Article

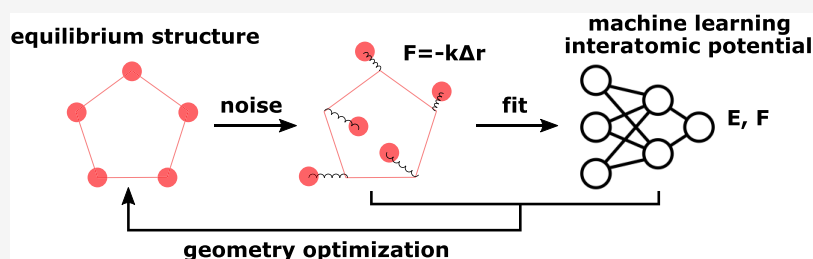# Response Matching for Generating Materials and Molecules

Bingqing Cheng*

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Diffusion models have recently emerged as powerful tools for the generation of new molecular and material structures. The key insight is that the noise in these models is related to the response of the atoms to displacement, and the denoising step is thus analogous to the geometry relaxation of atomistic systems starting from a random structure. Building on this, we present a generative method called Response Matching (RM), which leverages the fact that each stable material or molecule exists at the minimum of its potential energy surface. Any perturbation induces a response in energy and stress, driving the structure back to equilibrium. Matching this response is closely related to score matching in diffusion models. Another important aspect of state-of-the-art diffusion models is the incorporation of physical symmetries such as translation, rotation, and periodicity. RM employs a machine learning interatomic potential and random structure search as the denoising model, inherently respecting these symmetries and exploiting the locality of atomic interactions. RM handles both molecules and bulk materials under the same framework. Its efficiency and generalization are demonstrated on three systems: a small organic molecular data set, stable crystals from the Materials Project, and one-shot learning on a single diamond configuration.

INTRODUCTION

The exploration of new materials and molecules is crucial for technological advancements. Previous approaches rely on human intuition to propose and synthesize new molecules or high-throughput computational screening.[1] While these methods have led to some of the most important discoveries in history, they are often time-consuming and expensive, restricting the chemical space explored. This has encouraged the use of machine learning to generate new molecular[2] and material structures by training on data sets of equilibrium atomistic structures.[3,4] For example, the autoregressive G-SchNet[5] and equivariant normalizing flows (ENF)[6] have been employed to generate three-dimensional equilibrium structures of small organic molecules, while variational auto-encoders have been used for material generation.[7] More recently, diffusion models have been utilized for generating molecules[8−14] and materials.[15−17] Additionally, these models can be conditioned on specific chemical or biological properties through guidance mechanisms.[18−21]

A crucial aspect of recent generative models for molecules is the encoding of geometric symmetries such as translations and rotations. For materials, it is also essential to account for periodic boundary conditions (PBC) in the structures.[7,15] Moreover, efficiently accounting for different element types is vital since materials can contain over a hundred elements from the periodic table.

However, current generative models of atomistic structures do not incorporate a key inductive bias: the locality of atomic interactions. According to the locality assumption, the energy and forces acting on an atom depend solely on its neighboring atoms within a specific cutoff radius. This assumption is crucial for the success of most state-of-the-art machine learning interatomic potentials (MLIPs).[22−28]

Another key insight is that every stable material or molecule resides at the minimum of the potential energy surface (PES) of the system. As such, when slight perturbations are introduced to the atomic positions of the stable structure, the resulting forces guide the structure back to its equilibrium state during relaxation. This insight was used in Crystal Diffusion Variational AutoEncoder (CDVAE)[7] and generated new structures based on latent space representations.[29] One can also use this idea in

the reverse sense, for example, to learn a coarse-grained force field from the underlying distribution of atomistic configurations using a diffusion model.[30]

Here, we introduce a generative method for materials and molecules, Response Matching (RM), which leverages the locality of atomic interactions and the PES minimum while naturally incorporating permutation, translation, rotation, and periodic invariances. We highlight how RM is closely related to the Denoising Diffusion Probabilistic Model (DDPM).[31,32] Finally, we demonstrate the RM model across three systems: a small organic molecular data set, stable crystals from the Materials Project,[1] and a single diamond configuration training datum.

◼ METHODS

The method discussion begins by referencing foundational work on DDPMs,[31,32] which involves a noise model and a denoising neural network. The noise model corrupts a data point $\mathbf{x}$ to a sampled log signal-to-noise ratio, λ, as follows

$$\mathbf{x}_\lambda = \alpha_\lambda \mathbf{x} + \sigma_\lambda \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon}$ is a random noise. The denoising model learns to predict the clean input $\mathbf{x}$ from $\mathbf{x}_\lambda$, or equivalently, the added noise. The denoising neural network with parameters θ is trained on the score-matching objective over multiple noise scales

$$L_\lambda = \|\boldsymbol{\epsilon}^\theta(\mathbf{x}_\lambda) - \boldsymbol{\epsilon}\|^2 \tag{2}$$

Mirroring DDPM, Response Matching also includes a noise step and a denoising step. Noise is directly applied to the Cartesian coordinates of atomic structures. Consider the coordinates of an equilibrium structure as $\mathbf{R}_0 = (\mathbf{r}_1, ..., \mathbf{r}_N)$, where $\mathbf{r}_i$ denotes the position of atom i. Random atomic displacements are added to this initial structure, generating a sequence of increasingly amorphous structures, $\mathbf{R}_\lambda$. These noisy displacements can be determined all at once based on a continuum of distributions of λ, rather than being added step by step. The displacement for each atom i is represented as $\Delta \mathbf{r}_{i,\lambda}$ and for periodic atomic structures, this displacement follows the minimal image convention in PBC. For simplicity, we later omit λ in the subscripts.

As each equilibrium structure lies at the minimum of the PES, the forces on the atoms are nearly zero. When displacements are applied to these atomic positions, the atomic forces deviate from zero and tend to pull the atoms back to their original coordinates. The harmonic approximation, a simple yet widely used physical assumption in quantum mechanical calculations and atomistic modeling, states: the force on the atom i is

$$\tilde{\mathbf{F}}_i^H = -k\Delta \mathbf{r}_i \tag{3}$$

The tilde notation on the force indicates its fictitious nature. The force constant k is not a physical value but rather a hyperparameter of the RM model. This approximation effectively attaches a harmonic spring of stiffness k between the current position of atom i and its equilibrium position. In the present work, k is chosen as a constant throughout, which can be mapped to the choice of $\alpha_\lambda = 1$ and a constant $\sigma_\lambda = 1/k$ in eq 1 of DDPM during score matching. Alternatively, one can also use a k that is dependent on the magnitude of the atomic displacement (or equivalently, λ), which is akin to setting a specific different noising schedule in DDPM.

We also incorporate a physical inductive bias: atoms cannot approach too closely due to strong repulsive forces at short

distances. To account for this effect, we apply a short-range repulsive pairwise potential between atom pairs i and j, such as

$$g_c(r_{ji}) = \begin{cases} m\left(1 - \dfrac{r_{ji}^2}{r_c^2}\right)^n & \text{if } r_{ji} < r_c \\ 0 & \text{if } r_{ji} \geq r_c \end{cases} \tag{4}$$

where r_{ji} is the scalar distance between the pair, r_c is typically a fraction of an Angstrom, and m and n are hyperparameters dictating the strength of the repulsion. The resulting repulsive force on the atom i is

$$\tilde{\mathbf{F}}_i^R = -\sum_{j \in \mathcal{N}(i)} \frac{\mathrm{d}g_c(r_{ji})}{\mathrm{d}r_i} \tag{5}$$

summed over all the atoms j within the distance r_c from the atom i. Combining the harmonic and repulsive forces, the total fictitious force on atom i is $\tilde{\mathbf{F}}_i = \tilde{\mathbf{F}}_i^H + \tilde{\mathbf{F}}_i^R$.

For periodic systems, we also add distortions to their periodic cell to create the elastic strain $\boldsymbol{\gamma}$. The associated stress is approximated utilizing the stress−strain relationship for isotropic elastic materials

$$\tilde{\sigma} = \mathbf{C}\boldsymbol{\gamma} \tag{6}$$

where the components of $\boldsymbol{\epsilon}$ include normal and shear moduli that are treated as hyperparameters. For molecules without periodicity, the lattice strain step is skipped.

The general idea of RM is to use a denoising model with parameters θ to fit to the fictitious response properties, i.e., forces and stresses. The corresponding objective function is

$$L_\lambda = \sum_{i=1}^N \left\| \tilde{\mathbf{F}}_i^\theta - \tilde{\mathbf{F}}_i \right\|^2 + \beta \left\| \tilde{\sigma}^\theta - \tilde{\sigma} \right\|^2 \tag{7}$$

where β balances the relative weight between the force loss and the stress loss. Compared with the objective in DDPM eq 2, One can see that the mathematical framework is identical, with RM employing a physics-inspired response to the noise instead of the noise itself.

Using machine learning interatomic potentials (MLIPs) for this denoising model is advantageous because they inherently incorporate the translational, rotational, and permutational symmetries of atomic systems. Additionally, MLIPs exploit the nearsightedness of atomic interactions, expressing the system's total potential energy as the sum of the atomic energies for each atom, i.e.

$$E = \sum_i E_i \tag{8}$$

The forces can readily be computed by taking the derivatives of the total energy with respect to atomic coordinates, and the stress can be computed from the virial.

Any MLIP methods can be used for this purpose, e.g., Behler−Parrinello neural network potentials,[22] GAP,[23] Moment Tensor Potentials (MTPs),[24] Atomic Cluster Expansion (ACE),[25] proper orthogonal descriptors (POD),[33] NequIP,[26] MACE,[27] to name a few. Here, we use Cartesian Atomic Cluster Expansion (CACE)[28] without using message passing layers, due to its efficiency and alchemical learning capabilities, which are important for learning PES of materials with diverse elements. It is worth noting that CACE without message passing is mathematically equivalent to ACE,[25] MTPs,[24] POD,[33] and
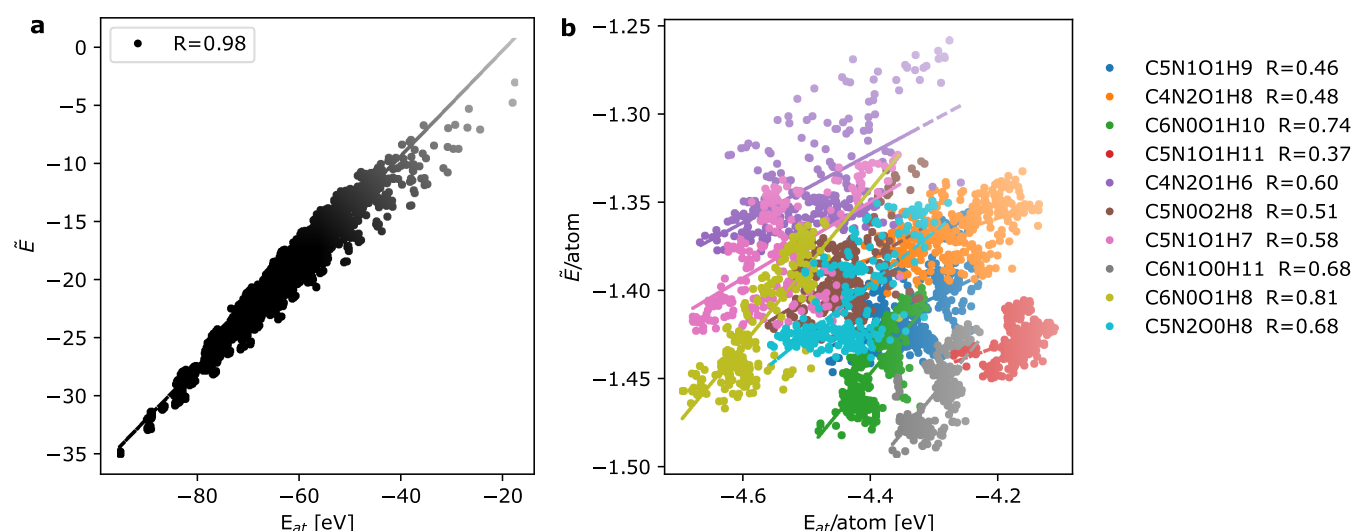
Figure 1. Comparison between the actual atomization energy (E_{at}) and the pseudo energy ($\tilde{E}$) predicted by the RM model for small molecules in the QM7b data set. (a) Shows a comparison of the energies per molecule. (b) Shows the comparison of the energies per atom for the molecules with the most common compositions in QM7b. The Pearson correlation coefficients R are provided in the legends.

MACE[27] with no message passing, so the results obtained in this paper are in principle transferable to these other MLIP methods.

In CACE, an atom is treated as a node on a graph, and edges connect atoms within a cutoff radius r_{cut}. Each chemical element is embedded using the learnable vector $\boldsymbol{\theta}$. The type of edge that connects two atoms, i and j, is encoded using the tensor product of the embedding vectors of the two nodes, $T = \boldsymbol{\theta}_i \otimes \boldsymbol{\theta}_j$. The length of the edge, r_{ji}, is described using a radial basis R. The angular component of the edge, $\hat{\mathbf{r}}_{ji}$, is encoded using an angular basis L. The edge basis combines all of this information

$$\chi_{cnl}(i, j) = T_c(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) R_n(r_{ji}) L_l(\hat{r}_{ji}) \quad (9)$$

The atom-centered representation is made by summing over all the edges of a node

$$A_{i,cnl} = \sum_{j \in \mathcal{N}(i)} \chi_{cnl}(i, j) \quad (10)$$

The orientation-dependent A features are symmetrized to get the rotationally invariant B features of different body orders ν, e.g., for $\nu = 2$[34]

$$B_{i,cnl}^{(2)} = \sum_{l} C(l) A_{i,cnl}^2 \quad (11)$$

Then, a multilayer perceptron (MLP) maps these invariant features to the target of the atomic energy of each atom i

$$E_i = \text{MLP}(B_i) \quad (12)$$

After the MLIP is trained using the objective in eq 7, it is then used for the denoising step in RM. As the denoising model has a pseudo potential energy surface $\tilde{E}$, rather than performing denoising with a fixed schedule, one can directly search for local minima on the $\tilde{E}$. This is effectively a random structure search process (RSS):[35] For each RSS run, we first chose a reasonable cell shape at random and added atoms of chosen elements and composition into the simulation cell at random positions while keeping the initial density of the cell close to the typical density range of this system. We set a lower bound on the interatomic distance for each pair of atomic species but otherwise imposed no additional constraints on the initial structures. We then relax both the atomic positions and simulation cell using the Fast

Inertial Relaxation Engine (FIRE) optimizer,[36] which continues until the pseudo forces on the atoms become negligible. FIRE dynamically adjusts step sizes for faster convergence,[36] making it a common choice in atomistic simulations due to its efficiency. Alternatively, one can use other methods such as Langevin dynamics, simulated annealing,[37] evolutionary algorithm,[38] particle-swarm optimization, and Bayesian optimization.[39] If Langevin dynamics is used here, the process can then be mapped to the denoising step using the reverse-time stochastic differential equations in diffusion models,[40,41] when the pseudo energy is the so-called energy function.[42] The choice of the optimization method will likely influence the probability of finding the global minimum on the PES versus the other local minima. For the purpose of finding new materials and molecules, it may not always be advantageous to maximize this probability as the local minima can also correspond to synthesizable structures.

■ EXAMPLES

Molecule Generation: QM7b. QM7b[43] is curated based on GDB-13[44] (a database of nearly 1 billion stable and synthetically accessible organic molecules), and it is composed of 7211 molecules of up to 23 atoms (including 7 heavy atoms of C, N, O, S, and Cl).

During training, random displacements with a maximum magnitude of 1.6 Å were added to all the atoms in the original molecules, and a CACE potential was employed to learn these pseudo forces. We used r_{cut} = 4.5 Å, l_{max} = 3, ν_{max} = 3, and $N_{embedding}$ = 3. The model has 16,490 trainable parameters. Training takes roughly 1 day on a laptop.

Notably, as training is solely based on the pseudo forces, atomization energy or other molecular properties were not included. However, as seen in Figure 1a, the atomization energy, E_{at}, is highly correlated with $\tilde{E}$, with a Pearson correlation coefficient of $R = 0.98$. This high correlation primarily results from the additivity of atomic energies eq 8, an important bias in the MLIP. Furthermore, the MLIP captures subtle energy differences between distinct atomic environments. We compared the actual atomization energy per atom with the pseudo energy per atom for molecules with the most common
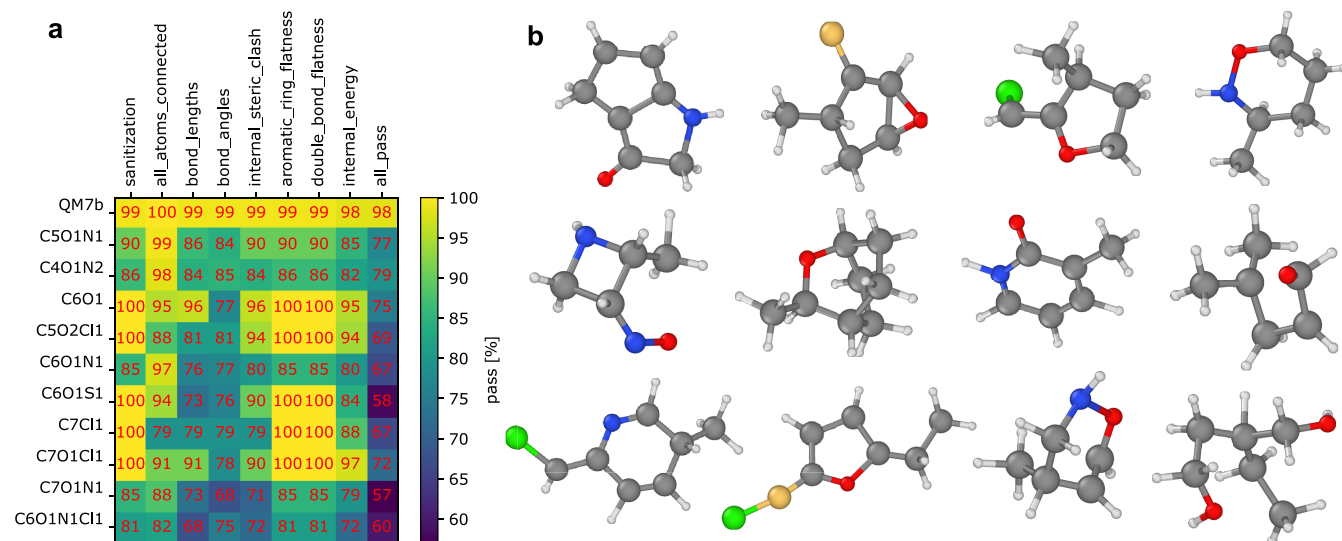
Figure 2. Illustrations of small molecules generated by using the RM model. (a) Shows the percentage of the molecules with given composition that passes the chemical feasibility checks using PoseBusters.[45] The last column indicates the percentage that passes all of the checks. (b) contains selected molecular configurations. The carbon, oxygen, nitrogen, sulfur, chlorine, and hydrogen atoms are colored using black, red, blue, yellow, green, and white, repectively.

compositions in QM7b. Each composition appears approximately 200−300 times in the training set. Figure 1a illustrates that E_{at} per atom is significantly correlated with $\tilde{E}$ per atom across these compositions. The MLIP identifies these energy differences because the training structures are PES minima.

To generate new molecules, we randomly placed atoms of specific compositions in an orthorhombic box without periodic conditions, with the sole constraint that the minimum interatomic distance be at least 0.7 Å. We conducted two sets of generation tasks: the first set used compositions well-represented in the training set, while the second set included compositions that were out-of-distribution (containing 8−9 heavy atoms). Generating a single structure by geometry relaxation generally takes less than 10 s on a laptop. Currently, the geometry relaxation is performed in serial mode, and it may be dramatically accelerated by relaxing many molecules together using one, sparsely connected, atomic graph. Figure 2 displays selected configurations generated using the denoising model.

We assess model performance by evaluating the chemical feasibility of the generated molecules and determining whether the model can learn chemical rules from data. Finding a rigorous and unbiased evaluation metric is challenging, and previous studies have used various criteria to assess the feasibility of a molecule represented in three-dimensional coordinates. For instance, a stability metric[46] checks whether all atoms in a molecule have correct valence based on specific bond distances. A common validity measure[47] assesses whether a molecule can be sanitized by RDKit[48] using default settings. However, a number of problems with these metrics were noticed,[14] including the inconsistent criteria in different studies and the discrepency between the criteria and actual chemical soundness. In Table 1, we compare the validity and stability results from recent models trained on small molecule data sets, including ENF,[46] the E(3) Equivariant Diffusion Model (EDM),[8] the MiDi diffusion model, which generates both two-dimensional (2D) molecular graphs and their corresponding three-dimensional (3D) coordinates,[49] the Geometric Latent Diffusion Model (GEOLDM),[10] the Geometry-Complete Diffusion Model (GCDM),[13] and SemlaFlow, an equivariant model

Table 1. Stability and Validity Metrics of Different Molecular Generation Models

| model and train set	stability (%)	validity (%)
ENF (14k QM9)[46]	4.9	40.2
EDM (100k QM9)[8]	82	91.9
MiDi (100k QM9)[49]	84	97.9
GEOLDM (100k QM9)[10]	89.4	93.8
GCDM (100k QM9)[13]	85.7	94.8
SemlaFlow (100k QM9)[14]	99.4	99.6
RM (7.2k QM7)	85.6	92.7

trained using flow matching along with scale optimal transport.[14]

Here, we use a set of comprehensive and stringent criteria offered by PoseBusters:[45] RDKit's chemical sanitization check (equivalent to the aforementioned validity), all atoms connected, bond lengths, bond angles, internal steric clash, aromatic ring flatness, double bond flatness, and the calculated energy of the input molecule based on a force field. The energy calculation step also includes a valency check for each atom, which is akin to the stability check. Figure 2 illustrates the percentage of molecules with given compositions that pass these chemical feasibility checks. For comparison, the first row displays statistics for the QM7b set. Each criterion provides different insights into the chemical structure, and the pass percentages vary but are correlated. The feasibility percentages are highly dependent on the specific compositions. Overall, the feasibility rate is quite high. The sanitization (validity) rate ranges from 80 to 100%, which is similar to rates reported in previous studies trained on the QM9 data set. The average values of stability and validity are reported in Table 1.

Material Generation: Materials Project Structures. We used the MP-20 structures as in ref 7, includes almost all experimentally stable materials from the Materials Project (MP)[1] with unit cells including at most 20 atoms. The training set contains 27,136 structures.

During the training, cell distortion of up to 0.1 Å was applied, and random displacements with a maximum magnitude of 0.8 Å
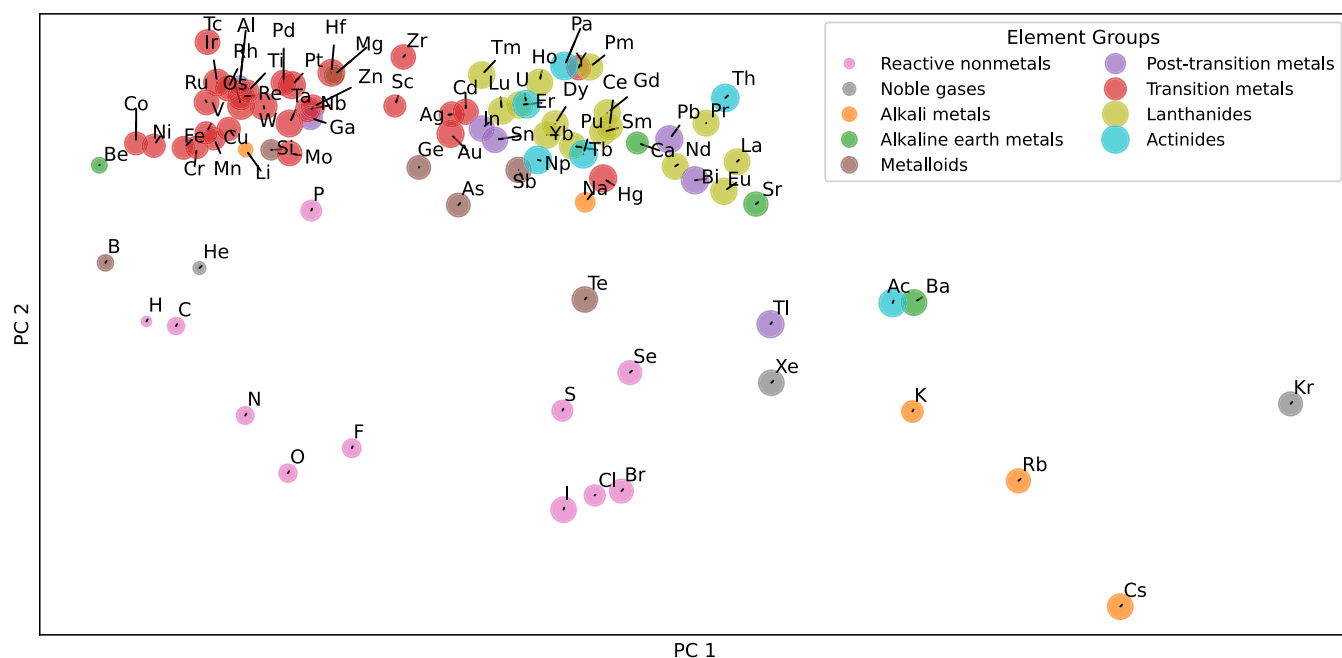
Figure 3. Similarity of chemical elements visualized using the first two principal components (PCs) of the CACE embedding matrix θ in the RM denoising model. Each element is colored according to its chemical group. The size of the symbol indicates the size of the elements. The noble gas element Ne is outside the plot.

were added to all atoms. A CACE potential was used to learn these pseudo forces and stresses. We used r_{cut} = 4 Å, l_{max} = 2, ν_{max} = 2, and $N_{embedding}$ = 4 for aggressive alchemical compression. The model has 16,312 trainable parameters. The training time took 2 days on an A10 card.

The alchemical learning capacity of CACE not only enhances learning efficiency but is crucial to developing a model that is applicable throughout the periodic table. The learnable embedding θ for each element type encodes its chemical information and can be visualized to provide insights into data-driven similarities. Given that $N_{embedding}$ = 4 in this case, we performed principal component analysis (PCA) and plotted the first two principal component axes in Figure 3. The elements are color-coded on the basis of their chemical groups, and the PCA map reveals that elements in the same group tend to cluster together. Elements within a group often have similar appearances and behaviors because they possess the same number of electrons in their outermost shell. This demonstrates that the element embedding scheme effectively captures the nature of the periodic table in a data-driven manner.

To generate new crystals, we randomly placed atoms of specific compositions in a box with periodic conditions with an initial molar volume that is from linear regression of the total volume with respect to the chemical elements from training structures. We then performed FIRE optimization for both the cell and atomic positions. We test our model on the same selected set as used in DiffCSP,[17] which contains 10 binary and 5 ternary compounds in the MP-20 test set. For each of these compositions, we generated 30 structures and compared them to the group truth using StructureMatcher from pymatgen.[50] For eight of these compounds (Ag_6O_2, Bi_2F_8, Co_2Sb_2, Co_4B_2, Cr_4Si_4, $KZnF_3$, Sr_2O_4, and YMg_3), we found the ground truth structures. Such match rate is similar to USPEX,[17,38] a crystal structure prediction software using the evolutionary algorithm to search for stable structures based on density functional theory

calculations, although worse than the 11/15 match rate using DiffCSP.[17]

To demonstrate how the RM model can help scientific applications, we used it for the Li−S battery system, which is an attractive candidate for numerous energy storage applications.[51] We searched for a number of stoichiometries of $Li_xS_{(1−x)}$. As sulfur is highly polymorphic and has complex structures, we used the known pure S structures from MP. To organize the generated structures, we plotted the pseudo convex hull using the pseudo excess energy, $\tilde{E}_{ex} = \tilde{E} − x\tilde{E}_{Li} − (1 − x)\tilde{E}_S$, in Figure 4.
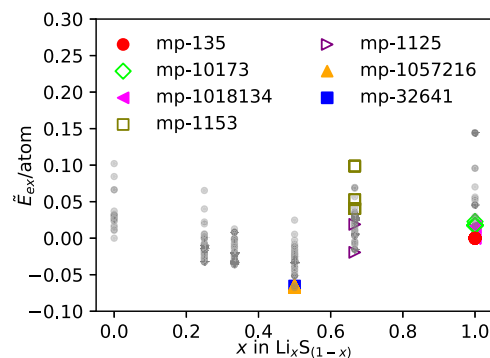

Figure 4. Pseudo convex hull of the generated Li−S structures at different Li fractions. The colored markers denote the generated structures that matched with the known structures in the Materials Project[1] using StructureMatcher from pymatgen,[50] and the material IDs are given in the legend. The filled (hollow) symbols indicate structures that are not in the training set.

In reality, Li_2S should be the stable phase on the energy convex hull[51] rather than LiS, but the overall trend in stability for LiS compounds is captured. It is worth cautioning that such results are not trustworthy, especially because energies did not enter the training of the model, and one always has to check against quantum mechanical calculations. Nevertheless, the trained RM

model can offer a quick and perhaps insightful first step in exploring a system. Several structures in the MP[1] are found, and they typically have lower pseudo energies, which suggests that one can use $\tilde{E}$ for additional screening. Among the seven matched MP structures, only the materials with IDs mp-10173 (Li, $P6_3/mmc$), mp-1125 (Li_2S, $Pnma$), and mp-1153 (Li_2S, $Fm\bar{3}m$) are in the training set.

One-Shot Learning: A Single Diamond Structure. To show the data efficiency and generalization of the RM, we trained on a single data point of a cubic diamond structure. In the noising stage, uniform random displacement with a maximum magnitude of 0.8 Å and lattice strain up to 0.1 were added to the original structure with a molar volume of 4.4 Å^3. For the CACE potential, we used r_{cut} = 4.5 Å, l_{max} = 3, ν_{max} = 3, and $N_{embedding}$ = 1. The model only has 2137 trainable parameters. Training takes less than 1 h on a laptop. During the relaxation stage, 2−12 carbon atoms were randomly placed in a simulation box with the molar volume set to be between 3.8 and 5.6 Å^3. Generating one structure takes a few seconds, depending on the system size.

We plot the pseudo energy $\tilde{E}$ against molar volume for the obtained structures in Figure 5. The cubic diamond structures
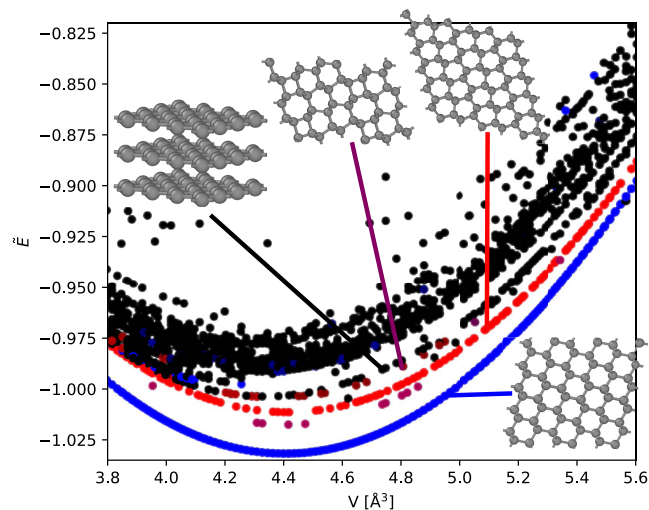
Figure 5. Pseudo energy of the generated carbon structures with different molar volumes. The blue markers represent cubic diamond structures; the red markers indicate hexagonal diamonds; and the purple markers denote diamonds with stacking faults. Graphite structures lie along the band pointed to by the black line.

(blue dots in Figure 5) consistently have the lowest $\tilde{E}$ at all molar volumes compared to the other structures found. The lowest-energy cubic diamond has a molar volume of 4.4 Å^3, which matches the training configuration exactly. Hexagonal diamonds (red dots in Figure 5) and diamonds with stacking faults (purple dots in Figure 5) are also frequently found. Notably, a set of graphite structures (in the inset of Figure 5) with varying volumes were identified. Graphite structures differ significantly from diamonds, consisting of stacked layers of carbon atoms in a hexagonal lattice. Within each layer, carbon atoms form strong covalent bonds with three neighboring atoms in a trigonal planar arrangement. The hexagonal diamonds, diamonds with stacking faults, and graphite structures are all local minima on the pseudo energy surface, while the cubic diamond is the true global minimum. This example thus demonstrates that tracing the local

minima can generate out-of-distribution, yet physically meaningful structures.

DISCUSSION

In summary, Response Matching (RM) is a generative method that has a noising step equivalent to other diffusion models and a denoising model that is effectively crystal structure prediction using a machine learning interatomic potential. Just as DDPMs[31,32] are trained to predict added noise, the MLIP in RM is trained to predict the response of the system to the added noise. These responses, in the form of pseudo forces and stress, are simply proportional to the noise. The benefits of using MLIP for denoising include: (i) Exploiting the locality of atomic interactions while naturally respecting permutation, translation, rotation, and periodic invariances. (ii) Allowing RM to simultaneously handle both molecules and bulk materials with and without periodic boundary conditions. (iii) Enabling advanced optimization methods such as FIRE,[36] simulated annealing,[37] particle-swarm optimization and Bayesian optimization[39] during denoising rather than adhering to a fixed schedule. (iv) The pseudo energies from the RM model are not directly trained but are somehow correlated with the real energies of the systems, which may offer physical insights and an additional way of screening the generated structures. Moreover, given that MLIPs are well-developed, their advances can be directly transferred to generative models of materials and molecules.

The current RM model can be extended in several ways: (i) Most crystals fall into a limited set of space groups. In crystal structure prediction, a common technique is to select a space group based on popularity and snap to it during relaxation.[35] This approach can also be applied to RM, which could enhance model efficiency by reducing the search space to realistic structures and facilitating the generation of experimentally relevant crystals. (ii) One can incorporate another term, $\|y^\theta - y\|^2$ in the objective function in eq 7, with y being the specific properties, such as bandgap, solubility, or thermal stability. This can enable RM to generate molecules and materials conditioned on specific properties, which is crucial for designing functional materials with targeted properties. (iii) Rather than fixing composition, alchemical swaps of elements using Monte Carlo moves or another generative model can dynamically determine atom types during denoising. (iv) One can further synergize RM and MLIPs, for example, training foundation models using both data from quantum mechanical calculations and unlabeled structural data. These enhancements would further expand the applicability and flexibility of RM in the generation of materials and molecular structures.

ASSOCIATED CONTENT

Data Availability Statement

The training scripts, trained models, and training data are available at https://github.com/BingqingCheng/cace-rm.

AUTHOR INFORMATION

Corresponding Author

Bingqing Cheng − *Department of Chemistry, University of California, Berkeley, California 94720, United States; The Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria;* ◉ orcid.org/0000-0002-3584-9632; Email: bingqingcheng@berkeley.edu

Complete contact information is available at:

https://pubs.acs.org/10.1021/acs.jctc.4c00998

Notes

The author declares no competing financial interest.

Code Availability The CACE package is publicly available at https://github.com/BingqingCheng/cace.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, No. 011002, DOI: 10.1063/1.4812323.

(2) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a datadriven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(3) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.

(4) Anstine, D. M.; Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736−8750.

(5) Gebauer, N.; Gastegger, M.; Schütt, K. In *Symmetry-Adapted Generation of 3d Point Sets for the Targeted Discovery of Molecules*, Advances in Neural Information Processing Systems; NeurIPS, 2019.

(6) Satorras, V. G.; Hoogeboom, E.; Fuchs, F.; Posner, I.; Welling, M. In *E(n) Equivariant Normalizing Flows*, Advances in Neural Information Processing Systems; NeurIPS, 2021; p 4181.

(7) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. S. In *Crystal Diffusion Variational Autoencoder for Periodic Material Generation*, International Conference on Learning Representations; ICLR, 2021.

(8) Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. In *Equivariant Diffusion for Molecule Generation in 3d*, International Conference on Machine Learning; PMLR, 2022; pp 8867−8887.

(9) Wu, L.; Gong, C.; Liu, X.; Ye, M.; Liu, Q. In *Diffusion-Based Molecule Generation with Informative Orior Bridges*, Advances in Neural Information Processing Systems; NeurIPS, 2022; p 36533.

(10) Xu, M.; Powers, A. S.; Dror, R. O.; Ermon, S.; Leskovec, J. In *Geometric Latent Diffusion Models for 3d Molecule Generation*, International Conference on Machine Learning; PMLR, 2023; pp 38592−38610.

(11) Peng, X.; Guan, J.; Liu, Q.; Ma, J. Moldiff: addressing the atom-bond inconsistency problem in 3d molecule diffusion generation, arXiv:2305.07508. arXiv.org e-Print archive, 2023 https://arxiv.org/abs/2305.07508.

(12) Huang, L.; Zhang, H.; Xu, T.; Wong, K.-C. Mdm: Molecular Diffusion Model for 3d Molecule Generation. *Proceed. AAAI Conf. Artifi. Intell.* **2023**, *37*, 5105−5112.

(13) Morehead, A.; Cheng, J. Geometry-complete diffusion for 3d molecule generation and optimization. *Commun. Chem.* **2023**, *7*, No. 150, DOI: 10.1038/s42004-024-01233-z.

(14) Irwin, R.; Tibo, A.; Janet, J.-P.; Olsson, S. Efficient 3d molecular generation with flow matching and scale optimal transport, arXiv:2406.07266. arXiv.org e-Print archive, 2024 https://arXiv.org/abs/2406.07266.

(15) Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Shysheya, S.; Crabbé, J.; Sun, L.; Smith, J.et al. Mattergen: a generative model for inorganic materials design, arXiv:2312.03687. arXiv.org e-Print archive, 2023. https://arXiv.org/abs/2312.03687.

(16) Sultanov, A.; Crivello, J.-C.; Rebafka, T.; Sokolovska, N. Data-driven score-based models for generating stable structures with adaptive crystal cells. *J. Chem. Inf. Model.* **2023**, *63*, 6986−6997.

(17) Jiao, R.; Huang, W.; Lin, P.; Han, J.; Chen, P.; Lu, Y.; Liu, Y. In *Crystal Structure Prediction by Joint Equivariant Diffusion*, Advances in Neural Information Processing Systems; NeurIPS, 2024.

(18) Ragoza, M.; Masuda, T.; Koes, D. R. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **2022**, *13*, 2701−2713.

(19) Gebauer, N. W. A.; Gastegger, M.; Hessmann, S. S.; Müller, K.-R.; Schütt, K. T. Inverse design of 3d molecular structures with conditional generative neural networks. *Nat. Commun.* **2022**, *13*, No. 973.

(20) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking, arXiv:2210.01776. arXiv.org e-Print archive, 2022. https://arXiv.org/abs/2210.01776.

(21) Pickard, C. J. Beyond theory driven discovery: hot random search and datum derived structures, arXiv:2407.06294. arXiv.org e-Print archive, 2024. https://arXiv.org/abs/10.48550/arXiv.2407.06294.

(22) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 146401.

(23) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, No. 136403.

(24) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(25) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99*, No. 014104.

(26) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, No. 2453.

(27) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csányi, G. In *Mace: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields*, Advances in Neural Information Processing Systems; NeurIPS, 2022; p 11423.

(28) Cheng, B. Cartesian atomic cluster expansion for machine learning interatomic potentials. *npj Comput. Mater.* **2024**, *10*, No. 157, DOI: 10.1038/s41524-024-01332-4.

(29) Pickard, C. J. Tbd, Personal Communication. 2024.

(30) Arts, M.; Satorras, V. G.; Huang, C.-W.; Zugner, D.; Federici, M.; Clementi, C.; Noé, F.; Pinsler, R.; van den Berg, R. Two for one: Diffusion models and force fields for coarsegrained molecular dynamics. *J. Chem. Theory Comput.* **2023**, *19*, 6151−6159.

(31) Ho, J.; Jain, A.; Abbeel, P. In *Denoising Diffusion Probabilistic Models*, Advances in Neural Information Processing Systems; NeurIPS, 2020; p 6840.

(32) Nichol, A. Q.; Dhariwal, P. In *Improved Denoising Diffusion Probabilistic Models*, International Conference on Machine Learning; PMLR, 2021; pp 8162−8171.

(33) Nguyen, N.-C. Fast proper orthogonal descriptors for manybody interatomic potentials. *Phys. Rev. B* **2023**, *107*, No. 144103.

(34) Zhang, Y.; Xia, J.; Jiang, B. Physically motivated recursively embedded atom neural networks: Incorporating local completeness and nonlocality. *Phys. Rev. Lett.* **2021**, *127*, No. 156002, DOI: 10.1103/physrevlett.127.156002.

(35) Pickard, C. J.; Needs, R. J. Ab initio random structure searching. *J. Phys.: Condens. Matter* **2011**, *23*, No. 053201.

(36) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbsch, P. Structural relaxation made simple. *Phys. Rev. Lett.* **2006**, *97*, No. 170201.

(37) Bertsimas, D.; Tsitsiklis, J. Simulated annealing. *Stat. Sci.* **1993**, *8*, 10−15.

(38) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **2006**, *124*, No. 244704, DOI: 10.1063/1.2210932.

(39) Cheng, G.; Gong, X.-G.; Yin, W.-J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **2022**, *13*, No. 1492.

(40) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. *Score-Based Generative Modeling through Stochastic Differential Equations*, International Conference on Learning Representations; ICLR, 2021.

(41) Song, Y.; Durkan, C.; Murray, I.; Ermon, S. In *Maximum Likelihood Training of Score-Based Diffusion Models*, Advances in Neural Information Processing Systems; NeurIPS, 2021; pp 1415−1428.

(42) Vincent, P. A connection between score matching and denoising autoencoders. *Neural Comput.* **2011**, *23*, 1661−1674.

(43) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, No. 095003.

(44) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

(45) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **2024**, *15*, 3130−3139, DOI: 10.1039/D3SC04185A.

(46) Satorras, V. G.; Hoogeboom, E.; Welling, M. In *E(n) Equivariant Graph Neural Networks*, International Conference on Machine Learning; PMLR, 2021; pp 9323−9332.

(47) Simonovsky, M.; Komodakis, N. In *Graphvae: Towards Generation of Small Graphs Using Variational Autoencoders*, Artificial Neural Networks and Machine Learning−ICANN 2018:27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4−7, 2018, Proceedings, Part I 27; Springer, 2018; pp 412−422.

(48) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An open source chemical structure curation pipeline using rdkit. *J. Cheminf.* **2020**, *12*, No. 51.

(49) Vignac, C.; Osman, N.; Toni, L.; Frossard, P. In *Midi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Springer, 2023; pp 560−576.

(50) 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. *Comput. Mater. Sci.* **2013**, *68*, 314−319.

(51) See, K. A.; Leskes, M.; Griffin, J. M.; Britto, S.; Matthews, P. D.; Emly, A.; Van der Ven, A.; Wright, D. S.; Morris, A. J.; Grey, C. P.; Seshadri, R. Ab initio structure search and in situ 7li nmr studies of discharge products in the li-s battery system. *J. Am. Chem. Soc.* **2014**, *136*, 16368−16377.