Virtual screening of inorganic materials synthesis parameters with deep learning

npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Virtual screening of inorganic materials synthesis parameters with deep learning

Edward Kim[1], Kevin Huang[1], Stefanie Jegelka[2] and Elsa Olivetti[1]

Virtual materials screening approaches have proliferated in the past decade, driven by rapid advances in first-principles computational techniques, and machine-learning algorithms. By comparison, computationally driven materials *synthesis* screening is still in its infancy, and is mired by the challenges of data *sparsity* and data *scarcity*: Synthesis routes exist in a sparse, high-dimensional parameter space that is difficult to optimize over directly, and, for some materials of interest, only scarce volumes of literature-reported syntheses are available. In this article, we present a framework for suggesting quantitative synthesis parameters and potential driving factors for synthesis outcomes. We use a variational autoencoder to compress sparse synthesis representations into a lower dimensional space, which is found to improve the performance of machine-learning tasks. To realize this screening framework even in cases where there are few literature data, we devise a novel data augmentation methodology that incorporates literature synthesis data from related materials systems. We apply this variational autoencoder framework to generate potential $SrTiO_3$ synthesis parameter sets, propose driving factors for brookite TiO_2 formation, and identify correlations between alkali-ion intercalation and MnO_2 polymorph selection.

npj Computational Materials (2017)3:53 ; doi:10.1038/s41524-017-0055-6

INTRODUCTION

To accelerate the design and realization of novel materials, a number of recent studies have screened promising candidates across a variety of categories, including light-emitting molecules,[1] perovskite compounds,[2–5] catalysts,[6,7] thermoelectrics,[8–12] and metal-organic frameworks.[13,14] Accordingly, the rise of virtual materials screening, along with high-throughput first-principles computations and experimentation, has resulted in the creation of numerous accessible databases for the materials science community.[15–22] There is, consequently, a pressing need for analogous virtual screening of inorganic materials *syntheses* to complement the growing volume of predicted and screened compounds.[23,24] Such synthesis screening approaches have indeed found recent success in organic chemistry, where a wealth of tabulated reaction data is available,[25–35] and synthesis parameter screening, driven by machine learning, has also been explored for the specific case of organically templated metal vanadium selenites.[20] These efforts have laid the groundwork for analogous large-scale *inorganic* synthesis screening. However, to the best of the authors' knowledge, no comprehensive approaches yet exist for computationally screening materials syntheses parameters across broad categories of inorganic materials systems.

Developing an approach toward virtual synthesis parameter screening introduces two primary computational challenges: data *sparsity* and data *scarcity*. We represent synthesis routes by constructing high-dimensional vectors consisting of synthesis parameters text-mined from the literature, including common solvent concentrations, heating temperatures, processing times, and precursors used.[36] Such canonical representations, however, are necessarily sparse as there are many more actions that one might perform during the synthesis of a material, compared to the number of actions actually used. Compressed, low-dimensional representations are typically more desirable than sparse, high-dimensional feature descriptors as low-dimensional representations are able to emphasize the most relevant dimensions (e.g., combinations of synthesis temperatures used) while also avoiding the so-called "curse of dimensionality."[37,38] Indeed, neural network-based dimensionality reduction has seen success in learning representations of meaningful word vectors,[39] hierarchical image filters,[40] representations of organic chemicals,[1,41] and quantum spin systems.[42]

While neural networks show broad potential for learning compressed data representations, they often consume large amounts of training data to achieve high accuracies,[43,44] and standard training sets often include millions of data points.[45,46] However, literature-reported inorganic materials syntheses are scarce by comparison, especially when considering the syntheses of a specific material system (e.g., $SrTiO_3$). To realize a deep learning approach to materials synthesis screening, it is, therefore, critical that a data augmentation method be used to increase the volume of available training data examples.

In this work, a computational synthesis screening framework is presented in which a variational autoencoder (VAE) neural network is used to learn compressed synthesis representations from sparse descriptors, and a novel data augmentation approach is developed to enable this framework for materials with uncommon syntheses. We perform synthesis screening on $SrTiO_3$ and $BaTiO_3$ syntheses, since these materials systems have only hundreds of text-mining-accessible published syntheses, and thus provide an environment for examining the advantages of data volume augmentation. We also visually explore two-dimensional learned VAE latent vector spaces to investigate potential driving

[1]Department of Materials Science and Engineering, MIT, Cambridge, MA, USA and [2]Department of EECS and CSAIL, MIT, Cambridge, MA, USA
Correspondence: Elsa Olivetti (elsao@mit.edu)

Received: 24 July 2017 Revised: 28 October 2017 Accepted: 2 November 2017
Published online: 01 December 2017

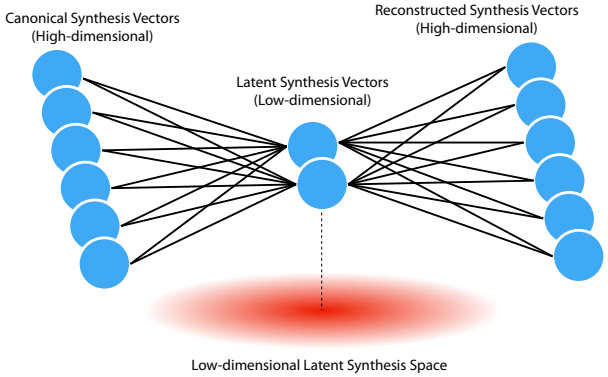

Fig. 1 Schematic set-up for variational autoencoder architecture. An overview of the architecture for the variational autoencoder. The canonical (and reconstructed) vector spaces are sparse, high-dimensional synthesis descriptors. The variational autoencoder minimizes data reconstruction error, while also learning to project the data into latent space points according to a continuous Gaussian distribution (diffuse red area)

factors for brookite TiO_2 formation and to understand ion intercalation effects in MnO_2 phase selection.

RESULTS AND DISCUSSION

To determine the effectiveness of VAE-driven dimensionality reduction from sparse data descriptors, we first describe three different encodings of synthesis parameters. In this study, we compare (1) the unmodified canonical synthesis features, which include descriptors such as heating temperatures or solvent concentrations, (2) canonical features modified by linear dimensionality reduction with principal component analysis (PCA), and (3) canonical features modified by non-linear dimensionality reduction using a VAE. The first of these techniques is an intuitive encoding of synthesis parameters, while the latter two techniques are compressed encodings which have lower dimensionality than the canonical descriptors. These compressed encodings automatically select combinations of the most informative synthesis parameters, and dimensionality reduction has been found to increase predictive performance in materials property prediction by improving the computational efficiency of training machine-learning algorithms for classification or regression.[47,48]

Autoencoders are a class of neural network algorithms that learn to reproduce the identity function, and thus reconstruct the training data, while "squeezing" the data through a low-dimensional inner layer, which acts as a bottleneck. This inner layer with lower dimensionality corresponds to a continuous "latent space," which aims to preserve information from the higher-dimensional input space. More formally, we may think of autoencoders as combinations of encoding and decoding functions f and g, which project data points x into and out of the latent space points x', with the combined goal of approximating the identity function:

$$f(x_i) = x_i' \; g\left(f(x_i)\right) = g\left(x_i'\right) \approx x_i \; x_i \in \mathbb{R}^n, \; x_i' \in \mathbb{R}^m \text{ and } m<n \qquad (1)$$

A *variational* autoencoder adds an additional constraint: the learned representations in the latent space must approximate a prior probability distribution, which improves the generalizability of the model by reducing the possibility of overfitting to the training data.[49] A common convention is to use a Gaussian function as the latent prior distribution, and we follow this convention in our work.[49,50] Beyond improving the performance of the model, this Gaussian prior provides a simple distribution from which to sample new data points, meaning that VAEs are also *generative* models that can produce virtual data. A schematic diagram of the VAE architecture is provided in Fig. 1.

Table 1. Prediction accuracies for determining correct synthesis target between syntheses of $SrTiO_3$ and $BaTiO_3$

Features used for classifier	Threefold cross-validation accuracy (%)	Threefold cross-validation standard deviation (%)
30-D Canonical	74	3
2-D PCA	63	3
10-D PCA	68	6
2-D Latent VAE	63	3
10-D Latent VAE	74	6

The canonical synthesis descriptors, 2-D and 10-D PCA features, and 2-D and 10-D VAE features were all used to train logistic regression classifiers on the task of correctly predicting a synthesis target given the text-mined synthesis parameter descriptors. The canonical features and the 10-D VAE features achieve the same prediction accuracy and are emphasized in boldface font. Additional details and comparisons to naive autoencoders are available in the Supplementary Methods

We compare the three aforementioned representations of synthesis data in the context of materials synthesis screening by using the different feature representations as inputs to a classifier that learns to solve the problem of synthesis target prediction between two closely related materials, $SrTiO_3$ and $BaTiO_3$. A similar task, involving synthesis target prediction from chemical reactions, has recently been used as a benchmark task for synthesis planning in organic chemistry.[27] In our supervised learning problem, a classifier is given synthesis descriptor vectors and must learn to differentiate between syntheses of $SrTiO_3$ and $BaTiO_3$, which are both perovskite-type materials exhibiting a variety of electronic properties with ferroelectric behavior as one example.[51–53] Traditionally, these materials are synthesized with the involvement of high-temperature heating steps to drive the formation of the final ternary compound from binary precursors.[54,55] As shown in Table 1, we find that a logistic regression classifier achieves an accuracy of 74% when differentiating between $SrTiO_3$ and $BaTiO_3$ syntheses using canonical feature input vectors. This suggests that distinguishing between $SrTiO_3$ and $BaTiO_3$ syntheses is neither impossible nor trivially easy, as such cases would tend to yield accuracies of 50% and 100%, respectively. Furthermore, this prediction accuracy is comparable to recent work in synthesis target prediction, where a machine-learned one-shot prediction accuracy of 72% is achieved for organic reaction outcomes.[27] As a baseline for desirable performance, human-intuition strategies achieve 78% accuracy when applied to the problem of predicting successful or failed reactions,[20] and this is again comparable to the accuracy achieved by our model.

As a representative method for linear dimensionality reduction, PCA is applied to this data set to explore the trade-off between data compression and prediction accuracy in the target prediction task. Two-dimensional PCA vectors, along with ten-dimensional PCA vectors, are able to capture approximately 33% and 75% of the variance in the data, respectively. Nonetheless, neither the prediction accuracies of the 2-D reduced features (accuracy = 63%) nor the 10-D PCA-reduced features (accuracy = 68%) match the prediction accuracy of the original canonical features (accuracy = 74%), as outlined in Table 1. This implies that the data compressed via PCA has lost information critical to predicting the target synthesized material associated with each set of synthesis parameters, and additionally provides us with a baseline performance against which to compare the non-linear VAE method for feature representation learning. Moreover, this suggests that information loss in compressed representations

can increase the difficulty of mapping between synthesis parameters and synthesized target materials.

To enable a VAE approach for learning compressed representations that reduce data *sparsity*, we first introduce a novel data augmentation algorithm to alleviate the problem of data *scarcity*. The total data set for $SrTiO_3$ syntheses is comprised of less than 200 total text-mined synthesis descriptors, and attempting to train a VAE on such a small data set is not likely to produce optimal results. To enable accurate training of a VAE, we devise a training data augmentation scheme based on ion-substitution material similarity functions (see Methods section). In brief, we apply context-based word similarity algorithms,[39] ion-substitution compositional similarity algorithms,[56] and cosine similarity between the canonical synthesis descriptor vectors to create an augmented data set, comprised of a neighborhood of similar materials, with an order of magnitude more data (1200 + text-mined synthesis descriptors). This augmented data set contains synthesis parameters drawn from a neighborhood of materials syntheses centered on the material of interest ($SrTiO_3$), and we train the VAE to learn feature representations using this larger data set with greater weighting placed on the most closely related syntheses. A schematic outline of this process is shown in Fig. 2a.

This data augmentation technique allows us to significantly boost data volume without resorting to artificial noise or interpolated data points. Moreover, we incorporate relevant domain knowledge via data-mined ion-substitution probabilities to ensure that the augmented data is pertinent to the original material system of interest. By training the VAE separately on both the non-augmented data set and the augmented data set, we find that the VAE attains reduced error in reconstructing the data when using the augmented data set, as shown in Fig. 2b. This weighted-error approach is easily incorporated into the iterative training of neural networks, whereas PCA cannot incorporate error-weighting parameters in a straightforward manner.

The performance of VAE features for differentiating $SrTiO_3$ and $BaTiO_3$ syntheses is reported in Table 1. The 10-D VAE features, which are compressed 67% compared to the canonical features, recover the prediction accuracy (74%) of the original features and appear to outperform PCA features at the same level of data compression. However, the authors do note that the standard deviations of these accuracies are fairly high (as reported in Table 1), and a more rigorous understanding of the general predictive capabilities of VAE-learned features will be explored in future work. Beyond data compression to a lower-dimensional continuous vector space, an additional advantage of the VAE is its nature as a generative model, which allows us to jointly produce entire sets of synthesis parameters (e.g., the entire set of reaction temperatures/times and solvent concentrations for a synthesis attempt). These generated virtual synthesis parameters represent plausible suggestions of synthesis conditions for planning novel syntheses.

We develop an additional machine-learning model to validate the quality of the learned $SrTiO_3$ VAE latent space. This model is a logistic regression binary classifier, which is trained to differentiate between virtual synthesis descriptor data, created by sampling from the Gaussian prior, and real data text-mined from the literature. Through this set-up, which is motivated by recent adversarial machine-learning techniques,[57] we verify the VAE's ability to learn accurate latent representations of $SrTiO_3$ synthesis parameters. Figure 3a shows a schematic of this model and its relation to the VAE. First, virtual data samples are drawn from the latent distribution and decoded using the VAE. Then, the decoded data are classified as real (i.e., text-mined from literature) or virtual by the binary classifier. Thus, data produced by the VAE which "tricks" the binary classifier into misclassifying it as text-mined data is, to an extent, indistinguishable from genuine literature-reported synthesis parameters.

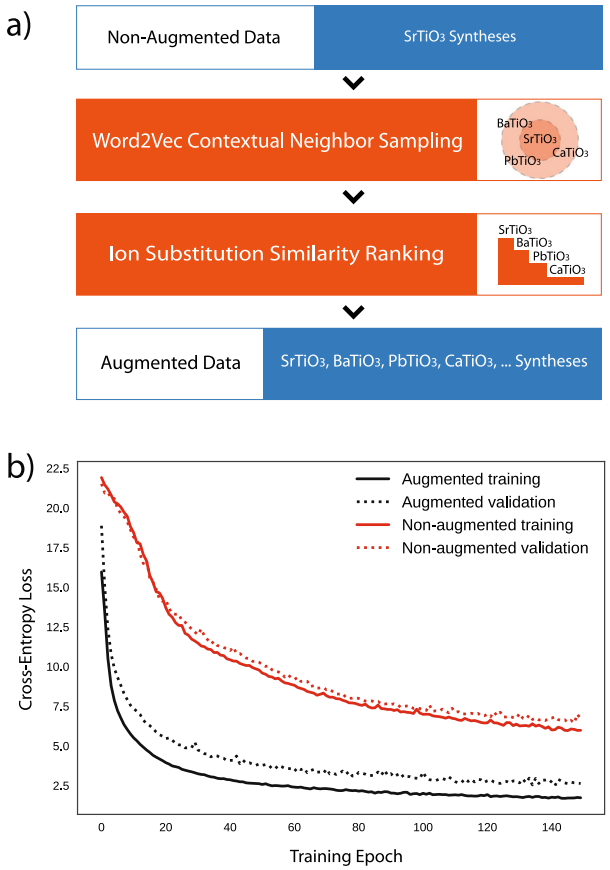

Fig. 2 Data augmentation for enabling deep learned synthesis parameter representations. **a** Schematic diagram outlining the process for data augmentation. A primary material of interest, $SrTiO_3$, is first chosen as the non-augmented data set. Then, the Word2Vec algorithm is used to find materials that appear in similar contexts across journal articles.[39] This list is then ranked by ion-substitution similarity scores with respect to $SrTiO_3$,[56] and selection by this ranking produces the final augmented data set. **b** The training and validation cross-entropy losses for the variational autoencoder are plotted against the number of training epochs for both the non-augmented and the augmented data sets. The cross-entropy loss is a standard classification training error function used in neural networks[27,75]

The virtual data is assessed by repeated trials of sampling new data from the Gaussian prior and recording the number of sample attempts needed until the classifier erroneously categorizes a virtual sample as a real sample. Across 50 total trials, we count the number of latent samples drawn in each trial until the classifier makes an error. Then, the probability of having at least one sample, which tricks the classifier is computed from these recorded counts and the cumulative probability distribution is displayed in Fig. 3b. Indeed, only five sampling attempts are needed to exceed a 95% chance of having produced at least one virtual data sample, which is sufficiently realistic to trick the classifier.

To provide examples of specific text-mined and virtual synthesis parameters for $SrTiO_3$ synthesis, Table 2 shows both text-mined and virtual data, demonstrating that the VAE is capable of jointly generating multivariable sets of realistic synthesis parameters. In each of the literature examples, only a subset of possible processing steps is used, as one would expect (e.g., calcination but not sintering). The virtual data from the VAE successfully mimics this aspect of the data, predicting that in any single synthesis only some subset of synthesis parameters should be used. Beyond this, the generated values for the synthesis

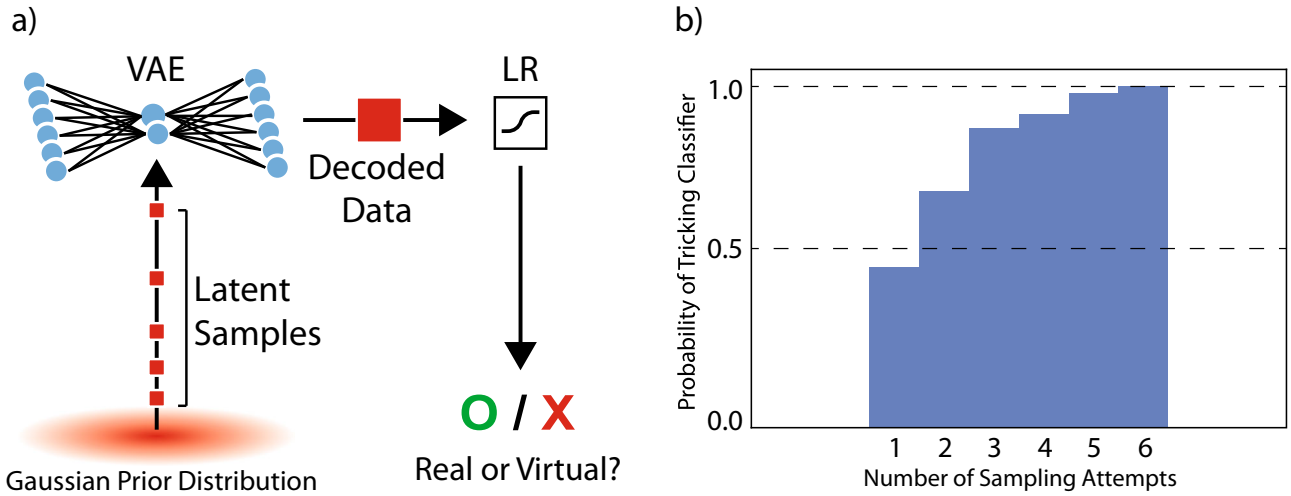

Fig. 3 Set-up and results for realistic synthesis data generation. **a** To assess the quality of virtual data produced by the variational autoencoder (VAE), random samples are drawn from its latent Gaussian prior distribution, and decoded into higher-dimensional canonical synthesis descriptors. The decoded data is passed to a binary logistic regression classifier (LR), which has been trained to differentiate between text-mined ("real") and VAE-generated ("virtual") synthesis descriptors. Latent samples are fed through this process until the trained classifier is "tricked" by erroneously classifying a virtual sample as a real sample, which signifies that the virtual sample is indistinguishable from a real one. **b** In 50 trials of the data quality assessment procedure, the number of sampling attempts needed to trick the classifier is recorded for each trial. From this data, the cumulative probability of tricking the classifier as a function of sampling attempts is computed. Dashed lines are drawn at 50 and 100% probabilities

parameters are comparable in magnitude to literature-reported values, without being trivially identical across all examples.

By using a generative model, such as a VAE, we screen entire sets of plausible and novel synthesis parameters, based upon those already reported in the literature. This technique can thus provide guidance for experimental synthesis planning or provide insight into driving factors for previously reported synthesis outcomes. This builds upon recently reported methodologies for sampling new synthesis parameters, where a discriminative model (i.e., classifier) is used to rank proposed synthesis routes where a single parameter is varied.[20] In particular, the model we have presented here allows for multivariate sampling in which each sample has a high probability of containing a realistic set of multiple synthesis parameters.

Having examined the ability of the VAE to compress data in the context of retaining predictive accuracy for synthesis target classification between $SrTiO_3$ and $BaTiO_3$, we now explore two additional materials of interest, TiO_2 and MnO_2. For each of these materials, we learn two-dimensional latent spaces with a VAE to maximize visual interpretability of the encoded synthesis parameters. We also produce augmented data sets for these materials using the same weighted neighborhood technique as previously described, and incorporate these larger data sets when training the VAE for TiO_2 and MnO_2 syntheses.

We first examine TiO_2, as an example of a frequently synthesized material with applications ranging from photocatalysis to lithium-ion battery electrodes.[58,59] However, although there are myriad reported syntheses of anatase and rutile-phase TiO_2, we focus here on uncovering synthesis patterns, which lead to the rarely reported brookite phase.[60] In Fig. 4a, latent synthesis vectors corresponding to text-mined syntheses are plotted for TiO_2, with darkened points denoting syntheses, which report the brookite phase. In this two-dimensional latent space, the VAE learns three broad clusters. Each are approximately outlined by dashed ovals, and these clusters are primarily separated by syntheses involving methanol, ethanol, and citric acid, which almost exclusively appear in each of their, respectively, labeled clusters (e.g., > 95% of ethanol-using syntheses appear in the ethanol-labeled cluster).

Brookite TiO_2 is poorly understood compared to anatase or rutile, and multiple synthesis parameters have been found to select for brookite formation, including pH,[61,62] particle size,[63] and phase-stabilizing anions.[64] From this latent 2-D space of text-mined articles, we, therefore, explore the various synthesis parameters, which lead to the formation of brookite by highlighting exemplar regions containing text-mined brookite syntheses, denoted as regions A and B, which contain 1120 and 312 total text-mined syntheses, respectively. These regions are chosen as two examples of latent space corresponding to brookite syntheses in areas of high data point density (region A) and low data point density (region B). Additionally, we note that the varied distribution of brookite-reporting syntheses throughout the latent space corresponds with the aforementioned knowledge that many different synthesis techniques have been utilized to selectively form the brookite phase. Consequently, the VAE thus provides a method for visualizing and exploring multiple valid paths for achieving a synthesized product.

In both of the highlighted regions A and B in Fig. 4a, the driving effect of pH is clear when examining the underlying data points: NaOH is commonly used to raise pH during brookite syntheses,[60,63] and is used by over 75% of the syntheses in both regions A and B. However, in region A, ethanol appears in 100% of the synthesis routes used (as one might expect by its location in the larger ethanol-dominated cluster), while in region B, no syntheses report the usage of ethanol. There is some existing discussion in the literature that alcoholysis may be another factor capable of selecting for brookite phases, but very few articles have considered this effect in detail,[60,65] and—to the best of the authors' knowledge—the specific influence of using ethanol as a solvent for brookite phase selection does not appear to be present in the literature. This difference in ethanol usage between regions A and B highlights the ability of the VAE-learned latent space in identifying diverse sets of synthesis parameters, which each yield valid pathways toward synthesizing a desired phase: the use of ethanol in the high data-point density regions (A) suggests dissolution in ethanol may often be a *sufficient* driving factor for phase selection, yet the existence of brookite-producing syntheses in other low density regions (B) suggests that this is not a *necessary* condition.

Table 2. Examples of literature-reported synthesis parameters and VAE-generated synthesis parameters for $SrTiO_3$ synthesis

Calcination conditions (°C, H)	Sintering conditions (°C, H)	Annealing conditions (°C, H)	NaOH concentration (M)	Synthesis type	Reference
800, 2	–	–	1.0	Hydrothermal	Ye et al., 2014[52]
800, 2	1250, 2	–	–	Solid State	Zhao et al., 2004[76]
1000, 12	–	500, 2	–	Hydrothermal	Zhao et al., 2015[77]
600–750, 4	–	–	–	Sol–gel	Puangpetch et al., 2008[54]
721, 1.8	–	**468, 0.4**	–	–	**N/A**
-	–	**450, 0.9**	**1.0**	–	**N/A**
955, 6.0	**1182, 7.5**	–	–	–	**N/A**

Four rows of literature data are followed by three rows of virtual generated data, selected from points which successfully "tricked" the classifier in Fig. 3. Virtual generated data rows are emphasized in boldface font

Our approach of using visualizable and explorable latent synthesis spaces allows us to generate reasonable synthesis hypotheses supported by the literature. In particular, the commonly used step of dissolving precursors in ethanol is selected as a strong major clustering feature (corresponding the central oval-marked cluster in Fig. 4a), and additionally indicates a potential driving factor for brookite phase selection. Since ethanol is a ubiquitous solvent, considering its usage as a phase-selecting solvent may not be an obvious hypothesis in the absence of the clustering learned by the VAE, and future work could test this hypothesis experimentally.

To further examine the VAE-learned latent spaces, we apply the VAE to the problem of phase selection in MnO_2, a system comprised of several polymorphs with notable applications in energy storage and catalysis.[66,67] As recently computed by Kitchaev et al.,[66] MnO_2 phase selection is strongly controlled by free energy differences resulting from alkali-ion intercalation: magnesium and lithium ions select the spinel (λ) phase across a wide range of ion concentrations, potassium selects strongly for hollandite (α), while sodium and potassium both select for birnessite (δ) at higher intercalated ion concentrations. Additionally, the ramsdellite phase (R) is not thermodynamically favored upon intercalating any of the aforementioned alkali ions.

In Fig. 4b, a latent phase diagram is computed using the VAE-learned representations of MnO_2 syntheses. Figure 4b shows the 2-D latent space, which has been divided into a uniformly distributed 250×250 grid of two-dimensional points (although grid lines are not shown for clarity). Each of these points is fed into the VAE decoder to generate a synthesis vector corresponding to the sampled point in latent space, analogous to the data generation set-up used in Fig. 3. Then, the phase regions and boundaries are generated by computing the maximally probable phase corresponding to each grid point (black curves in 4b).

Although the VAE has no explicit learning objectives other than reproducing training data and approximating a multivariate Gaussian distribution, Fig. 4b shows that the latent space captures the concepts of phase and synthesis-involved ions into broad, consistent regions. There are distinct boundaries separating where one polymorph dominates from another, and similarly distinct boundaries for separating ions. Thus, the VAE is capable of autonomously learning continuous two-dimensional neighborhoods of latent space where nearby syntheses produce similar phases and use chemically similar precursors or solvents. This result is somewhat serendipitous, since the VAE is not given an explicit objective related to clustering the data according to any particular variable.

Beyond this implicitly learned consistency, we find that the latent space corresponds well to the calculations performed by Kitchaev et al.[66] regarding intercalation-based phase stability. In Fig. 4b, the polymorph regions are overlaid with regions corresponding to syntheses, which are most probable to use particular alkali-ion bearing materials during synthesis (e.g., a precursor or dissolved salt). The spinel (λ) region entirely encompasses the magnesium and lithium regions, while the hollandite (α) phase lies entirely within the potassium region and the birnessite phase (δ) jointly spans the sodium and potassium regions. All these correspondences in the latent phase space are in good agreement with the first-principles-computed phase selection trends discussed earlier. Additionally, ramsdellite (R) encompasses only a small fraction of latent space, which again aligns with the previously computed result that this phase is not favored by any of the considered alkali-ion intercalations.[66]

In this work, we have presented an approach for synthesis screening, which combines deep learning and data augmentation techniques to address computational challenges around data *sparsity* and data *scarcity*, respectively. We find that this synthesis screening technique enables the generation of suggested synthesis parameters, accelerates positing of driving factors in forming rare phases, and identifies correlations across material polymorphs and intercalated ions. While this work has focused on the examples of $SrTiO_3$, TiO_2, and MnO_2, due to their technological relevance in applications ranging from energy storage to catalysis, these systems are intended primarily as representative cases to illustrate the scientific validity of this screening approach. We also show that using data-mined similarity functions for training data augmentation allows for tractable deep learning-based dimensionality reduction, even for specific materials with very few examples reported in the literature.

As part of the utility of this VAE approach lies in visualization, the general applicability is difficult to quantify rigorously; however, the authors believe that this VAE method should apply well to other inorganic materials, which are commonly made by solid-state, hydrothermal, and sol–gel synthesis routes, and thus should have similar canonical feature descriptors (e.g., calcination conditions). Moreover, we expect that the data augmentation methodology presented in this work will apply for many materials cataloged in materials databases, as the ion-substitution similarity function is constructed from querying such a data set.[16,17] As an approximate guideline for evaluating, which materials have a suitable neighborhood of similar materials for data augmentation, the ion-similarity matrix presented by Yang and Ceder could be used.[56] The authors do note that this data augmentation scheme will likely not extend to cases where the underlying first-principles assumption—that ion-substitution similarity is a relevant metric for considering similar syntheses—is false. Such cases may include highly amorphous materials, where crystal structure-based similarity may not prove very useful, or materials produced from recycled/waste material, where bulk chemical compositions are

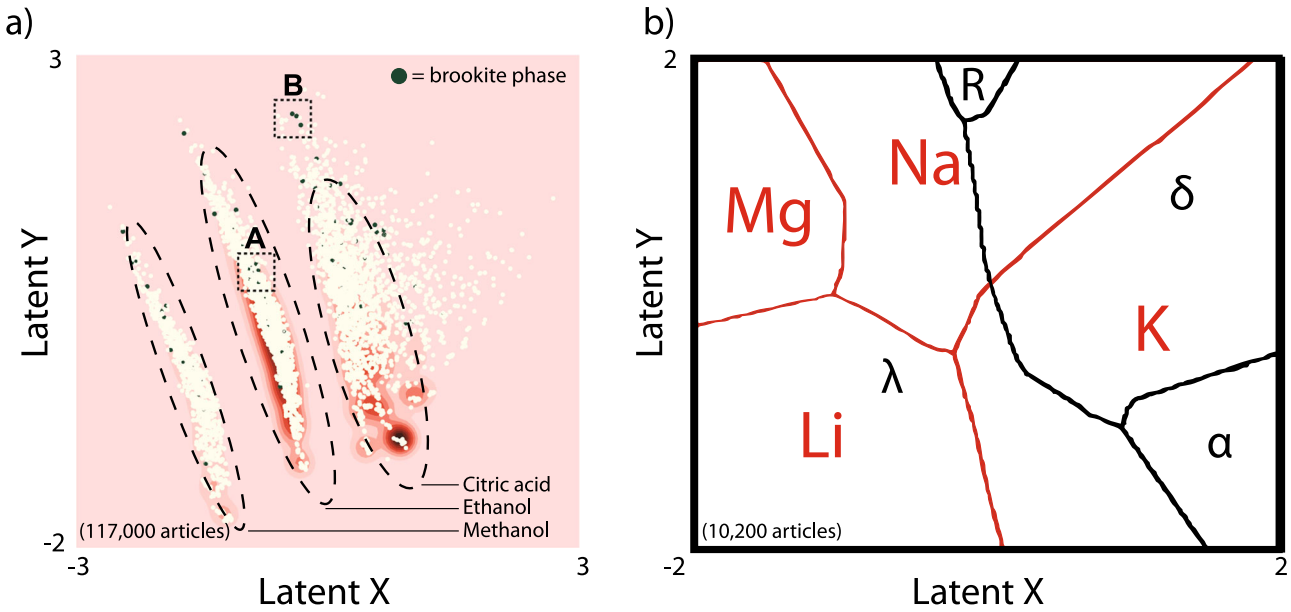

Fig. 4 Latent space for TiO_2 synthesis vectors and MnO_2 synthesis vectors. **a** Latent space for TiO_2, with each data point corresponding to the latent coordinates of a text-mined synthesis descriptor set. Darkened points contain reports of synthesized brookite phase titania. A kernel density estimate for data points is overlaid in the background by the red density map, with darker red indicating higher point density. The high- density regions of the primary clusters are highlighted with overlaid dashed ovals (which only approximate the exact shape of their, respectively, labeled clusters), and are labeled by synthesis parameters, which dominate the clustering behavior. Two example regions containing reports of brookite synthesis are highlighted with dashed squares, and are denoted as regions A and B. **b** Latent synthesis phase space for MnO_2 with phase boundaries of probabilistically dominant polymorph regions (black lines). Each region, labeled by an MnO_2 polymorph symbol, represents a continuous area where a single polymorph is most likely to be produced, as predicted by VAE decodings from latent space. Phase boundaries for probabilistically dominant ions used during synthesis are overlaid in the latent space (red lines)

poorly defined and so compositional similarity cannot easily be computed from underlying ion similarities.

While the VAE models presented in this study provide generative and exploratory capabilities for synthesis parameter screening, direct optimization of synthesis route parameters (e.g., to achieve a particular morphology or phase) is not addressed, and future work would benefit from the inclusion of techniques such as Bayesian optimization for this purpose, which may be performed in our learned latent space.[68,69] Overall, the authors believe that this VAE-based technique provides a first step towards virtual screening of inorganic materials syntheses.

METHODS

Text-mined synthesis data

A corpus of scientific literature is first text-mined using the methods described by Kim et al.[36] to produce a data set of machine-readable synthesis routes. In brief, experimental synthesis sections of journal articles are automatically identified and parsed for relevant synthesis keywords, including temperatures, synthesizing actions (e.g., heating), and material names. These keywords are then assembled into a database object, which can be queried programmatically in further data processing steps.

The text-mined synthesis routes are then converted to feature vectors by a user-specified list of features to consider per material system. Each material system has syntheses described by a list of features, including times and temperatures of common operations (e.g., "calcination at 800 degrees Celsius for 3 h"), indicator functions for these actions (e.g. "did not use annealing"), and indicators for common solvents and precursors. A full list of the features used in this study is provided in the Supplemental Information. These one-dimensional arrays of flattened synthesis parameters then serve as the canonical space for synthesis vectors.

Data augmentation with similar materials

Some materials systems are described by only a modest volume of published synthesis literature, and so to realize a robust methodology for synthesis exploration and screening, we make use of similarity metrics to boost the training data volume. That is, we consider for each material system a training set $\mathcal{X}$, composed of two data types, the non-augmented data and the augmented data. Each data point, $\boldsymbol{x}_i \in \mathcal{X}$, represents a set of synthesis parameters corresponding to a single synthesis route, along with an associated similarity value that measures how relevant the data point is to the original material system of interest. The non-augmented and augmented data refer to the original material of interest (e.g., $SrTiO_3$) and related materials (e.g., $BaTiO_3$, $PbTiO_3$), respectively:

$$\mathcal{X} = \mathcal{X}_{\text{non-aug}} \cup \mathcal{X}_{\text{aug}} \tag{2}$$

For each element of this data set, we wish to compute a similarity value that captures the relevance of the augmented data to the original, non-augmented data. Each data point $\boldsymbol{x}_i \in \mathcal{X}$ is composed of a real-valued similarity vector S_i, ranging from zero to one, along with a real-valued feature descriptor vector ϕ_i. A similarity value of one is only achieved for data points belonging to the non-augmented data set:

$$\forall \boldsymbol{x}_i \in \mathcal{X}, \boldsymbol{x}_i = \{S_i, \phi_i\}, \quad S_i \in [0,1], \phi_i \in \mathbb{R}^n, \quad S_i = 1 \text{ iff } \boldsymbol{x}_i \in \mathcal{X}_{\text{non-aug}} \tag{3}$$

Each similarity value S_i, is the product of two similarity measures: a material-based similarity S_i^m, and synthesis parameter-based similarity S_i^p:

$$S_i = S_i^m \times S_i^p, \; S_i^m \in [0,1], \; S_i^p \in [0,1] \tag{4}$$

The material-based similarity is measured between compositions of two material systems, such as $SrTiO_3$ and $BaTiO_3$, and denoted by $S_i^m(c_1, c_2)$. Each material system contains multiple literature-reported syntheses, and each instance of a literature-reported synthesis is denoted by a set of parameters ϕ_i. The synthesis parameter-based similarity is then measured at this more granular level, between individual sets of synthesis parameters reported into two different journal articles, and is denoted by $S_i^p(\phi_1, \phi_2)$.

To compute S_i^m, we first use the word2vec[70] algorithm to select the nearest neighbors of a material system of interest in a word-embedding vector space (e.g., representing "$SrTiO_3$" as a single word vector). Then we rank these neighbor compositions (e.g., "$PbTiO_3$") by their ion-substitution-based composition similarity[56] with respect to the original composition of interest, which generates values in the range [0,1]. More specifically, S_i^m is computed directly from composition similarity, where each composition c is a count vector of elements in a chemical formula unit:

$$S_i^m = \text{Sim}_{\text{ion}}\left(c_{\text{non-aug}}, c_{\text{aug}}\right) \tag{5}$$

As an example, $\mathrm{Sim}_{\mathrm{ion}}\left(c_{\mathrm{non-aug}}, c_{\mathrm{aug}}\right) = \mathrm{Sim}_{\mathrm{ion}}(\mathrm{SrTiO_3}, \mathrm{BaTiO_3}) \approx 0.892$ in the case of computing material-based similarities for (augmented) $BaTiO_3$ synthesis parameters as compared to (non-augmented) $SrTiO_3$ synthesis parameters. For the original material system of interest, which corresponds to the non-augmented data set, the value of $S_i^m = \mathrm{Sim}_{\mathrm{ion}}\left(c_{\mathrm{non-aug}}, c_{\mathrm{non-aug}}\right) = 1$. When building the augmented data neighborhood, we use a cutoff minimum compositional similarity of 0.80 for ternary compounds and 0.50 for binary compounds. Higher cutoffs are used for ternaries, compared to binaries, since a single ion-substitution in a ternary yields a higher overall relative compositional similarity (since two other ions are unchanged, vs. in binaries, where a single ion-substitution leaves only one other ion unchanged).

Following this, additional text-mined syntheses are sampled from our database corresponding to syntheses of these related materials, which have been selected by word2vec and ranked by ion-based similarity. Thus, for each material system, numerous corresponding journal articles are text-mined to produce synthesis descriptor vectors ϕ_i. The synthesis parameter similarity for an augmented data point S_i^p is computed by considering the mean cosine similarity between the augmented data point ϕ_i^{aug} and the five nearest neighboring non-augmented data points $\phi_j^{\mathrm{non-aug}}$, and this generates a value ranging from zero to one since all ϕ contain only positive values:

$$S_i^p = \frac{1}{N}\sum_{j=1}^{N} \cos\left(\phi_i^{\mathrm{aug}}, \phi_j^{\mathrm{non-aug}}\right),\ N = 5 \tag{6}$$

The number N of non-augmented syntheses over which we average similarities may be treated as a hyperparameter, and the authors found that $N = 5$ performed well for the materials discussed in this work. In principle, there is a trade-off between the risk of including too many outliers with very low values of N (since the augmented data points may cluster around a single outlier in the non-augmented data set), and never achieving any high similarity values with very large N (since any single augmented data point is unlikely to be similar to the entire distribution of non-augmented data points).

For data points $\phi_i^{\mathrm{non-aug}}$, which belong to the non-augmented data set, S_i^p is fixed to a value of one. It then follows that, as stated previously, $S_i = 1$ iff $\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{non-aug}}$ since both S_i^m and S_i^p each attain their maximal values only for non-augmented data points.

Finally, the similarities S_i, for each training data point $\boldsymbol{x}_i$, are incorporated into the training of the VAE by weighting each training sample ϕ_i by the similarity value in the computation of the overall training loss function.

The cross-entropy loss function, denoted by $l(\phi_i, \phi_i')$, measures how accurately the VAE can reconstruct the original data $\mathcal{X}$ by comparing the original and reconstructed synthesis descriptors.[27] The training of the VAE aims to find, via stochastic gradient descent, the neural network weights θ that minimize the weighted loss function $L(\theta, \mathcal{X})$, computed over n training data points:

$$\theta^* = \arg\min_{\theta} L(\theta, \mathcal{X}),\ L(\theta, \mathcal{X}) \propto \frac{1}{n}\sum_{i=1}^{n} S_i \times l(\phi_i, \phi_i') \tag{7}$$

Thus, training data with zero similarity do not contribute to representation learning at all, and training data with similarity one (i.e., belonging to the original queried data set) contribute maximally to representation learning.

The computed material system similarities for neighboring materials centered around $SrTiO_3$, TiO_2, and MnO_2 are presented in Supplementary Table S1.

Variational autoencoder

The VAE consists of input/output layers that match the dimensionality of the canonical data, along with an inner latent layer fixed to either two or ten dimensions to match the dimensionalities of the PCA models. All layers of the autoencoder are densely connected feed-forward layers, with the exception of the inner probabilistic layer, which samples from a multivariate Gaussian distribution. Validation and hyperparameter selection was performed by grid searches, where the latent layer dimension was varied from 2 through 30 dimensions and the standard deviation of the Gaussian prior was varied between 0.001 and 10.0. Hyperparameters were selected in each case based on minimizing validation loss while training the VAE, where the validation set was constructed by randomly selecting 10% of the training data to be held-out. Optimal latent dimensionality was found to be approximately ten dimensions, and optimal standard deviation for the Gaussian prior was found to be between 0.1 to 1.0 (i.e., the performance did not change appreciably between these values).

In the MnO_2 latent space in Fig. 4b, the entire latent region is divided into a 250×250 grid, and each grid point is sampled and inputted into the VAE decoder. The regions represent continuous sections of grid points with a consistent maximally probably decoded variable (e.g., α-MnO_2 being more probable than any other phase). The boundary lines represent transitions between regions where a different phase (or ion, or precursor) becomes the probabilistically favored one, as determined by the VAE.

Data availability

The code used to download journal articles for large-scale text-mining is available at [www.github.com/olivettigroup/article-downloader]. The trained word-embedding matrix, used for both text mining and materials similarity calculations, is available at [www.github.com/olivettigroup/materials-word-embeddings]. The VAE was written using Keras[71] and Tensorflow.[72] Chemical formula parsing was performed using pymatgen.[73] Logistic regression classifiers and PCA models were implemented using scikit-learn.[74] Any reasonable requests for additional data can be directed to the corresponding author.

ACKNOWLEDGEMENTS

We would like to acknowledge funding from the National Science Foundation Award #1534340, DMREF that provided support to make this work possible, support from the Office of Naval Research (ONR) under Contract No. N00014-16-1-2432, the MIT Energy Initiative, and NSF CAREER #1553284. Early work was collaborative under the Department of Energy's Basic Energy Science Program through the Materials Project under Grant No. EDCBEE. E. Kim was partially supported by NSERC. We would also like to acknowledge the tireless efforts of Ellen Finnie in the MIT libraries, support from publishers who provided the substantial content required for our analysis, and research input from Gerbrand Ceder, Daniil Kitchaev, Olga Kononova, and Matthew Staib. We thank Lusann Yang for providing useful Python scripts.

AUTHOR CONTRIBUTIONS

E.K. wrote the machine-learning algorithms and produced the figures. E.K., K.H., S.J., and E.O. discussed the machine-learning results and materials case studies. All authors wrote and commented on the manuscript and figures.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-017-0055-6).

Competing interests: The authors declare no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
2. Pilania, G., Balachandran, P. V., Gubernatis, J. E. & Lookman, T. Classification of *AB* O_3 perovskite solids: a machine learning study. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **71**, 507–513 (2015).
3. Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding New perovskite halides via machine learning. *Front. Mater.* **3**, 1–7 (2016).
4. Balachandran, P. V., Broderick, S. R. & Rajan, K. Identifying the 'inorganic gene' for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **467**, 2271–2290 (2011).
5. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
6. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. B. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **5**, 909–913 (2006).
7. Hong, W. T., Welsch, R. E. & Shao-Horn, Y. Descriptors of oxygen-evolution activity for oxides: A statistical evaluation. *J. Phys. Chem. C* **120**, 78–86 (2016).
8. Gaultois, M. W. et al. Data-driven review of thermoelectric materials: performance and resource considerations BT - chemistry of materials. *Chem. Mater.* **25**, 2911–2920 (2013).

9. Sparks, T. D., Gaultois, M. W., Oliynyk, A., Brgoch, J. & Meredig, B. Data mining our way to the next generation of thermoelectrics. *Scr. Mater.* **111**, 10–15 (2016).
10. Yan, J. et al. Material descriptors for predicting thermoelectric performance. *Energy Environ. Sci.* **8**, 983–994 (2015).
11. Seshadri, R. & Sparks, T. D. Perspective: Interactive material property databases through aggregation of literature data. *APL Mater.* **4**, 053206 (2016).
12. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
13. Wilmer, C. E. et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2011).
14. Lin, L. -C. et al. In silico screening of carbon-capture materials. *Nat. Mater.* **11**, 633–641 (2012).
15. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access. *JOM* **68**, 2031–2034 (2016).
16. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 1–11 (2013).
17. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *Nat. Publ. Gr.* **1**, 15010 (2015).
18. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
19. Hachmann, J. et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry—the Harvard Clean Energy Project. *Energy Environ. Sci.* **7**, 698 (2014).
20. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
21. Isayev, O. et al. Materials cartography: Representing and mining material space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2014).
22. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 16208 (2016).
23. Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *Npj Comput. Mater.* **1**, 15008 (2015).
24. Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big–deep–smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
25. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
26. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 (2009).
27. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
28. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **3**, 17 (2011).
29. Goodman, J. Computer software review: Reaxys. *J. Chem. Inf. Model.* **49**, 2897–2898 (2009).
30. Rocktäschel, T., Weidlich, M. & Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633–1640 (2012).
31. Guha, R. et al. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.* **46**, 991–998 (2006).
32. Murray-Rust, P. & Rzepa, H. S. Chemical markup, XML, and the world wide web. 4. CML schema. *J. Chem. Inf. Comput. Sci.* **43**, 757–772 (2003).
33. Pence, H. E. & Williams, A. Chemspider: An online chemical information resource. *J. Chem. Educ.* **87**, 1123–1124 (2010).
34. Kim, S. et al. PubChem substance and compound databases. *Nucl. Acids Res.* **44**, D1202–D1213 (2015).
35. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: March of the machines. *Angew. Chem. Int. Ed.* **54**, 3449–3464 (2015).
36. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4,** (2017).
37. Roweis, S. T. & Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
38. Kusne, A. G., Keller, D., Anderson, A., Zaban, A. & Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **26**, 444002 (2015).
39. Mikolov, T., Corrado, G., Chen, K. & Dean, J. Efficient estimation of word representations in vector space. *Proc. Int. Conf. Learn. Represent*. (2013).
40. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
41. Wu, Z. et al. MoleculeNet: A benchmark for molecularmachine learning. *ArXiv*. Preprint at https://arxiv.org/abs/1703.00564 (2017).
42. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
43. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. https://arxiv.org/abs/1704.01212 (2017).
44. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low Data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
45. Deng J. et al. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conf. Comput. Vis. Pattern Recognit*. 248–255 (2009).
46. Torralba, A., Fergus, R. & Freeman, W. T. 80 Millions tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Trans. Pattern. Anal. Mach. Intell.* **30**, 1958–1970 (2008).
47. Suh, C., Rajagopalan, A., Li, X. & Rajan, K. The application of principal component analysis to materials science data. *Data Sci. J.* **1**, 19–26 (2002).
48. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
49. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations*. https://arxiv.org/abs/1312.6114 (2013).
50. Gómez-Bombarelli, R., Hirzel, T. D., Duvenaud, D., Aguilera-Iparraguirre, J. & Adams, R. P. Automatic chemical design using variational autoencoders. *ArXiv*. Preprint at arxiv.org/abs/1610.02415 (2017).
51. Urban, J. J., Yun, W. S., Gu, Q. & Park, H. Synthesis of single-crystalline barium titanate and strontium titanate. *J. Am. Chem. Soc.* **124**, 1186–1187 (2002).
52. Ye, M. et al. Garden-like perovskite superstructures with enhanced photocatalytic activity. *Nanoscale* **6**, 3576 (2014).
53. Zhang, Q., Cagin, T. & Goddard, Wa The ferroelectric and cubic phases in BaTiO3 ferroelectrics are also antiferroelectric. *Proc. Natl Acad. Sci. U.S.A.* **103**, 14695–14700 (2006).
54. Puangpetch, T., Sreethawong, T., Yoshikawa, S. & Chavadej, S. Synthesis and photocatalytic activity in methyl orange degradation of mesoporous-assembled SrTiO3 nanocrystals prepared by sol-gel method with the aid of structure-directing surfactant. *J. Mol. Catal. A Chem.* **287**, 70–79 (2008).
55. Pavlovic, V. P. et al. Synthesis of BaTiO 3 from a mechanically activated BaCO 3 -TiO 2 system. *Sci. Sinter.* **40**, 21–26 (2008).
56. Yang, L. & Ceder, G. Data-mined similarity function between material compositions. *Phys. Rev. B.* **88**, 1–9 (2013).
57. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* (2014).
58. Ye, J. et al. Nanoporous anatase TiO_2 mesocrystals: Additive-free synthesis, remarkable crystalline-phase stability, and improved lithium insertion behavior. *J. Am. Chem. Soc.* **133**, 933–940 (2011).
59. Roy, P., Berger, S. & Schmuki, P. TiO2 nanotubes: Synthesis and applications. *Angew. Chem. Int. Ed.* **50**, 2904–2939 (2011).
60. Paola, A. Di, Bellardita, M. & Palmisano, L. Brookite, the least known TiO_2 photocatalyst. *Catalysts* **3**, 36–73 (2013).
61. Tomita, K. et al. A water-soluble titanium complex for the selective synthesis of nanocrystalline brookite, rutile, and anatase by a hydrothermal method. *Angew. Chem. Int. Ed.* **45**, 2378–2381 (2006).
62. Reyes-Coronado, D. et al. Phase-pure TiO(2) nanoparticles: anatase, brookite and rutile. *Nanotechnology* **19**, 145605 (2008).
63. Yanqing, Z., Erwei, S., Suxian, C., Wenjun, L. & Xingfang, H. Hydrothermal preparation and characterization of brookite-type TiO2 nanocrystallites. *J. Mater. Sci. Lett.* **19**, 1445–1448 (2000).
64. Pottier, A., Chanéac, C., Tronc, E., Mazerolles, L. & Jolivet, J. -P. Synthesis of brookite TiO_2 nanoparticles by thermolysis of $TiCl_4$ in strongly acidic aqueous media. *J. Mater. Chem.* **11**, 1116–1121 (2001).
65. Arnal, P., Corriu, R. J. P., Leclercq, D., Mutin, P. H. & Vioux, A. Preparation of anatase, brookite and rutile at low temperature by non-hydrolytic sol–gel methods. *J. Mater. Chem.* **6**, 1925–1932 (1996).
66. Kitchaev, D. A., Dacek, S. T., Sun, W. & Ceder, G. Thermodynamics of phase selection in MnO_2 framework structures through alkali intercalation and hydration. *J. Am. Chem. Soc.* **139**, 2672–2681 (2017).
67. Robinson, D. M. et al. Photochemical water oxidation by crystalline polymorphs of manganese oxides: Structural requirements for catalysis. *J. Am. Chem. Soc.* **135**, 3494–3501 (2013).
68. Ueno, T., Rhone, T. D., Hou, Z., Mizoguchi, T. & Tsuda, K. COMBO: An efficient Bayesian optimization library for materials science. *Mater. Discov.* **4**, 10–13 (2016).
69. Snoek, J., Larochelle, H. & Adams, R.P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst*. (2012).
70. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst*. (2013).
71. Chollet, F. Keras. (Github, 2015).
72. Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *USENIX Symposium on Operating Systems Design and Implementation* (2016).

73. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
74. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. & Thirion, B. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
75. Taigman, Y., Yang, M., Wolf, L., Aviv, T. & Park, M. DeepFace: Closing the gap to human-level performance in face verification. *In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2014).
76. Zhao, J., Wu, X., Li, L. & Li, X. Preparation and electrical properties of SrTiO3 ceramics doped with M2O3-PbO-CuO. *Solid State Electron.* **48**, 2287–2291 (2004).
77. Zhao, W. W. et al. Black strontium titanate nanocrystals of enhanced solar absorption for photocatalysis. *CrystEngComm* **17**, 7528–7534 (2015).