



# Improving *ab initio* diffusion calculations in materials through Gaussian process regression

Seyyedfaridoddin Fattahpour  and Sara Kadkhodaei *Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago, Chicago, Illinois 60607, USA*

(Received 3 July 2023; accepted 22 December 2023; published 24 January 2024)

Saddle point search schemes are widely used to identify the transition state of different processes, like chemical reactions, surface and bulk diffusion, surface adsorption, and many more. In solid-state materials with relatively large numbers of atoms, the minimum mode following schemes such as dimer are commonly used because they alleviate the calculation of the Hessian on the high-dimensional potential energy surface. Here, we show that the dimer search can be further accelerated by leveraging Gaussian process regression (GPR). The GPR serves as a surrogate model to feed the dimer with the required energy and force input. We test the GPR-accelerated dimer method for predicting the diffusion coefficient of vacancy-mediated self-diffusion in *body-centered cubic* molybdenum and sulfur diffusion in *hexagonal* molybdenum disulfide. We use a multitask learning approach that utilizes a shared covariance function between energy and force input, and we show that the multitask learning significantly improves the performance of the GPR surrogate model compared to previously used learning approaches. Additionally, we demonstrate that a translation-hop sampling approach is necessary to avoid overfitting the GPR surrogate model to the minimum-mode-following pathway and thus succeeding in locating the saddle point. We show that our method reduces the number of evaluations compared to a conventional dimer method.

DOI: [10.1103/PhysRevMaterials.8.013804](https://doi.org/10.1103/PhysRevMaterials.8.013804)

## I. INTRODUCTION

Transition state theory (TST) [1] is widely used to quantify the free-energy barrier (or activation free energy) of chemical reactions, such as molecular dissociation, as well as material processes including bulk diffusion, surface diffusion, or surface adsorption [2–4]. Within TST, the activated state is identified as the saddle point on the free-energy surface. Consequently, saddle-point search methods are crucial for quantifying the activated state, energy barrier, and rate of various kinetic processes in materials [5–11]. Among saddle-point search methods, minimum mode following methods [12,13], such as the dimer algorithm [8,14,15], have gained popularity due to their computational advantages, particularly for solid-state processes. Unlike alternatives such as the partitioned rational function optimization (P-RFO), these methods do not require energy Hessian calculations in the high-dimensional space of solid-state atomic systems [13]. However, utilizing the dimer algorithm can still be computationally prohibitive when combined with density functional theory (DFT) energy calculations. In this study, we show that we can further enhance computational efficiency by utilizing Gaussian process regression (GPR) as a surrogate model for inputting forces to the dimer algorithm. We implement a GPR-guided dimer algorithm, which we call the GPR-dimer, and apply it to investigate bulk diffusion in *body-centered cubic* (*bcc*) Mo as well as diffusion of sulfur in *hexagonal* MoS<sub>2</sub>. Building upon previous studies that combined GPR with saddle-point or minimum-energy-path search methods, this work provides two new insights for advancing the utility of GPR-dimer. First, we employ a multitask GPR learning approach, demonstrating a significant reduction in both training error and time compared to previously used GPR learning methods. Second, we introduce a translation-hop sampling approach designed to

reduce the computational effort and enhance the robustness of the search algorithm. Furthermore, this work extends the application of GPR-dimer to solid-state materials.

The application of machine learning in materials and molecular simulations, with the goal of improving computational efficiency, encompasses diverse approaches and objectives. Notable applications include global force field generation, such as GDML [16], sGDML [17,18], and BIGDML [19] for quantum machine learning force fields (see Discussion for details). Here, our focus is solely on reviewing studies that employ machine learning for the search of saddle points or minimum-energy paths. Previous studies have successfully employed GPR to accelerate the search for saddle points or minimum-energy paths [20–25]. For instance, Jónsson's group developed an adaptive GPR surrogate model of the potential energy surface (PES) [20]. They utilized this model to derive an initial interpolation of the minimum energy path, which was subsequently optimized using the nudged elastic band (NEB) method. Their investigations focused on 25 chemical reactions, primarily involving organic molecules, known as the Baker test systems [26]. The results demonstrated the superiority of the GPR-accelerated NEB search over the classical NEB optimizer. The GPR model was trained using the Matérn covariance function and a predetermined weighted combination of energy-based and force-based loss functions. Another study by Kästner's group combined a GPR-interpolated PES with the P-RFO method to identify transition states in the Baker test systems [21]. By providing the necessary Hessian information to the P-RFO optimizer, the surrogate GPR model rendered the method computationally efficient, comparable to force-based methods like the dimer algorithm. Subsequently, they introduced a GPR-based Hessian update scheme [25], where the GPR was employed

to update Hessian matrices using gradient-based information during the optimization procedure. This approach involved at least one initial Hessian, along with additional energies and gradients. Moreover, Denzel and Kästner [21] utilized the Matérn covariance function and the “derivative observation” GPR learning technique [27], which explicitly relates the learned forces to the negative partial derivatives of learned energies. In a subsequent study, the same group combined the GPR surrogate model, employing derivative observation learning, with the NEB optimizer [22]. In two additional studies, Jónsson’s group introduced the inverse-distance covariance function as an alternative to the previously employed covariance functions [23,24], resulting in a significantly enhanced GPR surrogate model. By utilizing this improved GPR model to guide the dimer and NEB saddle-point searches, they investigated the dissociative adsorption of an  $H_2$  molecule on the Cu(110) surface, three gas-phase chemical reactions, and the diffusion hop of an  $H_2O$  molecule on an ice  $I_h(0001)$  surface [23,24].

Building upon previous studies, we extend the application of GPR-dimer to investigate solid-state processes, moving beyond molecular processes. This study presents two examples of solid-state processes: vacancy-mediated self-diffusion in *bcc* Mo and sulfur diffusion in *hexagonal*  $MoS_2$ . We address the challenge of handling high dimensionality when applying GPR-dimer to solid-state processes through the use of the inverse-distance covariance function introduced by Jónsson’s group [23]. By only considering atoms in the vicinity of the diffusing atom, the inverse-distance covariance formulation significantly reduces the degrees of freedom in the high-dimensional space of the atomic systems considered in this study. More details are provided in Sec. II. Additionally, this work advances the GPR-dimer method in two key aspects. First, we introduce the use of multitask learning for the GPR surrogate model, resulting in a substantial improvement in the model’s performance and robustness compared to the previously employed derivative observation learning. The multitask GPR learning approach resembles the learning scheme used in Ref. [20], where the loss function represents a weighted average of energy and force losses. However, in the multitask approach, the contribution from force and energy losses is learned through a shared covariance function, unlike the approach in Ref. [20], which requires prior knowledge of the contribution of each loss. Second, we demonstrate that by employing a translation-hop sampling approach (defined below), the GPR-dimer search becomes both successful and robust. As detailed in Sec. II, the GPR-dimer method iteratively updates the GPR model as the dimer walker progresses, incorporating new DFT-calculated values from the energy surface into the training data. We show that a minimum number of dimer translation steps must be hopped over before updating the GPR to ensure successful guidance of the dimer to reach the saddle point. We refer to this approach as the *translation-hop sampling* approach. This sampling strategy strikes a balance between an overfitted and underfitted surrogate model. Sampling at every translation step leads to a GPR surrogate model that is overfitted to the dimer walk path on the PES, while skipping too many translation steps results in an underfitted model. A detailed discussion is provided in Sec. III. Denzel and Kästner discuss a similar balance between interpolation

and extrapolation with the use of an overshooting approach for sampling the GPR for geometry optimization [28] (not for saddle-point search). The remainder of this article is organized as follows: In Sec. II, we explain the GPR-dimer method developed in this study. In Sec. III, we validate the predictions of our GPR-dimer method for diffusivity coefficient of monovacancy diffusion in bulk *bcc* Mo and the activation energy for sulfur diffusion in  $MoS_2$ . Subsequently, in Sec. III, we elucidate the role of different factors in enhancing the performance of the GPR-dimer saddle-point search method. Finally, we compare the computational cost of the standard dimer method against our implementation of the GPR-dimer method. In Sec. IV, we provide a general interpretation of the numerical experiments using the GPR-dimer method within the context of GPR learning and the dimer search algorithm.

## II. METHOD

The approach to accelerate the dimer walk using GPR is based on a simple premise: GPR serves as a surrogate model for the computationally intensive sampling of the potential energy surface typically through methods like DFT. Once trained, the surrogate model can readily provide the energy values and their gradients (forces) at unsampled locations of the energy surface. The interaction between GPR and the dimer takes place through an iterative feedback loop: GPR provides estimates of the energy and its gradient along the dimer walk, while the dimer walker contributes new points (i.e., atomic configurations) on the energy surface. These new points are sampled through DFT and then used to update (retrain) the GPR. Through this iterative process, the dimer gradually converges towards the saddle point. The next two subsections delve into the design and training of the GPR, as well as the communication between the GPR and the dimer.

### A. Gaussian process regression (GPR) surrogate model

#### 1. Training data set

The GPR is initially trained on  $n_i$  atomic configurations. These atomic configurations are collected from the first  $n_i$  translation steps of a DFT-guided dimer walk (i.e., standard dimer). The standard dimer is launched from an atomic configuration which is estimated to be in the vicinity of the saddle point using a geometric interpolation (as detailed in Sec. III). The atomic configurations constitute the input space and the DFT-calculated atomic forces and energies constitute the target values in the training data set (see more details below). The training data set is expanded as the dimer progresses by adding a new DFT-calculated data at every  $n_h$  translation steps of the dimer walk. We call this approach the translation-hop sampling method. The effect of different  $n_h$  values are examined and explained in Sec. III. For both studies of *bcc* Mo self-diffusion and sulfur diffusion in *hex*  $MoS_2$ , we use three atomic configurations for the initial training of the GPR ( $n_i = 3$ ) and we hop over 10 translation steps before adding a new DFT calculation to the training data ( $n_h = 10$ ).

The DFT calculation of energy and forces are performed using the Vienna *Ab Initio* Simulation Package (VASP) [29], which employs the projector-augmented-wave (PAW) method [30] and the generalized gradient approximation (GGA) for

exchange-correlation energy in the Perdew-Burke-Ernzerhof (PBE) form [31]. For *bcc* Mo, we use a  $3 \times 3 \times 3$  supercell of the conventional *bcc* unit cell with 54 atoms. We use a Monkhorst-Pack *k*-point mesh of  $5 \times 5 \times 5$  and an energy cut-off of 520, respectively, within the PBE exchange-correlation functional. For *hex* MoS<sub>2</sub>, we use a  $2 \times 2 \times 2$  supercell of the convectional *hexagonal* unit cell with 48 atoms. We use a Monkhorst-Pack *k*-point mesh of  $5 \times 5 \times 1$  and an energy cut-off of 520, respectively, within the PBE exchange-correlation functional.

## 2. GPR covariance function

Choosing an appropriate covariance function is crucial for GPR performance [32]. Here, we use the inverse distance covariance function of Ref. [23], which demonstrates superior performance compared to the radial basis function (RBF) or its variants (e.g., Matérn) as shown in Ref. [24]. Compared to a stationary covariance function such as RBF, the inverse distance covariance function can better capture the asymmetry of interatomic forces, specifically the large repulsive forces caused when atoms get close to each other. This is because the inverse distance difference measure [i.e.,  $\mathcal{D}_{1/r}(\mathbf{x}, \mathbf{x}')$  term in Eq. (1)] stretches when atoms approach each other. This makes the covariance function nonstationary with respect to the atom coordinates and allows faster variation of energy in those directions (see more details in Ref. [24]). The inverse distance covariance function measures the similarity of two input atomic coordinates  $\mathbf{x}$  and  $\mathbf{x}'$  as [23]

$$k_{1/r}(\mathbf{x}, \mathbf{x}') = \sigma_c^2 + \sigma_m^2 \exp \left( - \underbrace{\frac{1}{2} \sum_{i \in A_m} \sum_{\substack{j \in A_m, j > i \\ j \in A_f}} \frac{\left( \frac{1}{r_{i,j}(\mathbf{x})} - \frac{1}{r_{i,j}(\mathbf{x}')} \right)^2}{l_{\phi(i,j)}^2}}_{\mathcal{D}_{1/r}(\mathbf{x}, \mathbf{x}')} \right). \quad (1)$$

Here,  $\mathbf{x}$  (or  $\mathbf{x}'$ ) denotes a  $3N$ -dimensional configuration vector including the Cartesian coordinates of the atomic system with  $N$  atoms,  $\mathbf{x} = [x_{11}, x_{12}, x_{13}, \dots, x_{N1}, x_{N2}, x_{N3}]^T$ .  $r_{i,j}$  is the distance between atoms  $i$  and  $j$ , defined as  $r_{i,j} = \sqrt{\sum_{d=1}^3 (x_{id} - x_{jd})^2}$ .  $l_{\phi(i,j)}$  denotes the length scale for the atom pair  $\phi(i,j)$ .  $\sigma_m$  controls the magnitude of the covariance function, and  $\sigma_c$  is the variance of a constant Gaussian prior distribution. The  $\mathbf{l}_{\phi}$  vector (with a size of the number of atomic pairs),  $\sigma_m$ , and  $\sigma_c$  are the training parameters of the inverse distance covariance function. As shown in Eq. (1), index  $i$  runs over moving atoms  $A_m$  and index  $j$  runs over other moving atoms and frozen atoms  $A_f$ . Therefore, atom pairs are only defined between moving atoms and the rest of the moving and frozen atoms. This construct reduces the total number of pairs (i.e., the size of vector  $\mathbf{l}_{\phi}$ ) from  $\frac{1}{2}N \times (N - 1)$  to  $\frac{1}{2}N_m \times (N_m - 1) + N_m \times N_f$ , where  $N$ ,  $N_m$ , and  $N_f$  denote the number of all atoms, moving atoms, and frozen atoms, respectively. In the examples of this study, we define the diffusing atom to be the moving atom ( $N_m = 1$ ), and frozen atoms

are those confined in a sphere of radius  $r_f$  centered around the moving atom. We call this spherical region the *active region*. We examine the effect of different  $r_f$  values on GPR performance in Sec. III. The partitioning of the atomic system into the moving and frozen atoms is specially advantageous in reducing the number of degrees of freedom (i.e., the number of atomic pairs or size of  $\mathbf{l}_{\phi}$  vector) for the solid-state phases in our study. Additionally, by only including the atomic pair distances between a moving atom and frozen atoms in the covariance function, we inform the GPR model with the most physically important atomic pairs. In other words, the atomic pair distances formed between nonmoving atoms carry less physically significant information in describing the potential energy surface. This physical knowledge embedded into the construct of the covariance function aids the model to learn the energy surface more effectively.

## 3. GPR training and prediction

For training the GPR, we adopt a multitask learning approach [33,34] as implemented in GPYtorch [35], which enables simultaneous learning of energy and forces by sharing information across the prediction tasks. As we show in Sec. III, multitask learning outperforms derivative observation learning by reducing the GPR training time and error and enhancing its performance and robustness. Through multitask GPR [33] the intertask dependencies are learned based solely on the task identities and the observed data for each task, unlike the derivative observation GPR which explicitly enforces the dependence of forces and energy values by equating forces to the negative derivative of energy. For details of the derivative observation GPR learning approach, see Eqs. (28) and (29) of Ref. [24]. For multitask GPR learning, a shared covariance function between tasks  $t_1$  and  $t_2$  is defined for two inputs  $\mathbf{x}$  and  $\mathbf{x}'$  as [33]

$$k([\mathbf{x}, t_1], [\mathbf{x}', t_2]) = k_{\text{input}}(\mathbf{x}, \mathbf{x}') \times k_{\text{task}}(t_1, t_2), \quad (2)$$

where  $k_{\text{input}}$  is the inverse distance covariance function defined in Eq. (1) and  $k_{\text{task}}$  is the intertask similarity measure describing the correlation between tasks. In this study, the related tasks are the prediction of energy and atomic force components. The prediction of each force component is a separate task, thus the total number of tasks is  $1 + 3N$  for  $N$  atoms in the system. Following the multitask learning approach of Ref. [33],  $k_{\text{task}}$  is defined as a “free-form” task-similarity matrix, instead of a parametric covariance function. Specifically,  $k_{\text{task}}$  is defined as a positive semidefinite matrix which is approximated by an incomplete-Cholesky decomposition of rank  $P$ . Here, we use rank 1 for approximating  $k_{\text{task}}$ , resulting in only one additional trainable parameter of the GPR. More details about the parameterization of the task-similarity matrix are given in Ref. [33].

Given a set of  $M$  tasks and  $D$  training data points (or observations), the shared covariance matrix  $\mathbf{K} \in \mathbb{R}^{DM \times DM}$  can be expressed as the Kronecker product of the input covariance matrix  $\mathbf{K}_{\text{input}} \in \mathbb{R}^{D \times D}$  and the task covariance matrix  $\mathbf{K}_{\text{task}} \in \mathbb{R}^{M \times M}$ :

$$\mathbf{K} = \mathbf{K}_{\text{task}} \otimes \mathbf{K}_{\text{input}}. \quad (3)$$

Here, the  $D$  distinct observations constitute the training data set  $\{\mathbf{X}, \mathbf{Y}\}$ .  $\mathbf{X}$  consists of input atomic configurations,  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_D\}$ , and  $\mathbf{Y}$  consists of DFT-evaluated energy and force components at  $\mathbf{X}$ ,  $\mathbf{Y} = (E_1, \dots, E_D, \mathbf{f}_1, \dots, \mathbf{f}_D)^T$ , where  $E$  and  $\mathbf{f} = [f_{1,1}, f_{1,2}, f_{1,3}, \dots, f_{N,1}, f_{N,2}, f_{N,3}]$  are the energy and force vector for each input atomic configuration, respectively. The set of trainable parameters  $\theta = \{\sigma_c, \sigma_m, \mathbf{l}_\phi\}$  and the single parameter of matrix  $\mathbf{K}_{\text{task}}$  are optimized by maximizing the log marginal likelihood over the shared covariance function [32]

$$\operatorname{argmax}_{\theta, \mathbf{K}_{\text{task}}} \left\{ \mathcal{L} = -\frac{1}{2} \mathbf{Y}^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{Y} - \frac{1}{2} \log |\mathbf{K} + \sigma_n^2 \mathbf{I}| - \frac{D}{2} \log 2\pi \right\}. \quad (4)$$

Here,  $\mathbf{K}$  is the shared covariance matrix of Eq. (3),  $\mathbf{I}$  is the identity matrix, and  $\sigma_n^2$  is the random noise variance, which we set to  $10^{-4}$ .

The GP approximation for the energy or each force component is then obtained as the mean prediction on a new data-point  $x^*$  for task  $l$  using the posterior distribution conditional on the optimized parameters:

$$f_l(x^*) = (k_{\text{task}}^l \otimes k_{\text{input}}^*) (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{Y}, \quad (5)$$

where  $k_{\text{task}}^l$  denotes the  $l$ th column of  $\mathbf{K}_{\text{task}}$  and  $k_{\text{input}}^*$  is the vector of covariances between the query point  $x^*$  and the training points.

### B. GPR-accelerated dimer

In this study, we use the dimer saddle-point search method as detailed in Ref. [6]. A dimer contains a pair of two auxiliary points in the atomic configuration space of dimension  $3N$  (or images), separated by a fixed distance of  $0.1 \text{ \AA}$ . Each dimer iteration is divided into a set of rotation steps and a translation step. During the rotation steps, the dimer is rotated around its midpoint to find the orientation that gives the lowest total energy of the two images. This gives the direction of the lowest curvature mode or the minimum mode. The dimer is then translated by reversing the force components in the minimum mode direction multiplied by a step size of  $0.1 \text{ \AA}$ . Details of the dimer algorithm used in this work are presented in Ref. [6]. We used Algorithm B1 and Algorithm B3 of Ref. [6], respectively, for rotation and translation of the dimer.

For a GPR-guided dimer, the GPR-approximation of the energy and force components [according to Eq. (5)] are used to provide the forces acting on the images of the dimer during rotation and translation. Rotational forces are then defined according to the projected atomic forces on the two images of the dimer as  $F_R$ . Rotational steps are carried out until  $F_R$  falls below a threshold ( $0.1 \text{ eV/\AA}$ ) or a maximum number of rotations are performed. The maximum number of rotations is set to 5 for the standard dimer (our code) and to 25 for the GPR-dimer. We use the conjugate gradient algorithm for determining the rotational plane of the dimer. As explained in the previous section, the GPR model is updated (retrained) after every  $n_h$  translation steps by using an expanded set of DFT-calculated training data points. The final convergence of the dimer to the saddle point is achieved when the

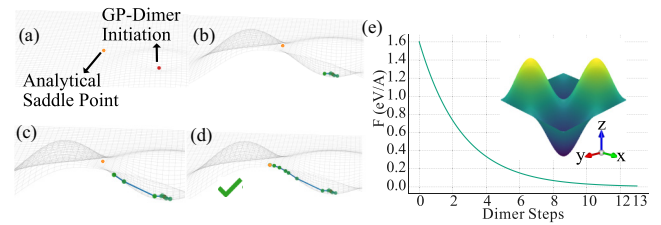


FIG. 1. The evolution of the GPR approximation of the potential energy surface and the dimer location on the GPR-approximated energy surface at (a) the initial point, and after (b) 4, (c) 5, and (d) 13 dimer translation steps. (d) The final step of the GPR-dimer, where it reaches the exact saddle point at  $x = 0$  and  $y = 0$ . (e) Force magnitude (i.e., derived from the negative gradient of the GPR-approximated potential energy) versus the GPR-dimer steps. The inset illustrates the 2D potential model,  $z = -\sin \pi x \sin \pi y$ .

maximum atomic force approximated by the GPR is below  $0.01 \text{ eV/atom}$ . An accurate DFT calculation of the force at the final point of dimer is performed to confirm the convergence to the saddle point.

## III. RESULTS

### A. Validation: Two-dimensional sinusoidal potential model

We first validate the GPR-dimer method of this work, as detailed in Sec. II, on a toy potential model of sinusoidal form. The model has the function form of  $z = -\sin \pi x \sin \pi y$ , where  $z$  denotes the potential energy value and  $x$  and  $y$  constitute the two coordinates of the input, mimicking the atomic coordinates in a two-dimensional space. We initiate the GPR-dimer from the minimum on the potential surface at  $x = -0.5$  and  $y = -0.5$ . The GPR-approximated energy surface  $\tilde{z}$  is updated (or retrained) after each dimer translation step (i.e.,  $n_h = 0$ ). The training set is expanded by a new data point,  $z_i(x_i, y_i)$ , at each dimer translation step, where  $(x_i, y_i)$  specify the 2D atomic coordinates of the new dimer location, and then the GPR is trained on the new training data. The threshold for training of the GPR is for the mean absolute error (MAE) of the force and energy to drop below  $0.01 \text{ eV/\AA}$  and  $0.01 \text{ eV}$ , respectively. Figure 1 illustrates the evolution of the GRP energy surface,  $\tilde{z}$ , and the dimer walker location at different dimer translation steps. Figure 1 also shows the decrease of the force magnitude,  $|\tilde{F}|$  (i.e.,  $|\tilde{F}| = \sqrt{\tilde{F}_x^2 + \tilde{F}_y^2}$  where  $\tilde{F}_x = -\frac{\partial \tilde{z}}{\partial x}$  and  $\tilde{F}_y = -\frac{\partial \tilde{z}}{\partial y}$ ) as the GPR-dimer progresses toward the saddle point. The GPR-dimer reaches the saddle point at  $x = 0$  and  $y = 0$  after 13 translation steps, with a total number of 60 dimer rotations. For the 2D potential model, we use the radial basis covariance function (RBF), as implemented in GPyTorch [35] and the derivative-observation learning approach [27].

### B. Validation: Self-diffusion in *bcc* Mo

To validate that the GPR-dimer method can successfully identify the transition state of a solid-state process, we apply it to calculate the energy barrier for a vacancy diffusive hop in the *bcc* phase of Mo. We initiate the GPR-dimer walker at an atomic configuration that is a linear interpolation between

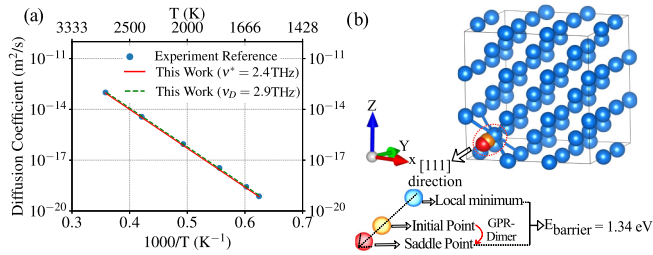


FIG. 2. (a) Diffusion coefficient versus temperature in *bcc* Mo. The predictions of this work, from DFT calculations and the GPR-dimer, are compared against experimental measurements [38]. (b) The evolution of the atomic configuration in the vicinity of the diffusing atom starting from the input atomic configuration (yellow) to the saddle point on the energy surface (red) output by the GPR-dimer. The atomic configuration of the local minimum state-vacancy residing on a *bcc* site- is also shown (blue).

the initial state (a local minimum state), where a *bcc* lattice site is vacant, and the final state (a symmetrically equivalent local minimum state), where the vacant *bcc* site has hopped to the nearest neighbor. The input configuration is a weighted average of the atomic coordinates with 3/4 contribution from the initial state and 1/4 from the final state [see Fig. 2(b)]. Supplemental Note 1, along with Supplemental Figs. S2–S5 [36], examine the GPR-dimer application using the local minimum configuration as the initial input. Given the input configuration, we perform the standard dimer for two translation steps to provide the data points for training the GPR. A total of three atomic configurations are used to train the GPR ( $n_i = 3$ ). The GPR-dimer is then launched to locate the saddle point. We use the inverse distance covariance function with an active region of size  $r_f = 3 \text{ \AA}$ . For training the GPR, we employ the multitask learning as detailed in Sec. II. The training data set is expanded by an additional atomic configuration at every 10th translation step of the dimer (i.e.,  $n_h = 10$ ), followed by an update (or retraining) of the GPR. The criterion of reaching the saddle point is for the total force magnitude to be less than  $0.01 \text{ eV/\AA}$ , where the GPR-dimer stops. The total force magnitude is calculated as  $F = \sqrt{\sum_{i=1}^N F_{i1}^2 + F_{i2}^2 + F_{i3}^2}$ . The energy difference between the final step of the GPR-dimer (or the transition state) and the initial state of the vacancy hop (the local minimum) is calculated to provide the energy barrier for the vacancy diffusive hop (or the enthalpy of vacancy migration),  $\Delta H_m$ . Energies of the transition and local minimum states are both calculated using DFT. The calculated enthalpy of vacancy migration is equal to  $1.34 \text{ eV}$  [see Fig. 2(b)], which is in good agreement of our previous calculation using NEB [37].

Using the calculated enthalpy of vacancy migration,  $\Delta H_m$ , we validate the diffusion coefficient of *bcc* Mo as a function of temperature with experimental measurements [38]. Figure 2(a) shows the calculated diffusion coefficient based on the located saddle point on the energy surface by the GPR-dimer method in comparison with experimental results [38]. We calculate the self-diffusion coefficient for monovacancy diffusive jumps according to  $D = C_v d^2 \Gamma$ , where  $d$  is the vacancy (or atom) jump distance,  $C_v$  is the equilibrium vacancy concentration, and  $\Gamma$  is the successful vacancy jump rate. Vacancy jump distance in *bcc* is equal to the

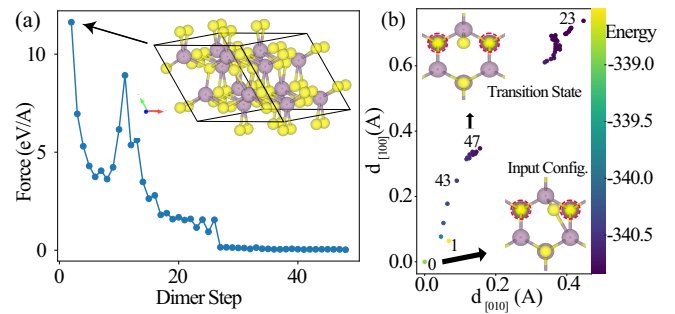


FIG. 3. Evolution of the GPR-dimer to locate the transition state of sulfur-monovacancy diffusive jump in *hex* MoS<sub>2</sub>. (a) The total force magnitude of the atomic configuration versus dimer translation step. The inset depicts the input configuration to the GPR-dimer. (b) Displacement of the diffusing sulfur atom during the GPR-dimer saddle-point search, projected along the [100] and [010] directions. The insets depict the atomic configuration in the vicinity of diffusing sulfur. Sulfur, molybdenum, and vacancy are colored by yellow, purple, and dotted red circles, respectively. The color bar maps the energy of different dimer steps.

nearest-neighbor distance or  $\frac{\sqrt{3}}{2}a_0$ , where  $a_0$  is the lattice constant. Vacancy concentration at temperature  $T$  is given by  $C_v = \exp(\frac{\Delta S_f}{k_B T}) \exp(-\frac{\Delta H_f}{k_B T})$ , where  $\Delta H_f$  and  $\Delta S_f$  are the formation enthalpy and entropy of vacancy, respectively, and  $k_B$  is the Boltzmann constant. We obtain the DFT-calculated values of  $a_0$  and  $C_v$  from our previous results in Ref. [37]. The vacancy jump rate  $\Gamma$  is obtained from the migration enthalpy,  $\Delta H_m$ , and the effective vibration frequency along the migration path,  $\nu^*$ , by  $\Gamma = \nu^* \exp(-\frac{\Delta H_m}{k_B T})$ . Here,  $\Delta H_m$  is the vacancy migration energy barrier calculated according to the saddle point located by the GPR-dimer. We obtain the DFT-calculated  $\nu^*$  from Ref. [37], which calculates  $\nu^*$  as the ratio of the product of normal vibration frequencies of the initial state of atomic migration,  $\nu_i$ , to that of the non-imaginary normal frequencies of the transition state,  $\nu'_j$ , i.e.,  $\nu^* = \frac{\prod_{i=1}^{3N-3} \nu_i}{\prod_{j=1}^{3N-4} \nu'_j}$ . The obtained value is  $2.4 \text{ THz}$ . Alternatively, we estimate  $\nu^*$  to be equal to the Debye frequency,  $\nu^* \approx \nu_D = 2.9 \text{ THz}$ , which is calculated from Debye temperature  $\Theta_D$  as  $\nu_D = \Theta_D \frac{k_B}{\hbar}$ , where  $\hbar$  is the reduced Planck's constant.

### C. Validation: Sulfur diffusion in *hex* MoS<sub>2</sub>

To further validate the accuracy of the GPR-dimer method in identifying transition states in solid-state processes, we apply it to calculate the energy barrier for the monovacancy-sulfur diffusive jumps in *hexagonal* MoS<sub>2</sub> (P6<sub>3</sub>/mmc space group with 2b and 4f wyckoff positions for Mo and S, respectively). The input atomic coordinates to the GPR-dimer are linearly interpolated with 3/4 contribution from the local minimum configuration, where one sulfur site is vacant, and 1/4 from the final state, where the vacancy and the nearest sulfur has exchanged their positions (see Fig. 3 for the input configuration). Subsequently, we launch the GPR-dimer method to locate the saddle point along the diffusion pathway. Like the example of *bcc* Mo, we use the first three atomic configurations from the translation steps of standard dimer to train the GPR (i.e.,  $n_i = 3$ ), and we consequently update

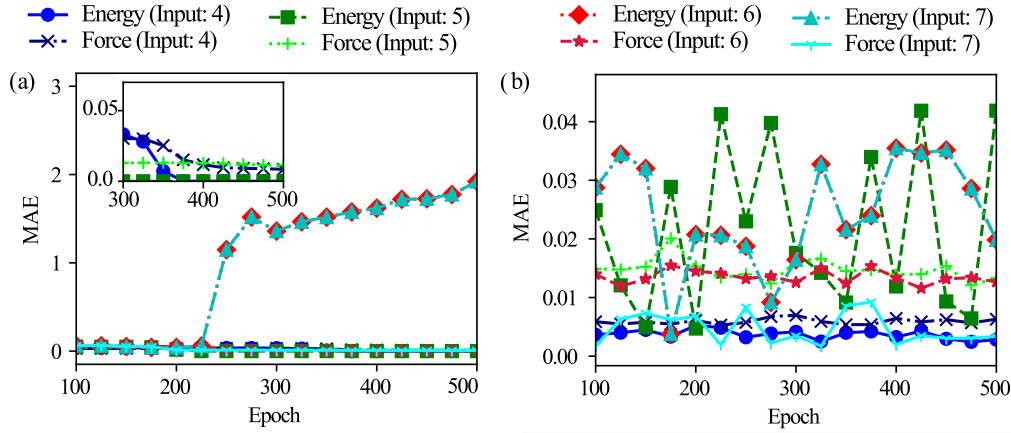


FIG. 4. Comparison of the gradient observation and the multitask learning on the GPR predictive performance. The mean absolute error (MAE) of the energy prediction (eV) and the atomic force magnitude prediction (eV/Å) in terms of the training epoch for (a) gradient observation and (b) multitask learning. Different line styles and markers indicate various sizes of the training data.

(or retrain) the GPR after every 10 translation steps (i.e.,  $n_h = 10$ ). We use an active region of radius  $r_f = 5$  Å centered at the diffusing sulfur atom for the inverse-distance covariance function of Eq. (1). Figure 3(a) shows the evolution of the total force magnitude of the atomic configuration at different translation steps of the GPR-dimer. Once the total force is below 0.01 eV/Å, we stop the GPR dimer. Figure 3(b) shows the displacement of the diffusing atom projected on the [100] and [010] directions during the GPR-dimer evolution. The energy of the output of the GPR-dimer (or the transition state) and the local minimum configuration are calculated using DFT. The calculated energy difference between these two configurations equals 2.3 eV, which is in good agreement with the DFT-NEB calculation of Ref. [39].

#### D. GPR-dimer performance analysis

In this section, we examine the impact of various parameters on the performance of the GPR-dimer method introduced in this study. These parameters include: (1) the learning approach or the GPR training method, (2) the energy surface sampling frequency or the number of dimer translation hops between DFT-calculations, (3) the size of the active region containing frozen atoms in the inverse-distance covariance function of Eq. (1), and (4) the number of preceding DFT-sampled data points included in GPR training set at each retraining step. We present our study on the diffusive jump of a monovacancy in *bcc* Mo.

To evaluate the impact of different GPR learning approaches, we compare multitask learning (explained in Sec. II) with the commonly used derivative-observation learning. Figure 4 illustrates the learning curves for multitask learning and derivative-observation learning. The GPR serves as the initial surrogate model for the energy surface, and the training data consist of  $n_i$  consecutive DFT-calculated dimer translation steps starting from the initial atomic configuration (as detailed in Sec. II). The mean absolute error (MAE) for energy and force predictions are presented over the training epoch for various numbers of input training data,  $n_i$ . The multitask learning method demonstrates superior learning performance in terms of prediction accuracy and stability

compared to derivative-observation learning. The multitask learned GPR consistently exhibits lower MAE across all epochs and input data sizes. Furthermore, the MAE remains low and relatively stable for different input data sizes, indicating the robustness of multitask learning and its lower sensitivity to observations (or training data). In contrast, the GPR trained with derivative-observation learning shows an increase in MAE as the number of input data increases. Specifically, the MAE for energy prediction starts to rise significantly above 1 eV after around 200 epochs, while the MAE for force prediction remains low. This behavior is likely attributed to overfitting of the derivative-observation GPR to a single task. The explicit relationship between forces and energies in derivative-observation learning constrains the optimization process, making it prone to issues such as overfitting to forces in this particular example. However, multitask learning employs an implicit regularization effect, mitigating the risk of overfitting to a single task. By simultaneously learning multiple correlated tasks, the model captures common latent features that generalize better to new data.

We assess the impact of the number of translation hops, denoted as  $n_h$ , in the translation-hop sampling approach described in Sec. II, on the performance of GPR-dimer. In Fig. 5, we present the behavior of the GPR-dimer for different values of  $n_h$  ( $n_h = 0, 1, 3, 5, 10$ ). The figure illustrates the DFT-calculated total atomic force magnitude and the force magnitude of the diffusing atom as a function of the dimer translation step. The force values are only displayed for the dimer translation steps where the GPR is updated, corresponding to the atomic configurations associated with the DFT-calculated steps. The GPR-predicted forces at the intermediate steps between DFT sampling points are not shown. Based on our examination, we observe that the GPR-dimer with zero, one, and three translation hops failed to locate the saddle point. In contrast, the GPR-dimer with five and ten translation hops successfully converged to the saddle point. This observation provides valuable insights into how the sampling frequency along the dimer path influences the GPR surrogate model, striking a balance between a localized and global representation of the energy surface. For zero, one, and three translation hops, the GPR becomes excessively

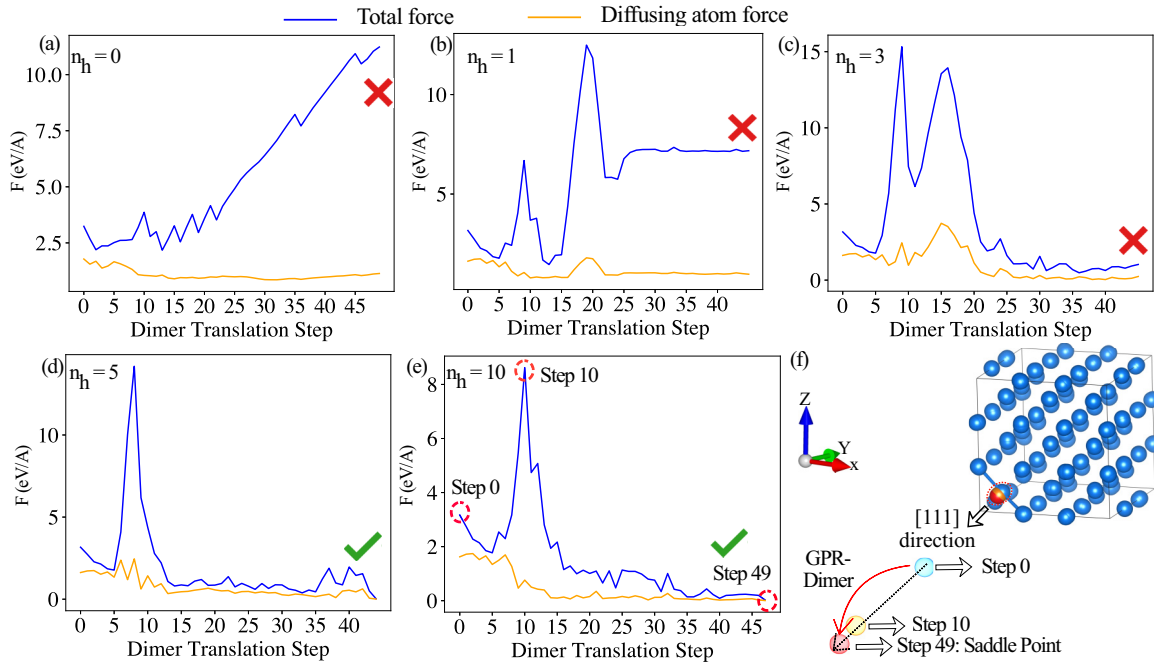


FIG. 5. Assessing the performance of the GPR-dimer for different translation-hop values in the translation-hop sampling approach. The evolution of the total atomic force and the diffusing atom force in terms of dimer translation steps for (a) zero, (b) one, (c) three, (d) five, and (e) ten translation hops. The  $x$  axis represents the GPR-dimer translation steps only where the GPR is updated with new DFT-calculated data points. (f) Evolution of the atomic configuration in the vicinity of the diffusing atom at different steps for the GPR-dimer with  $n_h = 10$ .

influenced by the energy surface in the vicinity of the dimer path. Consequently, it overfits to the minimum-mode following path while neglecting the broader energy landscape. However, delaying the sampling by 5 or 10 translation hops enables the GPR to capture a more balanced representation, encompassing both the vicinity of the path and the wider energy surface shape. In other words, the translation-hop sampling approach provides the opportunity to balance exploration and exploitation through tuning the  $n_h$  parameter. It is worth noting that in the case of a 2D potential model (depicted in Fig. 1), no translation hops are required. The GPR demonstrates robustness against overfitting due to the low dimensionality of the input space of the covariance function.

To further investigate the influence of translation-hop sampling frequency on the GPR model, we present the energy profile of the GPR-dimer for different translation hops in Fig. 6. The minimum energy pathway, obtained from NEB calculations implemented in VTST [7], is shown as a reference. The movement of the GPR-dimer is projected along the minimum energy path direction (or along the [111] lattice direction), which serves as the  $x$  axis in Fig. 6. In the case of zero hops [Fig. 6(a)], the GPR-dimer bypasses the saddle point and explores high-ridge regions of the energy surface. This is because the GPR is overfitted to the walker pathway and most likely to the noise in the initial dimer walker oscillations, which results in misguiding the walker. For one or three hops [Figs. 6(b) and 6(c)], the walker goes back and forth between lower and higher energy regions but fails to locate the saddle point. In contrast, for five and ten hops [Figs. 6(d) and 6(e)], the GPR-dimer walker deviates from the NEB path at the beginning, as the GPR's accuracy in predicting energies near the pathway is reduced. The walker takes larger steps

and explores a diverse range of points on the energy surface, eventually converging to the saddle point as it progresses and incorporates more sampled data points. Supplemental Fig. S1 [36] illustrates a 2D projection of the GPR-dimer's diffusing atom trajectory on the high-dimensional energy surface for different translation hops.

We investigate the impact of the active region radius, denoted as  $r_f$  (see Sec. II for details), on the overall performance of the GPR-dimer. Increasing the active region radius results in more frozen atoms in the inverse distance covariance function of Eq. (1). This leads to more pairs between the moving atom (or the diffusing atom) and the frozen atoms, providing the GPR with more information about the surrounding atomic configuration. Figure 7 illustrates the GPR-dimer's behavior for  $r_f$  values of 3, 5, and 7 Å (associated with the first, second, and third nearest neighbors of the moving atom, respectively), with a fixed translation hop of  $n_h = 10$ . The DFT-calculated total atomic force and diffusing atom force are shown as a function of the GPR-dimer step. In all three cases, the GPR-dimer successfully locates the saddle point. However, as shown in Fig. 7, the force evolution is smoother for  $r_f = 3$  Å compared to larger active regions. The force exhibits an early peak and a monotonic decrease as the GPR-dimer progresses towards the saddle point. In contrast, for  $r_f = 5$  Å and  $r_f = 7$  Å, the force demonstrates significant oscillations throughout the GPR-dimer progression, with  $r_f = 7$  Å displaying two force peaks before reaching the saddle point. The smooth evolution of the GPR-dimer observed at  $r_f = 3$  Å provides valuable insights into achieving an optimal balance between the number of atomic pairs incorporated in the GPR's covariance function and the captured physical information. By setting  $r_f = 3$  Å, the covariance function effectively captures

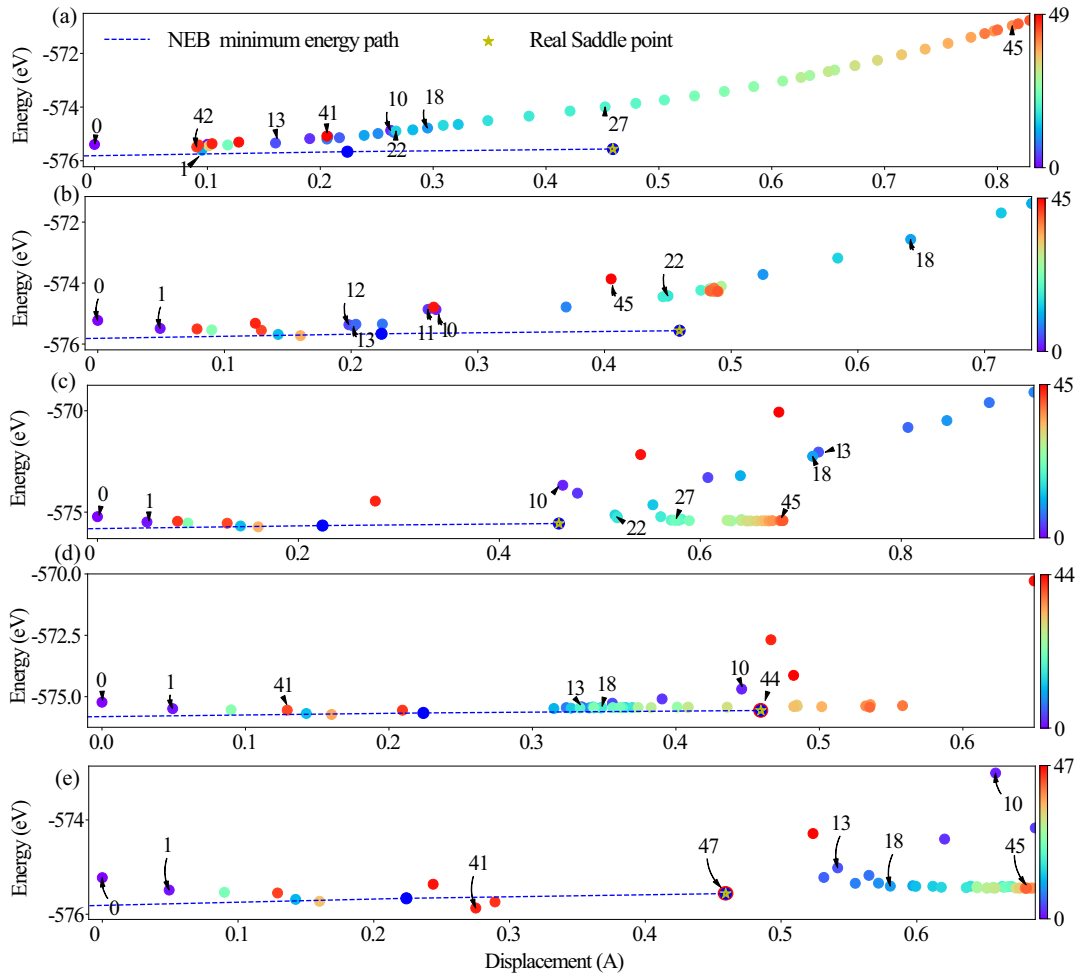


FIG. 6. The energy profile of the GPR-dimer for different translation hops. Energy of the GPR-dimer versus its displacement along [111] direction for (a) zero, (b) one, (c) three, (d) five, and (e) ten translation hops. The color bar depicts the progression of the GPR-dimer walker.

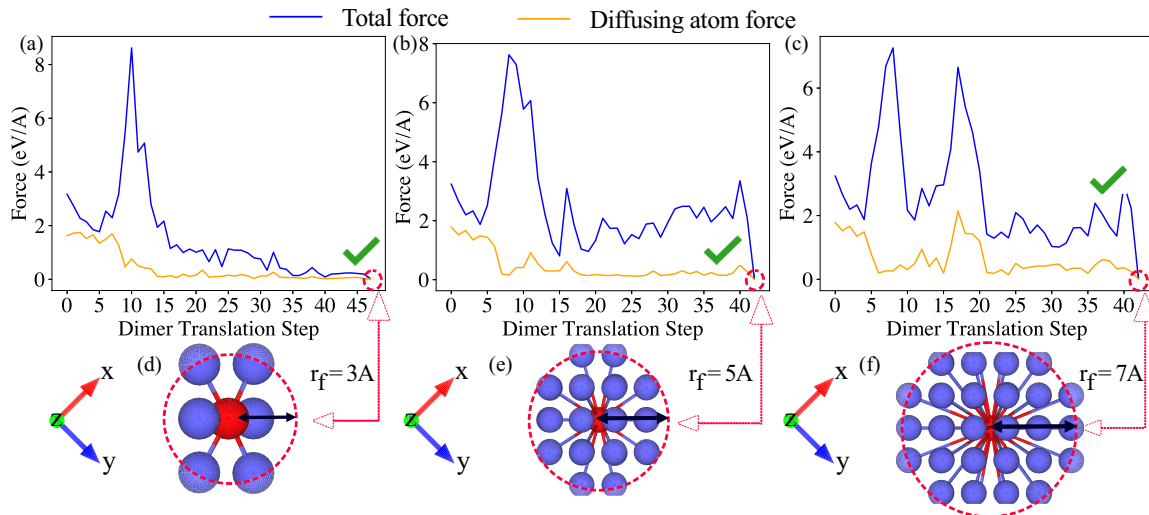


FIG. 7. The impact of active region radius on the GPR-dimer performance. DFT-calculated total atomic force and diffusing atom force for radii of (a) 3 Å, (b) 5 Å, and (c) 7 Å. The x axis represents the GPR-dimer translation steps at which GPR is updated (i.e., DFT calculation is performed). (d-f) The atomic configuration surrounding the diffusing atom, shown for the final GPR-dimer step in each case.



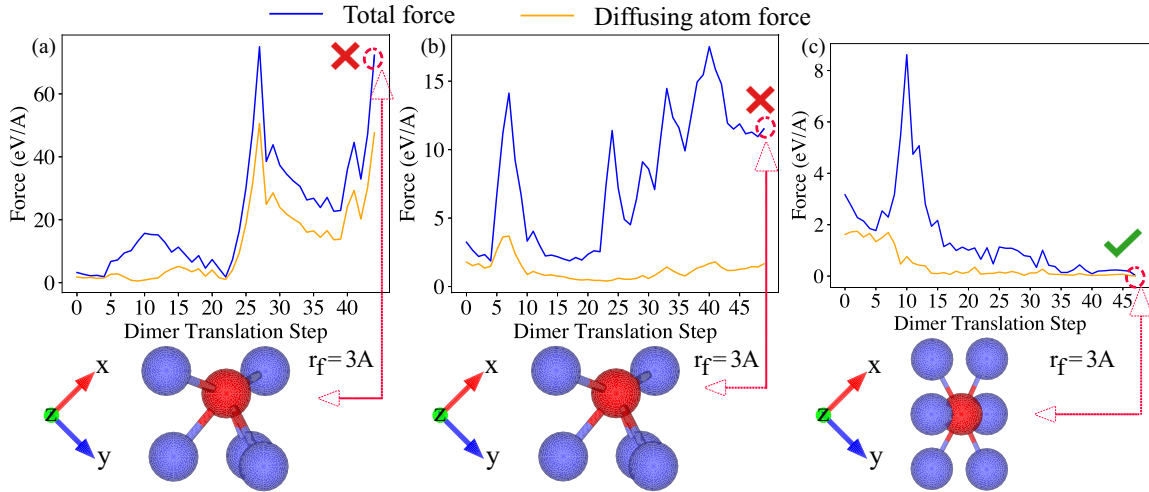


FIG. 8. Effect of GPR-dimer tail size (or the number of preceding dimer translation steps incorporated in the training set) on the GPR-dimer performance. The DFT-calculated total force and diffusing atom force are shown for a tail size of (a) 5, (b) 10, and (c) all preceding dimer steps. The x-axis shows the GPR-dimer translation steps at which the GPR is updated. The atomic configuration around the diffusing atom is shown for the last GPR-dimer step for each case.

variations in the pair distance between the diffusing atom and its first nearest neighbors throughout the progression of the GPR-dimer. Constraining the pair distance information to the first nearest neighbors proves to be the most effective approach for constructing a surrogate model. This is because the pair distances among the nearest neighbors contain the most relevant physical information while maintaining a relatively low number of pairs. Consequently, this results in a smaller size of the  $\mathbf{I}_\phi$  vector, reducing the risk of overfitting and improving the model's performance.

Last, we investigate the influence of the GPR-dimer's training data history on its performance. The GPR-dimer's tail size is defined as the last  $n_t$  translation steps preceding the current step, which are used as training data for updating the GPR surrogate model. We consider three tail sizes:  $n_t = 5$ ,  $n_t = 10$ , and  $n_t$  including all preceding DFT-sampled steps. The translation hop is set to  $n_h = 10$ , indicating 10 dimer translations between consecutive DFT-sampled steps. The active region radius is fixed at 3 Å ( $r_f = 3$  Å). Figure 8 presents the GPR-dimer evolution for different tail sizes  $n_t$ . The DFT-calculated total atomic force magnitude and diffusing atom force magnitude are shown as a function of the GPR-dimer progression. The GPR-dimer successfully locates the saddle point only when all preceding DFT-sampled steps are included in the training set. This observation aligns with our previous analysis of the translation-hop frequency. Limiting the GPR training data to the last 5 or 10 preceding steps results in a surrogate model representing a local view of the energy surface, causing the walker to bypass the saddle point and move towards high-energy ridges. This is evident from the large force magnitudes observed in Figs. 8(a) and 8(b). Conversely, updating the GPR using the entire history of the dimer walker allows the model to capture a broader view of the energy surface. Our examination reveals that excluding atomic configurations with large repulsive forces (force peaks in the early steps of the GPR-dimer) prevents the GPR from gaining a comprehensive understanding of the energy surface. Therefore, updating the GPR with a diverse set of low- and high-energy

points sampled along the walk is necessary for the successful identification of the saddle point by the GPR-dimer.

### E. Assessing computational efficiency

To assess the potential computational efficiency gains achievable by employing a GPR surrogate model with the dimer method, we compare the computational efforts required to locate the saddle point for self-diffusion in *bcc* Mo using three different approaches: the standard dimer as implemented in our work, the GPR-dimer as implemented in our work, and the dimer implementation of the transition state tools for VASP (VTST) [7]. Table I presents a comparison of the total number of DFT calculations of energy and forces necessary for each method to converge to the saddle point. In the standard dimer method (both in our implementation and the VTST code), DFT calculations are performed at every dimer step—during both the translation and rotation steps. For the GPR-dimer, however, forces are obtained from the GPR surrogate model during rotations, and DFT force calculations are only carried out at every six dimer translation steps, when the GPR is updated ( $n_h = 5$ ). The GPR surrogate model is initially trained on three DFT-calculated configurations (i.e.,  $n_i = 3$ ). Table I also includes the number of node-hours required for the DFT calculations by each method, conducted on 2 AMD EPYC 7742 CPUs with 64 cores. As illustrated

TABLE I. The number of DFT calculations and computational effort for standard dimer versus GPR dimer for vacancy diffusion in *bcc* Mo. For GPR-dimer, the active region cutoff radius is 3 Å and the translation hop is set to 5 ( $r_f = 3$  Å and  $n_h = 5$ ).

	DFT calculations	Computational time (node-hour)
Dimer (our code)	54	20.81
GPR-dimer (this work)	44	17.89
Dimer (VTST code)	60	23.1

in Table I, the GPR-dimer approach needs 44 DFT calculations (equivalent to 17.89 node hours), compared to 54 (20.81 node hours) in standard dimer (our code), showing around 15% enhancement in computational effort. The computational enhancement provided by GPR-dimer compared to the standard dimer becomes more pronounced with larger atomic systems (further details are provided below). As shown in Table I, the VTST implementation needs higher number of DFT calculations (60 calculations equivalent to 23.1 node hours) compared to our implementation of the standard dimer. This is because our implementation utilizes the more efficient conjugate gradient algorithm. Specifically, while the VTST code utilizes the original conjugate gradient (CG) algorithm [8], our code adopts a more efficient version of CG [6]. Supplemental Tables S1 and S2 [36] show the computational effort for the GPR-dimer method for different translation hops,  $n_h$ , and active region radii,  $r_f$ , respectively.

To assess the scalability of the GPR-dimer method with atomic system size and its computational efficiency, we compare the computational effort between GPR-dimer and standard dimer for the saddle-point search simulation of vacancy diffusion in *bcc* Mo for 53-atom, 127-atom, and 249-atom supercells. As shown in Supplemental Table S3 and Fig. S6 [36], for both the standard dimer and GPR-dimer, the number of required DFT calculations remains the same for different numbers of atoms (see Supplemental Table S3 [36]). However, the computational time associated with each DFT calculation by VASP grows by supercell size (see Supplemental Fig. S6 [36]). Additionally, the computational efficiency gain by GPR-dimer, compared to the standard dimer, increases for larger systems: For a 53-atom supercell, GPR-dimer reduces the computation time by 15% and 22% compared to the standard dimer in our code and VTST, respectively ( $\frac{20.81-17.89}{20.81} \approx 15\%$  and  $\frac{23.10-17.89}{23.10} \approx 22\%$ ). For 249-atom supercell, these computational enhancement grows to 18% and 26% ( $\frac{459.32-375.31}{459.32} \approx 18\%$  and  $\frac{509.87-375.31}{509.87} \approx 26\%$ ), respectively. Notably, the GPR-dimer method maintains approximately the same computational time for the non-DFT calculations (around 1 node-hour as shown in Supplemental Table S3 [36]) regardless of the size of the supercell, showing an  $\mathcal{O}(1)$  or constant time complexity. This is also evident by the horizontal line in Supplemental Fig. S6 [36]. The constant time complexity of the GPR-dimer is because of the active region approach used in defining the covariance function [see Eq. (1)]. The GPR covariance function only depends on the size of atoms in the active region surrounding the diffusing atom/vacancy (e.g., first- or second-nearest-neighbor atoms) rather than the full system. Therefore, the size of the active region is independent of the size of the full system, making the GPR-learning approach size-independent. Additionally, our results show that the presented GPR-dimer method can handle atomic systems of arbitrary sizes without any inherent limitation.

#### IV. DISCUSSION

We present a methodology that leverages Gaussian process regression (GPR) to develop a surrogate model for the *ab initio* energy surface. By integrating the dimer method with

GPR in an iterative feedback loop, we simultaneously sample the energy surface and converge to the saddle point. The versatility of our proposed GPR-dimer method is demonstrated through its successful application in identifying transition states of vacancy-mediated diffusion in both *bcc* molybdenum and *hexagonal* molybdenum disulfide. Our results indicate the promising potential of the GPR-dimer method in enhancing the efficiency of saddle-point search in solid-state materials characterized by a large number of atoms.

The presented GPR-dimer method is specifically tailored for efficient identification of transition states in solid-state materials. It excels in accurately and rapidly locating saddle points based on the regeneration of forces in the vicinity of the dimer pathway, derived from only a few *ab initio* data points (e.g., 40–50 data points as shown in Table I). Unlike methods such as GDML [16], sGDML [17,18], and BIGDML [19], which are designed for global force field generation over a wide range of atomic configurations, the GPR-dimer method focuses on local surrogate forces along the dimer walk path. Therefore, while sGDML is adept at generating accurate global force fields for molecular dynamics simulations, derived from large data sets and diverse structures (e.g., in the order of thousands), the GPR-dimer method is more suitable for detailed transition state analysis. Recognizing these distinctions and trade-offs is essential for choosing the most effective computational strategy in a given materials science investigation.

To establish a robust and computationally efficient GPR-dimer scheme, we introduced two key components: (1) multitask GPR learning and (2) translation-hop sampling of training data. The translation-hop sampling approach proves to be essential in striking a delicate balance between exploration and exploitation of the *ab initio* energy surface during the search for the saddle point. This approach enables effective utilization of the available training data while efficiently exploring the energy landscape. Furthermore, by applying the GPR-dimer method to solid-state materials with a high degree of atomic freedom, our findings offer valuable strategies to tackle the challenge of high-dimensionality when employing GPR.

The integration of a surrogate model, like GPR, with the dimer method not only enhances computational efficiency, as presented in this study, but also unlocks new avenues for simulating various physical aspects associated with the identification of transition states. Our ongoing research efforts are directed toward expanding the scope of this approach by incorporating temperature effects into the surrogate model for solid-state processes. By extending the functionality of the surrogate model, the presented GPR-dimer approach can be readily applied to transition state studies in high-temperature solid phases.

In summary, our methodology showcases the potential of GPR-dimer as a powerful tool for enhancing saddle-point search in solid-state materials. By integrating GPR with the dimer method and incorporating novel strategies, we pave the way for more efficient exploration of complex energy landscapes in the search for transition states.

The GPR-dimer source code is accessible upon request from the corresponding author.

## ACKNOWLEDGMENTS

This research is based upon work supported by the National Science Foundation (NSF) under Award No. DMR-1954621. This work used resources at the Extreme Science and Engineering Discovery Environment (XSEDE),

which was supported by National Science Foundation Grant No. 1548562 and from Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation Grants No. 2138259, No. 2138286, No. 2138307, No. 2137603, and No. 2138296.

- [1] H. Eyring, The activated complex in chemical reactions, *J. Chem. Phys.* **3**, 107 (1935).
- [2] G. H. Vineyard, Frequency factors and isotope effects in solid state rate processes, *J. Phys. Chem. Solids* **3**, 121 (1957).
- [3] W. F. K. Wynne-Jones and H. Eyring, The absolute rate of reactions in condensed phases, *J. Chem. Phys.* **3**, 492 (1935).
- [4] S. Kadkhodaei and A. van de Walle, A simple local expression for the prefactor in transition state theory, *J. Chem. Phys.* **150**, 144105 (2019).
- [5] H. Jónsson, G. Mills, and K. W. Jacobsen, Nudged elastic band method for finding minimum energy paths of transitions, in *Classical and Quantum Dynamics in Condensed Phase Simulations* (World Scientific, Singapore, 1998), pp. 385–404.
- [6] W. Ren and E. Vanden-Eijnden, A climbing string method for saddle point search, *J. Chem. Phys.* **138**, 134105 (2013).
- [7] G. Henkelman, B. P. Uberuaga, and H. Jónsson, A climbing image nudged elastic band method for finding saddle points and minimum energy paths, *J. Chem. Phys.* **113**, 9901 (2000).
- [8] G. Henkelman and H. Jónsson, A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives, *J. Chem. Phys.* **111**, 7010 (1999).
- [9] K. J. Caspersen and E. A. Carter, Finding transition states for crystalline solid–solid phase transformations, *Proc. Natl. Acad. Sci. USA* **102**, 6738 (2005).
- [10] D. Sheppard, R. Terrell, and G. Henkelman, Optimization methods for finding minimum energy paths, *J. Chem. Phys.* **128**, 134106 (2008).
- [11] G. Henkelman and H. Jónsson, Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points, *J. Chem. Phys.* **113**, 9978 (2000).
- [12] Y. Zeng, P. Xiao, and G. Henkelman, Unification of algorithms for minimum mode optimization, *J. Chem. Phys.* **140**, 044115 (2014).
- [13] R. A. Olsen, G. J. Kroes, G. Henkelman, A. Arnaldsson, and H. Jónsson, Comparison of methods for finding saddle points without knowledge of the final states, *J. Chem. Phys.* **121**, 9776 (2004).
- [14] P. Xiao, D. Sheppard, J. Rogal, and G. Henkelman, Solid-state dimer method for calculating solid-solid phase transitions, *J. Chem. Phys.* **140**, 174104 (2014).
- [15] M. Plasencia Gutiérrez, C. Argáez, and H. Jónsson, Improved minimum mode following method for finding first order saddle points, *J. Chem. Theory Comput.* **13**, 125 (2017).
- [16] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Machine learning of accurate energy-conserving molecular force field, *Sci. Adv.* **3**, e1603015 (2017).
- [17] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, *Nat. Commun.* **9**, 3887 (2018).
- [18] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, sGDML: Constructing accurate and data efficient molecular force fields using machine learning, *Comput. Phys. Commun.* **240**, 38 (2019).
- [19] H. E. Sauceda, L. E. Gálvez-González, S. Chmiela, L. O. Paz-Borbón, K.-R. Müller, and A. Tkatchenko, BIGDML—Towards accurate quantum machine learning force fields for materials, *Nat. Commun.* **13**, 3733 (2022).
- [20] O.-P. Koistinen, F. B. Dagbjartsdóttir, V. Ásgeirsson, A. Vehtari, and H. Jónsson, Nudged elastic band calculations accelerated with Gaussian process regression, *J. Chem. Phys.* **147**, 152720 (2017).
- [21] A. Denzel and J. Kästner, Gaussian process regression for transition state search, *J. Chem. Theory Comput.* **14**, 5777 (2018).
- [22] A. Denzel, B. Haasdonk, and J. Kästner, Gaussian process regression for minimum energy path optimization and transition state search, *J. Phys. Chem. A* **123**, 9600 (2019).
- [23] O.-P. Koistinen, V. Ásgeirsson, A. Vehtari, and H. Jónsson, Minimum mode saddle point searches using Gaussian process regression with inverse-distance covariance function, *J. Chem. Theory Comput.* **16**, 499 (2020).
- [24] O.-P. Koistinen, V. Ásgeirsson, A. Vehtari, and H. Jónsson, Nudged elastic band calculations accelerated with Gaussian process regression based on inverse interatomic distances, *J. Chem. Theory Comput.* **15**, 6738 (2019).
- [25] A. Denzel and J. Kästner, Hessian matrix update scheme for transition state search based on Gaussian process regression, *J. Chem. Theory Comput.* **16**, 5083 (2020).
- [26] J. Baker and F. Chan, The location of transition states: A comparison of cartesian, z-matrix, and natural internal coordinates, *J. Comput. Chem.* **17**, 888 (1996).
- [27] E. Solak, R. Murray-Smith, W. Leithead, D. Leith, and C. Rasmussen, Derivative observations in Gaussian process models of dynamic systems, in *Advances in Neural Information Processing Systems*, edited by S. Becker, S. Thrun, and K. Obermayer (MIT Press, Cambridge, MA, 2002), Vol. 15.
- [28] A. Denzel and J. Kästner, Gaussian process regression for geometry optimization, *J. Chem. Phys.* **148**, 094114 (2018).
- [29] G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, *Phys. Rev. B* **54**, 11169 (1996).
- [30] P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B* **50**, 17953 (1994).
- [31] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [32] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2005).

- [33] E. V. Bonilla, K. Chai, and C. Williams, Multi-task Gaussian process prediction, in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Red Hook, NY, 2007), Vol. 20.
- [34] K. Swersky, J. Snoek, and R. P. Adams, Multi-task bayesian optimization, in *Advances in Neural Information Processing Systems*, edited by C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2013), Vol. 26.
- [35] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration, in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2018).
- [36] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevMaterials.8.013804> for exploring saddle point searches in bcc-Mo using DFT-dimer and GPR-dimer methods. It features detailed figures and tables showcasing atom displacements, energy profiles, and computational efficiency. The document highlights challenges in escaping energy valleys with these methods and compares their scalability. It includes data on the impact of varying translation hops and active region radii on computational effort and efficiency.
- [37] S. Fattahpour, A. Davariashtiyani, and S. Kadkhodaei, Understanding the role of anharmonic phonons in diffusion of bcc metals, *Phys. Rev. Mater.* **6**, 023803 (2022).
- [38] C. Herzig and U. Köhler, Anomalous self-diffusion in bcc IVB metals and alloys, in *Vacancies and Interstitials in Metals and Alloys of Materials Science Forum* (Trans Tech Publications, Zurich, 1987), Vol. 15, pp. 301–322.
- [39] H.-P. Komsa, S. Kurasch, O. Lehtinen, U. Kaiser, and A. V. Krasheninnikov, From point to extended defects in two-dimensional MoS<sub>2</sub>: Evolution of atomic structure under electron irradiation, *Phys. Rev. B* **88**, 035301 (2013).