npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials

Pankaj Rajak[1,6], Aravind Krishnamoorthy[2,3,6], Ankit Mishra[2,3], Rajiv Kalia[2,3,4,5], Aiichiro Nakano[2,3,4,5] and Priya Vashishta[2,3,4,5]✉

Predictive materials synthesis is the primary bottleneck in realizing functional and quantum materials. Strategies for synthesis of promising materials are currently identified by time-consuming trial and error and there are no known predictive schemes to design synthesis parameters for materials. We use offline reinforcement learning (RL) to predict optimal synthesis schedules, i.e., a time-sequence of reaction conditions like temperatures and concentrations, for the synthesis of semiconducting monolayer MoS_2 using chemical vapor deposition. The RL agent, trained on 10,000 computational synthesis simulations, learned threshold temperatures and chemical potentials for onset of chemical reactions and predicted previously unknown synthesis schedules that produce well-sulfidized crystalline, phase-pure MoS_2. The model can be extended to multi-task objectives such as predicting profiles for synthesis of complex structures including multi-phase heterostructures and can predict long-time behavior of reacting systems, far beyond the domain of molecular dynamics simulations, making these predictions directly relevant to experimental synthesis.

npj Computational Materials (2021)7:108 ; https://doi.org/10.1038/s41524-021-00535-3

INTRODUCTION

Rapid development of technology based on advanced materials requires us to considerably shorten the existing ~20-year materials development timeline[1]. This long timeline results both from the empirical discovery of promising materials as well as the trial-and-error approach to identifying scalable synthesis routes for these material candidates. Over the last decade, we have made considerable progress in addressing the first of these challenges through data-driven materials science to perform large-scale materials screening for improved properties. The exponential explosion in available computing power and increase efficiency of ab initio and machine learning (ML) driven materials simulation software have enabled the high-throughput simulations of several tens of thousands of materials from multiple material classes[2–5]. These high-throughput simulations and the resulting rich databases are increasingly being mined and analyzed using emerging ML techniques to identify promising material compositions and phases[6–10]. These strategies have been successfully employed to identify ultrahard materials, ternary nitride compositions, battery materials, polymers[11], organic solar cells[12], OLEDs[13], thermoelectrics etc[14–16].

This identification of advanced materials is only one piece necessary towards the goal of reducing time to deployment of advanced materials[17]. An equally important component in this paradigm is the corresponding ability to synthesize these promising materials and compositions. However, techniques for experimental synthesis of materials have not kept pace with advances in computational materials screening[17,18]. As a result, materials synthesis is largely dominated by individual groups that can identify synthesis strategies for advanced materials based on empirically insights and materials intuition. There are several attempted strategies to identify and optimize synthesis routes prior to actual synthesis. The first strategy, common in chemical and biological synthesis of small molecules, uses high-throughput experimental synthesis to screen for optimal synthesis precursors for chemical synthesis of small molecules[19–22]. The effectiveness of such strategies is limited since an exhaustive search of synthesis strategies is prohibitively expensive and inefficient in regard to time and reagents, whereas a narrow search scheme that varies only a single synthesis parameter at a time will likely miss several promising synthesis strategies.

In contrast to the relatively widespread use of automated algorithms to optimize chemical reactions of molecular and organic systems[23], synthesis planning for bulk inorganic materials is still in its infancy[24,25]. Non-solution-based synthesis of quantum materials involves more complicated time-correlations between synthesis parameters, which are not amenable to experimental high-throughput synthesis[26]. This also requires considerably more refined models than previous efforts which only considered the combination of reactants to predict the outcome of chemical reactions[27,28]. Therefore, there are efforts to perform text-mining on published synthesis profiles from the literature, including common solvent concentrations, heating temperatures, processing times, and precursors used to understand common rules-of-thumb and identify synthesis schedules for materials[29–31]. However, even these upcoming ML techniques are limited by scarcity of data in terms of existing schedules and synthesized materials and therefore their extension to potentially unknown materials is problematic[30]. Finally, the identification of a synthesis schedule is the optimization of a time sequence of multiple synthesis parameters, which requires a new class of ML techniques. This problem is well-suited for Reinforcement Learning (RL), a branch of machine learning, where the goal of the RL agent is design an optimal policy to solve problems that involves sequential decision making in an environment consisting of thousands of tunable parameters and a huge search space[32,33].

[1]Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, IL, USA. [2]Collaboratory for Advanced Computing and Simulations, University of Southern California, Los Angeles, CA, USA. [3]Department of Chemical Engineering & Materials Science, University of Southern California, Los Angeles, CA, USA. [4]Department of Physics & Astronomy, University of Southern California, Los Angeles, CA, USA. [5]Department of Computer Science, University of Southern California, Los Angeles, CA, USA. [6]These authors contributed equally: Pankaj Rajak, Aravind Krishnamoorthy. ✉email: priyav@usc.edu

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

Due to this flexibility and ability of RL in handling complex tasks involving non-trivial decision making and planning under uncertainties imposed by the surrounding environment, it has been used in robotics, self-driving cars and in material science domain for problems such as designing drug molecules with desired proproteins, predict reaction pathways and construct optimal conditions for chemical reactions[19,34–41].

In this work, we describe a model-based offline reinforcement learning scheme to optimize synthesis routes for a prototypical member of the family of 2D quantum material, MoS_2, via Chemical Vapor Deposition (CVD). CVD, a popular scalable technique for the synthesis of 2D materials[42], has numerous time-dependent parameters such as temperature, flow rates, concentration of gaseous reactants, and type of reaction precursors, dopants and substrates (together referred to as the synthesis profile) that need to be optimized for the synthesis of advanced materials. Recent computational studies have identified several mechanistic details about the synthesis process[43,44], but there are no comprehensive rules for designing synthesis strategies for a given material. We use RL specifically to (1) Identify synthesis profiles that result in material structures that optimize a desired property (in our case, the phase fraction of the semiconducting crystalline phase of MoS_2) in the shortest possible time and (2) Understand trends and time-correlations in the synthesis parameters that are most important in realizing materials with desired properties. These trends and time-correlations effectively provide information about mechanism of the synthesis process. Experimental synthesis by CVD is time-consuming and not amenable to high-throughput synthesis and is therefore incapable of generating the significant amount of data on synthesis using multiple profiles required for RL training. Therefore, we train our RL workflow on data from simulated CVD performed using reactive molecular dynamics simulations (RMD), which were previously shown to accurately reflect the potential energy surface of the reacting system as well as capture important mechanisms involved in the CVD synthesis of MoS_2 from MoO_3, including MoO_3 self-reduction, oxygen-vacancy-enhanced sulfidation, SO/SO_2 formation, void formation and closure etc. identified in previous studies[44–49].

Below, we describe results from the molecular dynamics simulation of CVD, followed by a representation of the dynamics of this CVD-environment as a probability density function using a probabilistic deep generative model called Neural Autoregressive Density Estimator (NADE-CVD) and model-based Offline Reinforcement Learning to identify optimal synthesis strategies. We conclude with a discussion on applicability of RL + NADE-CVD models for prediction of long-time material synthesis.

RESULTS

Reactive MD for chemical vapor deposition

We perform RMD simulations to simulate a multi-step reaction of MoO_3 crystal with a sulfidizing atmosphere containing H_2S, S_2 and H_2 molecules. Each RMD simulation models a 20-ns long synthesis schedule, divided into 20 steps, each 1 ns long. At the beginning of each step, the gaseous atmosphere from the previous step is purged and replaced with a predefined number of H_2S, S_2 and H_2 molecules. These changes in RMD parameters reflect the time-dependent changes in synthesis conditions during experimental synthesis. The sulfidizing environment is then made to react with the partially sulfidized MoO_xS_y structure from the end of the previous step at a predefined temperature for 1 ns. Each step is characterized by 4 variables, the system temperature, and the number of S_2, H_2S and H_2 molecules in the reacting environment denoted as the quartet, $(T, n_{H_2}, n_{S_2}, n_{H_2S})$. While the initial structure for each RMD simulation at $t = 0$ ns is a pristine MoO_3 slab, the final output structure ($MoS_2 + MoO_{3-x}$) is a non-trivial function of its synthesis schedule, defined by 20 such quartets as shown in Fig. 1.

NADE for predicting output of synthesis schedules

RMD simulations can generate output structures for thousands of simulated synthesis schedules to overcome the primary problem of data scarcity common to experiments. RL-based optimization of synthesis schedules consists successive stages of policy generation by the RL agent and policy evaluation by the environment. However, using RMD simulations directly as the policy evaluation environment is infeasibly time-consuming since direct evaluation a single synthesis profile by RMD takes approximately 2 days of computing. To overcome this problem, we construct a probabilistic representation of the CVD synthesis of MoS_2 as a Bayesian Network (BN) which encodes a functional relationship between the synthesis conditions and generated output structures and can therefore predict output structures for an arbitrary input condition in a fraction of the time required by RMD simulations. The BN consists of two sets random variables, namely the (a) the unobserved variable Z given by the time dependent phase fractions of 2H, 1T phases and defects in the MoO_xS_y surface, and (b) the observed variables, X, given by the user-defined synthesis condition, namely the temperature and gas concentrations (Fig. 2a, b)[50]. Each node in the BN represents either the synthesis condition at time t as X_t or the distribution of different phases on MoO_xS_y surface as Z_t. Together, the BN represents the joint distribution of X and Z as $P(X, Z)$. Since, Z_1 (initial structure, pristine MoO_3) and X (synthesis condition) is known, we can convert $P(X, Z)$ into a conditional distribution $P(Z_{2:T}|X, Z_1)$ using chain rule. Further, using conditional independence between BN variables, $P(Z_{2:T}|X, Z_1)$ can be further simplified as the autoregressive probability density function, where each Z_{t+1} depends only upon

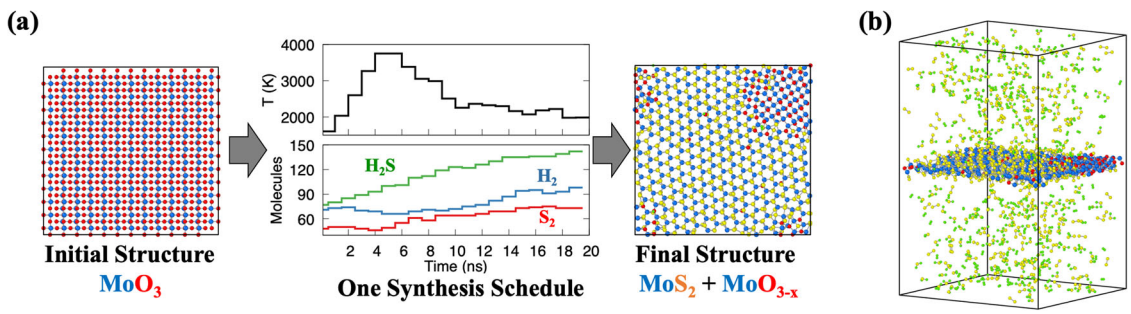

Fig. 1 Reactive MD for computational synthesis. a Schematic of the RMD simulation of a single 20-ns long synthesis schedule. The initial MoO_3 slab at $t = 0$ *ns* reacts with a time-varying sulfidizing environment to generate a final structure composed of MoS_2 and MoO_{3-x} at $t = 20$ *ns*. **b** Snapshot of RMD simulation cell for MoS_2 synthesis. The sulfidizing environment containing S_2, H_2 and H_2S gases reacts with the MoO_xS_y slab in the middle of the simulation cell (black lines).

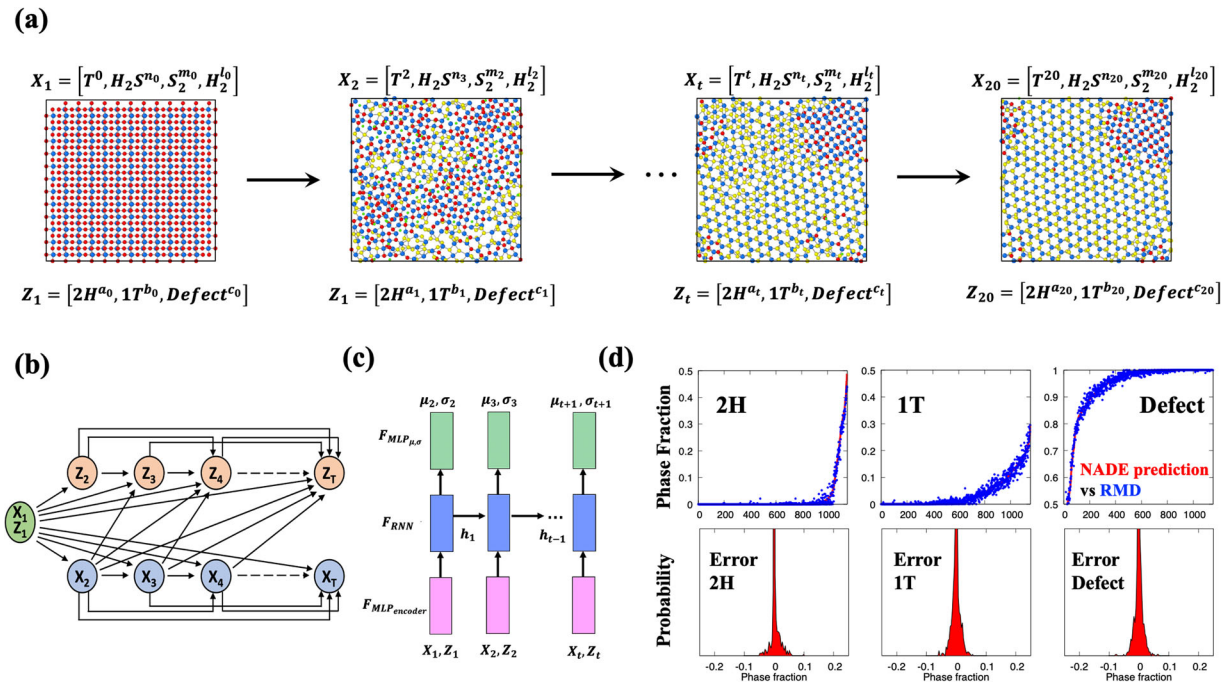

Fig. 2 NADE model of computational synthesis of MoS_2. a Each 1-ns step of the RMD simulation is characterized by an input vector X_i characterizing the synthesis conditions and the distribution of phases in the resulting structure, Z_i. **b** Bayesian Network representation of CVD synthesis of MoS_2 over T_{max}= 20 ns. The green and blue nodes are synthesis condition as observed variables (X_n), whereas orange nodes are unobserved (Z_n), which represents phase fraction of 2H, 1T and defect in MoO_xS_y surface as a function of time. **c** Schematic of the NADE-CVD, composed of two multi-layer perceptrons F_{MLP} as encoder and decoder networks and an intermediate recurrent neural network block, F_{RNN}. **d** Test accuracy of NADE-CVD with a mean absolute error <0.1 phase fraction.

the simulation history of observed and unobserved variables till time t (Fig. 2b).

$$P(Z_{2:T}|X, Z_1) = P(Z_2|Z_1, X_1) \ldots P(Z_{t+1}|Z_{1:t}, X_{1:t}) \ldots P(Z_T|Z_{1:T-1}, X_{1:T-1}) \quad (1)$$

In the BN, each of these conditional probabilities, $P(Z_{t+1}|Z_{1:t}, X_{1:t})$ is modeled as a multivariate Gaussian distribution $\mathcal{N}(Z_{t+1}|\mu_{t+1}, \sigma_{t+1})$, whose mean $\mu_{t+1} = \{\mu_{t+1}^{2H}, \mu_{t+1}^{1T}, \mu_{t+1}^{defect}\}$ and variance $\sigma_{t+1} = \{\sigma_{t+1}^{2H}, \sigma_{t+1}^{1T}, \sigma_{t+1}^{defect}\}$ is function of simulation history, $(Z_{1:t}, X_{1:t})$.

To learn the BN representation of the CVD process and capture the conditional distribution $P(Z|X, Z_1)$ compactly, we have developed a deep generative model architecture called a Neural Autoregressive Density Estimator (NADE-CVD; Fig. 2c), which consist of an encoder, decoder and recurrent neural network (RNN)[51–54]. The output of NADE-CVD function at time step $t + 1$ is μ_{t+1} and σ_{t+1} for three phases in MoO_xS_y surface which are functions of simulation history encoded by the RNN cell as h_t, where h_t is a function of h_{t-1} and synthesis condition (Z_t, X_t) at time t. Parameters of the NADE-CVD model are learned using maximum likelihood estimate using a training data of 10,000 RMD simulations of CVD using different synthesis conditions. The prediction error of the trained NADE-CVD model on test data (Fig. 2c) shows a RMSE error of merely 3.5 atoms and maximum prediction error on any phase of ≤30 atoms. The architecture of the NADE-CVD model is described in the Methods section and details about model training are provided the Supplementary Methods.

Offline model-based RL for optimal synthesis schedules

The NADE-CVD model accurately approximates a computationally expensive RMD simulation and provides a fast and probabilistic evaluation of the output structure from a given synthesis schedule. However, on its own, this model cannot be used to achieve the goal of predictive synthesis, which is to identify the most likely synthesis schedules that yield a material with optimal properties (such as high crystallinity, phase purity or hardness). For MoS_2 synthesis, one example of a design goal is to determine synthesis schedules that yield high quality MoS_2 (i.e., largest phase fraction of semiconducting 2H phase in the final product), in the shortest possible time. In other words, we wish to perform the non-trivial optimization of $X_{1:t}$ to maximize the value of $\sum_t Z_{1:t}$ (see Supplementary Methods). Mathematically, it can be written as

$$\arg\max_{X_{1:t}} \sum Z_{1:t} \text{ where}(Z_{1:t}, X_{1:t}) \sim P(Z_{1:t}, X_{1:t}) = P(Z_{1:t}|X_{1:t})P(X_{1:t}) \quad (2)$$

For this purpose, we construct a model-based offline reinforcement learning (RL) scheme, where the agent does not have access to the environment (RMD simulation) during training and learns the optimal policy from randomly sampled suboptimal offline data from the environment[55–59]. Here, the offline RL workflow consists of a RL agent coupled to NADE-CVD trained on offline RMD data as discussed in the previous section, (Fig. 3a). The RL agent (π_θ) is a multi-layer perceptron, where the input state (s_t) at time t is a 128-dimension embedding vector of the entire simulation history till t, $(Z_{1:t}, X_{1:t})$. At each time step t, the RL agent takes an action, a_t, which is the change in synthesis condition (i.e. reaction temperature and gas concentrations) at t, $a_t = \Delta Z = \{\Delta T, \Delta S_2, \Delta H_2, \Delta H_2S\}$. The synthesis condition for the next nanosecond of the simulation is defined as $X_{t+1} = X_t + a_t$. The corresponding action (a_t) to take at s_t is modeled using a Gaussian distribution $(a_t \sim \mathcal{N}(\mu(s_T), \sigma^2))$, whose parameters $\mu(s_T)$ – state dependent mean – is the output of the RL agent, $\mu(s_T) = \pi_\theta(s_T)$. The variance, σ^2 is assumed to be constant and is tuned as a hyperparameter of the RL scheme. Therefore, the RL scheme designs a 20 ns synthesis schedule (τ)starting with an arbitrary synthesis condition, $\{T^0, S_2^0, H_2^0, H_2S^0\}$, such that the

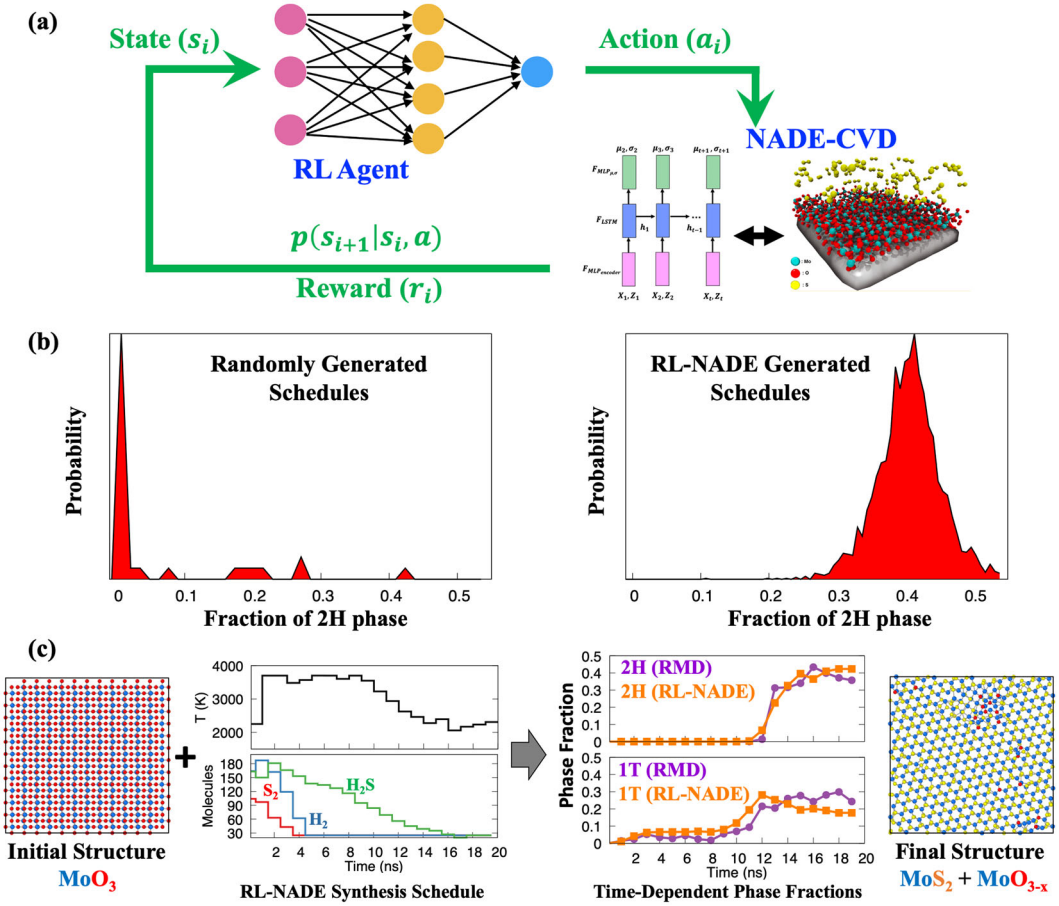

Fig. 3 Reinforcement Learning model for synthesis schedule design. a Schematic of the RL-NADE model for optimizing schedules for MoS_2 synthesis. **b** Comparison of structures generated by the RL-designed schedules against randomly generated schedules demonstrates that the RL-NADE model consistently identifies CVD synthesis schedules that generate highly crystalline products. **c** Validation of a promising RL-generated schedule using RMD simulations.

action proposed at each timestep t serves to convert the initial MoO_3 crystal into 2H-MoS_2 structure as quickly as possible.

During training, the RL agent learns the policy of designing the optimal synthesis condition *via* policy gradient algorithm informed by the NADE-CVD model[33,60–62]. At each time step t in an episode, the RL agent receives an input state s_t and proposes an action a_t that determines the synthesis condition at next time step, X_{t+1}. Using this, NADE-CVD predicts the distribution of various phases in the synthesized product Z_{t+1}. The NADE-CVD model also gives a reward (r_t) proportional to the concentration of 2H phase $Z_{t+1}[n_{2H}]$ and a new state s_{t+1} to the RL agent. During training, the goal of the RL agent is to use these reward signals and adjust its policy parameters (π_θ) so as to maximize its total reward, to produce 2H-rich MoS_2 structure in minimum time.

$$\text{Objective: } \arg\max_{\theta} \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{t}(s_t, a_t)\right] \text{where } r_t(s_t, a_t) = \begin{cases} 0.0 \text{ if } Z_{t+1}[n_{2H}]<0.4 \\ 0.2Z_{t+1} \text{ if } Z_{t+1}[n_{2H}] \geq 0.4 \end{cases} \quad (3)$$

The details of the network architecture, and the policy gradient algorithm is given in the Methods section and RL agent training is described in the Supplementary Methods.

The efficiency of the trained RL agent in identifying promising synthesis schedules is demonstrated in Fig. 3b, which compares the 2H phase fraction of the resulting structures from 3200 synthesis schedules generated by the RL agent against 3200 randomly generated schedules, similar to what is used for training NADE-CVD. The RL agent is able to consistently identify schedules that result in highly crystalline and phase-pure products, while the randomly generated schedules overwhelmingly yield poorly-sulfidized and/or poorly crystalline products. This shows that offline RL agent is able to learn a superior policy from the sub-optimal random RMD simulation used in its training. Also, from probabilistic viewpoint, the RL agent constructs a probability distribution function (pdf) of $X_{1:t}$ that places most of its probability mass on regions on $X_{1:t}$ that maximizes $\sum Z_{1:t}$. Figure 3c shows the validation of one RL-predicted synthesis schedule by subsequent RMD simulation, showing that the observed time-dependent phase fraction tracks the RL-NADE prediction closely.

Optimal synthesis schedules and mechanistic insights from RL

The RL agent is trained to learn polices that generate time-dependent temperatures, and concentrations of H_2S, S_2 and H_2 molecules to synthesize 2H-rich MoS_2 structures in least time. Closer inspection of these RL designed policies provides mechanistic insight into CVD synthesis and the effect of variations in temperature and gas concentration on the quality of the synthesized product. Figure 4 shows that the RL agent has learned to generate a two-part temperature profile consisting of an early high-temperature (>3000 K) phase spanning the first 7–10 ns followed by annealing to an intermediate temperature (~ 2000 K) for the reminder of the synthesis profile. This two-part synthesis profile identified by RL policy is consistent with the experiments and atomistic simulations, that is high temperature (> 3000 K) is necessary for both the reduction of MoO_3 surface and its sulfidation, whereas the subsequent lower temperature (~ 2000 K) is necessary for enabling crystallization in the 2H structure, while

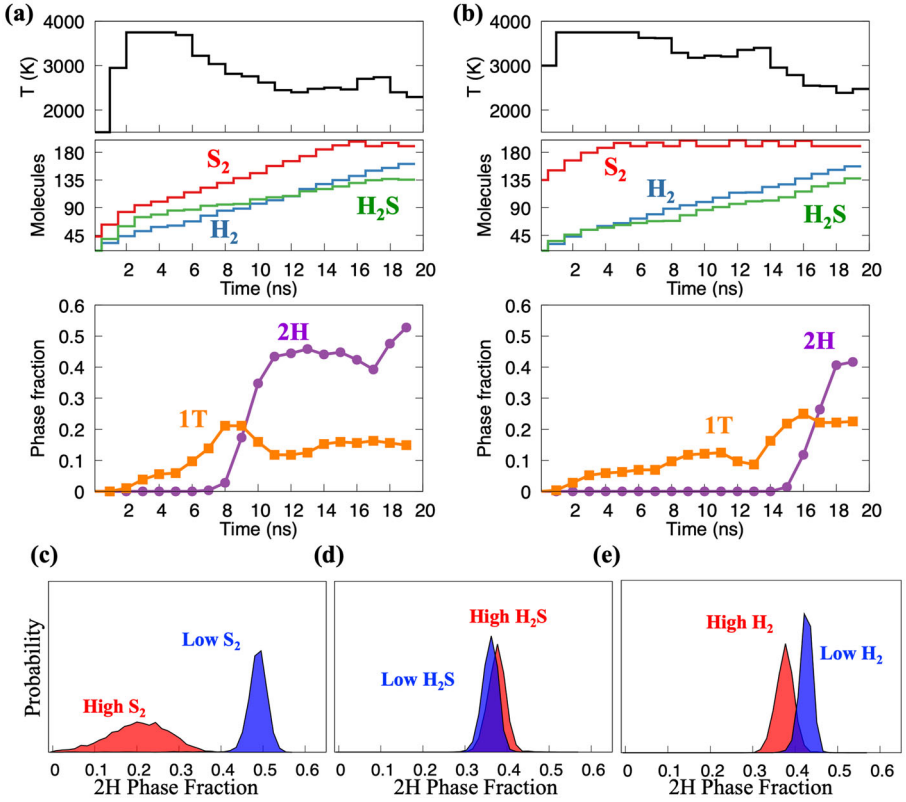

Fig. 4 Effect of synthesis conditions on products. a A generated synthesis profile starting from low temperature and low gas concentrations. The RL model quickly ramps up the temperature up to 7 ns to promote reduction and sulfidation and then lowers the temperature to intermediate values to promote crystallization. This profile generates significant phase fraction of 2H starting from 10 ns. **b** A generated synthesis profile starting from high temperature and high S_2 concentrations. The RL-NADE model retains the high temperature at early stages of synthesis and slowly anneals the system to intermediate temperatures after 10 ns. This schedule promotes relatively late crystallization and 2H phase formation. **c** Synthesis profiles with initially low S_2 concentrations yield significantly higher phase fraction of 2H in the final product compared to profiles containing higher S_2 concentrations at $t = 0$ ns. **d**, **e** Synthesis schedules are relatively insensitive to the initial concentration of reducing species, H_2S and H_2.

continuing to promote residual sulfidation. Consistent with previous reactive and quantum molecular dynamics simulations of material synthesis, a significantly elevated temperature is necessary to observe reaction event within the limited time domain accessible to atomistic simulations[44–46,49]. It is observed that the RL agent maintains this two-stage synthesis profile even if the provided initial temperature at $t = 0$ ns is low by quickly ramping up the synthesis temperature to the high-temperature regime (> 3000 K). The RL agent is also able to predict non-trivial mechanistic details about phase evolution, including the observation that the nucleation of the 1T phase precedes the nucleation of the 2H crystal structure (Fig. 4a, b). Similar trends were observed in previous mechanistic studies of MoS_2 synthesis[44].

Another important phenomenon identified by RL agent is the effect of gas concentrations on the quality of the final product (Fig. 4b). To analyse the effect of initial gas concentration, we compute the probability distribution of 2H phase in MoS_2 over the last 10 ns of the simulation for the synthesis conditions proposed by the RL agent under different initial conditions of gas conc. but with similar temperature profile. The mean (μ_{2H}) of the pdf is $\mu_{2H} = \mathbb{E}_{\tau\sim\pi_\theta}\left[\frac{1}{10}\sum_{t=20}^{t=20} Z_t[n_2H]\right]$, is the expected fraction of the 2H phase in over the last 10 ns of the synthesis simulation and a higher value of μ_{2H} provides an indication of the extent of sulfidation as well as the time required to generate 2H phases. The RL agent is found to promote synthesis profiles that have low concentration of gas molecules (particularly non-reducing S_2 molecules) at early stages (0–3 ns) of the synthesis, when the temperature is high. This partially evacuated synthesis atmosphere promotes the evolution of oxygen from and self-reduction of the MoO_3 surface. This can be clearly observed by comparing the histogram of 2H phase fractions in structures generated by synthesis profiles with low initial (i.e. $t = 0$ ns) concentration of S_2 molecules against those with higher concentration of S_2 molecules (Fig. 4c). Profiles with low initial S_2 concentrations enable greater self-reduction of the MoO_3 surface resulting in a significantly higher 2H phase fraction in the synthesized product at $t = 10$–20 ns. H_2S and H_2 molecules, which are more reducing than S_2, do not meaningfully affect the MoO_3 self-reduction rate, and the 2H phase fraction in the final MoO_xS_y product is largely independent of the initial H_2S and H_2 concentrations (Fig. 4d, e).

Multi-task RL-CVD: schedules for heterostructure synthesis

The outputs of the NADE-CVD model, each μ_{t+1} and σ_{t+1} is only function of simulation history up to time t. Similarly, each action a_t taken by the RL agent is a function only of the input state s_t, which is an encoded representation of simulation history up to time t. Hence, we can use RL + NADE-CVD to design policies for synthesis over time scales significantly longer than the 20 ns RMD simulation trajectories used for NADE-CVD training. Figure 5 shows a policy proposed by the RL + NADE-CVD model for a 30 ns simulation. This extended synthesis profile retains the design principles such as a two-phase temperature cycle and low initial gas phase concentrations that were learned from 20-ns trajectories. Further, the longer synthesis schedule also allows the RL agent to uncover synthesis design rules for improving 2H phase

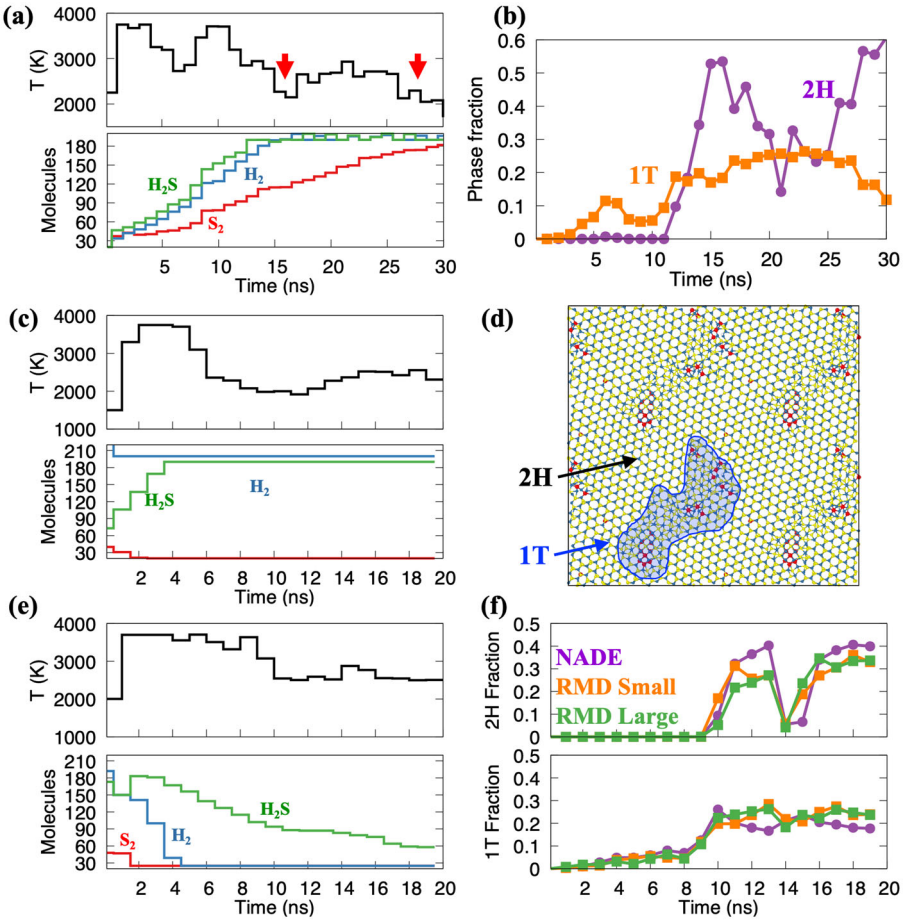

Fig. 5 Extensions of RL + NADE-CVD Method. a, b A 30-ns long synthesis profile predicted by RL + NADE-CVD retains design principles about two-phase temperature cycle and low initial gas phase concentrations learned from 20-ns RMD trajectories. In addition, the 30-ns profile also includes a temperature annealing step between 15–30 ns (arrows) that improves the 2H phase fraction beyond 60%. **c** RL + NADE-CVD generated synthesis schedule for optimizing 1T phase fraction. **d** Output structure from an RMD simulation of the 1T-optimized synthesis schedule reveals a heterostructure containing a 1T-rich region embedded in the 2H phase. **e, f** The robustness of RL-generated profiles against system size-scaling is validated by the identical fractions of 2H and 1T phases in laterally-small and laterally-large systems simulated using RMD using the same profile.

fraction. The RL profile in Fig. 5 includes a heating-cooling cycle between 15–30 ns what has previously been shown to improve the crystallinity and 2H phase fraction in the synthesized material[44].

The RL agent learns promising synthesis profile by adjusting its policy parameters (π_θ) to maximize a pre-defined reward function, that corresponds the material to be synthesized. Therefore, the RL agent can optimize synthesis schedules for other material structures, including multi-phase heterostructures, by constructing corresponding reward functions. The following reward function, $r_t(s_t, a_t)$ maximizes the phase fraction of 1T crystal structure over the 20 ns simulation.

$$\text{Objective: } \arg\max_\theta \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{t=20} r(s_t,a_t)\right] \text{where } r_t(s_t,a_t) = \begin{cases} 0.0 \text{ if } Z_{t+1}[n_{1T}]<0.17 \\ 0.35 Z_{t+1} \text{ if } Z_{t+1}[n_{1T}] \geq 0.17 \end{cases} \quad (4)$$

Figure 5c shows a RL-generated schedule to synthesize 1T-rich structures. The temperature profile is largely consistent with those observed for 2H-maximized synthesis schedules. The RL generated gas-phase concentrations optimized for 1T synthesis maximize H_2 and H_2S concentrations, while minimizing S_2 concentrations. This is consistent with experimental observations, where reducing environments were observed to produce more 1T phase fractions[63]. This is in contrast to schedules optimized for 2H MoS_2, where the concentration of all three gaseous species show correlated variations (Fig. 4a, b). Figure 5d shows a MoS_2 2H-1T heterostructure configuration generated at the end of MD simulations according to the RL-generated synthesis schedule. The synthesized heterostructure consists of an island of 1T-MoS_2 embedded in the 2H-MoS_2 matrix with an atomically sharp interface between the two phases. We note here that same RMD data is used to train the CVD dynamics (NADE) models followed by training the RL-agent for two different objectives (2H or 1T maximization) by simply modifying the reward function. This shows the capability of the model-based offline RL in learning policies for multiple-tasks/objective without generating additional data.

Finally, RL-predicted synthesis schedules are also extremely robust with respect to system-size scaling. Figure 5e shows the validation of a single RL-generated profile using RMD simulations on systems of two different sizes – 51 Å × 49 Å and 100 Å × 100 Å. Figure 5f shows that the observed fractions of 2H and 1T phases in RMD simulations of both the small and large systems are consistent with each other over the entire 20-ns simulation range. Further, these phase fractions are also quantitatively consistent with the values predicted by the NADE model used in the RL optimization loop (See Supplementary Figures 4 and 5 and Supplementary Discussion on accuracy and scale-independence of NADE-CVD predictions). This capability to optimize synthesis

schedules independent of system size is useful to extend this approach to experimental synthesis.

DISCUSSION

We have developed a machine learning scheme based on offline reinforcement learning for the predictive design of time-dependent reaction conditions for material synthesis. The scheme integrates a reinforcement learning agent with a deep generative model of chemical reactions to predict and design optimum conditions for the rapid synthesis of two-dimensional MoS_2 monolayers using chemical vapor deposition. This model was trained on thousands of computational synthesis simulations at different reaction conditions performed using reactive molecular dynamics. The model successfully learned the dynamics of material synthesis during simulated chemical vapor deposition and was able to accurately predict synthesis schedules to generate a variety of MoS_2 structures such as 2H-MoS_2, 1T-MoS_2 and 2H-1T in-plane heterostructures. Beyond mere synthesis design, the model is also useful for mechanistic understanding of the synthesis process and helped identify distinct temperature regimes that promote sulfidation and crystallization and the impact of a reducing environment on the phase purity of the synthesis product. We also demonstrate how the reinforcement learning scheme can be extended to predict the outcome of material synthesis over long time-scales for system sizes larger than those used for training. This flexibility makes the offline reinforcement learning based design scheme suitable for optimization of experimental synthesis of wide variety of nanomaterials, where the agent does not have to directly interact with the environment during training and can still learn optimal policy from the randomly data collected from the environment.

METHODS

Molecular dynamics simulation

All 10000 RMD simulations were performed using the RXMD molecular dynamics engine[64,65] using the reactive forcefield originally developed by Hong et al.[45] that is optimized for reacting Mo-O-S-H systems. RMD computational synthesis simulations were performed on a 51 Å × 49 Å × 94 Å simulation cell containing 1200-atom MoO_3 slab at z = 47 Å surrounded by a reacting atmosphere containing H_2, S_2 and H_2S molecules. During RMD simulations, a one-dimensional harmonic potential is applied to each Mo atom along the z-axis (i.e., normal to the slab surface) with the spring constant of 75.0 kcal/mol to keep the atoms in a two-dimensional plane at elevated temperatures. For each nanosecond of the computational synthesis simulation, the system temperature is maintained at the value specified in the synthesis profile by scaling the velocities of the atoms. MD trajectories are integrated with a timestep of 1 femtosecond and charge-equilibration is performed every 10 timesteps[66].

NADE-CVD

The NADE-CVD consists of an encoder, a LSTM block and a decoder (Fig. 2a). The encoder transforms (X_t, Z_t) into a 72-dimension vector, $e_t = F_{\text{encoder}}(X_t, Z_t)$. After that, the LSTM layer constructs an embedding of the simulation history till time t as $h_t = F_{\text{LSTM}}(h_{t-1}, e_t)$, where h_t is a 128 dimension vector. The decoder than uses the h_t to predict the mean and variance of various phases in MoO_xS_y surface as $\mu_{t+1}, \sigma_{t+1} = F_{\text{decoder}}(h_t)$. The encoder and decoder are fully connected neural network of dimensions 7 × 24, 24 × 48, 48 × 72 and 128 × 72, 72 × 24, 24 × 3, respectively. The parameters of the NADE-CVD (Θ) are learned via maximum likelihood estimate (MLE) of the following likelihood function

$$L(\Theta; D) = \prod_{j=1}^{j=m} P_\Theta(Z^j, X^j) = \prod_{j=1}^{j=m} \prod_{t=2}^{t=n} P_\Theta\left(Z_t^j | Z_{1:t-1}^j, X_{1:t-1}^j\right) \quad (5)$$

Here, $D = \left\{\left(X_{1:n}^1 Z_{1:n}^1\right), \left(X_{1:n}^2 Z_{1:n}^2\right), \ldots \left(X_{1:n}^m Z_{1:n}^m\right)\right\}$ is training dataset of m RMD simulation trajectories. Further details such as log-likelihood of training data during training and evaluation of the NADE-CVD on test data is given in Supplementary Methods.

RL agent architecture and policy gradient

The RL agent, π_θ, is constructed using a fully connected neural network with tunable parameters θ. It consists of an input layer of 128 nodes that is followed by two hidden layers with 72 and 24 nodes and then an output layer. The input s_t to π_θ is the embedding of the simulation history, $(X_{1:t}, Z_{1:t})$, generated by NADE-CVD, h_t. The output of the RL agent is the mean $\mu(s_t)$ of action a_t and value function $V(s_t)$ associated with s_t. The hyperparameters σ^2 associated with the variance of the Gaussian distribution of actions a_t is taken as 5. During training, the RL agent learns the optimal policy that maximize the total expected reward $\mathbb{E}$ (Eq. 1) using policy gradient algorithm by taking the derivative of $\mathbb{E}$ with respect to its parameter θ, $\nabla\mathbb{E} = \frac{\partial \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T} r(s_t, a_t)\right]}{\partial\theta}$, where trajectory $\tau = \{s_1, a_1, s_2, a_2, \ldots s_T, a_T\}$. This derivate reduces into the following objective function which is optimized via gradient accent.

$$\nabla_\theta E = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T_{max}} \nabla_\theta \log \pi_\theta(s_t, a_t)(G_t - V(s_t))\right]; \text{where } G_t = \sum_{t=1}^{t=t} r_t \quad (6)$$

Here, value function $V(s_t)$ is used as a variance reduction technique in the calculation of $\nabla_\theta\mathbb{E}$ via Monte Carlo estimate. Details of the above derivation and the policy gradient algorithm is given in Supplementary Methods.

DATA AVAILABILITY

Example profiles and structures from reactive molecular dynamics (RMD) simulations used for training RL models and the trained NADE model of CVD dynamics are distributed along with the code.

CODE AVAILABILITY

The deep learning code used in this study can be found at https://github.com/rajak7/RL_CVD.git.

Received: 19 October 2020; Accepted: 26 March 2021;
Published online: 14 July 2021

REFERENCES

1. Green, M. L. et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.* **4**, 011105 (2017).
2. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *NPJ Comput. Mater.* **5**, 99 (2019).
3. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).
4. Bartok, A. P., Payne, M. C., Kondor, R. & Csanyi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
5. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C.* **121**, 511–522 (2017).
6. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2**, 0121 (2018).
7. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
8. Gubernatis, J. E. & Lookman, T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2**, 120301 (2018).
9. Dai, C. & Glotzer, S. C. Efficient phase diagram sampling by active learning. *J. Phys. Chem. B* **124**, 1275–1284 (2020).
10. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
11. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *NPJ Comput Mater.* **3**, 54 (2017).
12. Tagade, P. M. et al. Attribute driven inverse materials design using deep learning Bayesian framework. *NPJ Comput. Mater.* **5**, 127 (2019).
13. Gomez-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 (2016).

14. Yan, J. et al. Material descriptors for predicting thermoelectric performance. *Energ. Environ. Sci.* **8**, 983–994 (2015).
15. Gaultois, M. W. et al. Data-driven review of thermoelectric materials: performance and resource considerations. *Chem. Mater.* **25**, 2911–2920 (2013).
16. Bassman, L. et al. Efficient discovery of optimal N-layered TMDC heterostructures. *MRS Adv.* **3**, 397–402 (2018).
17. de Pablo, J. J. et al. New frontiers for the materials genome initiative. *NPJ Comput. Mater.* **5**, 41 (2019).
18. Yang, Q., Sing-Long, C. A. & Reed, E. J. Learning reduced kinetic Monte Carlo models of complex chemistry from molecular dynamics. *Chem. Sci.* **8**, 5781–5796 (2017).
19. Zhou, Z. P., Li, X. C. & Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* **3**, 1337–1344 (2017).
20. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, 557 (2019).
21. McMullen, J. P. & Jensen, K. F. Integrated microreactors for reaction automation: new approaches to reaction development. *Annu. Rev. Anal. Chem.* **3**, 19–42 (2010).
22. Sanchez-Lengeling, B., Outeiral, C., Guimaraes, G. & Aspuru-Guzik, A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). Preprint at https://chemrxiv.org/articles/preprint/ORGANIC_1_pdf/5309668 (2017).
23. Fabry, D. C., Sugiono, E. & Rueping, M. Self-optimizing reactor systems: algorithms, on-line analytics, setups, and strategies for accelerating continuous flow process optimization. *Isr. J. Chem.* **54**, 341–350 (2014).
24. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
25. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73 (2016).
26. Jo, S. S. et al. Formation of large-area MoS2 thin films by oxygen-catalyzed sulfurization of Mo thin films. *J. Vac. Sci. Technol. A* **38**, 013405 (2019).
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. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
29. Kononova, O. et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
30. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *NPJ Comput. Mater.* **3**, 53 (2017).
31. Kim, E. et al. Data Descriptor: Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
32. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
33. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction, 2nd edition (MIT Press, 2018).
34. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360 (2018).
35. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
36. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
37. Kearnes, S., Li, L. & Riley, P. Decoding molecular graph embeddings with reinforcement learning. Preprint at https://arxiv.org/abs/1904.08915 (2019).
38. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).
39. Cova, T. F. G. G. & Pais, A. A. C. C. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Front. Chem.* **7**, 809 (2019).
40. Li, H. et al. Tuning the molecular weight distribution from atom transfer radical polymerization using deep reinforcement learning. *Mol. Syst. Des. Eng.* **3**, 496–508 (2018).
41. Betterton, J. R., Ratner, D., Webb, S. & Kochenderfer, M. Reinforcement learning for adaptive illumination with X-rays. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 328–334 (Paris, France, 2020), https://doi.org/10.1109/ICRA40945.2020.9196614.
42. Jin, G. et al. Atomically thin three-dimensional membranes of van der Waals semiconductors by wafer-scale growth. *Sci. Adv.* **5**, eaaw3180 (2019).
43. Hong, S. et al. Chemical vapor deposition synthesis of MoS_2 layers from the direct sulfidation of MoO_3 surfaces using reactive molecular dynamics simulations. *J. Phys. Chem. C.* **122**, 7494–7503 (2018).
44. Hong, S. et al. Defect healing in layered materials: a machine learning-assisted characterization of MoS_2 crystal phases. *J. Phys. Chem. Lett.* **10**, 2739–2744 (2019).
45. Hong, S. et al. Computational synthesis of MoS_2 layers by reactive molecular dynamics Simulations: initial sulfidation of MoO_3 surfaces. *Nano Lett.* **17**, 4866–4872 (2017).
46. Hong, S. et al. A reactive molecular dynamics study of atomistic mechanisms during synthesis of MoS_2 layers by chemical vapor deposition. *MRS Adv.* **3**, 307–311 (2018).
47. Hong, S. et al. Chemical vapor deposition synthesis of MoS_2 layers from the direct sulfidation of MoO_3 surfaces using reactive molecular dynamics simulations. *J. Phys. Chem. C.* **122**, 7494–7503 (2018).
48. Misawa, M. et al. Reactivity of sulfur molecules on MoO_3 (010) surface. *J. Phys. Chem. Lett.* **8**, 6206–6210 (2017).
49. Hong, S. et al. Sulfurization of MoO_3 in the chemical vapor deposition synthesis of MoS_2 enhanced by an H2S/H2 mixture. *J. Phys. Chem. Lett.* **12**, 1997–2003 (2021).
50. Koller, D. & Friedman, N. Probabilistic graphical models: principles and techniques (MIT Press, 2009).
51. Ou, Z. A review of learning with deep generative models from perspective of graphical modeling. Preprint at https://arxiv.org/abs/1808.01630 (2018).
52. Hugo, L. & Iain, M. The neural autoregressive distribution estimator. In *Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA*. 29–37 (2011).
53. Karol, G., Ivo, D., Andriy, M., Charles, B. & Daan, W. Deep autoregressive networks. In *31st International Conference on Machine Learning, Beijing, China*. 1242–1250 (2014).
54. Oord, A. V. D., Kalchbrenner, N. & Kavukcuoglu, K. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. 1747–1756 (New York, NY, USA, 2016).
55. Wu, Y., Tucker, G. & Nachum, O. Behavior regularized offline reinforcement learning. Preprint at https://arxiv.org/abs/1911.11361 (2019).
56. Levine, S., Kumar, A., Tucker, G. & Fu, J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. Preprint at https://arxiv.org/abs/2005.01643 (2020).
57. Kidambi, R., Rajeswaran, A., Netrapalli, P. & Joachims, T. MORel: model-based offline reinforcement learning. *Adv. Neural. Inf. Process. Syst.* **33**, 21810–21823 (2020).
58. Yu, T. et al. MOPO: model-based offline policy optimization. *Adv. Neural. Inf. Process. Syst.* **33**, 14129–14142 (2020).
59. Mandlekar, A., Xu, D., Martín-Martín, R., Savarese, S. & Fei-Fei, L. Learning to generalize across long-horizon tasks from human demonstrations. Preprint at https://arxiv.org/abs/2003.06085 (2020).
60. Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. Preprint at https://arxiv.org/abs/1506.02438 (2015).
61. Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* **12**, 1057–1063 (2000). 12.
62. Duan, Y., Chen, X., Houthooft, R., Schulman, J. & Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. 1329–1338 (New York, NY, USA, 2016).
63. Liu, L. N. et al. Phase-selective synthesis of 1T ' MoS_2 monolayers and heterophase bilayers. *Nat. Mater.* **17**, 1108 (2018).
64. Nomura, K.-i, Kalia, R. K., Nakano, A., Rajak, P. & Vashishta, P. RXMD: a scalable reactive molecular dynamics simulator for optimized time-to-solution. *SoftwareX* **11**, 100389 (2020).
65. Nomura, K.-I., Kalia, R. K., Nakano, A. & Vashishta, P. A scalable parallel algorithm for large-scale reactive force-field molecular dynamics simulations. *Comput. Phys. Commun.* **178**, 73–87 (2008).
66. Nomura, K., Small, P. E., Kalia, R. K., Nakano, A. & Vashishta, P. An extended-Lagrangian scheme for charge equilibration in reactive molecular dynamics simulations. *Comput. Phys. Commun.* **192**, 91–96 (2015).

ACKNOWLEDGEMENTS

This work was supported as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award Number DE-SC0014607. This research was partly supported by Aurora Early Science programs and used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Computations were performed at the Argonne Leadership Computing Facility under the DOE INCITE and Aurora Early Science programs and at the Center for Advanced Research Computing of the University of Southern California.

AUTHOR CONTRIBUTIONS

P.R. and A.K. contributed equally to this work. P.R., A.K., R.K.K, A.K. and P.V designed the research. P.R. and A.K. performed the molecular dynamics simulation as well as reinforcement learning framework. A.M. processed the simulation data for the machine learning model. All authors contributed to analyzing the results and writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00535-3.

Correspondence and requests for materials should be addressed to P.V.

Reprints and permission information is available at http://www.nature.com/reprints

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