

# Data-Driven Virtual Material Analysis and Synthesis for Solid Electrolyte Interphases

Deepalaxmi Rajagopal,\* Arnd Koeppe,\* Meysam Esmailpour, Michael Selzer, Wolfgang Wenzel, Helge Stein, and Britta Nestler


Solid electrolyte interphases (SEIs) form as reduction products at the electrodes and strongly affect battery performance and safety. Because SEI formation poses a highly nonlinear, complex multi-physics problem over various lengths and time scales, traditional modeling approaches struggle to characterize SEI evolution solely with existing physical properties. To improve the characterization of SEIs, it proposes a data-driven strategy for a virtual material design that learns to represent and characterize SEI formation with physical and data-driven properties from kinetic Monte Carlo simulations. A Variational AutoEncoder with a property regressor learns data-driven properties, which represent SEI configurations and correlate with physical target properties. This new neural network design encodes the high-dimensional structural and reaction spaces into a lower-dimensional latent space, while the property regressor orders the latent space by physical target properties. The model achieves high correlation scores between target and predicted properties from latent representations, thereby proving that the data-driven properties enrich the expressiveness of SEI characterizations.

## 1. Introduction

Over the past decades, lithium-ion batteries (LIBs) have become part of daily life.<sup>[1]</sup> The applications of LIBs are very diverse, ranging from portable devices like smartphones, smartwatches, and

D. Rajagopal, A. Koeppe, M. Selzer, B. Nestler  
Institute of Applied Materials (IAM-MMS)  
Karlsruhe Institute of Technology (KIT)  
Straße am Forum 7, 76131 Karlsruhe, Germany  
E-mail: deepalaxmi.rajagopal@kit.edu; arnd.koeppe@kit.edu

D. Rajagopal, A. Koeppe, M. Esmailpour, M. Selzer, W. Wenzel, B. Nestler  
Institute of Nanotechnology (INT)  
Karlsruhe Institute of Technology (KIT)  
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen,  
Germany  
H. Stein  
Helmholtz Institute Ulm  
Lise-Meitner Str. 16, 89081 Ulm, Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aenm.202301985>

© 2023 The Authors. Advanced Energy Materials published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

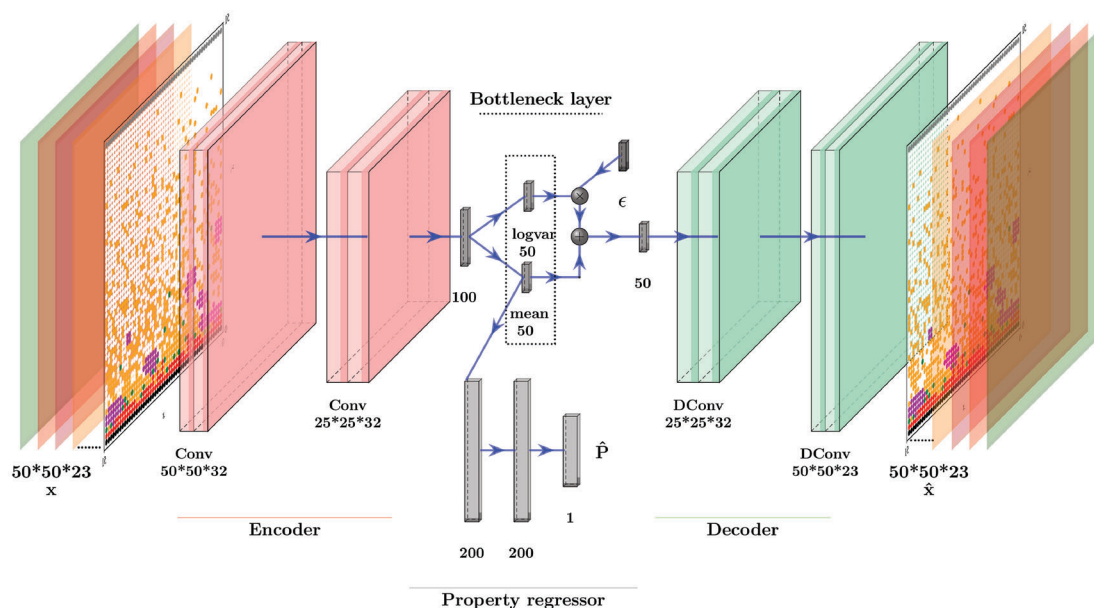
DOI: 10.1002/aenm.202301985

laptops to heavy-duty ones like electric vehicles, large-scale utility storage, and space exploration. Even though the LIB has made remarkable advancements, a better understanding of the solid electrolyte interphase (SEI) formed during the initial cycles of the lithium-ion battery is required.<sup>[2]</sup> The SEI is a passivation layer formed on the electrode surfaces as a result of the decomposition of the electrolyte. This passivation layer blocks electrons from the further decomposition of the electrolyte and allows Li<sup>+</sup> transport. The formation of thick and stable SEI plays an important role in the battery's performance as it deals with the consumption of active Li metal and electrolytes, which results in capacity fading, reduced power density, and increased resistance.

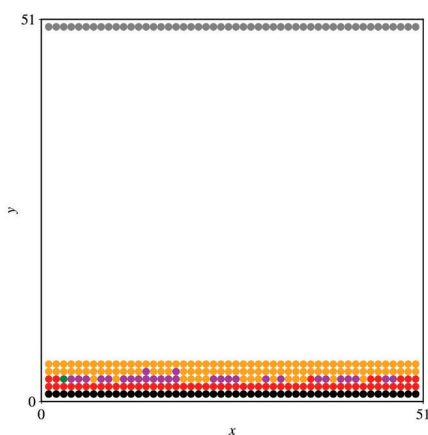
On the other hand, preventing electron tunneling improves the electrochemical

stability of the battery, which is responsible for the battery life and safety.<sup>[3]</sup> Due to the importance of the SEI, numerous studies were conducted to understand its chemistry, physical properties, and formation process. To study the interphasial chemistry and morphology of the SEI, ex situ techniques like X-ray photo spectroscopy (XPS), Fourier transform infrared, or Raman spectroscopy are used. However, these techniques significantly damaged the sensitive SEI due to environmental exposure and high-energy beams.<sup>[4–7]</sup> In situ characterization techniques like secondary ion mass spectroscopy and atomic force microscopy (AFM) offer a way to obtain information about the evolution of SEI under realistic battery operating conditions. With the development of these in situ characterization techniques over the years, the structure and formation of SEI have been studied in detail. In situ transmission electron microscopy (TEM) and scanning electron microscopy (SEM) are used to study the morphological evolution of the SEI, such as volume expansion and crack formation. However, due to their complex heterogeneous structure and the lack of reliable in situ characterization techniques, the formation and growth mechanisms of this passivation layer remain elusive, and the probes used in the characterization technique struggle to reach the specific SEI locations due to the high sensitivity of SEI.<sup>[8]</sup>

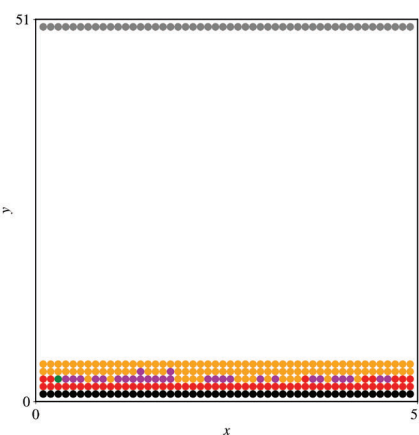
Because SEI growth involves effects interconnected across different time and length scales, evolution cannot be modeled using methods limited to certain length scales. The quantum



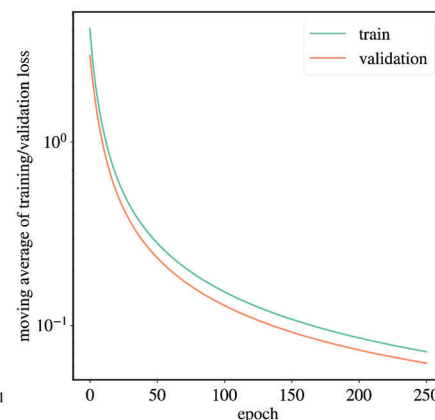
(a)



(b)



(c)

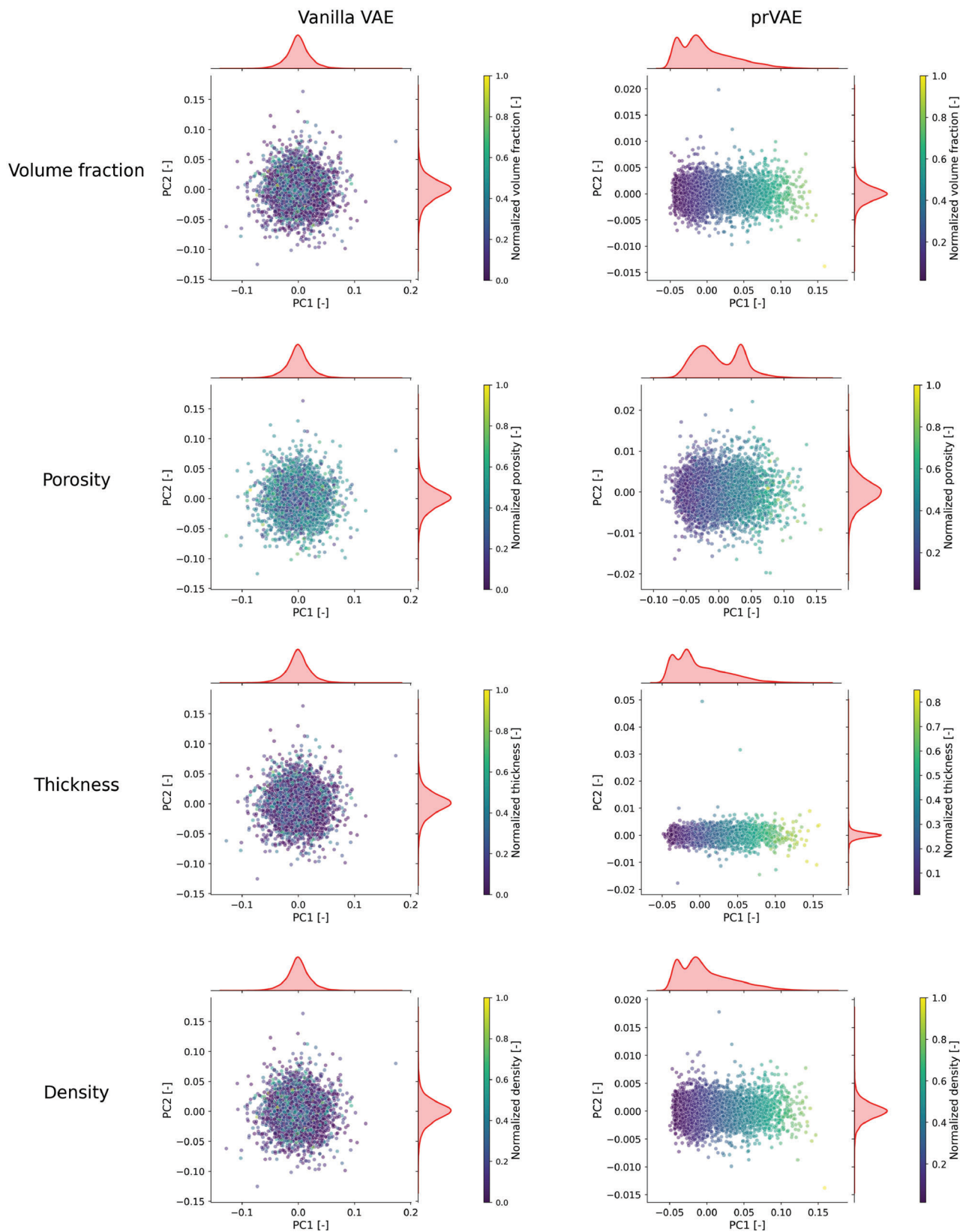


(d)

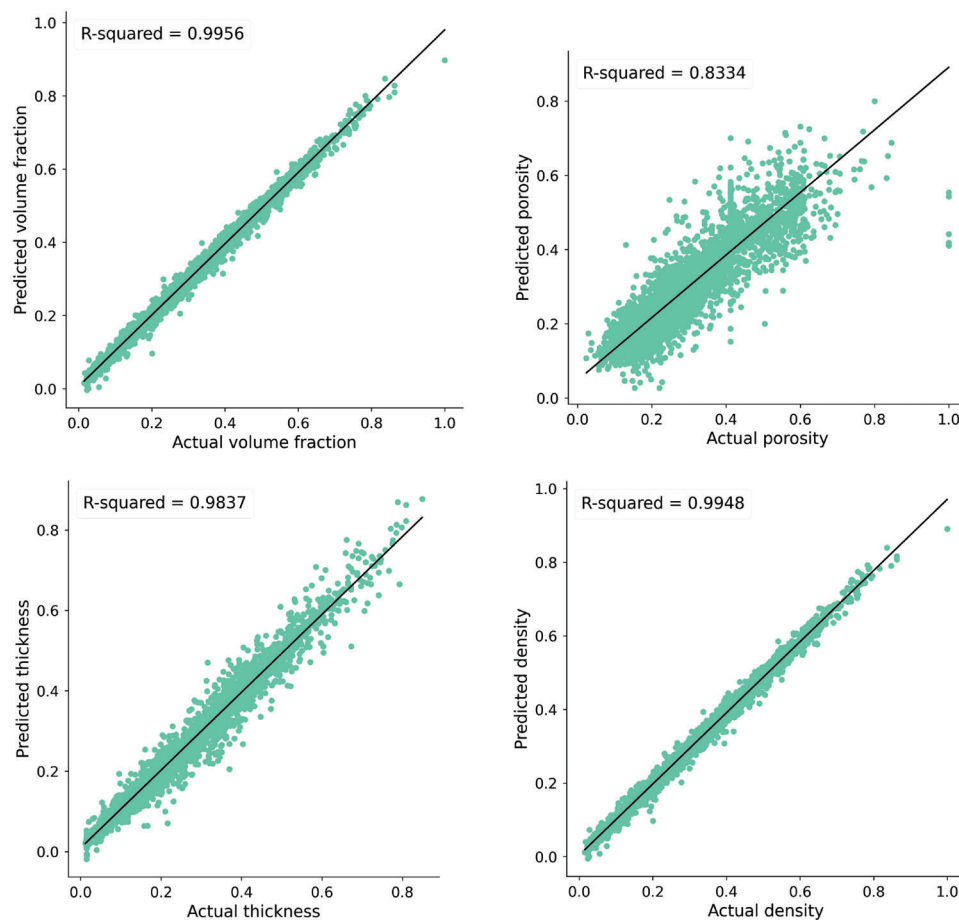
**Figure 1.** a) The architecture of the proposed prVAE model with mean and variance represents the parameters of the encoded distribution learned during training, and  $\epsilon$  represents the small noise added to reparameterize the model for backpropagation during training; b) the randomly selected SEI configuration of the test dataset; c) the corresponding prediction of the SEI configuration of the selected sample of the test dataset. The prVAE model shows consistent prediction accuracy among the test dataset; d) Loss curves during model training.  $\hat{P}$  shows the prediction of chosen physical properties for training such as volume fraction, thickness, porosity, and density of the SEI.

chemical (QC)<sup>[3,9]</sup> and molecular dynamic (MD)<sup>[10,11]</sup> approaches are widely used to understand the initial electrolyte reduction reaction and decomposition mechanisms. But these methods cover only the early stages of SEI formation. The continuum models are used beyond the atomic scale to cover the growth of SEI on a larger scale.<sup>[12]</sup> Due to the lack of microscopical understanding of the growth of SEI on the mesoscale, certain aspects of SEI, such as the electronically insulating extent of SEI and the structure of the evolving SEI layer, are still unclear. To overcome this problem, the bottom-up multiscale approach was formulated to understand the system-specific characterization of microscopic SEI formation processes.

The Kinetic Monte Carlo (KMC) protocol uses reaction rates obtained from quantitative chemical calculations to determine the magnitude of the SEI in a mesoscopic model with molecular resolution.<sup>[13]</sup> The above-described proposed model includes spatial and temporal information about the evolving SEI, governed by a series of chemical reactions, diffusion, and aggregation mechanisms, with kinetic data obtained for specific electrolyte-anode reactions. The process for synthesizing such an SEI configuration will be discussed in detail in the following sections. By combining the KMC protocol with active learning, the configuration-property linkage can be understood for further optimization of SEI configurations, and the generation



**Figure 2.** Principal Component Analysis on latent space of the vanilla VAE and prVAE with the selected physical property such as volume fraction, porosity, thickness, and density of the SEI configurations. Compared to the vanilla VAE model (Left), the prVAE (Right) shows ordered information-rich reduced dimensional space.



**Figure 3.** Prediction of selected properties of the SEI by the property regressor. The property regressor trained along with VAE achieved a good  $R^2$  score for each selected property and high prediction accuracy.

of SEI configuration and its corresponding property prediction can be made faster than a simulation method with numerous computations.<sup>[14,15]</sup>

The conventional trial-and-error-based optimization of materials starts from known material configurations and requires human effort to direct the material optimization for better performance. In contrast, the inverse material design allows the definition of the system's desired performance targets to determine the material configuration that meets these targets.<sup>[16]</sup> So far, the pioneering battery interphase design has been very unidirectional, starting from known structure to known properties, resulting in a lack of mapping between structure and system performance. The main challenges in battery phase design are incorporating different lengths, time scales, and complex chemical structures.<sup>[17]</sup> The inverse design of materials helps to overcome this challenge by taking advantage of deep generative models. These deep generative models allow the incorporation of different domains and time scales without any restrictions, as in MD simulations.

Deep generative models can discover new reliable material configurations by learning the underlying essential information obtained from large datasets.<sup>[16]</sup> Variational Auto Encoder (VAE)<sup>[18]</sup> and Generative Adversarial Neural networks (GAN)<sup>[19]</sup> are the most widely used deep generative models that use an unsupervised learning approach to understand the unique repre-

sentation of the data. Besides learning the underlying key features of the input battery interphase data, these models can generate reliable battery systems by utilizing their updated prior knowledge.<sup>[20,21]</sup> The semi-supervised VAEs are used to boost the representational learning and classification of the input data. Conditional VAEs<sup>[22]</sup> with proper training can be used to generate favorable SEI compositions for better battery performance.

Batteries are interphasial systems with numerous phases that require optimization of several properties.<sup>[23]</sup> These observed properties are controlled by the battery interphase configurations at different lengths and timescales.<sup>[24]</sup> Therefore, training a generative model with a multitask setting is required to extract latent representations for each considered data at multiple scales.<sup>[25,26]</sup>

We propose a data-driven strategy for virtual material analysis and synthesis that learns to represent, characterize, and generate SEI configurations with physical and data-driven properties from kinetic Monte Carlo simulations. A VAE model with a property predictor is established to learn the key features of 2D SEI configurations of 50000 samples obtained at the end of the KMC simulation. The key features known from the SEI configuration are studied at the bottleneck of the VAE to understand the influence of observable properties of the SEI, such as thickness, porosity, density, and volume fraction, on the learned data-driven properties. To further improve the classification of 2D SEI



**Figure 4.** Heatmap showing the correlation between different target physical properties. The low variance and weak correlation of porosity with other properties explain the low  $R^2$  score in predicting the porosity information of SEI configurations.

configurations concerning their influential properties, the inputs to the variational autoencoder model were conditioned with a reaction barrier set responsible for specific SEI conditions. Therefore, this data-driven strategy generates SEI configurations with tailored physical properties for given reaction barrier sets.

## 2. Data-Driven Strategy for Virtual SEI Analysis and Synthesis

### 2.1. Data Generation

The virtual SEI configurational data utilized for our study is obtained through a 2D KMC scheme from Esmailpour et al. 2023.<sup>[13]</sup> This KMC formulation follows a rejection-free algorithm called the BKL algorithm.<sup>[27]</sup> This new formulation of KMC simulations developed by Esmailpour et al. follows a bottom-up multiscale approach to simulate SEI's growth based on system-specific characterizations of microscopic processes that lead to the formation of SEI. The current data-driven study uses the results based on this new formulation of KMC simulation. Kinetic Monte Carlo (KMC) simulations can model the growth of solid electrolyte interphase on the spatial scale of nanometers and on a time scale of microseconds with a molecular resolution of  $\approx 1$  nm. KMC is a variant of Monte Carlo simulations that captures the evolution of mesoscale processes where the considered system is spatially and temporally discretized according to a set of reactions. The results of a 2D model on a square lattice in which space is discretized at a scale proportional to the size of the molecular components, such as  $\text{Li}_2\text{EDC}$ ,  $\text{Li}_2\text{CO}_3$ , and  $\text{C}_2\text{H}_4\text{OCOOLi}$  are used to generate meta solid electrolyte interphase configuration. The simulations are based on a variation of the rejection-free KMC algorithm,<sup>[27]</sup> which selects a reaction from the gathered reaction sets of the previous state to the next state according to the transition state theory. The KMC algorithm simulates SEI growth on a mesoscopic scale.<sup>[13]</sup> The SEI formation is governed by a series of reactions that start with electron reduction to generate inorganic or organic components, followed by an aggrega-

tion of inorganic components to form the inorganic part of the SEI and diffusion of organic components to form SEI cluster or organic part of the SEI. The reaction rates used to generate the SEI configurations are sampled using Latin hypercube sampling by randomly selecting the initial reaction rate set adopted from published literature<sup>[28–30]</sup> on the rates of SEI growth. Fifteen possible reaction rates responsible for SEI growth are considered, and 50000 15D reaction rate vectors are sampled using Latin hypercube sampling.<sup>[31,32]</sup> Each 15D reaction rate vector generates a spatiotemporal SEI configuration. The components formed during the SEI growth simulation for the given input reaction rates are color-coded. The color coding of the final snapshot of the SEI growth simulation is used to identify the inorganic, organic, and intermediate precursor components of the SEI. This study defines the volume fraction, thickness, density, and porosity as observables of the SEI. Each simulation takes up to 30 min of CPU hours on a single core. The sequential dependence of the reactions makes it challenging to implement the KMC algorithm in parallel. The computational cost of the simulation depends on the size of the lattice for the constant list of reactions; that is, the CPU hours increase with an increase in lattice size. To accelerate the time-consuming process of Monte Carlo simulations, surrogate models can learn to sample additional large datasets from smaller initial subsets of data.<sup>[33,34]</sup> In the following, we propose a deep generative model to replace the additional sampling by directly predicting tailored SEI configuration and observables through screening the corresponding reaction barrier space. Thus, the deep generative model can support the KMC simulation in terms of computational cost and faster discovery.

### 2.2. Preprocessing

The source dataset for training consists of 50000 2D SEI configurations obtained at the last step of KMC simulations,<sup>[13]</sup> along with its observable/physical properties such as volume fraction, thickness, density, and porosity, and the corresponding reaction barrier of the 15D vector. The 2D SEI configurations are categorically encoded, followed by one-hot encoding according to the color codes for each reaction product in the considered configuration. The physical properties or observable of the SEI configuration and reaction barrier set are preprocessed to the normalized range for better performance and stability of the implemented machine learning model. The preprocessed data of each SEI configuration is written as a TFRecord file<sup>[35]</sup> for effective serialization of structured data and to prepare the data for the machine learning study.

### 2.3. Deep Generative Models

Recent developments in deep generative models have provided access to efficient representation learning of large datasets and the discovery of new material configurations in the material science field. VAE<sup>[18]</sup> and GAN<sup>[19]</sup> are two widely used deep generative models for understanding the underlying structure of a high-dimensional dataset by sampling over lower-dimensional latent space. The GAN is identified by training a pair of competing neural networks, namely, the generator and discriminator,

to generate new samples similar to the observed data. In contrast, VAE uses parametric data encoding to a normal distribution over a continuous latent space, and each point in the latent space is decoded back into the input data space. VAE adds randomness to its encoded samples to create a generalized latent space. This forces the decoder to decode a wide range of points, resulting in robust representations. In contrast, standard autoencoders often have discontinuous and irregularly bounded distributions in their learned latent representations, which are unsuitable for generation and optimization applications. As opposed to GANs, the trained VAE allows continuous mapping of the latent space from the input data, which is tractable. Therefore, the VAE can produce meaningful latent-space representations.<sup>[36,37]</sup> The encoder component of VAE transforms each input sample into a normal distribution across the continuous latent space using the statistical mean and variance of the data. The normal distribution used in this context is defined by two parameters: the mean and the variance (logvar). Each data point is mapped to a vector consisting of a mean and variance, which define a multivariate normal distribution around that point. A random point is then sampled from this distribution and returned as the latent variables, which form the sampling layer of VAE. The decoder uses this latent variable to generate the output. The encoding and decoding components of the VAE are represented as neural networks trained to optimize the lower bound on the likelihood of the input samples. This is achieved by minimizing a loss function that has two terms: a reconstruction term that pushes the decoder to correctly rebuild the input given its latent representation and a regularization term that is the Kullback-Leiber divergence function (KL) between the conditional distribution defined by the encoder and the defined prior of the dataset. During training, updating the parameters of VAE using backpropagation can be challenging due to the stochastic nature of the sampling layer. One way to overcome this is to compute the gradient of the sampling layer during backpropagation with respect to the mean and log-variance vectors. In addition, maintaining the stochasticity of the model is possible by multiplying an extra parameter called epsilon ( $\epsilon$ ) as follows:  $z = \mu + \sigma^* \epsilon$ , where  $\epsilon \approx \mathcal{N}(0, I)$  and  $\sigma = e^{\log \text{var}/2}$ . The epsilon variable remains a random variable sampled from a standard normal distribution with a very low value, ensuring that the model does not shift away from the true distribution. This reparameterizes the model for end-to-end training. Designing a desirable material configuration involves optimizing various properties correlated with each other. Optimizing de novo material configurations concerning a single target property can cause undesirable changes in other properties. To avoid this, the Conditional Variational AutoEncoder (CVAE) model offers a way to generate material configurations by manipulating a given set of target properties. The prime difference between the VAE and CVAE can be observed in their objective functions.<sup>[38]</sup> The objective function of the VAE and CVAE are as follows:

$$E[\log P(X|z)] - D_{KL}[Q(z|X) \parallel P(z)] \quad (1)$$

$$E[\log P(X|z, c)] - D_{KL}[Q(z|X, c) \parallel P(z|c)] \quad (2)$$

where  $P(z)$  is the prior distribution of latent variables,  $P(X|z)$  is the approximated distribution of  $X$  (input material configuration) conditioned on latent variables  $z$  by the decoder part,  $Q(z|X)$  is the

approximated distribution of the latent variables  $z$  conditioned on input variable  $X$  and  $c$  is the conditioning vector. The structure of the conditional VAE can be improved by using the conditional vector only on the decoder network to allow the encoder network of conditional VAE to initialize the network parameters. The evidence lower bound or objective function of this improved conditional VAE is given by:

$$E[\log P(X|z, c)] - D_{KL}[Q(z|X) \parallel P(z|c)] \quad (3)$$

As we can see from this equation, the log marginal likelihood is also given as a function of hidden representation, but the outputs are also conditioned on input  $X$  and latent variable  $z$ . Given these additional conditions over the output distributions, we can train the model using the backpropagation algorithm used originally in CVAE.<sup>[39,40]</sup> This improved CVAE allows the guided generation of new samples of a specified category.

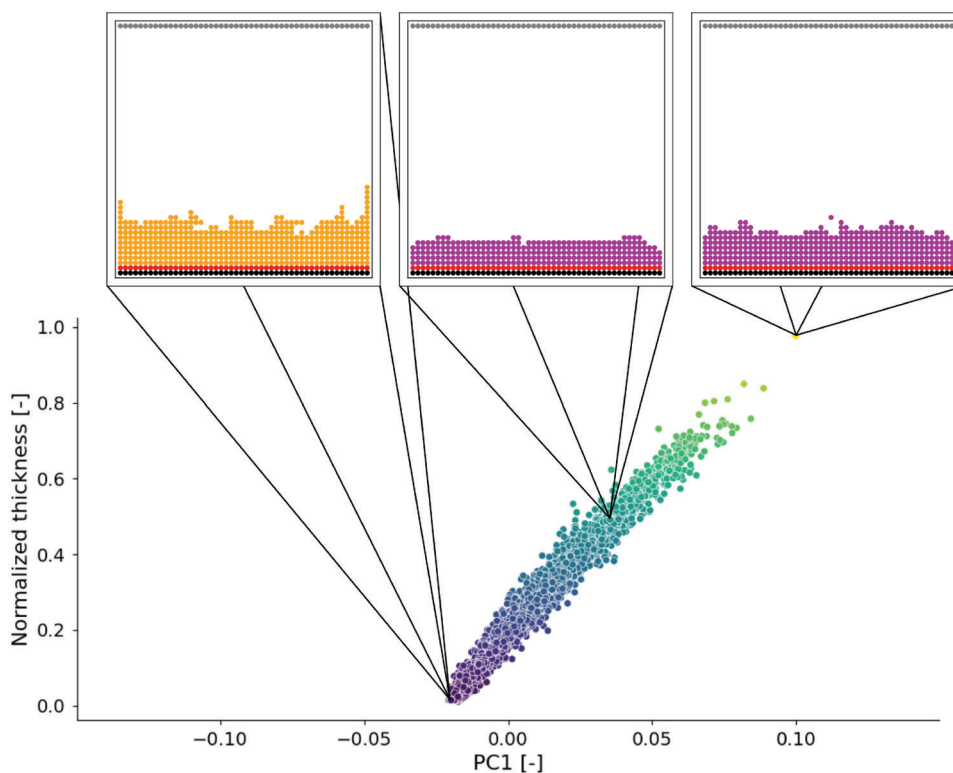
#### 2.4. Proposed Model Architecture and Training

As discussed in the previous section, the VAE learns the continuous latent space by focusing only on the spatial SEI configuration. To enable the inverse design of SEI configurations and characterize the SEI growth given its barrier set, the SEI configuration embedded at the bottleneck of the VAE should correspond to the target physical properties of the SEI. Therefore, as shown in **Figure 1**, we jointly trained a VAE and property regressor. The property regressor is a deep neural network that predicts the property of the SEI ( $\mathbf{P}$ ) propagates the learned physical information back into the latent space of the encoder network. **Figure 1a** shows the model architecture used for training. This type of model architecture introduces regression loss in addition to the VAE loss terms so that the model trains on SEI spatial configuration and target physical properties. The property regressor uses a series of connected neural networks to analyze data. The encoder and decoder components of the VAE are defined using convolutional neural network layers. During training, the decoder receives a latent vector sampled from the approximated posterior distribution as input, while the property regressor only takes the mean value  $\mu$  of the encoded distribution as input. **Figures 1b,c** show the ground truth and corresponding prediction of SEI configuration from the test dataset outside the training dataset. The trained prVAE accurately predicts SEI configuration even in the test dataset. Furthermore, **Figure 1d** demonstrates the model has better generalization between the training and validation sets.

### 3. Results and Discussion

#### 3.1. Characterization of SEI in Latent Space

To determine how well prVAE can encode the complex SEI feature in a reduced-dimensional latent space, we randomly selected 7500 test samples outside the training dataset. The resulting property information-rich latent space is continuous and decodes into a valid SEI configuration with target properties. For better visualization, we conducted a dimensional reduction



**Figure 5.** Latent space exploration along the first principal component: The first principal component holds the majority of the variational information of the property thickness and is conditioned by the reaction barrier set. The decoding of minimum (left), mean (middle), and maximum (right) values of the first principal component to the corresponding SEI configuration is shown here.

using Principal Component Analysis (PCA) analysis on the latent spaces obtained from encoding the test dataset to compare the performance of prVAE and vanilla VAE. As shown in **Figure 2**, prVAE performs better than vanilla VAE in accurately capturing the physical information associated with the SEI feature. The result contains valuable SEI property information, making exploring, optimizing, and interpolating target SEI configurations easier. The prVAE captures the property information into the encoded dimension of latent space; each point in the latent space decodes into an SEI configuration with valid physical properties.

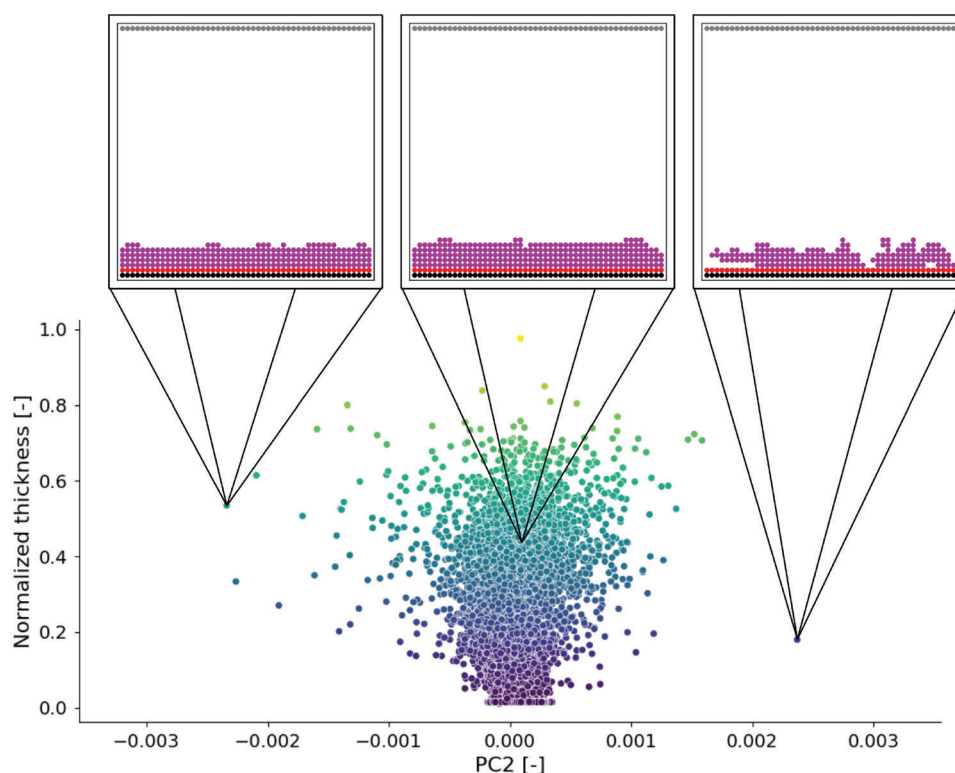
### 3.2. Property Prediction of SEI Configuration

To facilitate the design of SEI with target properties, we jointly trained fully connected dense neural network layers with VAE. The VAE as a generative model can generate new data by decoding points randomly sampled from the learned lower dimensional latent space. The ability of VAE to generate new data helps to find a new material configuration better for the application. The extended VAE with property regressor (prVAE) predicts the property from the encoded dimension of each SEI configuration. By jointly training the property regressor, the selected property values order the encoded distribution as a gradient of the property. **Figure 2** illustrates this representation for each property. **Figure 3** shows the  $R^2$  score between the actual and predicted values of the selected property values of the randomly se-

lected test dataset outside the training dataset. With high prediction accuracy, the property regressor captured almost all property information to order the encoded latent space. From **Figure 3**, we can see that the  $R^2$  score in the case of porosity is less compared to other properties; this explains the low variance and weak correlation of porosity value (**Figure 4**) with other selected properties for the study. The extension VAE model with property regressor solved not only its primary purpose to order the latent space automatically but also the obtained information-rich lower dimensional latent space, which can be later used to optimize the SEI configurational space with reaction barrier values and the corresponding properties for better performance of the battery.

### 3.3. Guided SEI Generation by Walking the Latent Space

The SEI configurations generated by decoding points on the latent space can be conditioned by adding additional inputs to the decoder of the prVAE. In our study, we used the reaction barrier space as input to the decoder in addition to sampled latent vectors from the encoded distribution of lower dimensional latent space. By training this prVAE with conditional inputs at the decoder, the model learns to generate SEI configurations with reaction products of SEI growth based on given reaction barrier spaces. Results in Sections 3.1 and 3.2 show that the prVAE can order its learned latent space based on property values; similarly, prVAE with conditional inputs also automatically organizes



**Figure 6.** Latent space exploration along the second principal component: The second principal component does not carry any variational information about the thickness. The explainability of this principal component to describe the SEI configuration according to SEI configuration is insignificant compared to the first principle component. The decoding of minimum (left), mean (middle), and maximum (right) values of the second principal component to the corresponding SEI configuration is shown here.

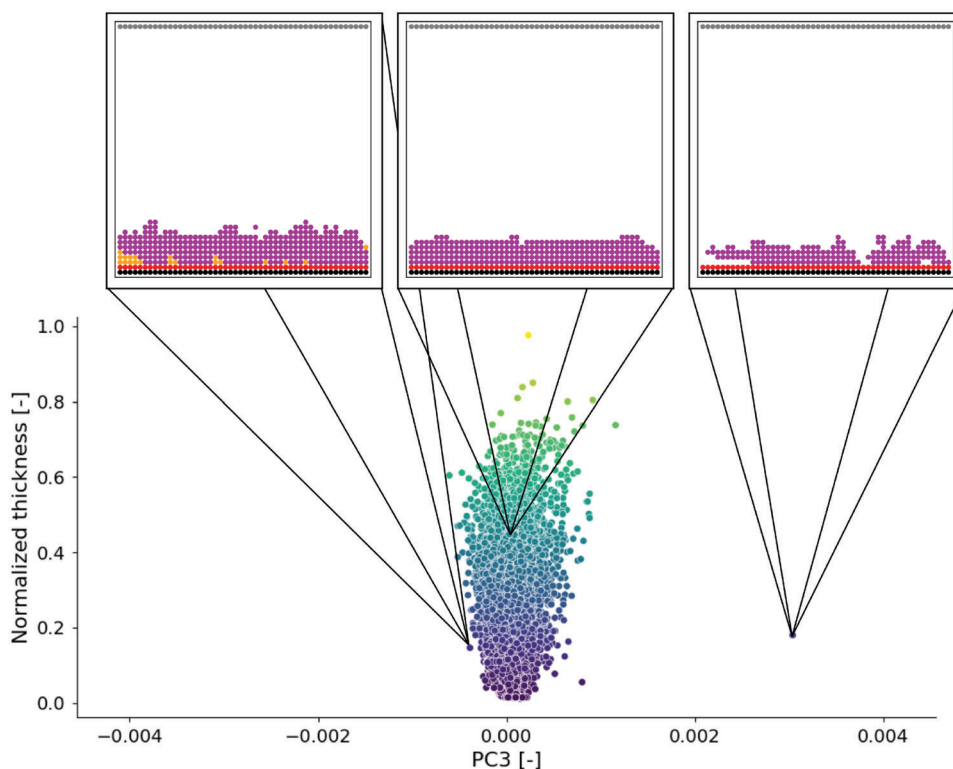
the latent space with property values. For this study, we considered the latent space learned by the conditional prVAE with thickness information. When training the prVAE model with a conditional barrier set, each dimension in the latent space attempts to learn the encoded distribution in an ordered fashion. Principal component analysis is utilized to evaluate the effectiveness of the learned latent space of conditional prVAE in describing the SEI configurations and property information. The information presented in **Figure 5** indicates that the first principal component effectively captures the variations in thickness and presents SEI configurations in a well-ordered manner along its axis. On the other hand, the second and third principal components, as shown in **Figures 6** and **7**, remain almost constant despite changes in thickness. Here, the rise in purple and red layers signifies the increase in thickness in the organic and inorganic SEI layers, respectively, while the orange color represents the intermediate product required for organic SEI formation. To explore the learned latent space with property information and their effects on SEI configuration, the minimum and maximum of all latent dimensions and reaction barrier space are used to interpolate the latent variables. Walking this latent space along the given interpolation direction, we can generate SEI configurations for the given reaction barrier space. In this study, we use **Figures 8a,b** to demonstrate how SEI configurations can be generated by varying the input barrier or latent variables while keeping the other constant and vice versa. By having conditional inputs as reaction barrier set, we can define the re-

action product type to generate the SEI configuration of a corresponding thickness (**Figure 8a**). In other words, the reaction barrier as conditional inputs control the chemical space required to form SEI.

#### 4. Conclusion and Outlook

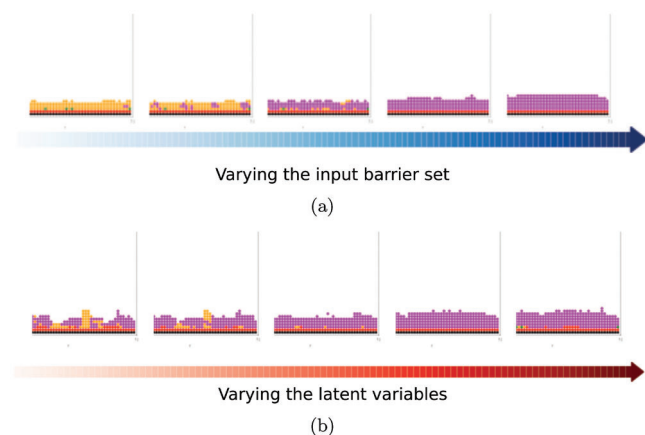
The proposed prVAE model architecture captures the data-driven properties of the SEI configurations essential for its characterization. The property regressor added to the standard VAE model effectively extracts the key features of virtual SEI configurations to encode a continuous and information-rich latent space. Training the VAE jointly with the property regressor significantly enhances its generation quality. Furthermore, the property regressor demonstrates higher accuracy in predicting the physical properties. The obtained information-rich continuous latent space constitutes the data-driven properties organized according to the physical property values. By walking the encoded latent space, we can decode the SEI configuration of target physical properties. Adding the selected reaction barrier set responsible for SEI configurational space as conditional input to the decoder of prVAE, we can direct the model to generate the SEI configuration of specific properties and reaction product range—this guided generation of SEI help to identify the new SEI configurations and target properties for further optimization. The prVAE architecture allows customization for different applications, such as exploration, interpolation, and optimization. The obtained contin-





**Figure 7.** Latent space exploration along the third principal component: The third principal component also shows a similar trend to the second principal component. The decoding of minimum (left), mean (middle), and maximum (right) values of the third principal component to the corresponding SEI configuration is shown here.

ous latent space can be used as an input optimization model to identify property combinations and SEI configurational space with better battery performance and safety. As further work in this direction, we are extending the model to facilitate the inverse design of SEI configuration and reaction barrier space by using predicted properties from the property regressor as input to the decoder of prVAE.



**Figure 8.** Latent space exploration of the SEI configuration: a) Guided SEI generation from the sampled mean of the latent variables while varying the conditional input barrier; b) SEI generation from the sampled mean of the input barrier while varying the latent variables.

## 5. Tools and Methods

### 5.1. Data Handling

The implemented machine learning workflow in this work is defined and handled with the help of an open-source data platform called Kadi4Mat (Karlsruhe Data Infrastructure for Materials Science).<sup>[41]</sup> Kadi4Mat functions as both a communal repository and Electronic Lab Notebook (ELN).<sup>[42]</sup> Here, the data platform collects and organizes the data and metadata from the source simulation. The ELN of Kadi4Mat provides access to a wide range of tools to handle, preprocess, and analyze data. For our data-driven study, we used KadiAI and CIDS (Computational Intelligence and Data Science tools)<sup>[43]</sup> extension of the Kadi4Mat ecosystem to define and execute machine learning processes.

### 5.2. Model Hyperparameters

For the proposed prVAE architecture, the encoder used 2D convolution layers of filter sizes 32, 64, and 128, respectively, followed by a fully connected dense layer of size 100. The decoder used 2D deconvolutional layers with 128, 64, 32, and 23 filter sizes. Every convolutional layer has a batch normalization layer following it to stabilize the activation values and improve the model performance. After some trials, the latent dimension or the bottleneck of prVAE is set to 50. The stride values for the convolutional layer network are two and one. The activation function used by the

convolutional layers is the rectified linear unit. The property regressor uses two fully connected dense layers of 200 neurons for predicting SEI properties from the VAE bottleneck.

### 5.3. Training Hyperparameters

After the necessary preprocessing, the dataset is split into training, validation, and test samples to generalize the model knowledge. The split ratio is 0.7 for training, 0.15 for validation, and 0.15 for testing. The model's training involved 250 epochs, with a learning rate of  $3e^{-6}$  and a batch size of 64. ADAM optimizer is used to minimize the loss function of the model during training.

### Acknowledgements

The research was made possible by the generous support of several organizations. The authors thank the Deutsche Forschungsgemeinschaft's "Cluster of Excellence" POLIS (project number 390874152), the BMBF's "FestBatt" competence cluster (project number 03XP0174E), the Ministry of Science, Research, and Art Baden-Württemberg (MWK-BW) for their contributions to the MoMaF-Science Data Center project, which received funding from the state digitization strategy digital@bw (project number 57). The authors are also grateful for the financial support of the German Federal Ministry of Education and Research (BMBF) for the AQuaBP project under grant number 03XP0315B. Lastly, The authors would like to acknowledge the Helmholtz Association's support through the program MTET, no: 38.02.01, and KNMFI, no. 43.31.01. The training data and model architecture are stored, handled, and processed within the Kadi4Mat ecosystem. KadiStudio workflows<sup>[42]</sup> and KadiAI's machine learning workflow<sup>[44,45]</sup> achieve reproducibility and track data provenance.

Open access funding enabled and organized by Projekt DEAL.

### Conflict of Interest

The authors declare no conflict of interest.

### Data Availability Statement

The data supporting this study's findings are available from the corresponding author upon reasonable request.

### Keywords

kinetic Monte Carlo simulations, solid electrolyte interphases, variational autoencoder, virtual material design

Received: June 24, 2023  
Revised: August 14, 2023  
Published online:

- [1] C. P. Grey, D. S. Hall, *Nat. Commun.* **2020**, *11*, 1.  
[2] B. Dunn, H. Kamath, J.-M. Tarascon, *Science* **2011**, *334*, 928.  
[3] A. Wang, S. Kadam, H. Li, S. Shi, Y. Qi, *npj Comput. Mater.* **2018**, *4*, 1.

- [4] M. Nie, D. P. Abraham, D. M. Seo, Y. Chen, A. Bose, B. L. Lucht, *J. Phys. Chem. C* **2013**, *117*, 25381.  
[5] I. A. Shkrob, Y. Zhu, T. W. Marin, D. Abraham, *J. Phys. Chem. C* **2013**, *117*, 19255.  
[6] P. Lu, S. J. Harris, *Electrochem. Commun.* **2011**, *13*, 1035.  
[7] P. Lu, C. Li, E. W. Schneider, S. J. Harris, *J. Phys. Chem. C* **2014**, *118*, 896.  
[8] D. Liu, Z. Shadike, R. Lin, K. Qian, H. Li, K. Li, S. Wang, Q. Yu, M. Liu, S. Ganapathy, et al., *Adv. Mater.* **2019**, *31*, 1806620.  
[9] J. Wu, M. Ihsan-Ul-Haq, Y. Chen, J.-K. Kim, *Nano Energy* **2021**, *89*, 106489.  
[10] L. Alzate-Vargas, S. M. Blau, E. W. C. Spotte-Smith, S. Allu, K. A. Persson, J.-L. Fattebert, *J. Phys. Chem. C* **2021**, *125*, 18588.  
[11] S. Bertolini, P. B. Balbuena, *J. Phys. Chem. C* **2018**, *122*, 10783.  
[12] J. Christensen, J. Newman, *J. Electrochem. Soc.* **2004**, *151*, A1977.  
[13] M. Esmaeilpour, S. Jana, H. Li, M. Soleymanibrojeni, W. Wenzel, *Adv. Energy Mater.* **2023**, *13*, 2203966.  
[14] R. Kulagin, P. Reiser, K. Truskovskiy, A. Koeppel, Y. Beygelzimer, Y. Estrin, P. Friederich, P. Gumbsch, *Adv. Eng. Mater.* **2023**, *25*, 2300048.  
[15] Y. Zhao, P. Altschuh, J. Santoki, L. Griem, G. Tosato, M. Selzer, A. Koeppel, B. Nestler, *Acta Mater.* **2023**, *253*, 118922.  
[16] A. Bhowmik, I. E. Castelli, J. M. Garcia-Lastra, P. B. Jørgensen, O. Winther, T. Vegge, *Energy Storage Mater.* **2019**, *21*, 446.  
[17] R. Younesi, G. M. Veith, P. Johansson, K. Edström, T. Vegge, *Energy Environ. Sci.* **2015**, *8*, 1905.  
[18] D. P. Kingma, M. Welling, *arXiv preprint arXiv:1312.6114* **2013**.  
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Commun. ACM* **2020**, *63*, 139.  
[20] R. Singh, V. Shah, B. Pokuri, S. Sarkar, B. Ganapathysubramanian, C. Hegde, *arXiv preprint arXiv:1811.09669* **2018**.  
[21] X. Y. Lee, J. R. Waite, C.-H. Yang, B. S. S. Pokuri, A. Joshi, A. Balu, C. Hegde, B. Ganapathysubramanian, S. Sarkar, *Nat. Comput. Sci.* **2021**, *1*, 229.  
[22] S. Kang, K. Cho, *J. Chem. Inf. Model.* **2018**, *59*, 43.  
[23] G. Li, C. W. Monroe, *Annu. Rev. Chem. Biomol. Eng.* **2020**, *11*, 277.  
[24] A. A. Franco, A. Rucci, D. Brandell, C. Frayret, M. Gaberscek, P. Jankowski, P. Johansson, *Chem. Rev.* **2019**, *119*, 4569.  
[25] L. Maaløe, M. Fraccaro, V. Liévin, O. Winther, *Adv. Neural Inf. Process. Syst.* **2019**, *32*.  
[26] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, A. Courville, *arXiv preprint arXiv:1611.05013* **2016**.  
[27] A. B. Bortz, M. H. Kalos, J. L. Lebowitz, *J. Comput. Phys.* **1975**, *17*, 10.  
[28] K. Ushirogata, K. Sodeyama, Y. Okuno, Y. Tateyama, *J. Am. Chem. Soc.* **2013**, *135*, 11967.  
[29] Y. Wang, S. Nakamura, M. Ue, P. B. Balbuena, *J. Am. Chem. Soc.* **2001**, *123*, 11708.  
[30] K. Miyabe, R. Isogai, *J. Chromatogr. A* **2011**, *1218*, 6639.  
[31] R. L. Iman, *Encyclopedia of quantitative risk analysis and assessment* **2008**, *3*.  
[32] A. Olsson, G. Sandberg, O. Dahlblom, *Structural safety* **2003**, *25*, 47.  
[33] A. Koeppel, F. Bamer, B. Markert, *Acta Mech.* **2019**, *230*, 3279.  
[34] F. Bamer, D. Thaler, M. Stoffel, B. Markert, *Front. built environ.* **2021**, *7*, 679488.  
[35] E. Bisong, E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, Apress, New York, **2019**, P. 347.

- [36] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press, Cambridge, **2016**.
- [37] L. Wang, Y.-C. Chan, F. Ahmed, Z. Liu, P. Zhu, W. Chen, *Comput. Methods Appl. Mech. Eng.* **2020**, *372*, 113377.
- [38] J. Lim, S. Ryu, J. W. Kim, W. Y. Kim, *J. Cheminf.* **2018**, *10*, 1.
- [39] Y. Yang, K. Zheng, C. Wu, Y. Yang, *Sensors* **2019**, *19*, 2528.
- [40] A. Tevosyan, L. Khondkaryan, H. Khachatryan, G. Tadevosyan, L. Apresyan, N. Babayan, H. Stopper, Z. Navoyan, *J. Cheminf.* **2022**, *14*, 1.
- [41] N. Brandt, L. Griem, C. Herrmann, E. Schoof, G. Tosato, Y. Zhao, P. Zschumme, M. Selzer, *Data Sci. J.* **2021**, *20*, 1.
- [42] L. Griem, P. Zschumme, M. Laqua, N. Brandt, E. Schoof, P. Altschuh, M. Selzer, *Data Sci. J.* **2022**, *21*, 1.
- [43] A. Koeppe, the CIDS team, Cids and KadiAI, <https://gitlab.com/intelligent-analysis/cids>, **2023**.
- [44] A. Koeppe, F. Bamer, M. Selzer, B. Nestler, B. Markert, *PAMM* **2021**, *21*, e202100238.
- [45] A. Koeppe, F. Bamer, M. Selzer, B. Nestler, B. Markert, *Front. Mat.* **2022**, *8*, 636.