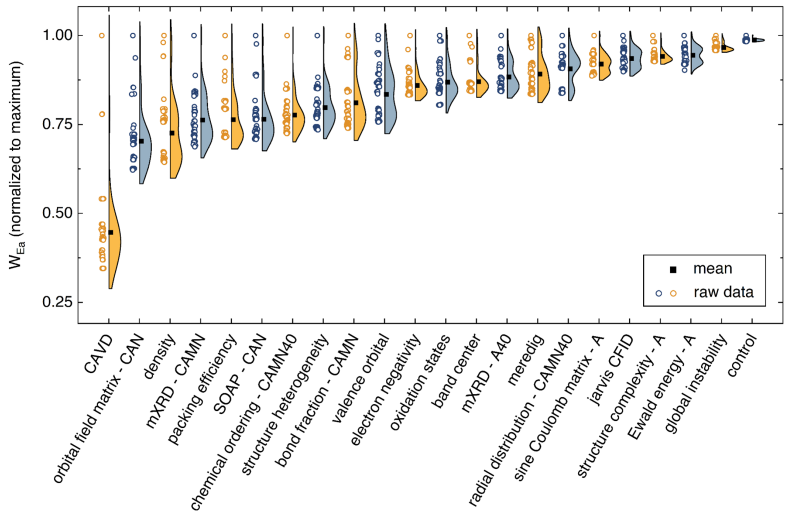

Figure A.2: The W_{E_a} for the first 50 clusters generated using each descriptor. Half-violin plots show the raw W_{E_a} score for each cluster as symbols next to the violin distribution. Simplification-descriptor combinations are sorted in order of ascending mean. The control is a random assignment of clusters, with W_{E_a} values averaged over 100 randomly assigned sets.

each atom. A mXRD–CAMN descriptor-simplification performs well on the BVSE label set; however, the mXRD representation used by Toyota (mXRD–A40) drops from to 14$^{\text{th}}$ best on the E_a label set. The result may suggest that the mXRD–A40 pairing does not generalize well. When comparing the top 10 descriptors for each label set, 6 descriptors are common to both approaches: SOAP, density, mXRD, structure heterogeneity, orbital field matrix, and bond fraction.

A.3 Second-order SOAP descriptor

Semi-supervised ML models may be further improved by merging descriptors and clustering on the union representation. Second order descriptor unions are examined by combining the best-performing descriptors with all other descriptors. The two input descriptor vectors (d_A and d_B) were combined with a mixing ratio (α) to yield the union representation (d_{AB}):

$$d_{AB} = d_A \cup \alpha d_B$$

The ideal mixing ratio is unknown for each union and we find that incremental changes to the mixing ratio do not result in continuous changes to the W_σ. Thus, outcomes are manually screened for mixing ratios from 10^{-6} to 10^{6} (see supplemental information – section VI). Most descriptor unions result in no improvement to the W_σ across all mixing ratios. However, the W_σ for SOAP when mixing with the non-simplified sine Coulomb matrix descriptor (for $\alpha = 2 \cdot 10^{-6} - 4 \cdot 10^{-6}$) is lowered by 2–3%, with the exact percentage depending on the depth of clustering.

Almost no descriptor combinations are successful in reducing the W_σ. Excluding combinations that include the SOAP–CAN descriptor, no combinations outperform the 1$^{\text{st}}$–order SOAP–CAN representation. For combinations that include SOAP–CAN, some mixing ratios with the sine Coulomb matrix and the Ewald energy descriptors resulted in modest improvements in the W_σ. The best improvement is found when mixing SOAP–CAN with the sine Coulomb descriptor for $\alpha = 2 \cdot 10^{-6}, 3 \cdot 10^{-6},$ and $4 \cdot 10^{-6}$. All three combinations result in the same improved curve, plotted below in Figure A.3.

The agglomerative dendrogram in the main text shows that the 2$^{\text{nd}}$–order SOAP–CAN descriptor facilitates aggregation of high-conductivity labels. In the simplified 9-cluster representation, most of the high-conductivity ($\sigma_{RT} > 10^{-5}$ S cm^{-1}) labels are contained within the 2$^{\text{nd}}$ "mega cluster". The 2$^{\text{nd}}$ mega cluster accounts for only 15% of the input structure. By clustering further, increasingly dense representations are found. For example, at the 241$^{\text{st}}$ clustering depth, the 21 high-conductivity labels have been sorted into five subclusters (Figure A.4. Taken together, the five subclusters account for 52.5% of the high conductivity labels while containing only 2.2% of the input structures. We note that the control (random clustering) exhibits a Ward Variance 214% greater than the 2$^{\text{nd}}$–order SOAP–CAN model at the 241$^{\text{st}}$ clustering depth. The difference in Ward Variance illustrates that the 2$^{\text{nd}}$–

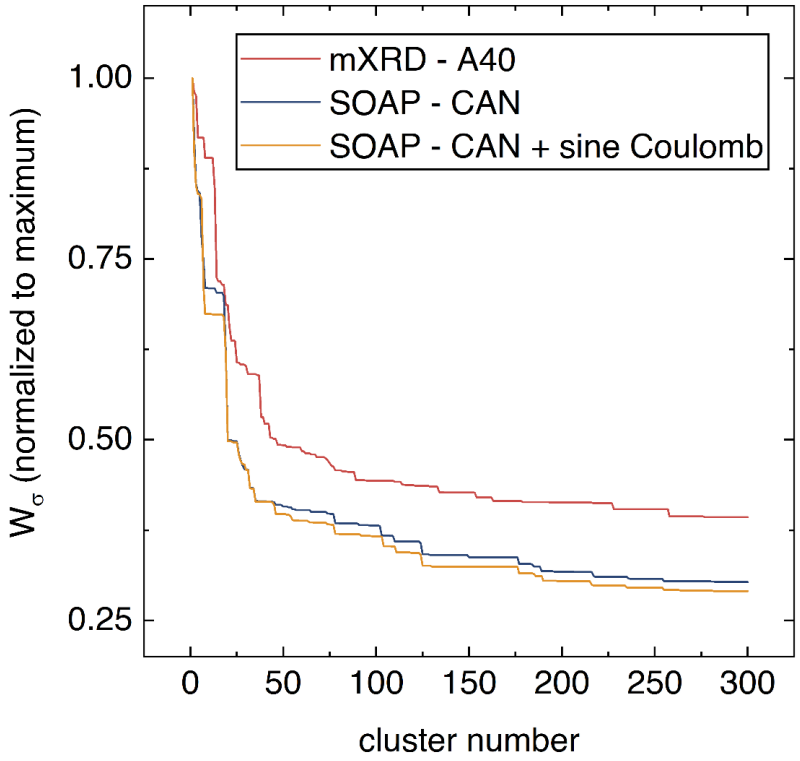

Figure A.3: The best performing 2nd order descriptor: SOAP–CAN mixed with the sine Coulomb descriptor. The clustering performance is shown for the full label set of 219. Since the mXRD–A40 representation is also compatible with the full label set, it is shown for reference. The 2nd order descriptor outperforms the 1st–order SOAP–CAN descriptor at most depths of clustering.

order SOAP–CAN model is much better at identifying high-conductivity structures, relative to random selection.

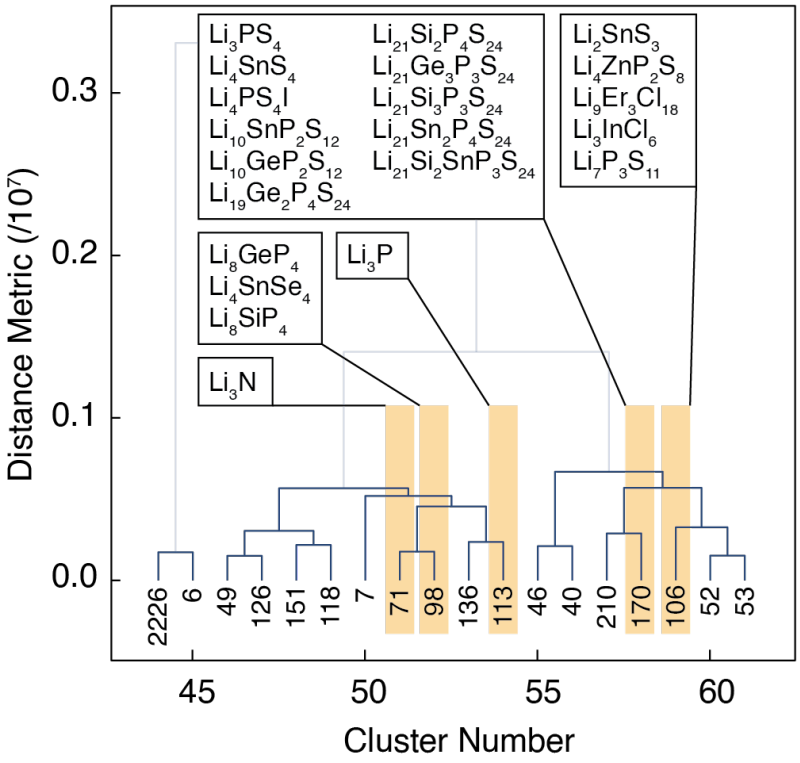

Figure A.4: The partial agglomerative dendrogram generated for the 2nd–order SOAP–CAN descriptor-simplification. The area shown is the 2nd mega cluster taken from Figure 3 of the main text. At a clustering depth of 241, the 21 high-conductivity labels are sorted into 5 clusters which account for 2.2% of the input structures.

A.4 Climbing Image – Nudged Elastic Band

Migration barriers for Li ion hopping are evaluated with the Climbing Image – Nudged Elastic Band (CI–NEB) method as implemented in the QuantumESPRESSO PWneb software package [4–7]. Density-functional theory (DFT) calculations are performed using the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation functional and projector-augmented wave (PAW) sets [8, 9]. Convergence testing for the kinetic-energy cutoff of the plane-wave basis and the k-point sampling is performed for each structure to ensure an accuracy of 1 meV per atom. The lattice parameters and atomic positions of the as-retrieved structure are optimized. Supercells are created for each structure that are a minimum of 10 Å in each lattice direction to minimize interactions between periodic images of the mobile ion. To study the migration barrier in the dilute limit, a single Li vacancy is created in the boundary endpoint structures of each studied pathway. A uniform background charge is used to balance excess charge. Each boundary configuration is relaxed until the force on each atom is less than 3×10^{-4} eV/Å. Images are created by linearly interpolating framework atomic positions between the initial and final boundary configurations. The initial pathway for the mobile ion is generated from the BVSE output minimum energy pathway to promote faster convergence of the NEB calculation. An NEB force convergence threshold of 0.05 eV/Å is used. The calculation is first converged using the default NEB algorithm and then restarted with the CI scheme to allow for the maximum energy of the pathway to be determined.

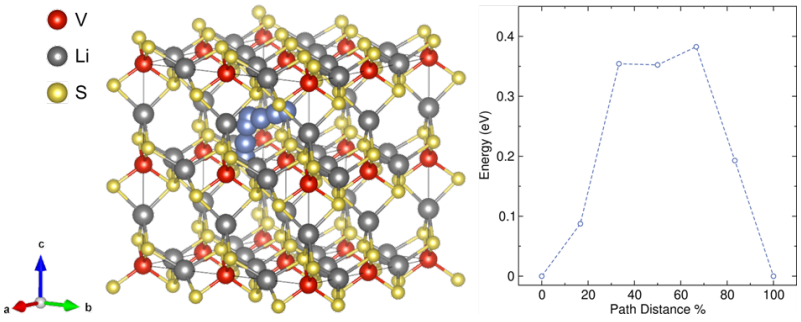

Figure A.5: The 2×2×2 supercell of Li_3VS_4 used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

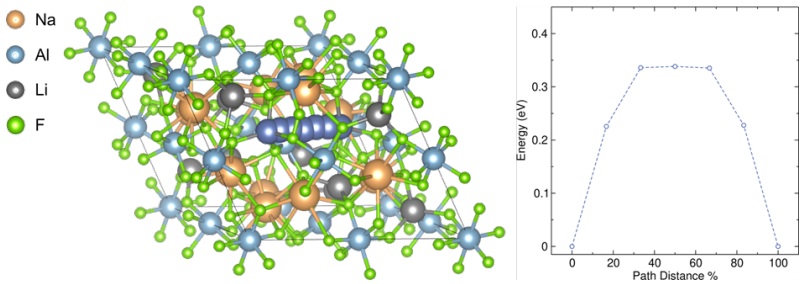

Figure A.6: The primitive cell of $Na_3Li_3Al_2F_{12}$ used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

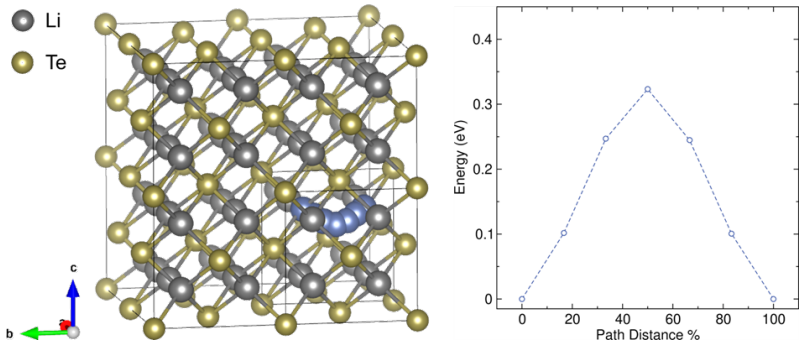

Figure A.7: The 2×2×2 supercell of Li_2Te used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

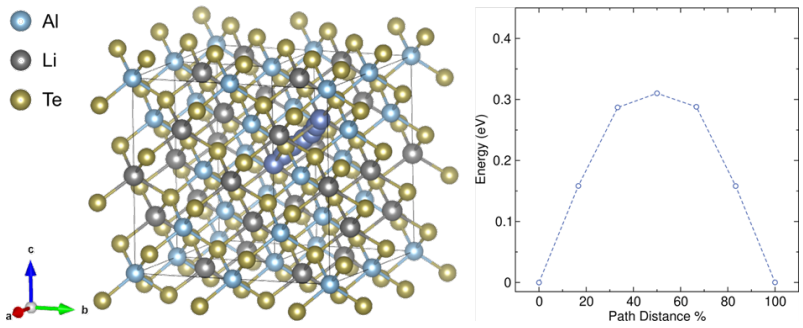

Figure A.8: The 2×2×1 supercell of $LiAlTe_2$ used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

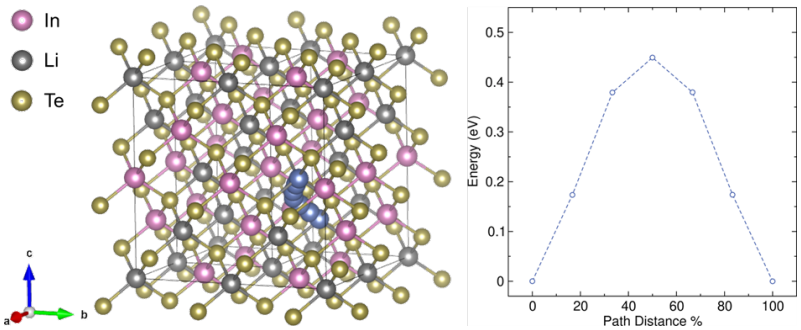

Figure A.9: The 2×2×1 supercell of $LiInTe_2$ used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

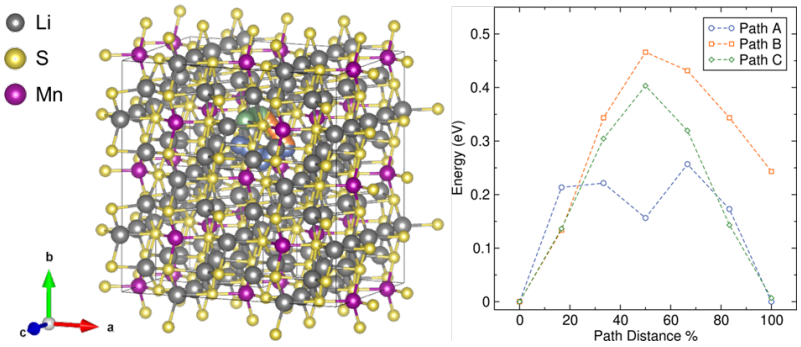

Figure A.10: The 2×2×2 supercell of Li_6MnS_4 used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

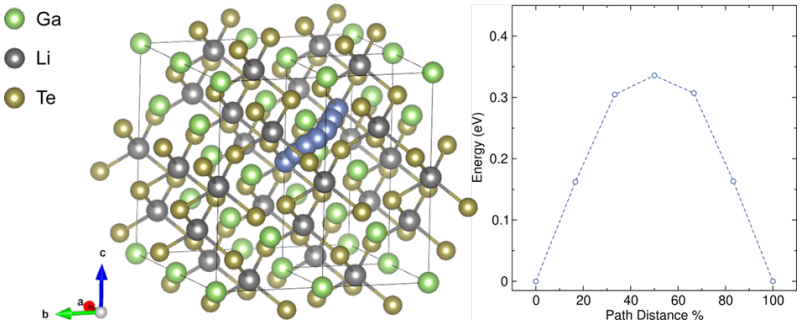

Figure A.11: The 2×2×1 supercell of $LiGaTe_2$ used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

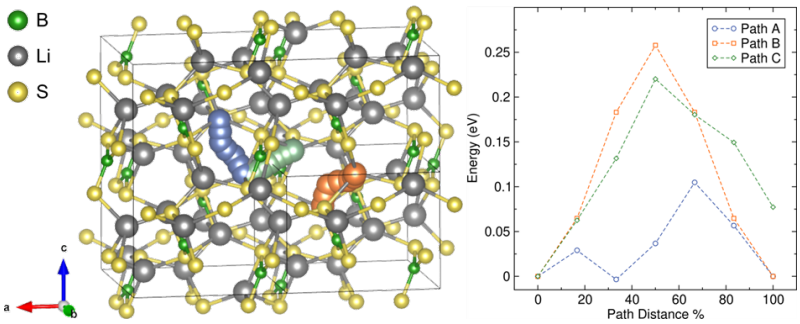

Figure A.12: The 2×1×2 supercell of Li_3BSS_3 used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

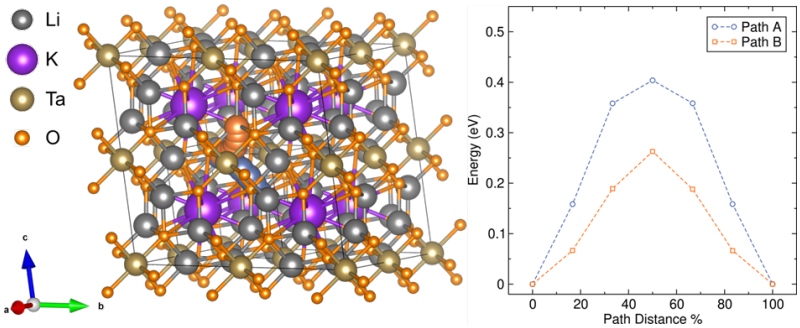

Figure A.13: The 2×2×2 supercell of KLi_6TaO_6 used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

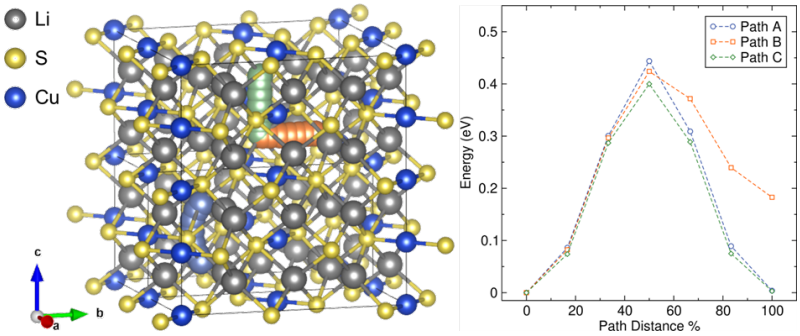

Figure A.14: The 2×1×2 supercell of Li_3CuS_2 used for the CI–NEB calculation of Li migration energy. Blue atoms represent the Li position from the CI–NEB output images.

A.5 a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ impedance

Electrochemical impedance data for the amorphized Si-substituted $LiBS_3$ (a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$) suggests the presence of two RC features. The VSP-300 potentiostat can supply a maximum sinusoidal frequency of 3 MHz, sufficient to resolve a partial semicircle in the Nyquist impedance plot (Figure A.15). Attempted fits to the partial semi-circle reveal that it would not intersect the origin at higher frequencies, suggesting the presence of an additional RC feature. It is plausible that two RC features exist, describing the bulk and grain-boundary transport of Li^+. A more conservative estimate of the conductivity (σ_{tot}) can be derived by extrapolating a linear of the Warburg tail to the x intercept. While the more conservative estimate is used in the main manuscript, we note here that the actual bulk conductivity is likely higher.

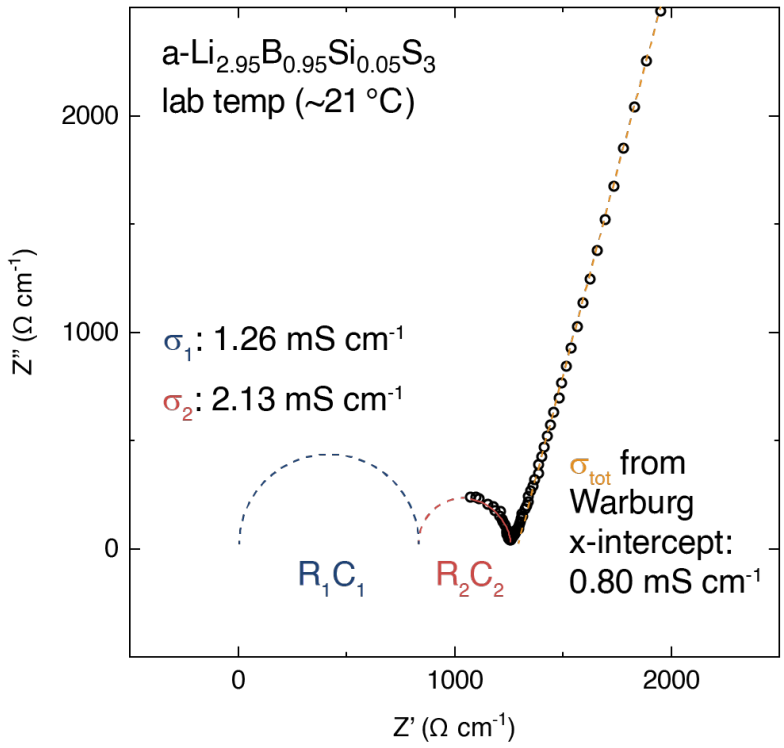

Figure A.15: Nyquist data for a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$ near room temperature. The partially resolved semi-circular features suggests the presence of at least two RC circuit elements.

A.6 Full list of promising structures

Excluding the labeled dataset, there are 50 compounds that are predicted to be stable and to exhibit a Li-hopping activation energy below 600 meV. Ten of the predicted compounds have already been experimentally examined and are hereafter excluded: Li_2O, Li_2S, LiCl, LiI, LiBr, Li_6AsS_5I, $Li_4Ti_5O_{12}$, Li_2InCl_3, $LiInI_4$, and Li_6NiCl_8. Another nine are excluded because they are used in cathodes, anodes, or glassy electrolyte formulations: $LiFeCl_4$, Li_2CO_3, Li_2PtO_3, $Li_2NiGe_3O_8$, Li_2CrO_4, Li_2SeO_4, Li_4AIS, $Li_2Mn_3NiO_8$, and $LiInSe_2$. The remaining 31 promising structures are discussed below and plotted by ascending activation energy in Figure A.16.

a. Stable compounds

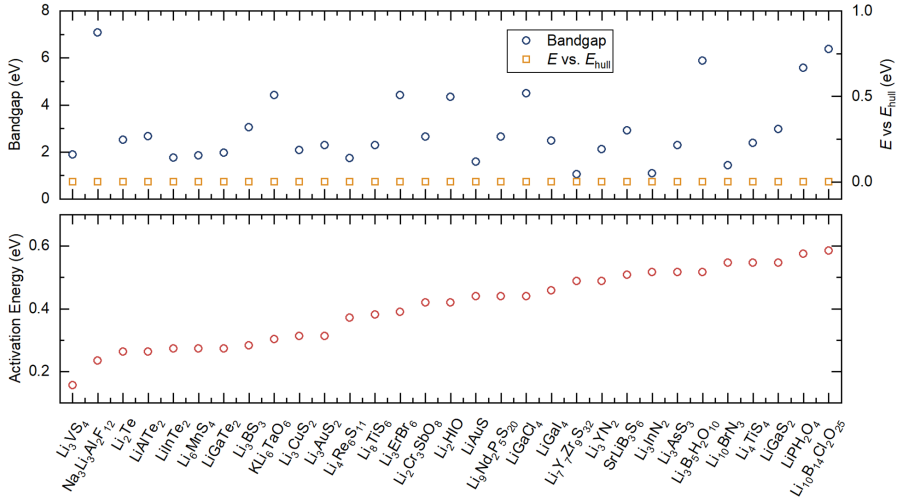

Figure A.16: The 31 promising structures that are predicted to be stable and to exhibit Li-hopping activation energy below 600 meV.

b. Quasi-stable compounds (E_{hull} below 15 meV)

Excluding the labeled dataset, there are 34 compounds that are predicted to be within 15 meV of the convex hull (E_{hull}) and to exhibit a Li-hopping activation energy below 600 meV. Ten of the predicted compounds have already been experimentally examined and are hereafter excluded: Li_3SbS_4, Li_6AsS_5I, Li_6PS_5I, Li_3ScCl_6, Li_2MnBr_4, Li_3N, $LiTi_2P_3O_{12}$, $Li_{10}SiP_2S_{12}$, Li_2ZnCl_4, and Li_3InO_3. Another three are currently being excluded because they are used in cathodes: Li_3NbS_4, Li_3CuS_2, Li_6VCl_8. The remaining 21 promising structures are discussed below and plotted by ascending activation energy in Figure A.17.

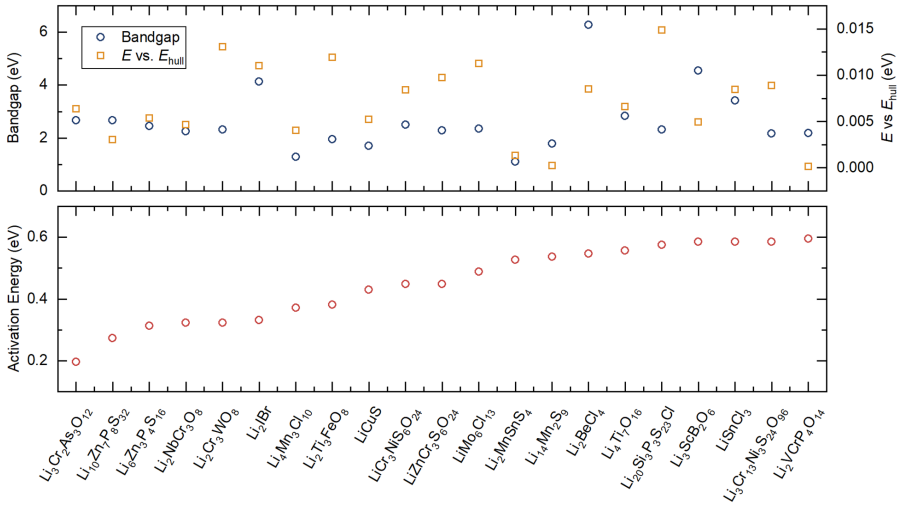

Figure A.17: The 21 promising structures that are predicted to be within 15 meV of E_{hull} and to exhibit Li-hopping activation energy below 600 meV.

c. Unknown-stability compounds (sans Materials Project entry)

There are 18 predictions that have no associated Material's Project entry. These structures lack stability data. Seven of the predicted compounds have already been experimentally examined and are hereafter excluded: Li_2O, Li_2S, $Li_7Y_2Zr_9S_{32}$, $Li_4SnSe_4O_{13}$, Li_2MnBr_4, Li_5AlS_4, and $Li_3Fe_2P_3O_{12}$. Another five are currently being excluded because they are used in cathodes: $Li_2Mn_3NiO_8$, $Li_2Mn_3CoO_8$, $Li_5Mn_{16}O_{32}$, $Li_2Mn_{15}AlO_{32}$, and $Li_3V_2P_3O_{12}$. The remaining 6 promising structures are discussed below and plotted in order of ascending activation energy in Figure S17.

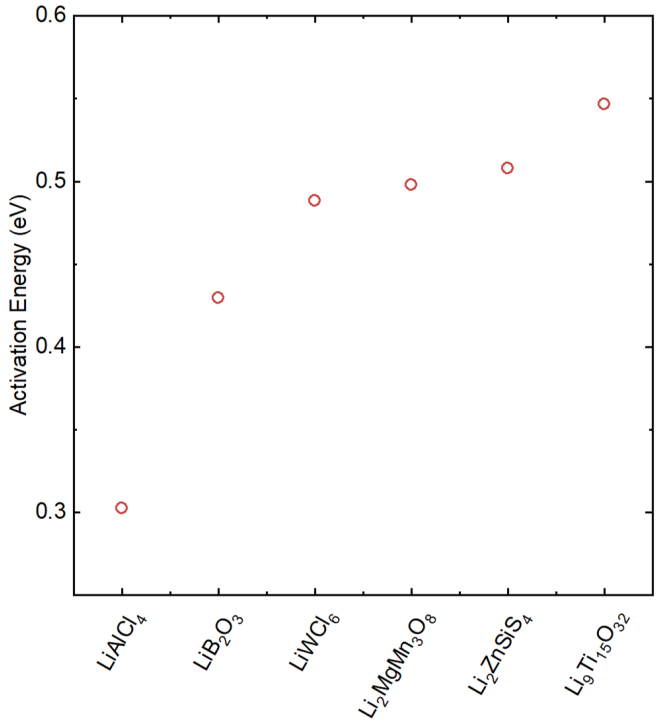

Figure A.18: The six promising structures that lack Materials Project data but are predicted to exhibit Li-hopping activation energy below 600 meV.

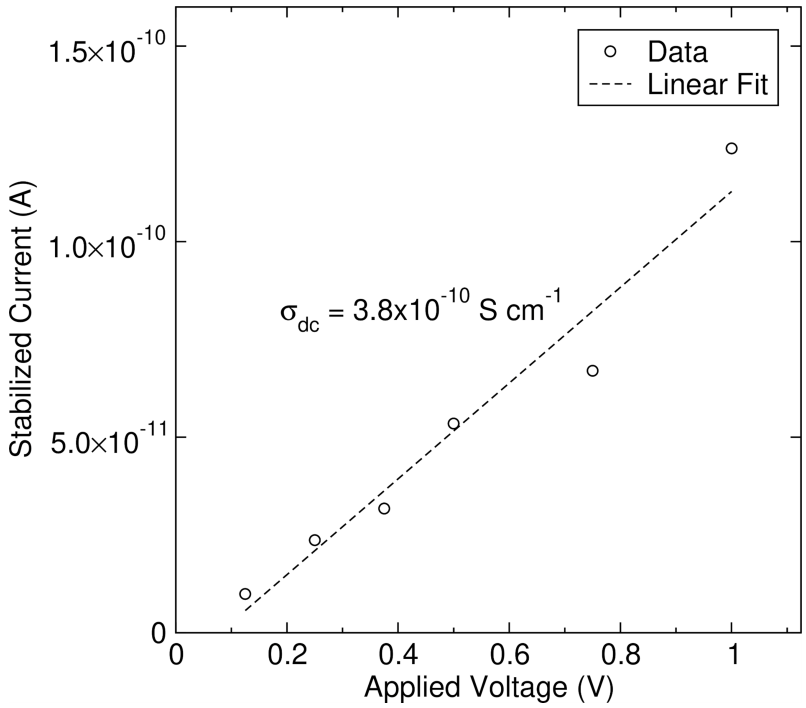

Figure A.19: Steady-state current of Au/a-$Li_{2.95}B_{0.95}Si_{0.05}S_3$/Au cell for different voltage polarizations. Measurements were done at 25°C with applied voltages of 0.125 V, 0.25 V, 0.375 V, 0.5 V, and 1.0 V.

A.7 Bibliography

[1] Chen, H.; Adams, S. Bond Softness Sensitive Bond-Valence Parameters for Crystal Structure Plausibility Tests. *IUCrJ* **2017**, *4*, 614–625.

[2] Adams, S. *Bond Valences*; Springer, 2013; pp 91–128.

[3] He, B.; Mi, P.; Ye, A.; Chi, S.; Jiao, Y.; Zhang, L.; Pu, B.; Zou, Z.; Zhang, W.; Avdeev, M.; others A Highly Efficient and Informative Method to Identify Ion Transport Networks in Fast Ion Conductors. *Acta Materialia* **2021**, *203*, 116490.

[4] Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *The Journal of chemical physics* **2000**, *113*, 9901–9904.

[5] Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *The Journal of chemical physics* **2000**, *113*, 9978–9985.

[6] Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; others QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum-simulations of Materials. *Journal of physics: Condensed matter* **2009**, *21*, 395502.

[7] Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M. B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; others Advanced Capabilities for Materials Modelling with Quantum ESPRESSO. *Journal of physics: Condensed matter* **2017**, *29*, 465901.

[8] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical review letters* **1996**, *77*, 3865.

[9] Dal Corso, A. Pseudopotentials Periodic Table: From H to Pu. *Computational Materials Science* **2014**, *95*, 337–350.

Appendix B

SUPPORTING INFORMATION FOR CHAPTER 3: SUBSTITUTION OF Li_3BS_3: REVEALING NEW SUPERIONIC CONDUCTOR PHASES AND THE SIGNIFICANCE OF CRYSTALLINITY

[This chapter is temporarily embargoed.]

Appendix C

SUPPORTING INFORMATION FOR CHAPTER 4: CLASSIFICATION OF (DIS)ORDERED STRUCTURES AS SUPERIONIC LITHIUM CONDUCTORS

Table C.1: Statistics of Test and Training/Validation splits. Ten percent of the initial database of 548 ionic conductivity and structure pairs is designated as the test set. The remaining training/validation set is used for model and feature evaluation under *k*-fold and leave-one-cluster-out cross validation schemes.

Set	No. entries	Max $\log_{10}(\sigma_{exp})$	Min $\log_{10}(\sigma_{exp})$	Mean $\log_{10}(\sigma_{exp})$	No. positive class	No. negative class
Test	55	-2.03	-20.80	-5.46	25	30
Training and Validation	493	-1.55	-30.57	-6.31	186	307

Table C.2: Statistics of Training and Validation splits for *k*-fold cross validation rounded to two decimal places.

Validation Fold	Set	No. entries	Max $\log_{10}(\sigma_{exp})$	Min $\log_{10}(\sigma_{exp})$	Mean $\log_{10}(\sigma_{exp})$	No. positive class	No. negative class
0	Training	394	-1.62	-30.57	-6.32	149	245
	Validation	99	-1.55	-22.40	-6.27	37	62
1	Training	394	-1.55	-30.57	-6.33	149	245
	Validation	99	-1.62	-25.78	-6.23	37	62
2	Training	394	-1.55	-25.78	-6.25	148	246
	Validation	99	-1.84	-30.57	-6.55	38	61
3	Training	395	-1.55	-30.57	-6.29	149	246
	Validation	98	-1.76	-24.40	-6.40	37	61
4	Training	395	-1.55	-30.57	-6.36	149	246
	Validation	98	-1.92	-19.74	-6.12	37	61

Table C.3: Statistics of Training and Validation splits for leave-one-cluster-out cross validation rounded to two decimal places.

Validation Fold	Set	No. entries	Max $\log_{10}(\sigma_{exp})$	Min $\log_{10}(\sigma_{exp})$	Mean $\log_{10}(\sigma_{exp})$	No. positive class	No. negative class
0	Training	422	-1.55	-30.57	-6.54	152	270
	Validation	71	-2.86	-17.36	-4.98	34	37
1	Training	410	-1.55	-25.83	-6.07	159	251
	Validation	83	-2.25	-30.57	-7.49	27	56
2	Training	408	-1.55	-30.57	-5.86	171	237
	Validation	85	-2.49	-25.78	-8.48	15	70
3	Training	419	-1.72	-30.57	-6.60	153	266
	Validation	74	-1.55	-14.00	-4.70	33	41
4	Training	451	-1.55	-30.57	-6.17	172	279
	Validation	42	-2.83	-20.81	-7.81	14	28
5	Training	475	-1.55	-30.57	-6.27	183	292
	Validation	18	-2.74	-17.01	-7.46	3	15
6	Training	446	-1.55	-30.57	-6.44	166	280
	Validation	47	-1.72	-13.54	-5.13	20	27
7	Training	457	-1.55	-30.57	-6.55	155	302
	Validation	36	-2.38	-9.26	-3.26	31	5
8	Training	478	-1.55	-30.57	-6.34	182	296
	Validation	15	-2.96	-12.64	-5.52	4	11
9	Training	471	-1.55	-30.57	-6.26	181	290
	Validation	22	-3.02	-25.83	-7.49	5	17

Table C.4: Statistics of compositional similarity between training and validation sets for k-fold validation. Compositional similarity is determined by identifying validation set entries that have at least one training set entry where the atomic fraction of each constituent element differs by no more than 5%. The difference is computed as $\frac{|x_1-x_2|}{(x_1+x_2)/2}$, where x_1 and x_2 are the atomic fractions of a given element in the validation and training entries, respectively. Rounded to two decimal places.

Validation fold	No. validation entries	No. similar compositions in training set	Percentage similar compositions in training set
0	99	15	15.15%
1	99	14	14.14%
2	99	11	11.11%
3	98	20	20.41%
4	98	21	21.43%

Table C.5: Statistics of compositional similarity between training and test sets. Compositional similarity is determined by identifying validation set entries that have at least one training set entry where the atomic fraction of each constituent element differs by no more than 5%. The difference is computed as $\frac{|x_1-x_2|}{(x_1+x_2)/2}$, where x_1 and x_2 are the atomic fractions of a given element in the validation and training entries, respectively. Rounded to two decimal places.

Fold	No. test entries	No. similar compositions in training set	Percentage similar compositions in training set
Test	55	13	23.64%

Table C.6: Statistics of compositional similarity between training and validation sets for leave-one-cluster-out cross validation. Compositional similarity is determined by identifying validation set entries that have at least one training set entry where the atomic fraction of each constituent element differs by no more than 5%. The difference is computed as $\frac{|x_1-x_2|}{(x_1+x_2)/2}$, where x_1 and x_2 are the atomic fractions of a given element in the validation and training entries, respectively. Rounded to two decimal places.

Validation fold	No. validation entries	No. similar compositions in training set	Percentage similar compositions in training set
0	71	0	0.00%
1	83	0	0.00%
2	85	0	0.00%
3	74	1	1.35%
4	42	0	0.00%
5	18	0	0.00%
6	47	1	2.13%
7	36	12	33.33%
8	15	0	0.00%
9	22	0	0.00%

Table C.7: Hyperparameter values explored with HyperOpt under leave-one-cluster-out cross validation.

Hyperparameter	Values
Batch size	32, 64, 128
Dropout probability initial	0.1, 0.15, 0.2, 0.25, 0.3
Dropout probability final	0.1, 0.15, 0.2, 0.25, 0.3
L2 kernel regularization factor initial	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}
L2 kernel regularization factor final	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}
No. layers	2, 3, 4
No. neurons	32, 64, 128, 256
Initial learning rate	2×10^{-5}, 4×10^{-5}, 6×10^{-5}, 8×10^{-5}, 1×10^{-4}
Maximum learning rate	1×10^{-4}, 1×10^{-3}, 3×10^{-3}, 5×10^{-3}, 7×10^{-3}

Table C.8: Optimal hyperparameter values for each choice of validation cluster.

Validation Cluster	Batch size	Dropout probability initial	Dropout probability final	L2 kernel regularization factor initial	L2 kernel regularization factor final	No. layers	No. neurons	Initial learning rate	Maximum learning rate
0	32	0.2	0.1	10^{-3}	10^{-1}	2	32	6×10^{-5}	5×10^{-3}
1	128	0.15	0.3	10^{0}	10^{-2}	3	128	4×10^{-5}	1×10^{-3}
2	32	0.1	0.15	10^{-2}	10^{-4}	4	64	2×10^{-5}	7×10^{-3}
3	32	0.2	0.2	10^{-4}	10^{-1}	2	128	6×10^{-5}	3×10^{-3}
4	128	0.15	0.25	10^{-2}	10^{-2}	3	128	2×10^{-5}	1×10^{-3}
5	128	0.1	0.15	10^{-2}	10^{-4}	3	64	4×10^{-5}	5×10^{-3}
6	32	0.1	0.2	10^{0}	10^{-4}	3	32	6×10^{-5}	1×10^{-3}
7	32	0.1	0.1	10^{0}	10^{-2}	2	32	8×10^{-5}	5×10^{-3}
8	128	0.15	0.25	10^{0}	10^{-2}	2	256	6×10^{-5}	5×10^{-3}
9	64	0.15	0.1	10^{-3}	10^{-4}	2	256	6×10^{-5}	1×10^{-3}

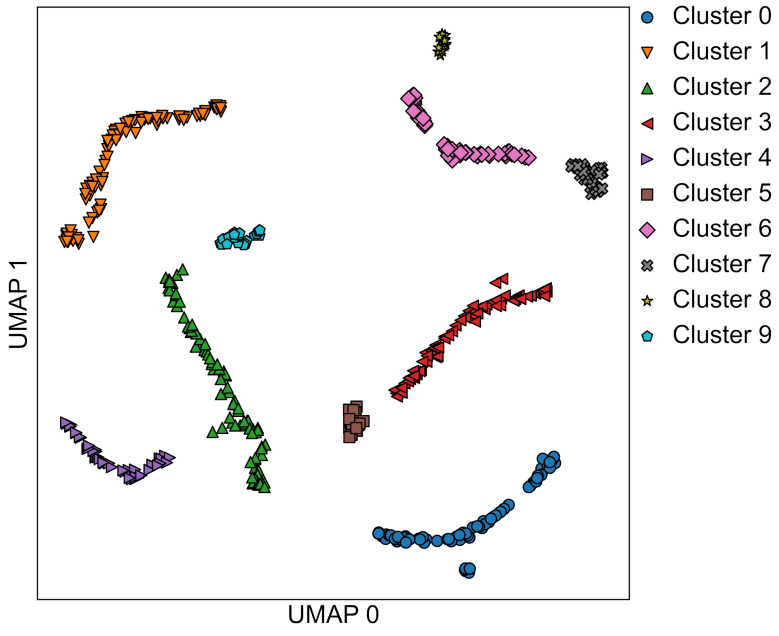

Figure C.1: UMAP projection of database ElMD features

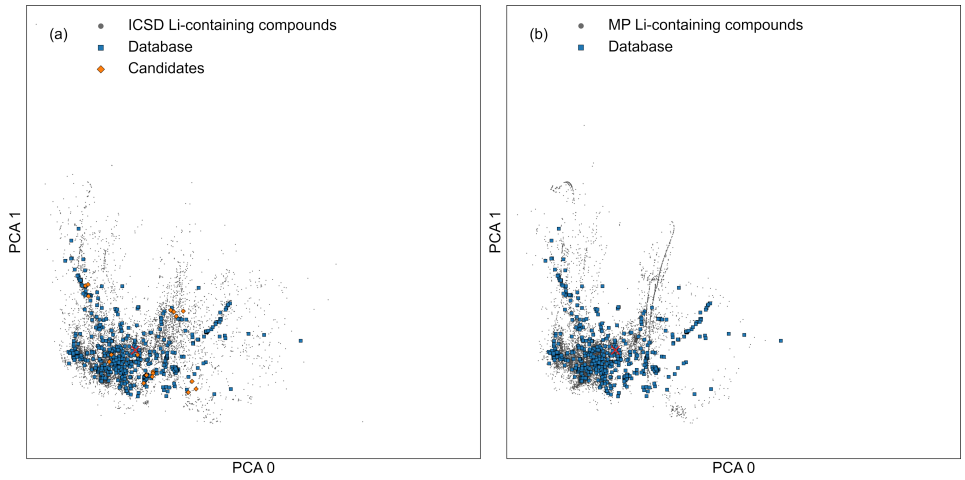

Figure C.2: Principal Component Analysis (PCA) of Li-containing compounds. (a) ICSD Li-containing compounds compared to database structures and candidate materials. (b) Materials Project (MP) Li-containing compounds compared to database structures. Each point represents a compound projected onto the first two principal components using the graph-based atom features.

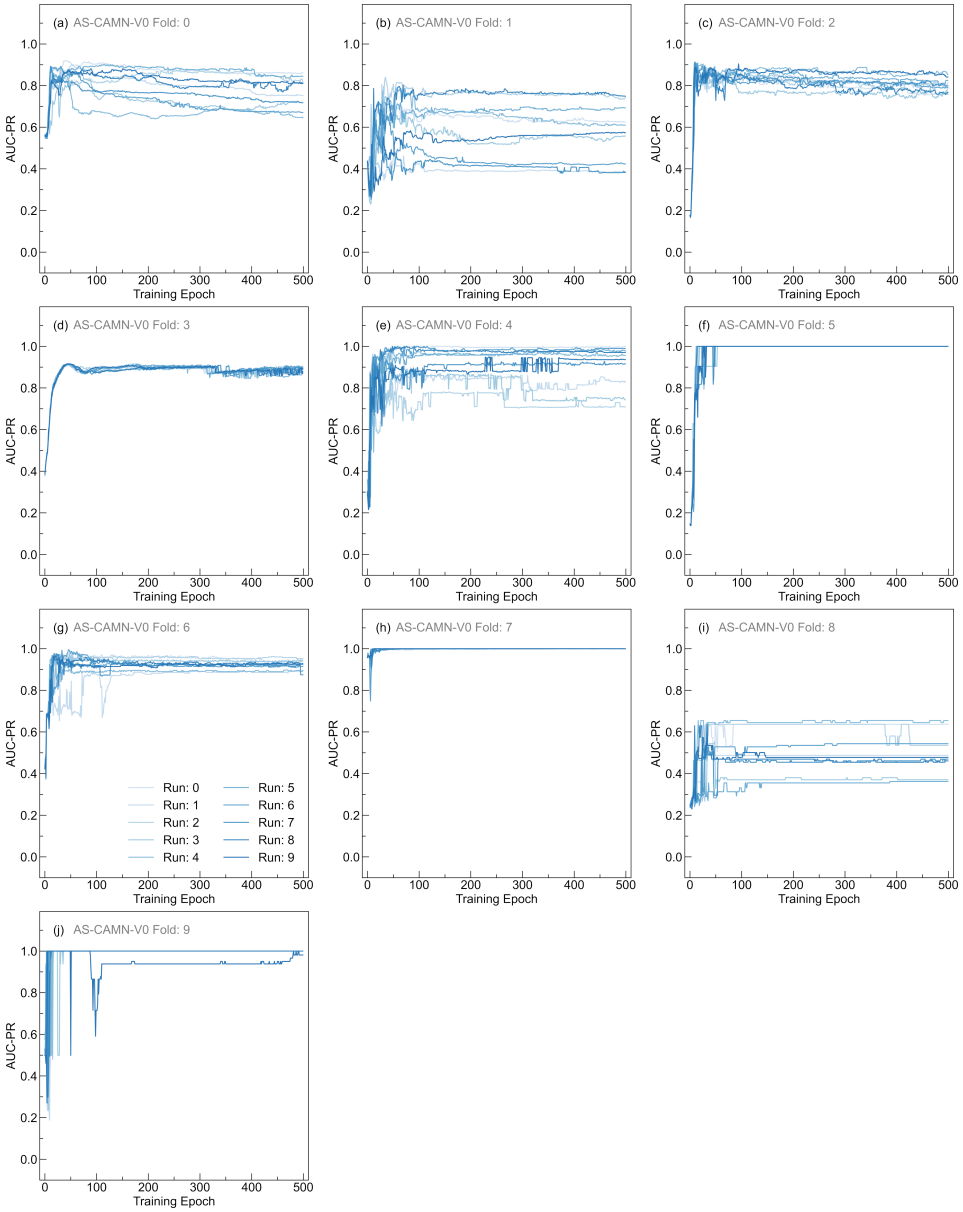

Figure C.3: Area under the precision-recall curve (AUC-PR) of the AtomSets CAMN-V_0 (AS-CAMN-V_0) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

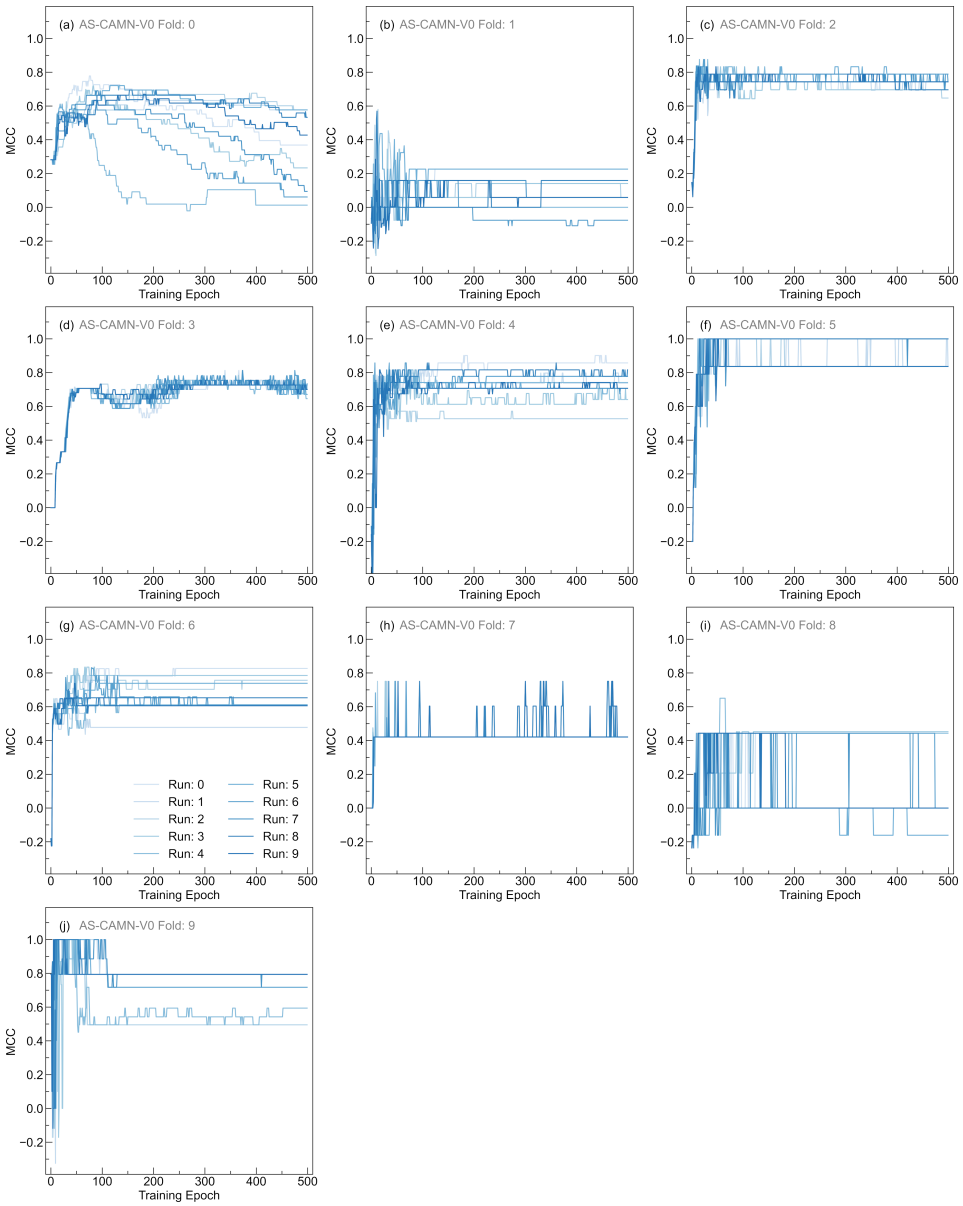

Figure C.4: Matthew's correlation coefficient (MCC) of the AtomSets CAMN-V_0 (AS-CAMN-V_0) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

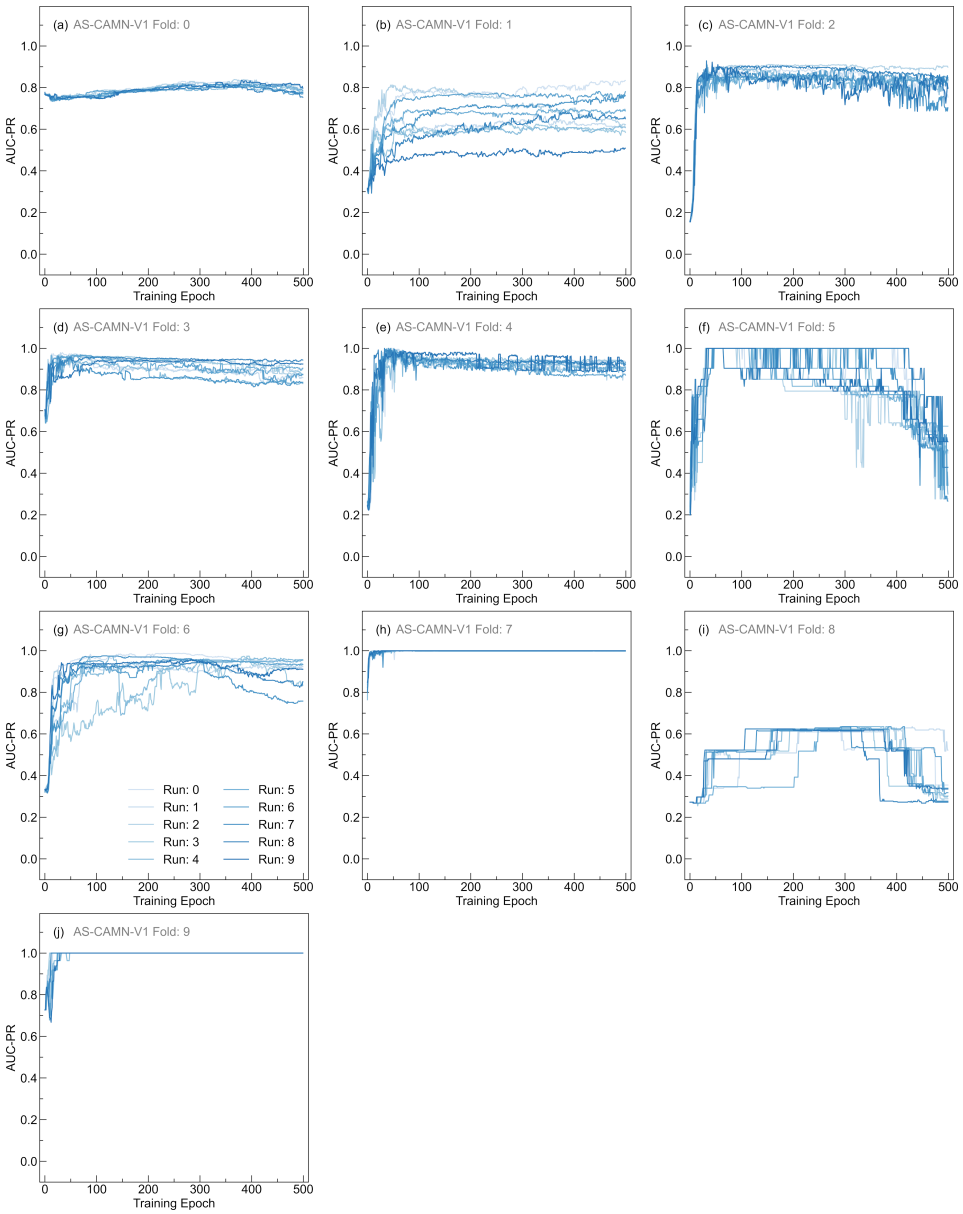

Figure C.5: Area under the precision-recall curve (AUC-PR) of the AtomSets CAMN-V_1 (AS-CAMN-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

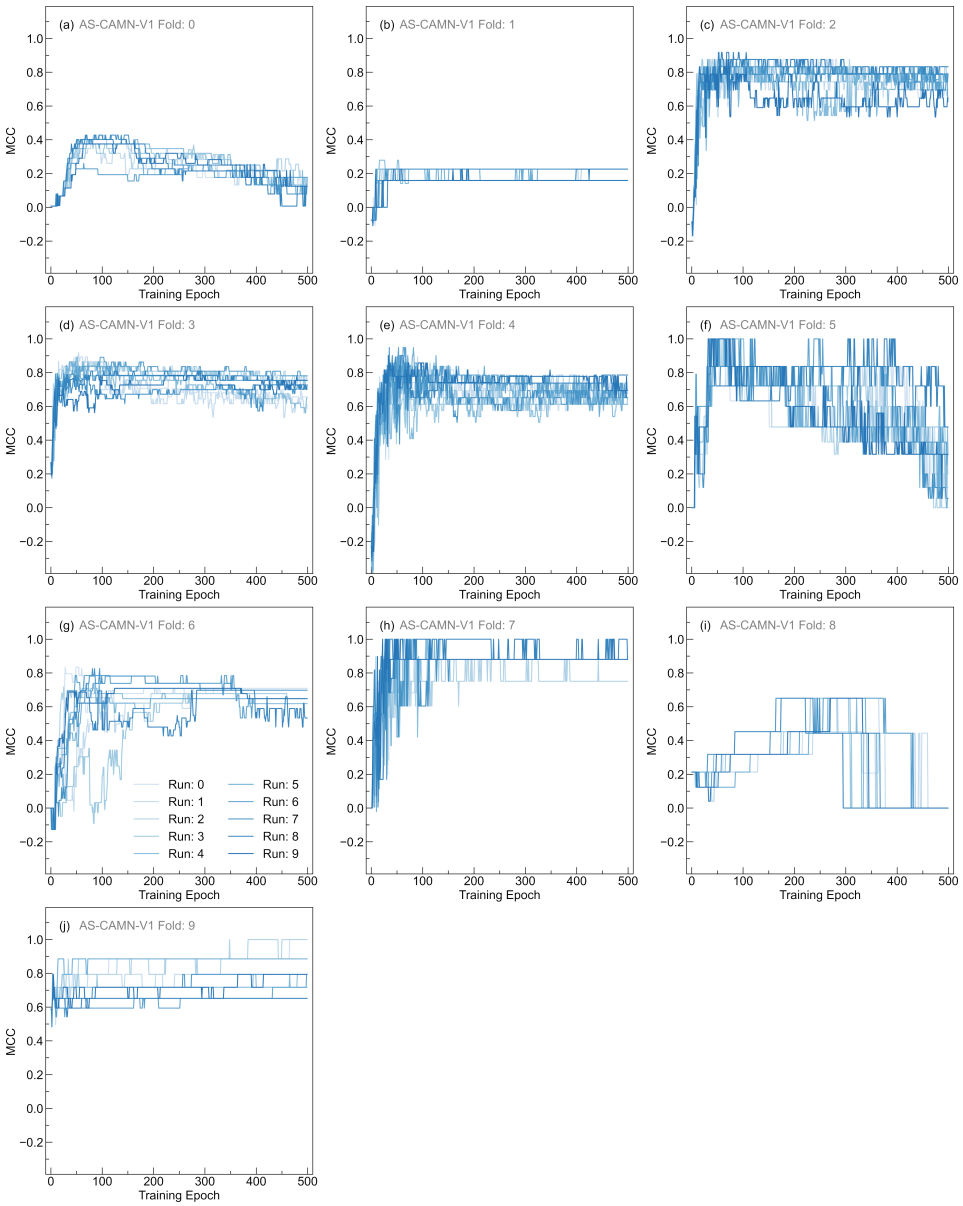

Figure C.6: Matthew's correlation coefficient (MCC) of the AtomSets CAMN-V_1 (AS-CAMN-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

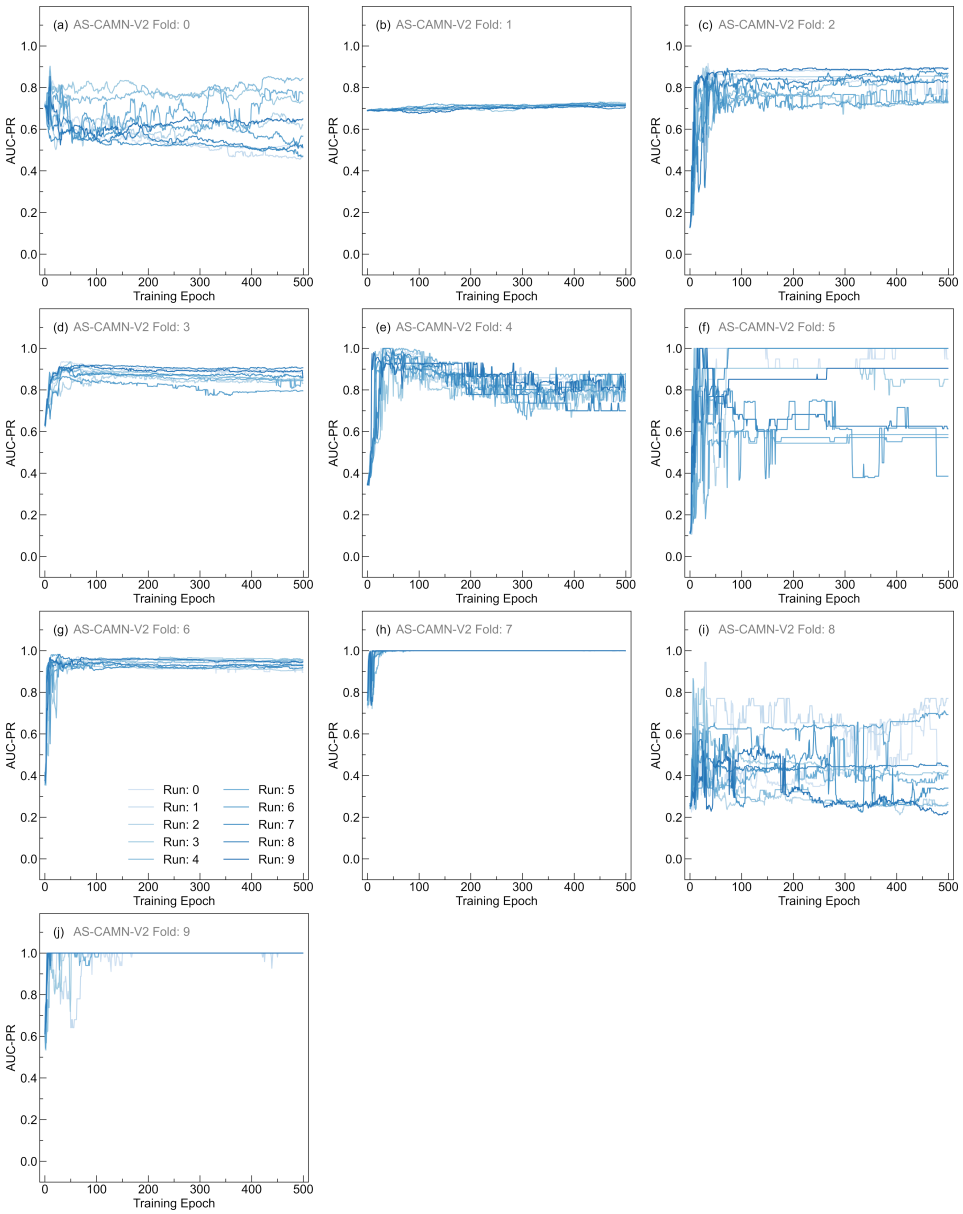

Figure C.7: Area under the precision-recall curve (AUC-PR) of the AtomSets CAMN-V_2 (AS-CAMN-V_2) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

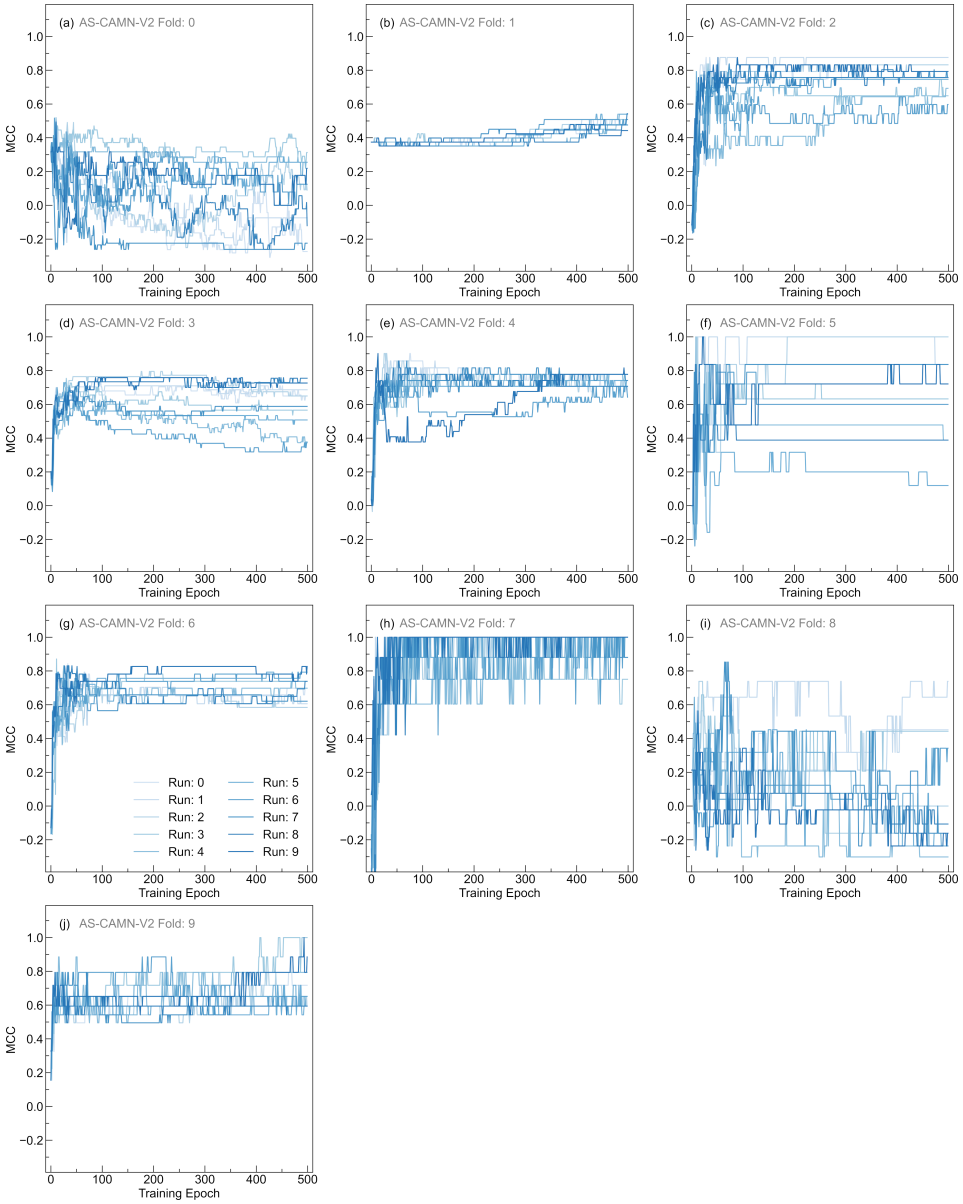

Figure C.8: Matthew's correlation coefficient (MCC) of the AtomSets CAMN-V_2 (AS-CAMN-V_2) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

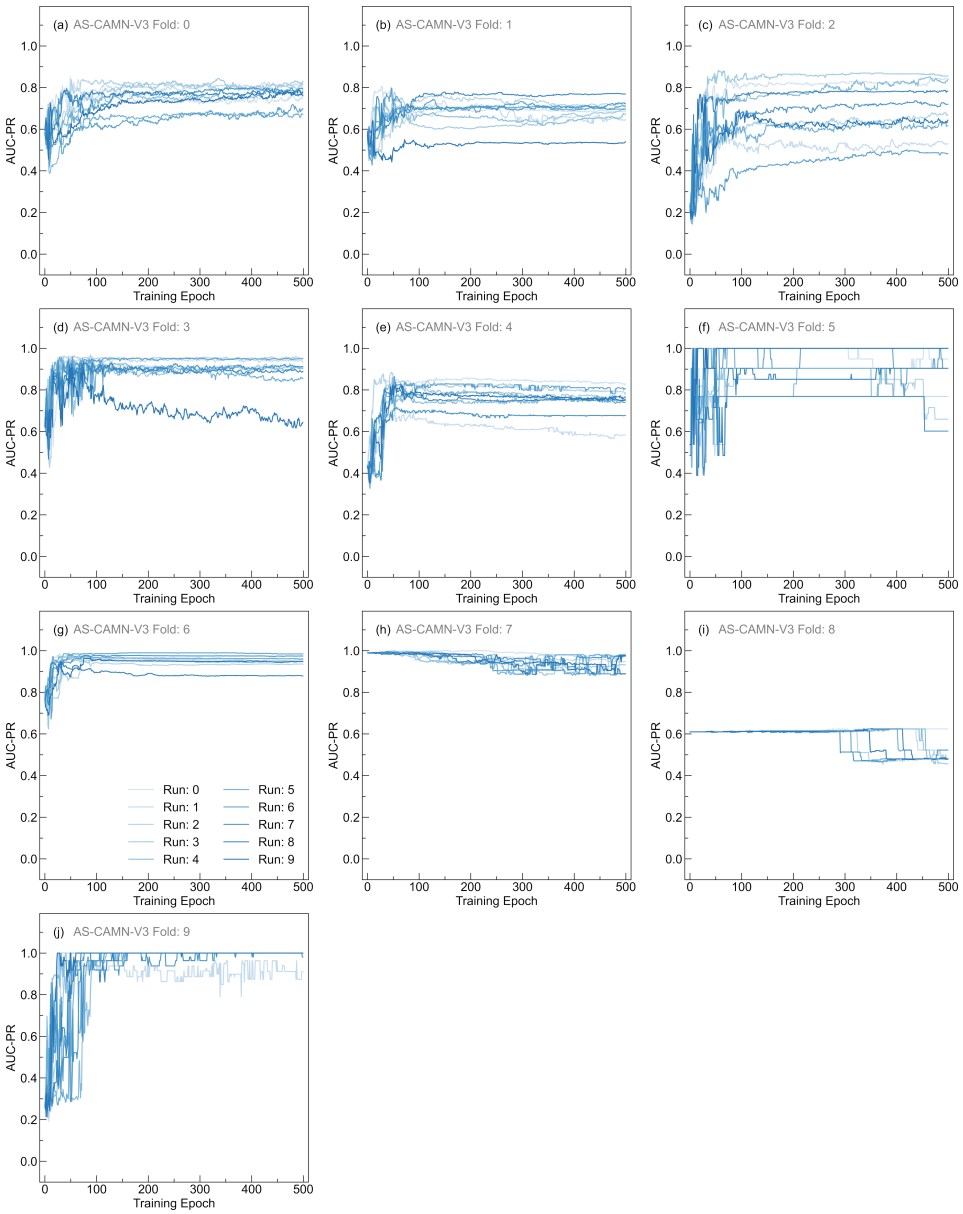

Figure C.9: Area under the precision-recall curve (AUC-PR) of the AtomSets CAMN-V_3 (AS-CAMN-V_3) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

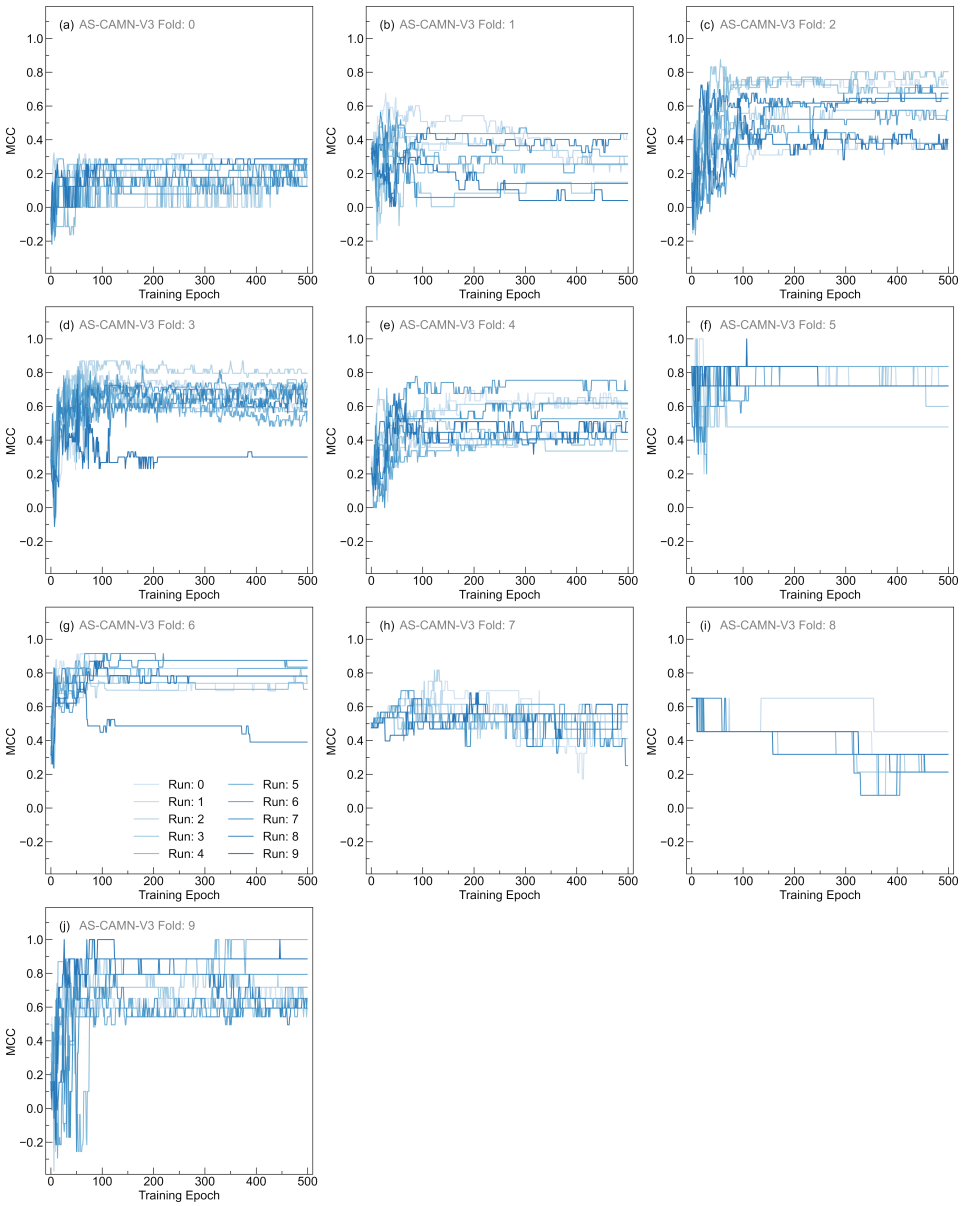

Figure C.10: Matthew's correlation coefficient (MCC) of the AtomSets CAMN-V_3 (AS-CAMN-V_3) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

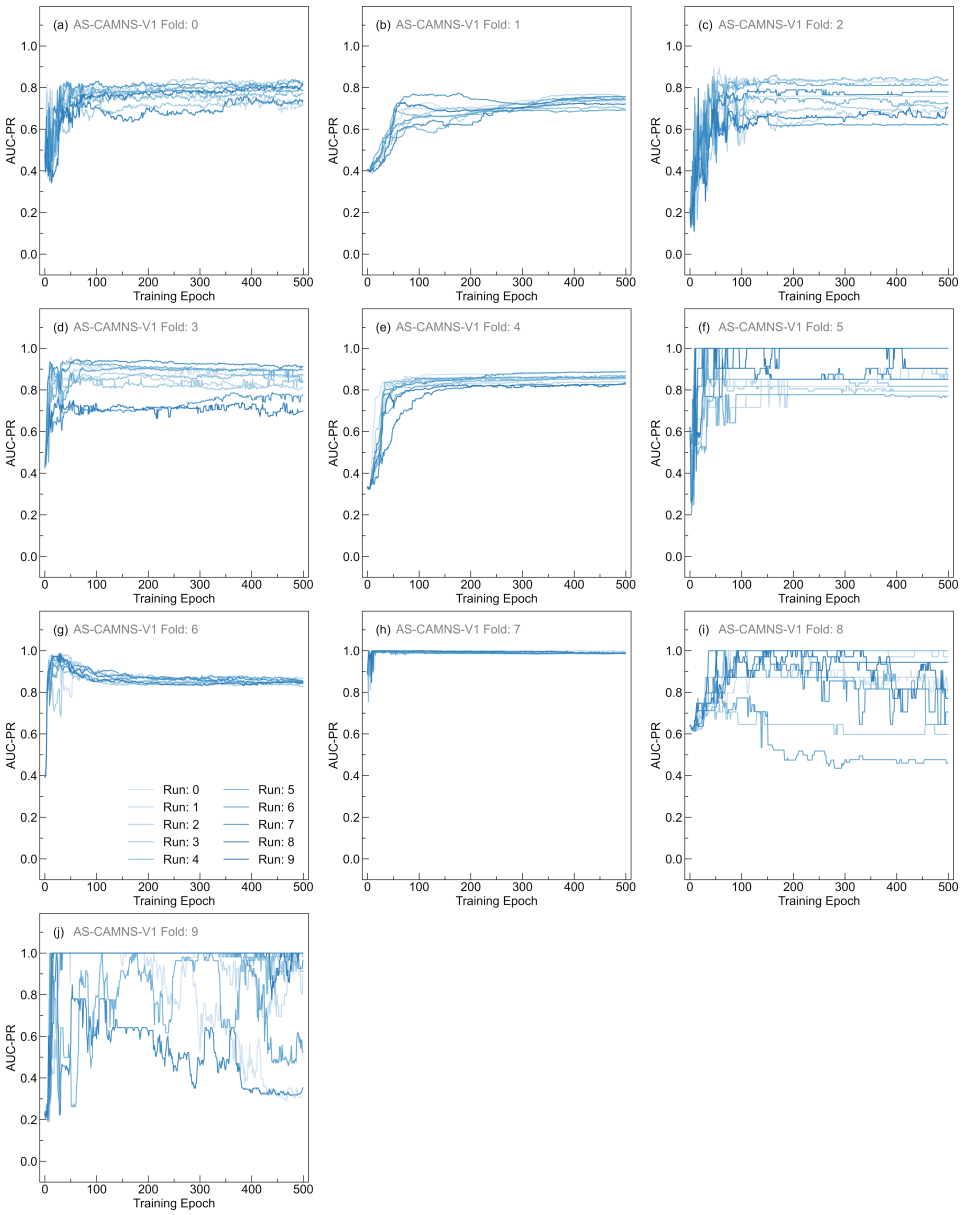

Figure C.11: Area under the precision-recall curve (AUC-PR) of the AtomSets CAMNS-V_1 (AS-CAMNS-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

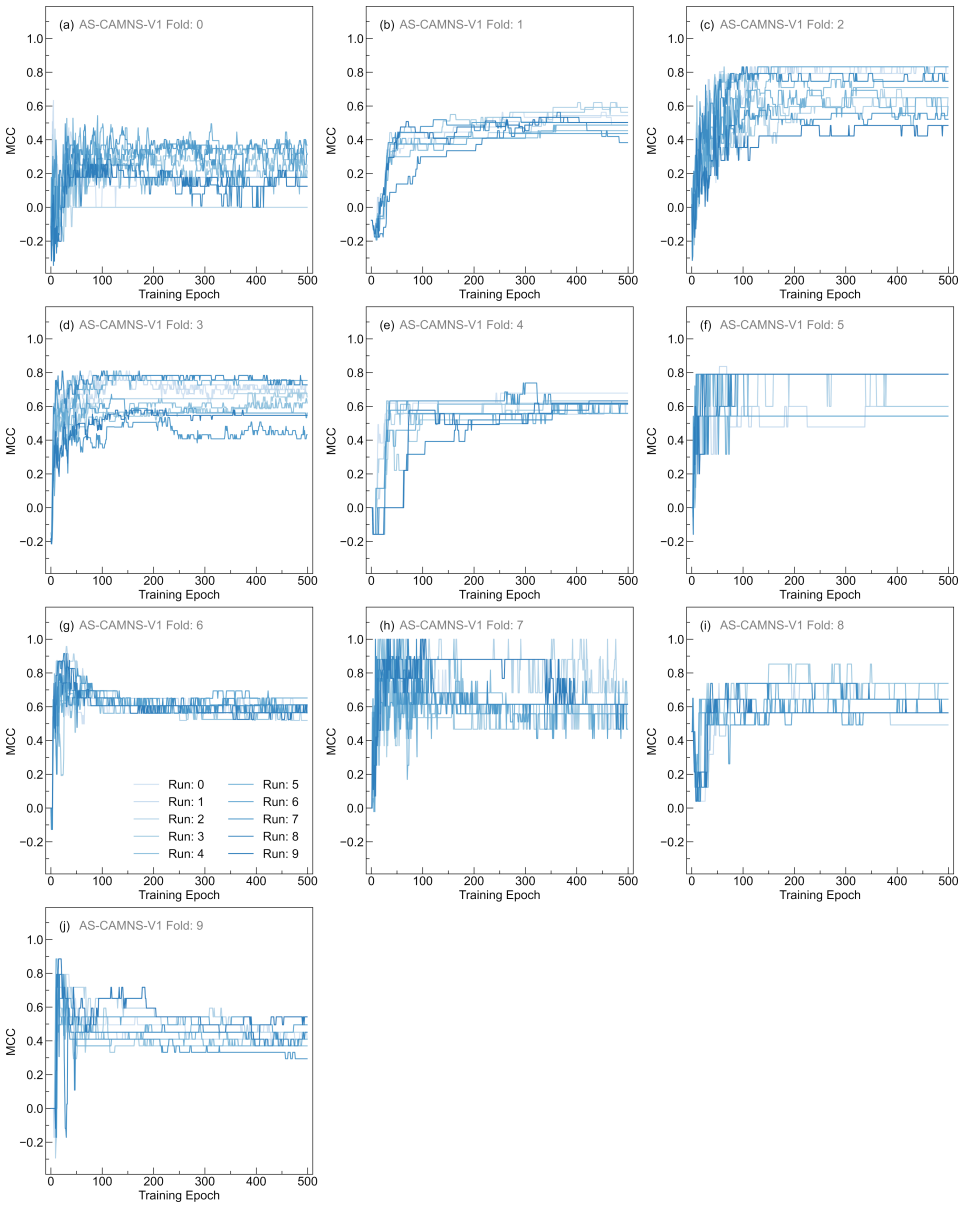

Figure C.12: Matthew's correlation coefficient (MCC) of the AtomSets CAMNS-V_1 (AS-CAMNS-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

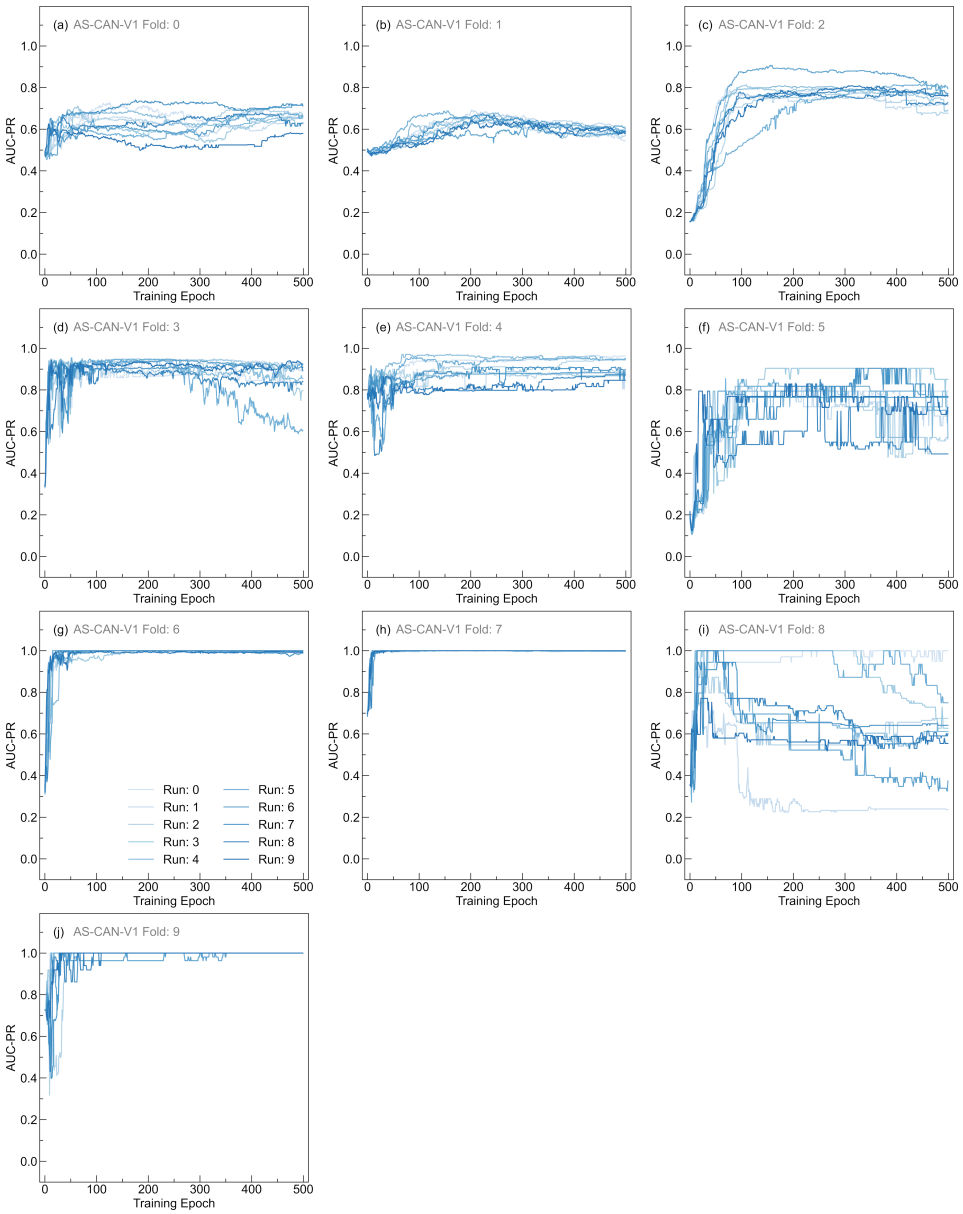

Figure C.13: Area under the precision-recall curve (AUC-PR) of the AtomSets CAN-V_1 (AS-CAN-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

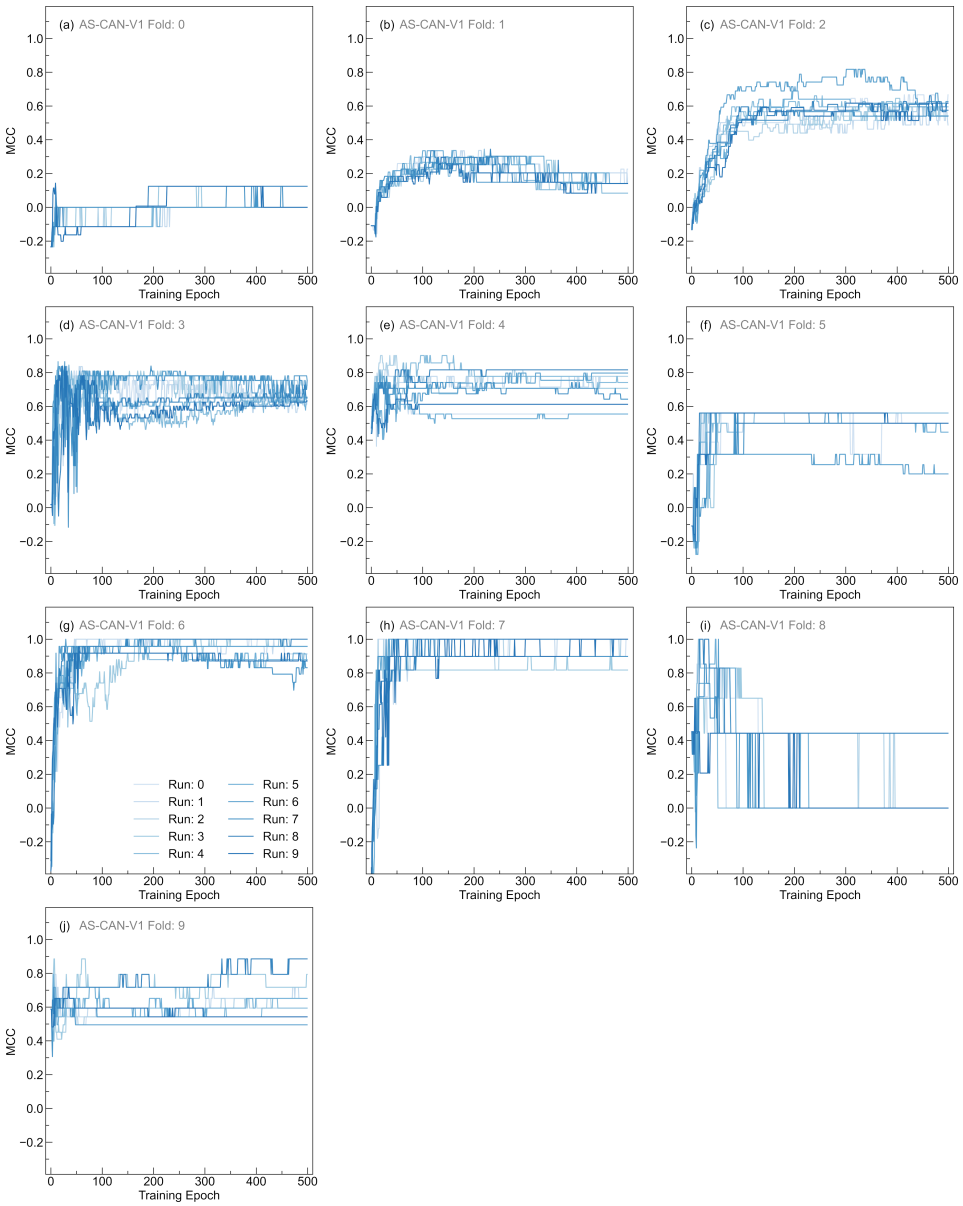

Figure C.14: Matthew's correlation coefficient (MCC) of the AtomSets CAN-V_1 (AS-CAN-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

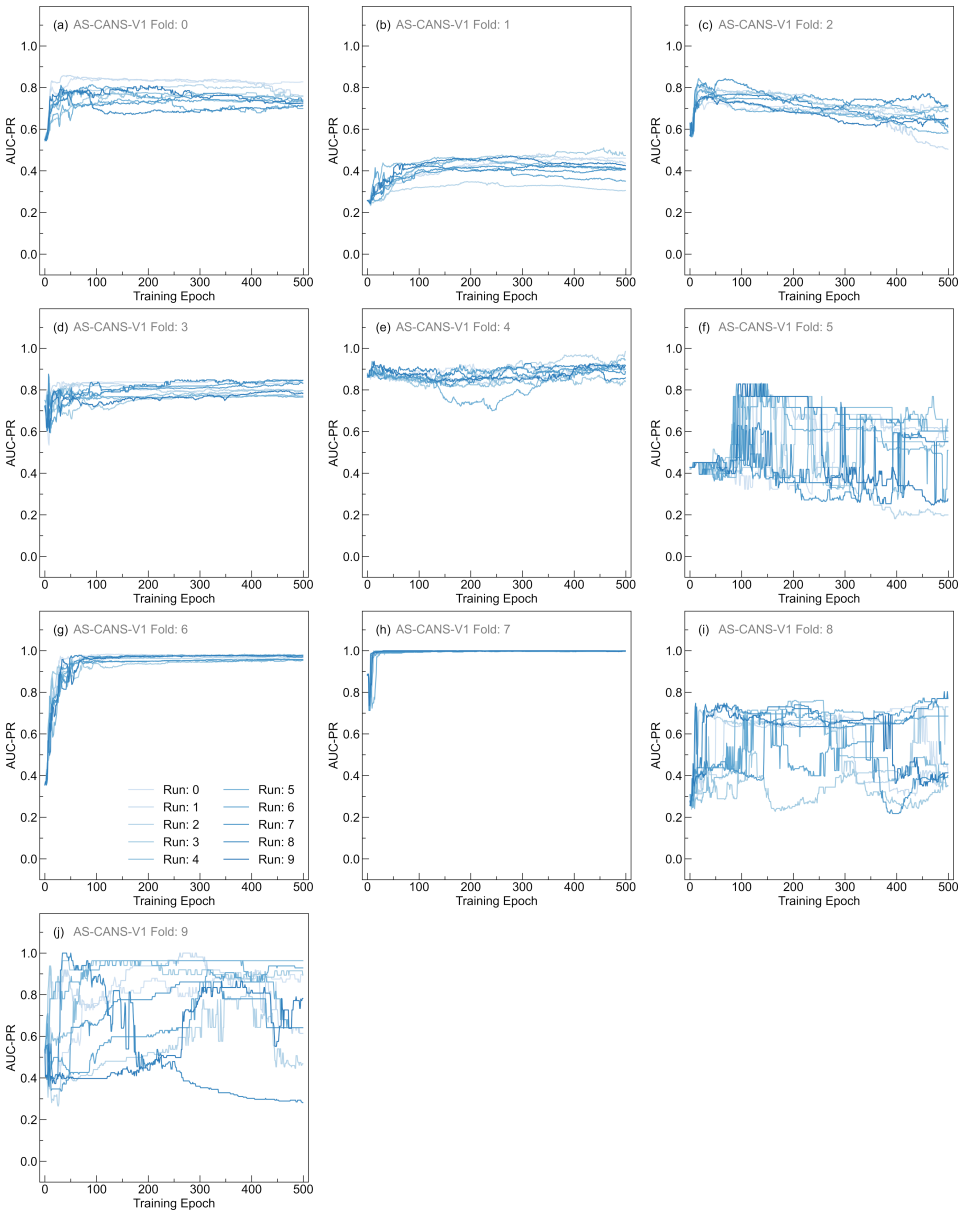

Figure C.15: Area under the precision-recall curve (AUC-PR) of the AtomSets CANS-V_1 (AS-CANS-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

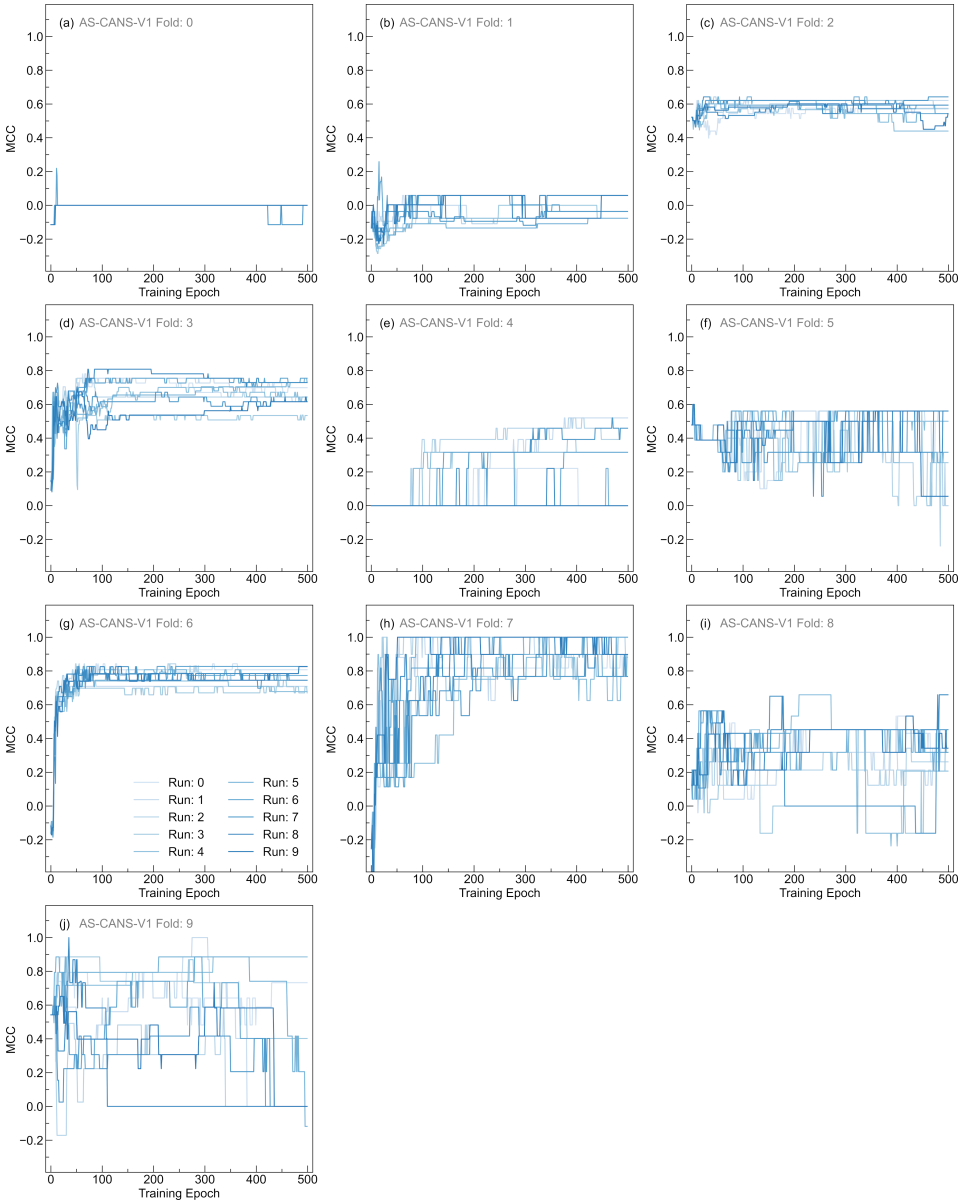

Figure C.16: Matthew's correlation coefficient (MCC) of the AtomSets CANS-V_1 (AS-CANS-V_1) model-feature combination. Panels (a)-(j) show the training curves of the model using clusters 0-9 as the validation set, respectively. Ten repeated runs are displayed to demonstrate model variation. The best performing hyperparameter configuration is shown for each choice of validation cluster.

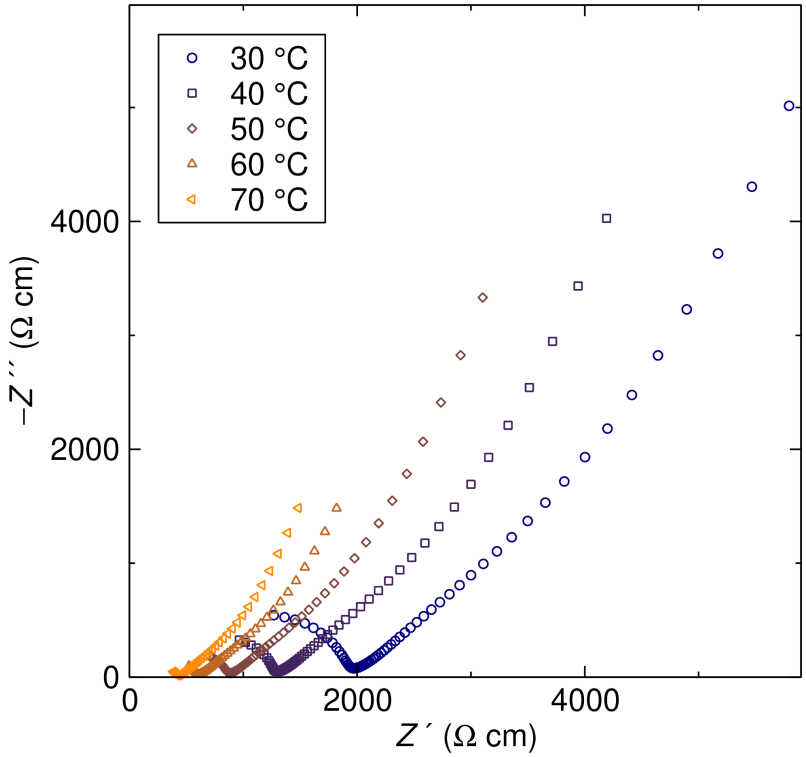

Figure C.17: Nyquist plots from temperature-dependent electrochemical impedance spectroscopy of $Li_9B_{19}S_{33}$. The impedance is multiplied by the ratio of the contact area (0.28 cm^2) and the pellet thickness (0.95 cm).