Batteries & Supercaps

Research Article
doi.org/10.1002/batt.202500309

Chemistry Europe
European Chemical Societies Publishing

www.batteries-supercaps.org

# Mining Chemical Space with Generative Models for Battery Materials

Chiku Parida, Diptendu Roy, Juan Maria García Lastra, and Arghya Bhowmik*

Generative models represent a powerful new paradigm for accelerating the discovery of novel materials across vast chemical space. To evaluate the viability of deploying generalized crystal generative models for application specific discovery tasks, here Li-ion battery (LIB) materials are chosen as a case study. The pretrained MatterGen model is used to generate diverse crystalline structures, conditioned for stability and tested for uniqueness and novelty for promising Li-containing compositions. An unsupervised clustering analysis is performed using atomic neighborhood fingerprints to compare the distribution of generated structures against the training dataset and materials project (MP) data in the chemical space. The multitired workflow for LIB materials combines a universal crystal generative model with foundational machine learning potential to identify the most promising stable (close to convex hull with respect to MP data) candidates for final density functional theory-based stability calculations. Open circuit voltage (OCV) and specific capacity calculations on selected stable materials highlight their potential as LIB materials. Among 91 identified Li-containing stable ($\leq$0.03 eV/atom above MP convex hull) materials, two novel cathode materials are identified useful for LIB, considering average OCV, OCV at highest and lowest state of charge, and the specific capacity.

## 1. Introduction

Large foundational deep learning models are fundamentally reshaping the landscape of scientific discovery. Within the realm of inorganic material discovery, the emergence of sophisticated computational models and atomistic simulation techniques holds immense promise to significantly accelerate the identification of novel high-performance materials.[1] These advancements offer the potential to identify promising candidates for various applications, but are limited by their computational cost. Machine learning force fields (MLFF) over the last decade have emerged as a viable alternative with much lower computational cost but similar accuracy.[2] Going away from materials specific MLFFs, that require data collection and model training for every new projects, the recently developed foundational simulation models, like MACE0,[3] MatterSim,[4] CHGNet,[5] and MEGNet,[6] have enabled atomic simulations across a broad range of materials without the need for retraining. However, it has been recognized that for downstream applications targeting specific material classes or properties, the general models tend to have limited reliability and accuracy.[7]

Similarly to the success story of MLFFs that enable faster and larger simulations, generative models are routinely used for fast and efficient inverse design of molecules.[8] These generative AI models using various architectures, such as variational autoencoders (VAEs), generative adversarial networks (GANs), transformer-based models, and diffusion-based models, can generate new molecules by learning from vast curated databases.[9] Atomic graphs for materials have added complexities of periodicity and crystal symmetries. Recently, generative models have been developed specifically for crystal systems such as the crystal diffusion variational autoencoder (CDVAE),[10] DiffCSP,[11] MatterGen,[12] and MatterGPT,[13] each leveraging distinct underlying learning architectures and different data types for representations. Such crystal generative models have been trained with curated chemical space-specific crystal structure data and used successfully in the sampling of novel and stable materials.[14–16] Very recently, apart from the subspace-specific models, foundational crystal structure generative models have been trained for broad chemical spaces with data sets like MP,[17] Alexandria,[18] and so on. Their applicability toward a specific materials discovery campaign, such as the high-voltage Li-ion battery (LIB) cathode, is an open question and a necessary one before wide-scale deployment.

In this regard, we have employed MatterGen, a recently developed geometric crystal structure diffusion-based universal generative model, to evaluate its capability for discovering LIB cathode materials. We use a hybrid workflow, integrating MatterGen[12] with universal MLFF MatterSim[4] and density functional theory (DFT) simulations to test the possibilities of targeted materials discovery without new data collection for model retraining. After Li-containing materials are sampled with MatterGen, MLFF-based screening is done for initial identification of stable candidates, and finally, it is validated using accurate DFT simulations. In addition, we have screened the exceptionally stable candidates which

C. Parida, D. Roy, J. M. G. Lastra, A. Bhowmik
Department of Energy Storage and Conversion
Technical University of Denmark—DTU Kgs.
2800 Lyngby, Denmark
E-mail: arbh@dtu.dk

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202500309

are close to or below the existing MP convex hull. Among them, 10 interesting candidates are selected for the calculations of open-circuit voltage (OCV) and specific capacity. Finally, we performed gamma-point phonon calculations to check the dynamic stability of all the potential LIB cathodes and found that, except for three materials, all are dynamically stable. Considering OCV, specific capacity, thermodynamic, and dynamic stability, two potential LIB cathodes are found. We also compared the chemical space traversed with this workflow with that of the materials project (MP) dataset and the generative model training data. This comparative analysis provided insight into MatterGen's strengths and weaknesses in narrow chemical space exploration, targeted sampling, and the stability of the generated structures.

## 2. Results and Discussion

MatterGen is the first openly available foundational generative model[12] for crystal structures that is trained on two large DFT-validated solid-state materials datasets.[17,18] MatterGen is based on a diffusion-based generative model that operates on atomic nodes as 3D point objects and can be conditioned on target composition constraints as well as stability with an aim to discover stable, unique, and novel (S.U.N) materials beyond the training datasets. Similar to other generative resampling methods, MatterGen learns the underlying distribution of atomic arrangements in crystalline solids. Through an iterative denoising diffusion process, it generates new candidate structures. Unlike conventional structure prediction methods that often rely on random sampling or expensive global optimizations, MatterGen efficiently explores vast chemical spaces based on learning the data distribution. MatterGen's ability to access the full chemical space of unknown materials gives it a distinct advantage as part of a materials discovery pipeline over methods that are limited to candidate library-based screening. It has demonstrated the capacity to generate a good number of stable and novel candidate materials.[12] Thus, in this study, we have used the capacity of the MatterGen model to explore S.U.N battery materials.

### 2.1. Workflow

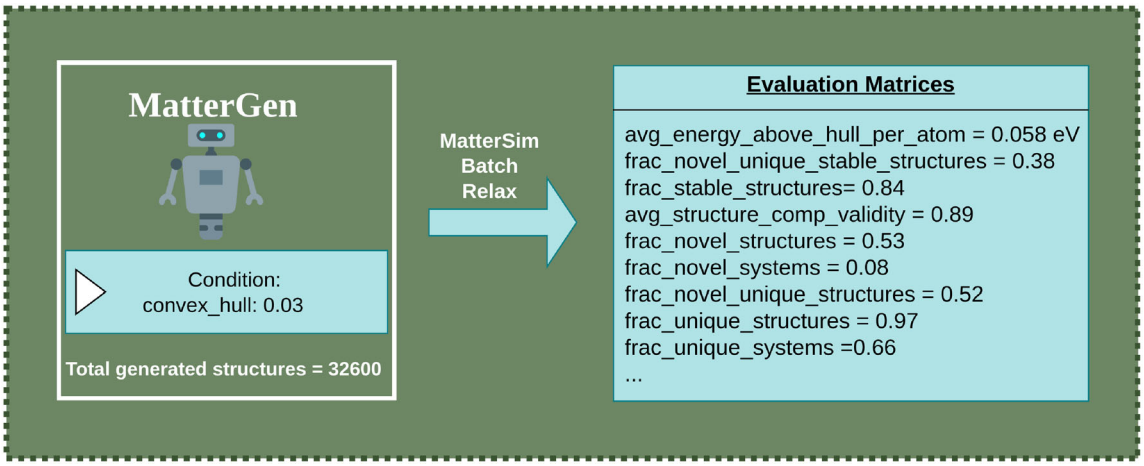
Our LIB material discovery workflow uses MatterGen for structure sampling,[12] MatterSim for initial screening of stable materials,[4] and DFT for final validation. As both MatterGen and MatterSim are trained on datasets that span the whole crystalline material space, our workflow does not include any data generation or model (re)training, making it a true test of an easy-to-assemble and low computational cost accelerated discovery process. Success in the form of discovery of novel and stable (below or near the convex hull) materials with desirable properties through our framework can establish the usefulness of off-the-self foundational generative models in targeted material discovery. Similarly to material library preparation in traditional screening pipelines, the workflow uses MatterGen to generate new candidate structures. In this regard, confirmation of the thermodynamic stability is crucial for the synthesizability of a material. It is typically assessed by comparing the formation energy of a compound with the most stable phases available in the same chemical system. In practice, this is done using a convex hull construction, where a compound is considered thermodynamically stable if its formation energy lies on (or very near) the convex hull. In this study, candidates are generated with a conditional criterion of validity for structures below 0.03 eV/atom above the MP convex hull. This threshold ensures that the generated materials are likely to be thermodynamically stable. The total energies of the generated structures are then computed using MatterSim,[4] a generalized MLFF implemented within the evaluation script of the MatterGen repository. By evaluating with respect to the MP energy convex hull, we exclude candidates that exceed the 0.1 eV/atom threshold, thus filtering out structures with poor stability. Candidates are also screened for uniqueness and novelty following the (S.U.N) approach. For stability, the distance from the convex hull (energies below the threshold $\leq$ 0.1 eV/atom) is considered. To maintain uniqueness, redundant structures are identified and removed. Structures that are not present in existing trained database (MP) are considered novel, and those structures that maintain all these criteria are taken for further studies. The evaluation strategies using the MatterGen and MatterSim pipeline have been shown in **Scheme 1**.

Next, we narrow down the candidates on the basis of elemental composition. Since the focus is on LIB materials, only structures containing lithium are retained. Structures that include heavy elements are eliminated from the data set because they are generally not useful for battery applications due to their low gravimetric capacity. In this case, we have excluded the elements $Z > 70$ and Tc due to its radioactivity. Although we have excluded elements $Z > 70$, Pt, Pb, and Bi are included due to their presence in existing LIBs. With these Li-containing materials, we applied dimensionality reduction and clustering techniques to gain insights into the chemical variations using chemiscope software.[19] In the final stage of the workflow, selected candidates (referred to as S.U.N structures containing Li in our pipeline) are subjected to full DFT relaxation. The energy of these DFT-relaxed structures is compared with the MP energy convex hull. For a subset of the most promising materials, the OCV is calculated, as this property is critical for evaluating the performance as a LIB cathode. The entire workflow has been illustrated in **Figure 1**. In the following sections, a detailed discussion and analysis of each step of the screening workflow have been demonstrated.

### 2.2. Screening Outcome

Here, we first conditionally sample structures with a target convex hull energy of 0.03 eV/atom with respect to MP data within the MatterGen framework, and 32,600 structures are generated in this process. MatterSim-based batch relaxation has been employed on these 32,600 structures, among which 27,227 and 5373 structures are found below and above 0.1 eV/atom energy convex hull of MP data, respectively (**Figure 2**). 27,227 structures are considered stable and further tested for uniqueness and novelty. A total of 12,550 novel and unique structures are identified among them, i.e., stable,

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing




**Scheme 1.** Evaluation metrics of generated materials using MatterGen and MatterSim pipeline.

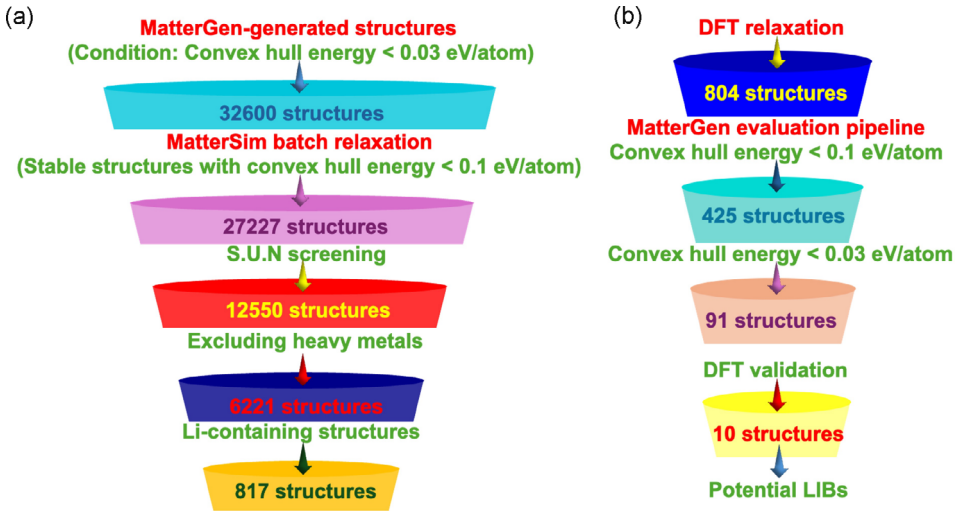



**Figure 1.** a) Screening workflow involving MatterGen generation, MatterSim relaxation, S.U.N filtering to find out Li-containing S.U.N structures generated, and b) final screening involving DFT relaxation and stability validation and DFT evaluation of functional property (OCV).

unique, and novel (S.U.N). Since our goal is to find out LIB materials, we only select the Li-containing materials and, among those, exclude structures with heavy metals in the compositions. Finally, we arrive at 817 structures (Figure 2). Notably, from Scheme 1, it is clear that MatterGen-generated $\approx 32,000$ structure sample was largely novel and unique. While greater sampling could produce more promising candidates, we prioritize a swift evaluation of MatterGen's ability to generate S.U.N materials for targeted discovery, which can be robustly tested within this sample size. Increasing the sample size would significantly raise the computational cost of MatterSim batch relaxation and DFT validation. Future longer-term discovery campaigns should generate significantly larger S.U.N structures libraries in the chemical space of interest for high chances of encountering high-performance materials.

Next, we perform DFT relaxation for all these 817 structures. However, 13 structures among them have shown convergence error. So, we proceed with 804 structures. To validate stability, convex hull energy is estimated (based on DFT energy) with respect to MP data. 425 candidates are below 0.1 eV/atom with respect to the MP energy convex hull, which is more than 50% of the total DFT-relaxed structures (**Figure 3**). The selected structures with below 0.1 eV/atom above MP convex hull from MatterSim energies and DFT energies differ possibly due to the small differences in energies predicted by the surrogate and actual DFT simulations (Figure S1, Supporting Information). On the set of 425 candidates, DFT energies being more reliable, we implement a more stringent criterion of below 0.03 eV/atom above MP energy hull, resulting in 91 candidates. From these, we select a subset of promising LIB materials and perform the calculations to determine average OCV ($US_{eq}$), OCV at highest ($U_{HSOC}$), lowest state of charge ($U_{LSOC}$), and specific capacity to evaluate the possibility of discovering LIB materials of good functional properties.

### 2.3. Sampling Diversity of MatterGen

In this study, we have utilized the potential of the MatterGen model to generate valid stable chemically diverse structures that are of interest to the LIB community. MatterGen uses a generative

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing

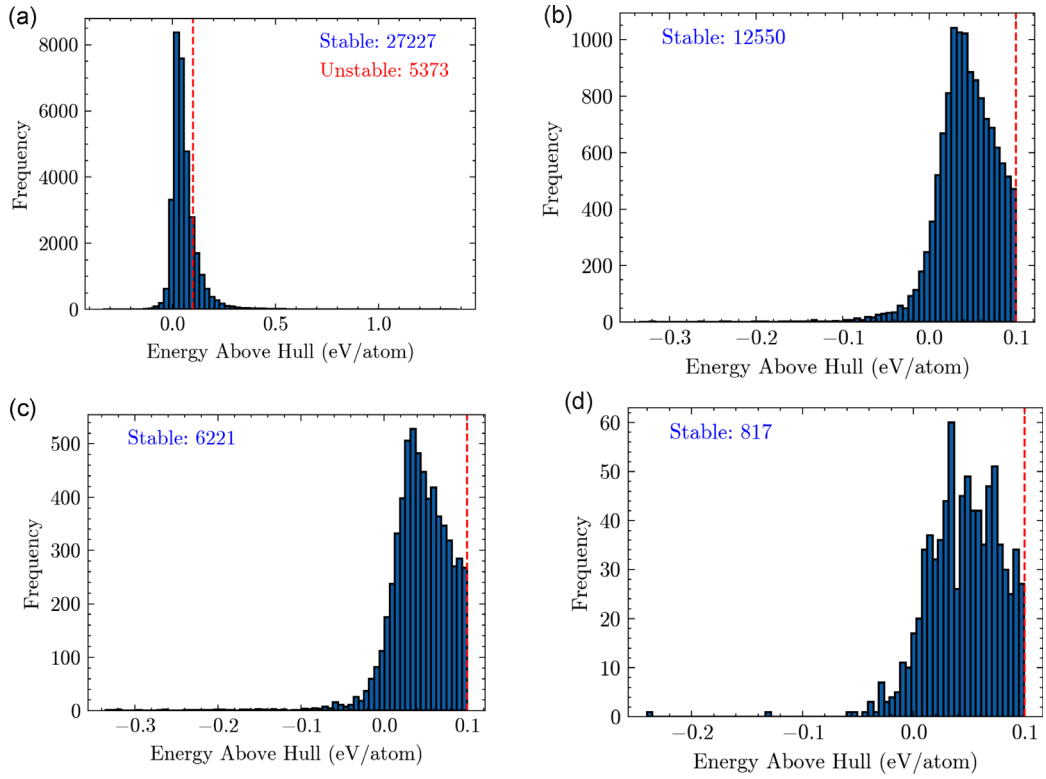



**Figure 2.** a) MatterGen-generated raw structures, b) screened structures with S.U.N filtering using evaluation metrics, c) screened structures with excluding heavy elements, and d) screened Li-containing structures.

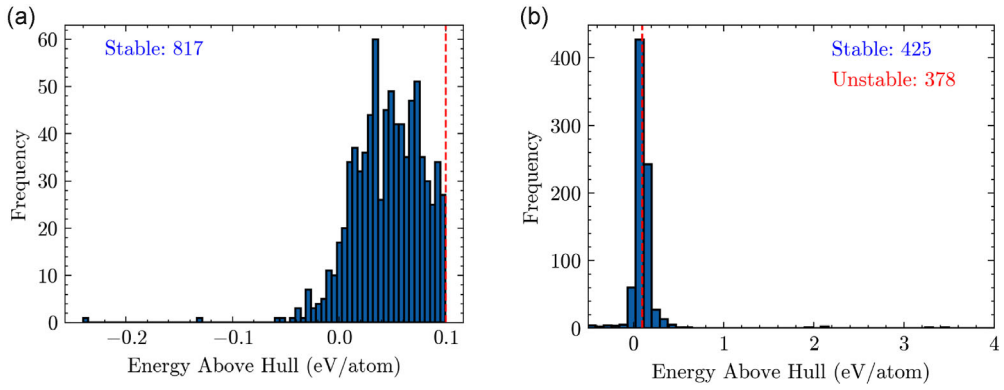



**Figure 3.** Distribution of energy of a) MatterSim-relaxed structures and b) DFT-relaxed structures with respect to MP data based convex hull.

diffusion model defined on the 3D atomic structure. The model has two components, a forward corruption process and a reverse denoising process.[12] In the forward corruption stage, the model takes existing material structures from a large training dataset, which is constructed based on a subset from Alexandria and MP dataset containing over 600,000 structures and progressively adds noise. This noise is introduced into three key aspects of the crystal structure such as atom types, atomic coordinates, and the dimensions and shape of the periodic lattice. The reverse denoising process is where the model learns to reverse the corruption process, i.e., reconstruct valid crystal structures from noise. This is accomplished by training an equivariant score network on the extensive materials dataset (MP + Alexandria). The score network learns to predict the "score" or the gradient of the data distribution, which essentially indicates the direction in which to move from a noisy, random structure toward one that is more realistic and stable. MatterGen implements a classifier-free diffusion guidance strategy, where one key parameter that directly affects the diversity of generated samples is the diffusion guidance factor. In this study, we have considered the default diffusion guidance factor value of 2.0. After the generation of the structures, to understand their diversity compared to the training chemical space, we performed several data analysis. Here, to compare the chemical diversity of the training data and model-generated data, we have

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing

plotted the kernel density plot and compared the Li-containing MP data and Li-containing MatterGen-generated data (**Figure 4**). It can be observed that the density distribution of training MP data and generated data containing Li metal is almost similar. Moreover, the peak close to 0 eV/atom indicates the high frequency of stable materials in Li-containing MP data. Notably, here we have not included the Alexandria dataset for comparison, as it contains many hypothetical structures.

We analyzed the distribution of individual elements in both the MatterGen-generated dataset (**Figure 5**a) and the MatterGen training dataset (Figure 5b). For both datasets, the elemental distribution is presented as the proportion of each element relative to all others. Notably, the MatterGen training data, a subset of the MP + Alexandria dataset, exhibits the highest frequency for Li, a trend that is also reflected in the MatterGen-generated dataset. This suggests that our training dataset contains a higher proportion of Li-containing materials compared to the popular MP dataset, where O is the most frequent element (Figure S2, Supporting Information). This observation is encouraging for our specific demonstration of LIB material discovery.

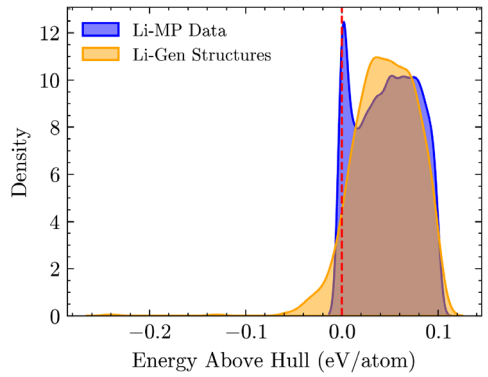



**Figure 4.** Comparison of the density distribution of MatterGen model generated data and MP data.

The overall distribution of elements is broadly similar between the two datasets, although the range of elements is somewhat narrower in the generated data compared to the training data. This can be explained by the general oversampling of diffusion and other generative models in areas of high-training data density.[20] To better understand the elemental distribution within our working data, we also examined the distribution in both datasets after excluding heavy elements (Figure S3, Supporting Information). This analysis revealed a similar trend, with a narrowed range of elements in the generated data.

These observations indicate that while MatterGen effectively captures the general distribution of many elements, there are subtle differences in the representation of certain elements compared to the original training data. This provides the insight that for target tasks involving elements that are sparsely represented in the typical training datasets, either will require synthetic data rebalancing or generative architectures that work well with imbalanced datasets.

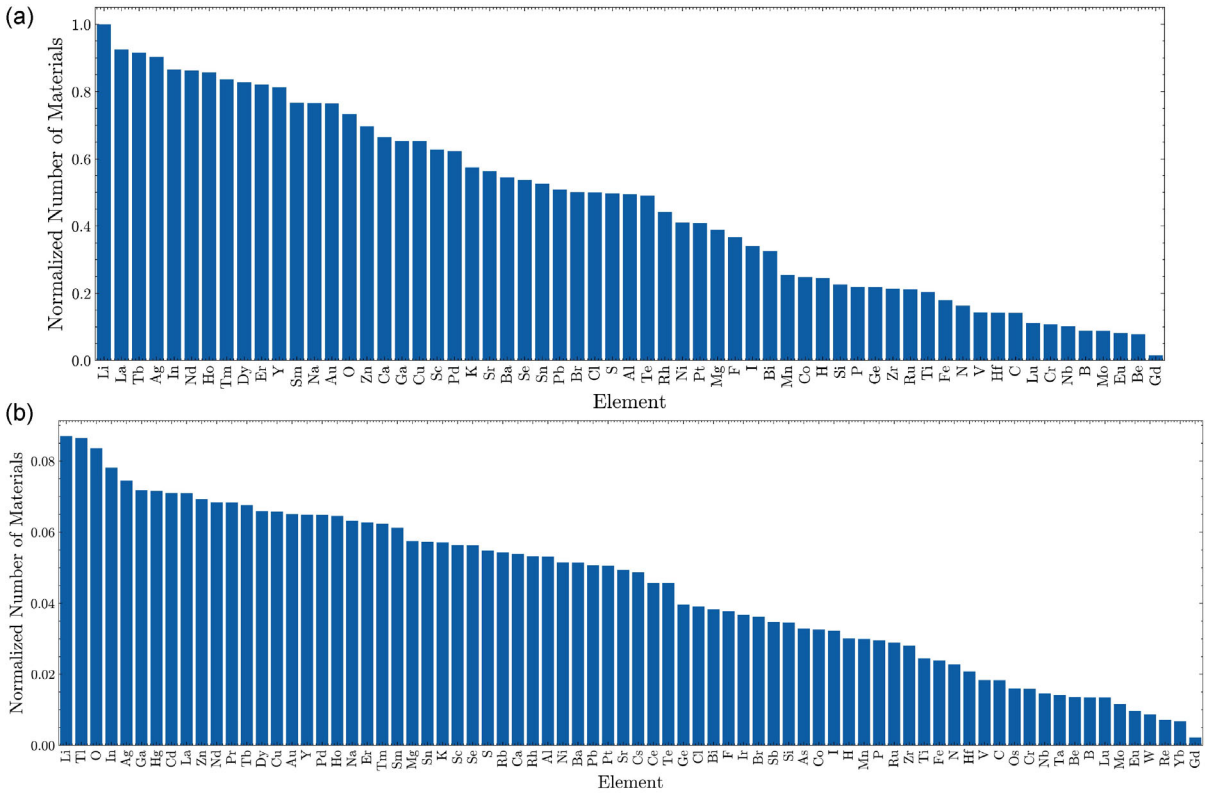



**Figure 5.** Normalized elemental distribution of elements in a) MatterGen generated data and b) MatterGen training data.

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing

To visualize the chemical space of the generated dataset, we have extracted structural descriptors using Smooth Overlap of Atomic Positions (SOAP), followed by the dimensionality reduction of the data using t-distributed Stochastic Neighbor Embedding (t-SNE) clustering technique implemented in chemiscope software.[19] Here, for the extraction of SOAP descriptors, we choose a cutoff radius of 4.5 Å, eight radial basis functions as spherical Gaussian-type orbitals, and six as the maximum degree of spherical harmonics. Then these SOAP-generated high-dimensional descriptors are reduced to a 2D space using t-SNE for better visualization. At first, this analysis was performed on the combined dataset of MatterGen-generated data and training MP data, which interestingly revealed a diverse set of crystal structures grouped into various distinct clusters based on their chemical functional groups (**Figure 6**a). The clustering analysis successfully identified and separated various classes of compounds, including selenides, tellurides, borates, phosphates, oxides, fluorides, oxyfluorides, hydroxides, chlorides, nitrides, and metal alloys, each occupying unique regions within the clustered data. This clustering behavior suggests that the SOAP descriptor effectively captures the underlying chemical characteristics of the materials. In addition, the Li-containing generated structures are nicely distributed in each of the clustered regions of the MP data (Figure 6b).

We take the same approach for the entire space of Li-containing training data (MP + Alexandria) and Li-containing MatterGen-generated data and found a similar pattern (**Figure 7**a), confirming that the generated data roughly follow the same distribution as the training data. In this case also, we have observed that the structures for training data containing Li are nicely clustered according to their functional groups, such as borates, phosphates, selenides, tellurides, oxyfluorides, and so on. This indicates that the pretrained MatterGen model has effectively learned to represent diverse chemical environments from the trained crystal systems. We have also compared the Li-containing MP data with the final set of stable structures considered for DFT relaxation and found that this set of data is also distributed in most of the clustered regions of Li-containing MP data (Figure 7b). This indicates that the stable structures obtained after screening are also significantly diverse in their chemical space but have class imbalance. Notably, our analysis reveals a discernible bias toward the generation of structures of metallic alloys. Although these materials are not suitable for battery applications, this bias may be attributed to the training data imbalances and higher thermodynamic stability, potentially making them more likely to be sampled during conditional resampling.

To validate the structures obtained after screening, we have compared the energy of the DFT-relaxed structures with respect to the convex hull energy of MP data, and more than 50% DFT-relaxed structures have been found as stable structures. Among them, 91 structures have been chosen considering more rigorous criteria of below 0.03 eV/atom above the MP energy convex hull. These 91 structures include metal oxides, selenides, tellurides, sulfides, fluorides, oxyfluorides, etc. We have also checked materials that are below the convex hull energy of the MP data and demonstrated all these materials in Table S1, Supporting Information. As we imposed a test for Li intercalation and deintercalation possibility depending on the variable oxidation state of the metals, 10 interesting LIB materials are chosen among them, and we have performed DFT calculations to determine the highest possible delithiated states, the OCV at the highest ($U_{HSOC}$) and lowest state of charge ($U_{LSOC}$), average OCV ($U_{eq}$), and specific capacity for all these systems. Moreover, we have also performed gamma-point phonon calculations for all of these 10 potential LIB cathode materials, and their corresponding all possible delithiated states sequentially. Out of 31 structures (including fully lithiated and corresponding to all possible delithiated states for 10 potential LIBs), we found phonon modes with high imaginary frequencies in 16 structures. For those 16 structures, we distorted the structures by following the eigenvector modes with high imaginary

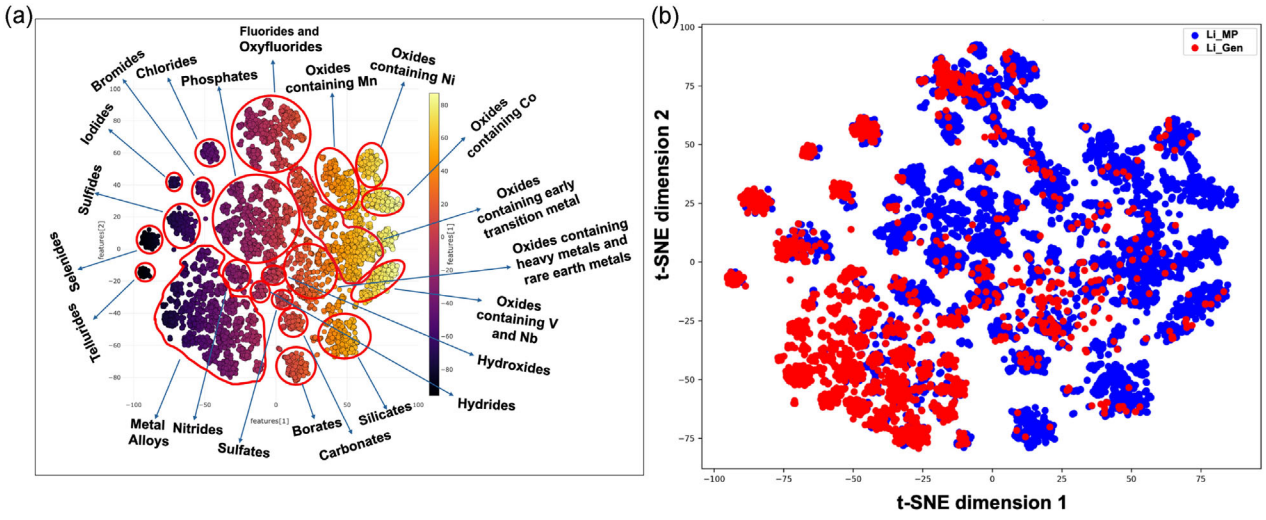



**Figure 6.** a) Comparison between MP training data containing Li and the Li containing structures generated using MatterGen model in the various clustered regions where Li-MP data and generated data are indicated by blue and red color, respectively. b) The clustered regions of combined Li-MP and Li-containing data, where the clustered regions are identified on the basis of their special chemical compositional variants.

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing

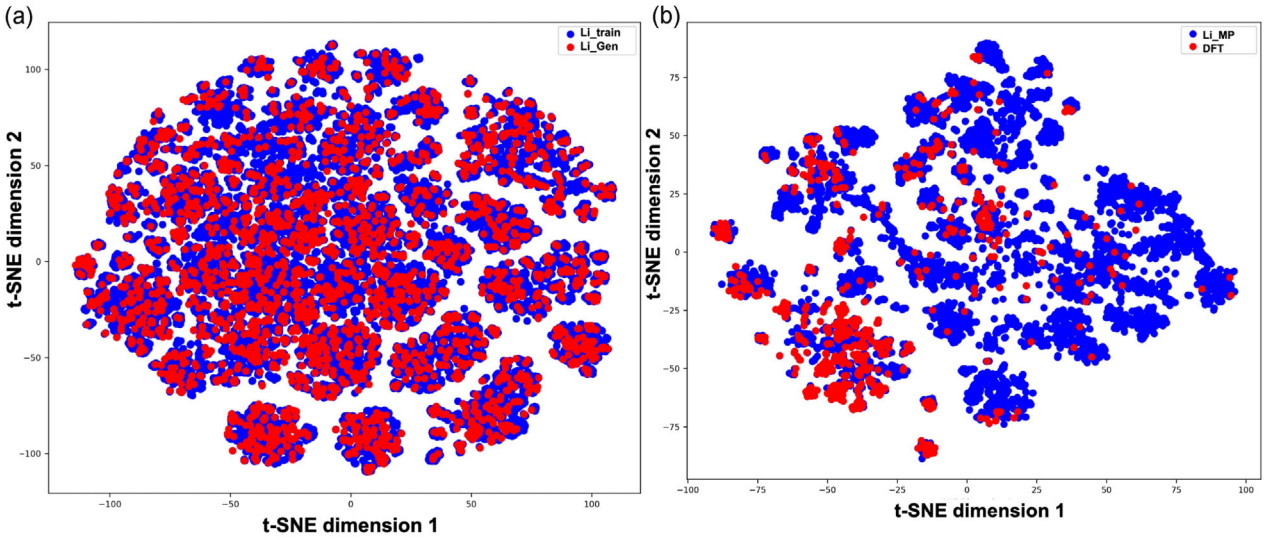



**Figure 7.** a) Comparison between Li containing training data (MP + Alexandria) of pretrained MatterGen model and Li-containing MatterGen-generated structures in the various clustered regions where Li MP + Alexandria training data and Li MatterGen-generated data are indicated by blue and red color, respectively. b) Comparison between Li-MP data and DFT-relaxed structures in the various clustered regions, where the Li-MP data and DFT data are indicated by blue and red color, respectively.

frequency. Out of those 16 structures, 10 structures got relaxed toward lower energy with no imaginary frequency modes. However, there are six structures from four different materials ($Li_2V_2TiCrO_9$/$Li_3Ni_5O_7$/$Li_2Ni_2F_8$/$Li_4MnFeO_5F$, $Li_3MnFeO_5F$, $Li_2MnFeO_5F$), that do not relax toward lower energy structures with no imaginary frequency.

We observe that materials similar to these 10 materials have been either explored as potential LIBs or synthesized for other purposes, where different materials have different advantages. Vanadium oxyfluorides are already synthesized and reported as Li-intercalated cathode materials with high capacity and high rate capability.[21] With our discovery pipeline, we have found vanadium oxyfluoride and vanadium fluoride-based cathode materials with significant OCV and relatively lower capacity (**Table 1**). These specific materials are novel and have not been synthesized. The DFT-relaxed structures of all these 10 materials are shown in **Figure 8**. Ni fluoride-based cathode materials have also been found with high OCV, whereas, in previous literature, transition metal fluorides have been reported to provide high voltage and capacity.[22,23] Among two Ni fluoride-based cathode materials found, $Li_2Ni_2F_8$ shows significantly high OCV with reasonable capacity. Although its fully delithiated state can be stable as $Ni^{+4}$ can exist in the presence of fluoride ions due to their strong crystal field effect, no electrolytes can withstand such a high potential. On the other hand, significantly high voltage has been found for $Li_2Mn(CO_3)_3$ with significant specific capacity. While Mn oxides are established cathode materials,[24] Mn carbonate has been reported as an anode material with Mn in the +2 state.[25] In contrast, in the present study, fully delithiated $Li_2Mn(CO_3)_3$ features Mn in the +6 state, which could be the reason for showing its potential as a cathode. We have also found $Li_2TiV_2CrO_8$ and $LiNi_3F_8$ with significant voltage, which are unlikely

**Table 1.** The table reporting final LIB candidates describing their fully lithiated state and highest possible delithiated state, their OCV at lowest state of charge ($U_{LSOC}$), and at hightest state of charge ($U_{HSOC}$), average OCV for all the possible subsequent delithiated states ($U_{eq}$) and their specific capacities.

| Fully discharged compound | Maximally charged compound | $U_{LSOC}$ [V] | $U_{HSOC}$ [V] | $U_{eq}$ [V] | Specific capacity [mAh $g^{-1}$] |
|---|---|---|---|---|---|
| $Li_5MnFeO_5F$ | $Li_2MnFeO_5F$ | 0.76 | 3.54 | 2.47 | 328.46 |
| $Li_3V_4O_9$ | $V_4O_9$ | 2.03 | 2.90 | 2.53 | 218.03 |
| $Li_2Mn(CO_3)_3$ | $Mn(CO_3)_3$ | 3.09 | 4.87 | 3.98 | 215.32 |
| $Li_2Ni_2F_8$ | $Ni_2F_8$ | 5.65 | 6.01 | 5.83 | 189.14 |
| $Li_3Ni_5O_7$ | $Ni_5O_7$ | 2.39 | 3.01 | 2.70 | 188.52 |
| $Li_2TiV_2CrO_8$ | $TiV_2CrO_8$ | 2.42 | 4.06 | 3.24 | 155.94 |
| $Li_3V_3F_4O_6$ | $LiV_3F_4O_6$ | 3.04 | 3.04 | 3.04 | 155.00 |
| $Li_2Ti_3S_6$ | $Ti_3S_6$ | 1.96 | 2.16 | 2.06 | 153.16 |
| $K_2Li_2V_2F_{10}$ | $K_2V_2F_{10}$ | 2.86 | 3.07 | 2.97 | 139.56 |
| $LiNi_3F_8$ | $Ni_3F_8$ | 3.97 | 3.97 | 3.97 | 132.61 |

Research Article
doi.org/10.1002/batt.202500309
Chemistry Europe
European Chemical Societies Publishing

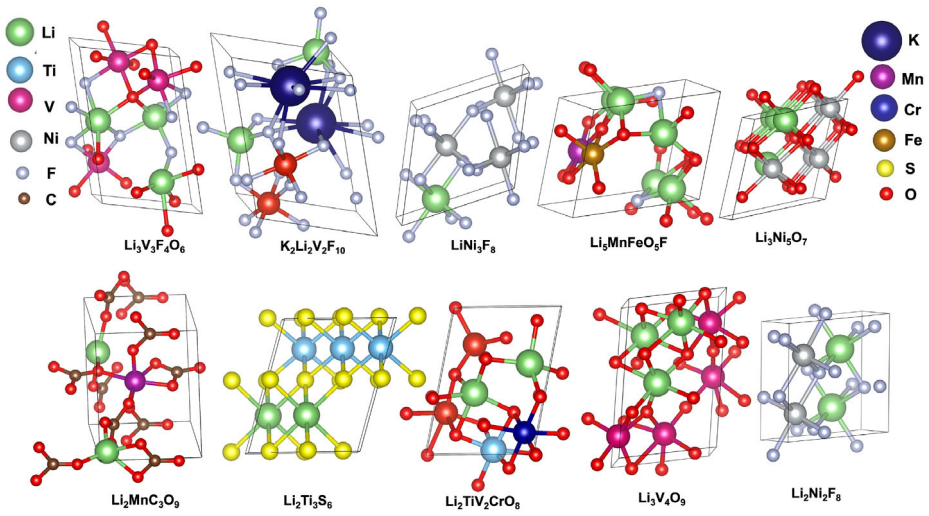



**Figure 8.** DFT-relaxed structures of the potential LIB materials found.

to be good cathode materials due to low gravimetric capacities. Despite moderate voltages, $Li_3V_4O_9$ and $Li_5MnFeO_5F$ exhibit high specific capacities; notably, $Li_5MnFeO_5F$ offers the highest among all others, with three delithiated states enhancing its appeal. Although vanadium oxides are well-known LIB materials,[26–28] the specific compound identified here is novel. Similarly, though Fe–Mn oxyfluorides are known electrocatalysts for oxygen evolution reaction,[29] they have not been reported for LIBs and structure, and composition is different from the compound we found. However, individual Fe and Mn-based oxyfluorides show significant OCV and capacity,[30,31] which strengthens the future potential of this material as LIB cathode. Among newly identified materials, $Li_2Mn(CO_3)_3$, $Li_3V_4O_9$, and $Li_5MnFeO_5F$ stand out for LIB applications due to their favorable $U_{LSOC}$, $U_{HSOC}$, $U_{eq}$, and specific capacities. Their positions below the MP convex hull (Table S1, Supporting Information) further confirm their exceptional stability. However, among these three materials, the delithiated states of $Li_5MnFeO_5F$ are not dynamically stable, and thus, the other two materials are worthwhile for the attempt to synthesize and characterize to explore their potential as battery materials.

## 3. Conclusion

MatterGen represents a significant leap forward in the application of generative AI to accelerate the discovery and inverse design of inorganic crystalline materials. Its unique capabilities in generating novel inorganic materials across the chemical space create new ways to establish a material discovery pipeline, as demonstrated here. We can now achieve materials discovery over full chemical space with very limited DFT calculations by leveraging foundational crystal generative models such as the MatterGen model along with universal interatomic potentials. In this preliminary demonstration, we have generated a small set of ≈32000 samples and among them, 12,550 structures are found to be stable, unique, and novel. Considering only Li-containing materials and excluding materials containing heavy elements, we get 817 structures that are relaxed using DFT. Next, we have compared the Li-containing materials generated with existing trained MP data and performed an unsupervised clustering analysis using atomic neighborhood fingerprints. Although, MatterGen does generate structures following the training data distribution, we observe class imbalances in the final screened stable Li-containing materials. Finally, we have identified two interesting potential LIB materials with exceptional thermodynamic (thermodynamically more stable than MP data convex hull) as well as dynamic stability, significant OCV, and high specific capacity (DFT-based). In this study, focused on testing the validity of using foundational generative models in materials discovery tasks, we generated a small test sample of ≈32,000 structures. Given the high novelty and uniqueness of those, generating a larger dataset would likely yield many more potential candidates. Scaling up the sampling would substantially increase the computational resources required for MatterSim batch relaxation and subsequent DFT validation, which is beyond the scope of the current study. Developing and utilizing screening surrogates that directly target specific battery performance metrics, such as energy density or cycle life, will be useful for maximizing the model's impact. The continued development and application of generative AI models like MatterGen promise to usher in a new era of accelerated innovation in battery technology, paving the way for a more sustainable and energy-efficient future, and this work offers a promising step forward in this direction.

## 4. Experimental Section

### MatterGen

Crystal structures are characterized by infinitely repeating unit cells. Atoms are arranged following specific symmetry rules to form a crystal lattice within the unit cell. Therefore, any material can be

represented by the atom types arranged in the unit cell, their fractional coordinates, and the lattice, which decides the shape of the unit cell.

The representation of material M is as follows:

$$M = (A, X, L) \tag{1}$$

where $A = [a^0, a^1, a^2, \ldots, a^N]$ is the atomic species of the total atoms (N) arranged in the unit cell; $L = [l^1, l^2, l^3] \in R^{3\times3}$ is the lattice; and $X = [x^1, x^2, \ldots, x^N) \in [0, 1)^{3\times N}$ is the fractional coordinates of the atoms.[10,12,32] MatterGen is based on a generative 3D crystal structure diffusion model that is trained on data encompassing a vast chemical space. Diffusion models learn the underlying data distribution with data denoising processes. MatterGen diffusion algorithm adheres to the periodicity in crystal structures of materials.[32]

MatterGen utilizes a joint diffusion process on the material representation $M$ by using a Markov chain forward process via a transition kernel. This forward process of diffusion diffuses the atom types, fractional coordinates, and lattice independently.[32]

We perform conditional generation of crystals using the pretrained model checkpoint "**chemical _system_energy _above_hull**". This checkpoint has the capacity to generate materials with multiple conditions over chemical systems and energy above the hull. To keep our experiment simple and general, we have used one condition, i.e., energy above the hull. This has been done by setting the condition keyword "**energy _above _hull"** to **"0.03"**.

For evaluation of the generated structures, we first used the original MatterGen evaluation pipeline with small necessary tweaks and performed detailed analysis on the individual metrics. It uses MatterSim (a universal MLFF)[4] to relax structures in order to perform the initial thermodynamic stability (the energy above the hull with respect to MP) evaluation.[32] To perform accurate validation of screened structures, we have performed DFT calculations for the final lithium-containing candidates and manually looked into the structures having energy below 0.03 eV/atom above the convex hull. With this condition, 91 structures are considered to be verified manually with respect to the oxidation state of transition metals, charge neutrality, possible delithiation, etc., and the best verified LIB candidates are reported.

### DFT Methods

In this study, we have conducted the structural relaxations using DFT for screening 804 newly generated unique and stable structures. All DFT calculations are performed using the projector-augmented wave (PAW) method as implemented in the Vienna ab initio simulation package (VASP).[33,34] The exchange-correlation potential was described using the generalized gradient approximation (GGA) of the Perdew–Burke–Ernzerhof (GGA-PBE) functional.[35,36] The plane-wave basis set was employed with a kinetic energy cutoff of 550 eV. The energy and force convergence thresholds were set to $10^{-6}$ eV and $10^{-2}$ eV/Å, respectively, while the k-point sampling was controlled by a KSPACING parameter of 0.15. The average OCV of the reaction has been determined from the Gibbs free energy change ($\Delta G$) per transferred Li ion. In the context of intercalation reactions, the average voltage, V, is given by the relationship between the Gibbs free energy change and the Li ion transfer process using the following equation,

$$V = \frac{\Delta G}{\mathrm{nF}} \tag{2}$$

here $\Delta G$ represents the Gibbs free energy change per formula unit, n is the number of electrons transferred, and F is Faraday's constant. Given the typically minor entropic contributions in solids compared to liquids or gases, $\Delta G$ can be effectively approximated by the change in formation energy ($\Delta E$). So, the approximated equation is as follows

$$V = \frac{\Delta E}{\mathrm{nF}} \tag{3}$$

where $\Delta E$ denotes the change in formation energy between lithiated and delithiated states of a material which can be expressed as

$$\Delta E = E_{\mathrm{LiMX_2}} - E_{\mathrm{MX_2}} - E_{\mathrm{Li}} \tag{4}$$

here $E_{\mathrm{LiMX_2}}$ and $E_{\mathrm{MX_2}}$ are the DFT calculated energies of the lithiated and delithiated states, respectively, whereas $E_{\mathrm{Li}}$ is per atom energy of bulk Li. Moreover, here X represents anions and M represents any suitable metal from the periodic table other than Li. In this study, we have also calculated the specific capacity using the following equation

$$\text{Specific Capacity } (\mathrm{mAh/g}) = \frac{\mathrm{nF}}{3.6 * M} \tag{5}$$

here n = Number of $Li^+$ ions involved in the reaction, F = Faraday constant (96485 C $mol^{-1}$), M = Molecular weight of the fully lithiated material, and 3.6 is the conversion factor for converting C/g to mAh/g. We have also performed gamma point phonon calculations using the finite displacement method implemented in VASP to check the dynamical stability of the potential LIB cathodes.

## Acknowledgements

C.P. and D.R. contributed equally to this work. The authors thank the Det Frie Forskningsråd under Project "Data-driven quest for TWh scalable Na-ion battery (TeraBatt)" (Ref. Number 2035-00232B) and "Autonomous agents of Discovery for earth-Abundant Na-ion battery cathodes (ADANA)" (Ref. Number 3164-00297B). The authors also thank the Novo Nordisk Foundation Data Science Research Infrastructure 2022 Grant: A high-performance computing infrastructure for data-driven research on sustainable energy materials, Grant no. NNF22OC0078009. C.P. acknowledges financial support from the Technical University of Denmark (DTU) through the Alliance Ph.D. scholarship.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are openly available in [github] at [https://github.com/chiku-parida/genAI4LIBs], reference number [0].

**Keywords:** crystal structure generative models · generative inverse design · li-ion battery materials · machine learning workflow · universal machine learning potential

[1] N. Marzari, A. Ferretti, C. Wolverton, *Nat. Mater.* **2021**, *20*, 736.

[2] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, *Nat. Mater.* **2021**, *20*, 750.
[3] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, F. Bigi, S. M. Blau, V. Cărare, M. Ceriotti, S. Chong, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, J. L. A. Gardner, M. J. Gawkowski, A. Genreith-Schriever et al., arXiv preprint arXiv:2401.00096v2, **2024**.
[4] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, Z. Lu, arXiv preprint arXiv:2405.04967v2, **2024**.
[5] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, G. Ceder, *Nat. Mach. Intell.* **2023**, *5*, 1031.
[6] T. W. Ko, B. Deng, M. Nassar, L. Barroso-Luque, R. Liu, J. Qi, E. Liu, G. Ceder, S. Miret, S. P. Ong, arXiv preprint arXiv:2503.03837, **2025**.
[7] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, J. S., etal Billinge, *Npj Comput. Mater.* **2022**, *8*, 59.
[8] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[9] D. M. Anstine, O. Isayev, *J. Am. Chem. Soc.* **2023**, *145*, 8736.
[10] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, T. Jaakkola, arXiv preprint arXiv:2110.06197v3, **2022**.
[11] R. Jiao, W. Huang, P. Lin, J. Han, P. Chen, Y. Lu, Y. Liu, *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 17464.
[12] C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, Z. Wang, A. Shysheya, J. Crabbé, S. Ueda, R. Sordillo, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C. W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, C. Yang, W. Li, R. Tomioka, T. Xie, *Nature* **2025**, *639*, pp. 624–632.
[13] Y. Chen, X. Wang, X. Deng, Y. Liu, X. Chen, Y. Zhang, L. Wang, H. Xiao, arXiv preprint arXiv:2408.07608v1, **2024**.
[14] D. Wines, T. Xie, K. Choudhary, *J. Phys. Chem. Lett.* **2023**, *14*, 6630.
[15] L. Guo, Y. Liu, Z. Chen, H. Yang, D. Donadio, B. Cao, *Npj Comput. Mater.* **2025**, *11*, 97.
[16] P. Lyngby, K. S. Thygesen, *Npj Comput. Mater.* **2022**, *8*, 232.
[17] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *Apl Mater.* **2013**, *1*, 011002.
[18] J. Schmidt, T. F. Cerqueira, A. H. Romero, A. Loew, F. Jäger, H.-C. Wang, S. Botti, M. A. Marques, *Mater. Today Phys.* **2024**, *48*, 101560.
[19] G. Fraux, R. Cersonsky, M. Ceriotti, *J. Open Source Softw.* **2020**, *5*.
[20] V. A. Fajardo, D. Findlay, C. Jaiswal, X. Yin, R. Houmanfar, H. Xie, J. Liang, X. She, D. B. Emerson, *Expert Syst. Appl.* **2021**, *169*, 114463.
[21] M. A. Cambaz, B. Vinayan, O. Clemens, A. R. Munnangi, V. S. K. Chakravadhanula, C. Kubel, M. Fichtner, *Inorg. Chem.* **2016**, *55*, 3789.
[22] X. Zeng, C. Zhan, J. Lu, K. Amine, *Chem* **2018**, *4*, 690.
[23] A. Liu, H. Yuan, Y. Wang, Y. Liu, J. Luo, J. Nai, X. Tao, *Inf. Funct. Mater.* **2024**, *1*, 26.
[24] B. Ammundsen, J. Paulsen, *Adv. Mater.* **2001**, *13*, 943.
[25] L. Xiao, S. Wang, Y. Wang, W. Meng, B. Deng, D. Qu, Z. Xie, J. Liu, *ACS Appl. Mater. Interfaces* **2016**, *8*, 25369.
[26] F. Mattelaer, K. Geryl, G. Rampelberg, J. Dendooven, C. Detavernier, *ACS Appl. Mater. Interfaces* **2017**, *9*, 13121.
[27] Y. Liu, E. Uchaker, N. Zhou, J. Li, Q. Zhang, G. Cao, *J. Mater. Chem.* **2012**, *22*, 24439.
[28] H. Li, P. He, Y. Wang, E. Hosono, H. Zhou, *J. Mater. Chem.* **2011**, *21*, 10999.
[29] K. Lemoine, Z. Gohari-Bajestani, R. Moury, A. Terry, A. Guiet, J.-M. Grenèche, A. Hémon-Ribaud, N. Heidary, V. Maisonneuve, N. Kornienko, J. Lhoste, *ACS Appl. Energy Mater.* **2021**, *4*, 1173.
[30] L. Zhang, D. Dambournet, A. Iadecola, D. Batuk, O. J. Borkiewicz, K. M. Wiaderek, E. Salager, M. Shao, G. Chen, J.-M. Tarascon, *Chem. Mater.* **2018**, *30*, 5362.
[31] S.-W. Kim, N. Pereira, N. A. Chernova, F. Omenya, P. Gao, M. S. Whittingham, G. G. Amatucci, D. Su, F. Wang, *ACS Nano* **2015**, *9*, 10076.
[32] C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C. W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, R. Tomioka, T. Xie, arXiv preprint arXiv:2312.03687v2, **2024**.
[33] G. Kresse, J. Furthmüller, *Phys. Rev. B 54*, **1996**, 11169.
[34] P. E. Blöchl, *Phys. Rev. B 50*, **1994**, 17953.
[35] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett. 77*, **1996**, 3865.
[36] J. P. Perdew, *Phys. Rev. Lett.* **1997**, *77*, 3868.

Manuscript received: April 24, 2025
Revised manuscript received: June 30, 2025
Version of record online: