Automated calculation and convergence of defect transport tensors

T D Swinburne, Danny Perez

npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Automated calculation and convergence of defect transport tensors

Thomas D. Swinburne[1]✉ and Danny Perez[2]

Defect diffusion is a key process in materials science and catalysis, but as migration mechanisms are often too complex to enumerate a priori, calculation of transport tensors typically have no measure of convergence and require significant end-user intervention. These two bottlenecks prevent high-throughput implementations essential to propagate model-form uncertainty from interatomic interactions to predictive simulations. In order to address these issues, we extend a massively parallel accelerated sampling scheme, autonomously controlled by Bayesian estimators of statewide sampling completeness, to build atomistic kinetic Monte Carlo models on a state-space irreducible under exchange and space group symmetries. Focusing on isolated defects, we derive analytic expressions for drift and diffusion coefficients, providing a convergence metric by calculating the Kullback–Leibler divergence across the ensemble of diffusion processes consistent with the sampling uncertainty. The autonomy and efficacy of the method is demonstrated on surface trimers in tungsten and Hexa-interstitials in magnesium oxide, both of which exhibit complex, correlated migration mechanisms.

npj Computational Materials (2020)6:190 ; https://doi.org/10.1038/s41524-020-00463-8

INTRODUCTION

The migration and transformation of intrinsic and extrinsic crystal defects play a central role in numerous materials science and chemistry phenomena, such as postirradiation annealing[1,2], plasma surface interactions[3], or active site formation for heterogeneous catalysis[4,5]. The atomistic mechanisms by which defects migrate are generally impossible to divine a priori due to the routine presence of complex multiatom transformations[6–12]. While these mechanisms self-average with increasing defect size, allowing the use of suitably parametrized coarse-grained models[13–15], the smallest defects are both the least predictable and most mobile[12], with the greatest influence on transport behavior.

In the general case, atomistic mechanisms must be discovered through unbiased dynamic[10,16–20] or static[11,21,22] sampling approaches. When the true dynamics can be characterized as rare transitions between metastable basins on the energy landscape[23], the basin-to-basin dynamics can be mapped to a continuous-time Markov chain[24,25], which forms the theoretical foundation of atomistic kinetic Monte Carlo (akMC) methods[26]. The resulting model can then be stochastically or in some cases (such as that presented here) analytically integrated to extract observables of interest.

A well-recognized problem is that an akMC model will in general have an incomplete catalog of available mechanisms due to a finite amount of sampling, and this can produce catastrophically erroneous predictions if important mechanisms are omitted[19,20,27,28]. Sampling adequacy is often assessed *qualitatively* using the domain expertise of simulation practitioners. Whilst this approach has undoubtedly yielded significant successes, it requires significant end-user analysis for each system under study. This has a punitive impact on the feasibility of automating complex materials simulations on massively parallel computational resources, where the required decision frequency rapidly exceeds practical human limits. Further, in absence of *quantitative* uncertainty quantification approaches, assessing the reliability of mesoscale or macroscale predictions is extremely challenging.

For example, the system sizes required even for small defect clusters render ab initio calculation unfeasible. As a result, interatomic potential models must be used which induces an additional model form uncertainty. The development of interatomic potentials has been revolutionized in recent years through the use of linear-in-descriptor or neural network regression techniques[29–31]. These approaches offer a natural encoding of model form uncertainty through isosurfaces of the cost function used for potential parametrization. High-throughput calculations are essential in this context to enable the systematic propagation of this uncertainty on interatomic interactions to the observables of scientific interest, without prohibitive hours of end-user analysis.

In this contribution, we present an autonomous, highly scalable sampling scheme to efficiently calculate defect diffusion tensors with quantified uncertainty on the sampling completeness. We demonstrate the ability to discover complex migration behaviors of defects in tungsten and magnesium oxide and show how the quantified uncertainty can be used to rapidly yield well-defined convergence measures. Our approach enables high-throughput workflows to rapidly discover, converge, and analyze complex kinetic properties of defect structures in surrogate energy landscapes, with minimal end-user involvement.

RESULTS

Isomorphic compression of state-space

In previous work[28] we introduced `TAMMBER`, a massively parallel accelerated sampling scheme whose formal objective is to exhaustively sample the set of all metastable minima $\mathcal{M}$ and all inter-minima transitions $\mathcal{T}$ for a given system. `TAMMBER` builds a matrix K_{ji} of $j \leftarrow i$ rates in the known state space and a vector k_i^u, which estimates the "unknown" (as yet undiscovered) escape rate

[1]Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, Marseille 13288, France. [2]Theoretical Division T-1, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ✉email: swinburne@cinam.univ-mrs.fr

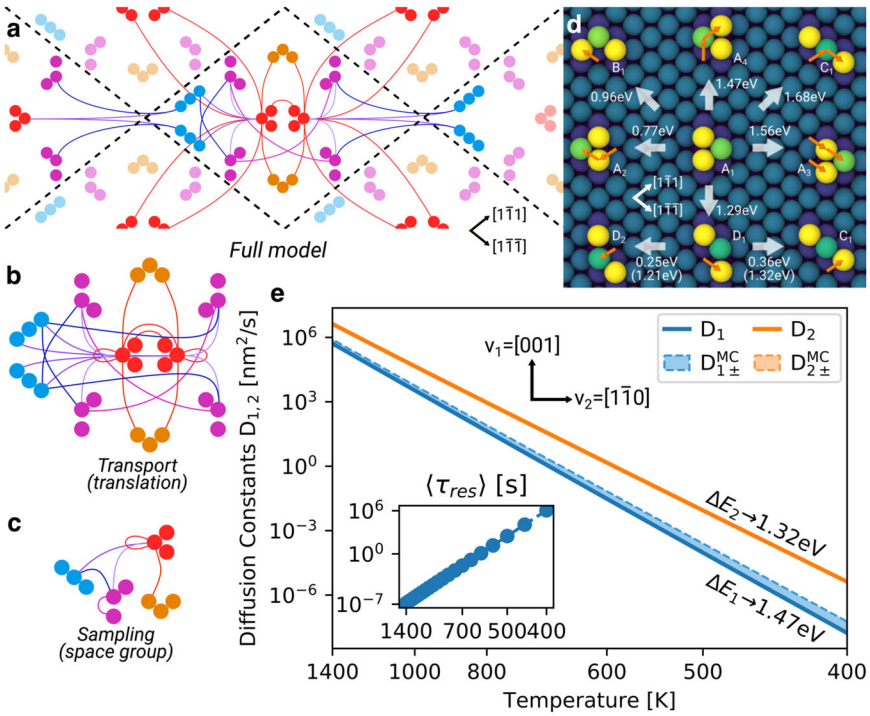

Fig. 1 Compressed sampling methodology applied to trimer migration on W(110). a Cartoon of the full akMC state space and those irreducible under space and translation symmetries. Isomorphic states have the same color. **b** One primitive unit cell with self transitions (closed loops) is sufficient to build a Markov model for transport. **c** A set of states irreducible under all space group symmetries is optimal for discovering new states and transitions through accelerated sampling. **d** Atomic structure of the lowest energy irreducible states and found transition mechanisms. Atoms colored by centrosymmetry[51]. **e** Diffusion tensor eigenvalues D_1, D_2, as determined by (5), over a range of temperatures. The Monte Carlo bounds $D^{MC}_{1\pm}$, $D^{MC}_{2\pm}$ are described in the main text. Effective Arrhenius slopes at low temperature are given. Inset: $\langle\tau_{res}\rangle$ with temperature.

from each state i, encoded as transitions to an absorbing sink[19,20,28,32,33] (Methods). States are identified by constructing a connectivity graph from a minimized configuration, which is then hashed to produce a pseudo-unique integer label for each configuration[11,34]. The presence of nonzero k^u causes trajectories to leave the known state space, giving a natural measure of model validity, the expected residence time[28] in the known state space

$$\langle\tau_{res}\rangle = \mathbf{1}[\mathbf{K}^{tot} - \mathbf{K}]^{-1}\mathbf{P}_0, \tag{1}$$

where $\mathbf{K}^{tot}_{ji} = \delta_{ji}\left[k^u_i + \sum_l K_{li}\right]$, $\mathbf{P}_0$ is the initial distribution, and $\mathbf{1}$ is a row vector of ones. In practice, truly comprehensive sampling of $\mathcal{M}$ and $\mathcal{T}$ is often impractical as their size grows exponentially with the number of atoms in the system. However, $\mathcal{M}$ and $\mathcal{T}$ are both highly reducible under exchanges of indistinguishable atoms $\mathcal{R}$, lattice vector translations $\mathcal{S}$ and space group symmetries $\mathcal{G}$, especially in the case of quasi-zero-dimensional defects. It is known that configurations that are degenerate (in the periodic minimum image sense) under these operations will have isomorphic connectivity graphs[11,34], with isomorphisms that can be efficiently calculated[35]. We exploit this reducibility to make a partition of $\mathcal{M}, \mathcal{T}$ into equally sized subsets

$$\mathcal{M} = \bigcup_{p\in[1,P]} \mathcal{M}_p, \qquad \mathcal{T} = \bigcup_{p,q\in[1,P]^2} \mathcal{T}_{pq}, \tag{2}$$

where all states in $\mathcal{M}_p$ are identical to all others up to a translation $\mathbf{t} \in \mathcal{S}$ and reindexing $\mathcal{R}$. For a given set of primitive lattice vectors, $\mathcal{M}_p$ is thus completely determined by a single local minimum, meaning we can focus all sampling effort on a single state per subset. All of the distinct exit pathways from the chosen $\mathcal{M}_p$ state that end in $\mathcal{M}_q$, or vice versa, are sufficient to fully determine $\mathcal{T}_{pq}$. This already produces a space of P states irreducible under translation that will be used to generate transport coefficients, as illustrated in Fig. 1.

To exploit space group symmetries, consider transforming every member of some subset $\mathcal{M}_p$ with a symmetry operation $\mathbf{G} \in \mathcal{G}$ of the host crystal, producing a new subset of states $\mathcal{M}_q = \mathbf{G}\mathcal{M}_p$. There are two possible outcomes—if $q = p$ we have identified operations $\mathbf{G} \in \mathcal{G}_p$ of that subset's symmetry group $\boldsymbol{\mathcal{G}}_p$, which contains at least the identity operation. As all members of $\boldsymbol{\mathcal{M}}_p$ are related by translation, they share the same space group symmetries. If $q \neq p$ we have further reduced the size of our state space to $P_G \leq P$, as we can generate all members of the subset $\mathcal{M}_q$ from the subset $\mathcal{M}_p$ through the action of $\mathbf{G}$.

However, $\mathcal{M}_p$ is left unchanged under the action of any operation $\mathbf{G}_p \in \mathcal{G}_p$, i.e., $\mathbf{G}_p\mathcal{M}_p = \mathcal{M}_p$. This means the symmetry operation which takes $\mathcal{M}_p$ to $\mathcal{M}_q$ is not unique. To resolve this ambiguity we generate the left cosets[36] with respect to $\mathcal{G}_p$, defined for any operation $\mathbf{G}$ as $\{\mathbf{G}\mathbf{G}_p : \mathbf{G}_p \in \mathcal{G}_p\}$, i.e., the set of operations formed by postmultplying $\mathbf{G}$ with every operation in $\mathcal{G}_p$. It can be shown that each operation in $\mathcal{G}$ appears in exactly one coset[36].

Acting on $\mathcal{M}_p$ with any operation from a given coset will give the same result. For example, the subset's symmetry group $\mathcal{G}_p$ forms a coset, whose action is the identity operation. We can then index the nonequivalent partitions $\{\mathcal{M}_q\}$ which can be generated from $\mathcal{M}_p$ by the coset which generates them; drawing some operation $\mathbf{G}_{qp}$ from each coset we can thus make the further compression $\mathcal{M} = \bigcup_{p=1}^{P_G}\bigcup_q \mathbf{G}_{qp}\mathcal{M}_p$, with analogous partitioning of $\mathcal{T}_{pq}$. This maximally compressed representation is used for sampling, as illustrated in Fig. 1, and we then partially "decompress" into the partitioning (2) of P distinct states irreducible under translation for the calculation of transport tensors.

We, therefore, only need to sample the exit transitions from a maximum of P_G configurations, as all others can be generated by the known symmetric relations. By focusing sampling resources on a much smaller number of possible states, the unknown rates decrease at a much faster rate, with a subsequent increase in the validity timescales for the constructed Markov chain. As in TAMMBER the unknown rates at low temperatures decrease as an inverse power of the MD time invested at high temperatures, state-space compression can yield extremely large benefits in practice. As an example of the efficiencies this approach affords, in previous work we investigated the breakup of two interstitial defects in bcc iron[28], obtaining a Markov model on the uncompressed (reducible) state-space with $\langle\tau_{\text{res}}\rangle \sim 80$ s at 300 K, an insufficient duration to make confident predictions on the breakup mechanism. Despite the relatively low symmetry of this defect system, we show in the Supplementary Information that the compressed sampling scheme described above yields a Markov model with $\langle\tau_{\text{res}}\rangle \sim 5\times10^6$ s with only 75% of the computational effort, allowing convergence in the model predictions. The isolated point defects we consider in this work are typical of even higher symmetry, giving correspondingly greater efficiencies.

Evaluation of transport tensors

To define the drift and diffusion tensors, we first require a defect position in the supercell for each configuration. To remove ambiguities from periodic boundary conditions, we carve out a defective region of some configuration in each partition by thresholding some structural descriptor, here the centrosymmetry, then take a descriptor-weighted center of mass. As we allocate sampling to increase the residence time, the threshold value can be freely determined in post-processing; typically a single value is suitable for the same material system.

A periodic minimum image displacement vector $\mathbf{d}_{lm}$ for any $l \in \mathcal{M}_p \to m \in \mathcal{M}_q$ transition is found through the application of the known rigid transformation from the sampled configurations. Importantly, transitions between states within the same space group partition are then represented as *self transitions* $l \in \mathcal{M}_p \to m \in \mathcal{M}_p$ (closed loops in Fig. 1) and $\mathbf{d}_{lm}$ is a lattice vector.

We are then able to construct an irreducible group of states from the space group partitioning of Eq. (2), with one state indexed by $p \in [1, P]$ for each set $\mathcal{M}_p$. All the transitions from a state p to states q in the uncompressed model are mapped to their compressed states, forming a (possibly empty) set $\mathcal{C}_{qp} = \{\mathrm{k}_l, \mathbf{d}_l\}$ of transitions and displacement vectors. The compressed rate matrix is then defined as $\mathrm{K}^c_{qp} = \sum_{l\in\mathcal{C}_{qp}} \mathrm{k}_l$, with a statewise total escape rate matrix $\mathrm{K}^{tot,c}_{qp} = \delta_{qp}(\mathrm{k}^{u,c}_p + \sum_r \mathrm{K}^c_{rp})$, where the superscript c indicates we consider rates in the compressed model. The unknown rates $\mathrm{k}^{u,c}_p$ are taken from the maximally compressed representation, with the same unknown rate for symmetrically equivalent partitions. One could then generate akMC trajectories[37] with a matrix and vector of branching probabilities $\mathbf{B}^c \equiv \mathbf{K}^c[\mathbf{K}^{tot,c}]^{-1}$ and $\mathbf{B}^{u,c} \equiv \mathbf{k}^{u,c}[\mathbf{K}^{tot,c}]^{-1}$, accumulating a total displacement $\mathbf{x} = \sum \mathbf{d}_l$ in a residence time $\tau_{\text{res}} = \sum \delta\tau_l$ before absorption. In the Supplementary Information we show that the drift and diffusion coefficients $\boldsymbol{\mu}$, $\mathbf{D}$ can then be extracted from the relations $\langle\mathbf{x}\rangle = \langle\tau_{\text{res}}\rangle\boldsymbol{\mu}$ and $\langle\mathbf{x}\otimes\mathbf{x}\rangle = 2\langle\tau_{\text{res}}\rangle\mathbf{D} + \langle\tau^2_{\text{res}}\rangle\boldsymbol{\mu}\otimes\boldsymbol{\mu}$ for large $\langle\tau_{\text{res}}\rangle$. However, it is also possible to analytically evaluate averages over all possible pathways using a "displacement generating function"

$$Z(\boldsymbol{\lambda}, \mathbf{P}_0) = \mathbf{B}^{u,c}\mathbf{G}^c(\boldsymbol{\lambda})\mathbf{P}_0, \quad [\mathbf{B}^c(\boldsymbol{\lambda})]_{qp} \equiv \sum_{l\in\mathcal{C}_{qp}} \frac{\mathrm{k}_l e^{-\boldsymbol{\lambda}\cdot\mathbf{d}_l}}{\mathrm{K}^{tot,c}_{pp}}, \tag{3}$$

where $\mathbf{G}^c(\boldsymbol{\lambda}) = \sum_{n=0}^{\infty}[\mathbf{B}^c(\boldsymbol{\lambda})]^n = [\mathbb{I} - \mathbf{B}^c(\boldsymbol{\lambda})]^{-1}$. Moments of the total displacement can then be written $\langle\mathbf{x}\rangle = -\partial_{\boldsymbol{\lambda}} Z(\boldsymbol{\lambda}, \mathbf{P}_0)|_{\boldsymbol{\lambda}=\mathbf{0}}$ and $\langle\mathbf{x}\otimes\mathbf{x}\rangle = \partial_{\boldsymbol{\lambda}}\otimes\partial_{\boldsymbol{\lambda}} Z(\boldsymbol{\lambda}, \mathbf{P}_0)|_{\boldsymbol{\lambda}=\mathbf{0}}$ (Supplementary Information). While (3) could be used for any choice of the initial condition, we note that in the well-sampled limit $\mathrm{k}^{u,c}_p \ll \sum_r \mathrm{K}^c_{rp}$, which is necessary but not sufficient for global convergence, the matrix $[\mathbf{K}^{tot,c} - \mathbf{K}^c]$ will have a spectral gap, with one eigenvalue $0 < v_0 \ll v_1 < v_2 \ldots$ much smaller than all others[24]. The right eigenvector for v_0 is the *quasistationary* distribution (QSD) $\boldsymbol{\pi}^{QSD}$ in the known state space[24], the limiting distribution conditional on not absorbing for an arbitrarily long time, which as $\mathbf{k}^{u,c} \to \mathbf{0}$ becomes the Boltzmann distribution, $\boldsymbol{\pi}^{\text{QSD}} \to \hat{\boldsymbol{\pi}}$. As the QSD is the longest-lived mode and transport coefficients are defined as the limit of infinitely long trajectories, it is natural to set $\mathbf{P}_0 = \hat{\boldsymbol{\pi}}^{\text{QSD}} = \boldsymbol{\pi}^{\text{QSD}}/(\mathbf{1}\boldsymbol{\pi}^{\text{QSD}})$ to eliminate the influence of initial conditions. With this choice, it is simple to show that $\langle\tau^n_{\text{res}}\rangle = n!/v_0^n$ and the expected drift and diffusion coefficients emerge as (Supplementary Information)

$$\boldsymbol{\mu}(\mathbf{k}^{u,c}) \equiv \langle\tau_{\text{res}}\rangle^{-1}\partial_{\boldsymbol{\lambda}}\left[1/Z(\boldsymbol{\lambda}, \hat{\boldsymbol{\pi}}^{\text{QSD}})\right]\Big|_{\boldsymbol{\lambda}=\mathbf{0}}, \tag{4}$$

$$\mathbf{D}(\mathbf{k}^{u,c}) \equiv \frac{1}{2}\langle\tau_{\text{res}}\rangle^{-1}\partial_{\boldsymbol{\lambda}}\otimes\partial_{\boldsymbol{\lambda}}\left[1/Z(\boldsymbol{\lambda}, \hat{\boldsymbol{\pi}}^{\text{QSD}})\right]\Big|_{\boldsymbol{\lambda}=\mathbf{0}}. \tag{5}$$

Equations (4) and (5) are a central result of this contribution, expressions for the drift and diffusion tensors of an arbitrarily complex diffusion process in a periodic system, autonomously constructed in a massively parallel sampling scheme, which crucially are dependent on 'unknown' rates $\mathbf{k}^{u,c}$ that robustly quantify sampling incompleteness. In the Supplementary Information we show the limiting expressions $\boldsymbol{\mu} \equiv \lim_{\mathbf{k}^{u,c}\to\mathbf{0}}\boldsymbol{\mu}(\mathbf{k}^{u,c})$ and $\mathbf{D} \equiv \lim_{\mathbf{k}^{u,c}\to\mathbf{0}}\mathbf{D}(\mathbf{k}^{u,c})$ reduce to expressions obtained in previous derivations on multistate diffusion in periodic media[38,39], with all uncorrelated and correlated contributions that are essential to capture complex diffusion pathways. We also show that the error in our estimators scales inversely with $\langle\tau_{\text{res}}\rangle$, and we confirm that for systems obeying detailed balance, which we focus on in the remainder, $\boldsymbol{\mu} = \mathbf{0}$ and the correlated contribution to the diffusivity is always nonpositive[40].

The convergence of the diffusivity

The key feature of (5) is that $\mathbf{D}(\mathbf{k}^{u,c})$ estimates the diffusion tensor over all possible trajectories in the known state-space before exit due to the unknown rates. Possible changes to the diffusivity under the discovery of additional transition rates between *known* states can then be bounded; the rate matrix $\mathbf{K}^c$ can be modified by an additional rates matrix $\delta\mathbf{K}^c$, which must satisfy detailed balance and not increase the total exit rate from each known state p by more than the unknown rate $\mathrm{k}^{u,c}_p$, meaning $[\mathbf{1}\delta\mathbf{K}^c]_p \leq \mathrm{k}^{u,c}_p$[1]. We have designed a Monte Carlo procedure to sample the space of permissible $\delta\mathbf{K}^c$ (see Supplementary Information) which typically requires less than a core-minute for the systems studied here and is trivially parallelizable. We discuss the sensitivity to the discovery of additional states at the end of this section.

Anisotropic diffusion is most conveniently expressed using the eigenvalues and eigenvectors $\{\mathrm{D}_l, \mathbf{v}_l\}$, $l \in [1, 3]$ of the symmetric 3×3 diffusion matrix. In the limit of comprehensive sampling, we have found that the ensemble of diffusion tensors produced in our Monte Carlo procedure can be simultaneously diagonalized to high accuracy, allowing a direct measure of convergence by comparing the spread $\mathrm{D}_{l+} - \mathrm{D}_{l-}$ in eigenvalues, where $\mathrm{D}_{l\pm}$ are the maximum and minimum values obtained by Monte Carlo, as can be seen in figure 1. However, in the general case, the eigenbasis

[1]We note that the calculated transition rates presently employ harmonic transition state theory, which can be in error at high temperatures[50]. However, as the present method is generally applicable for any employed rate theory we do not consider such errors.

can vary significantly over the ensemble of diffusion tensors, requiring a more robust measure of convergence.

As diffusive processes are characterized by a probability distribution that broadens with time, we build a convergence measure by comparing the Kullback–Leibler divergence[41] $\mathcal{R}_\pm \equiv \int \mathrm{d}\mathbf{x}\rho \ln \rho/\rho_\pm$ between fundamental solutions $\rho(\mathbf{x}, t)$, $\rho_\pm(\mathbf{x}, t)$ to the three-dimensional diffusion equation with a diffusion tensor $\mathbf{D}$, $\mathbf{D}_\pm$, where $\mathcal{R}_\pm = 0$ if and only if $\rho = \rho_\pm$. The matrices $\mathbf{D}_\pm$ are chosen such that $\mathcal{R}_\pm$ are the maximum and minimum values obtained during Monte Carlo. It can be shown that, for this choice of distribution, $\mathcal{R}_\pm$ is time-independent for identical initial conditions (Supplementary Information). For our convergence measure we take the spread $\delta\mathcal{R} \equiv \mathcal{R}_+ - \mathcal{R}_- \geq 0$, reading

$$\delta\mathcal{R} = \frac{1}{2}\mathrm{Tr}\left(\mathbf{D}^{-1}[\mathbf{D}_+ - \mathbf{D}_-]\right) + \ln\sqrt{\mathrm{Det}\left(\mathbf{D}_-^{-1}\mathbf{D}_+\right)}. \quad (6)$$

In the limit of comprehensive sampling is is simple to show that $\delta\mathcal{R} \to \sum_I (\mathrm{D}_{I+} - \mathrm{D}_{I-})/2\mathrm{D}_I$, meaning that $\delta\mathcal{R}$ reduces to the above convergence definition, the relative spread in eigenvalues. We thus identify $\delta\mathcal{R}$ as a well defined dimensionless convergence metric ideal for autonomous implementation, which has an informative limit, namely the relative spread in eigenvalues consistent with the sampling uncertainty. In the Supplementary Information, we show how $\delta\mathcal{R}$ monotonically decays with sampling effort, with asymptotic rate $1/\langle\tau_{\mathrm{res}}\rangle$, even when starting from poorly converged samples.

It is clear that the diffusivity cannot be globally bounded against the discovery of some new set of states which are free to possess arbitrary transport properties. However, as transition rates to such a set of states are bounded by the $\mathbf{k}^u$, the Monte Carlo procedure outlined above does characterizes transport behavior in the known state-space over timescales of order $\langle\tau_{\mathrm{res}}\rangle$. Convergence to the large $\langle\tau_{\mathrm{res}}\rangle$ limit can be accelerated by seeding the TAMMBER procedure with as many states as possible, which are free to be completely disconnected (in this case we have independent QSDs $\boldsymbol{\pi}_s^{\mathrm{QSD}}$ for each connected sub-network s). The integration of automated structural search algorithms[23,42] into the present workflow will be the subject of future work. Of course, as TAMMBER generates thousands of high-temperature molecular dynamics trajectories across the known state space, it is an automated global minimum search method, with $\langle\tau_{\mathrm{res}}\rangle$ acting as a quantitative measure of sampling quality when the system is ergodic. As a result, whilst in common with all theoretical studies on high dimensional landscapes, we cannot provide bounds on global minima, we can provide a key uncertainty quantification on the validity of our findings, namely a rigorous prediction timescale from an arbitrary distribution on the known state space with a corresponding bound on transport coefficients. We emphasize that the inability to assign bounds on global minima searches applies equally to human-guided sampling or the present approach. The convergence metrics we provide is thus a valuable analytical tool that removes ambiguities inherent to traditional methods in addition to enabling a fully automated workflow suitable for high-throughput computation.

Application to trimer diffusion on W(110)

The diffusion of adatom clusters is a fundamental process in surface science and has recently been conjectured to play a crucial role in the formation of complex "fuzzy" surface morphologies during plasma exposure in nuclear fusion reactors[43]. We focus here on the trimer defect a demonstrative case study; comprehensive high-throughput investigations for which the present approach is designed will be presented elsewhere. In the present case, TAMMBER was initialized with state A_1 in Fig. 1, then run for 8 h on 144 cores using the EAM4 embedded atom method potential by Marinica et al.[44], covering a temperature range of 400–1400 K. The corresponding values of $\langle\tau_{\mathrm{res}}\rangle$ as a function of temperature are also shown in Fig. 1.

The resultant diffusion behavior is highly correlated, with many $A_1 \leftrightarrow A_2$ and $D_1 \leftrightarrow D_2$ transitions in particular. However, the overall system does not exhibit a clear "superbasin-to-superbasin" diffusion mechanism, meaning access to the full highly correlated trajectory ensemble is essential to extract accurate transport coefficients. As shown in Fig. 1, the eigenvalues D_1, D_2 ($\mathrm{D}_3 \to 0$ for surface diffusion) show a high degree of convergence under the Monte Carlo sensitivity procedure, with $\delta\mathcal{R} \leq 2$ at all temperatures. The eigenvectors for D_1 and D_2 were found to be $\mathbf{v}_1 = [001]$ and $\mathbf{v}_2 = [1\bar{1}0]$. To look for dominant pathways, we form a weighted graph from the connectivity of four primitive unit cells, where the graph edges are weighted by the corresponding saddle point energy[23,28]. Dijkstra's shortest path algorithm[45] was then used to identify the dominant pathways. In agreement with the found Arrhenius slopes, migration along $\mathbf{v}_2$ is dominated by $A_1 \to D_1 \to C_1 \to D_2 \to A_5$ paths, with a well-defined activation energy of $\Delta E_1 = 1.32$ eV at the lower temperatures. Migration along $\mathbf{v}_1$ is similarly dominated by $A_1 \to A_4$ paths at lower temperatures, but $A_1 \to A_3$ paths have a growing contribution with temperature, giving a weak nonlinearity to the Arrhenius gradient $\Delta E(\beta) = -\partial_\beta \ln|\mathrm{D}|$. An in-depth study of this procedure, and its role in a fully automated workflow, will be the subject of future work. We note that a recent study[43] of surface island diffusion on $W(110)$ at 1000K using a different interatomic potential[46] reported trimer migration via $A_1 \to A_3$, which is accounted for in the present study but is not found to be the dominant mechanism. The ability to efficiently resolve such ambiguities, without prohibitive person-hours, is a key advantage of the methodology presented here.

Application to Hexa-interstitial in bulk MgO

To conclude this contribution, we investigate the diffusion of a stochiometric Hexa-interstitial in MgO, whose lattice has a space group $Fm\bar{3}m$ with point group $\mathcal{O}_h$. In a cubic supercell with axes aligned with ⟨100⟩ directions, we thus retain the full symmetry of the host lattice before the introduction of any defective structures. Connectivity graphs are constructed with vertices colored to indicate specie. Due to the high degree of symmetry, the maximally reduced state space contained only four states, while the state space irreducible under translation contained 56 states. TAMMBER was initialized in a relatively high energy state (red circle) that was found by Uberuaga et al.[7,8] to possess very low migration barriers; a much lower energy state was rapidly found, upon which sampling was subsequently concentrated, demonstrating the ability of TAMMBER to act as a massively parallel global minimum search routine. Using the same modified Buckingham potential as in that work, TAMMBER was run for 8 h on 144 cores targeting a temperature range of 500–1000 K. While we find the very low migration barriers (<0.1 eV) in agreement with[7,8], the full diffusion tensor was found to be isotropic $\mathrm{D}_1 = \mathrm{D}_2 = \mathrm{D}_3$, with characteristic activation energy that converged to 0.887 eV at the lower temperatures, which slowly increases at higher temperatures, as found for the above trimer example. Analyzing the dominant pathways with temperature reveals that this activation barrier corresponds to self-migration of the lowest energy state, with little effect of correlation on the diffusion paths; analyzing the state-by-state contribution to the uncorrelated diffusivity, shown in Fig. 2, we see that at higher temperatures the self-migration of other states have an increasingly large contribution. $\langle\tau_{\mathrm{res}}\rangle$ as a function of temperature are also shown in Fig. 2. The Monte Carlo procedure indicated a high degree of convergence, with $\delta\mathcal{R} < 0.001$ over all temperatures considered.

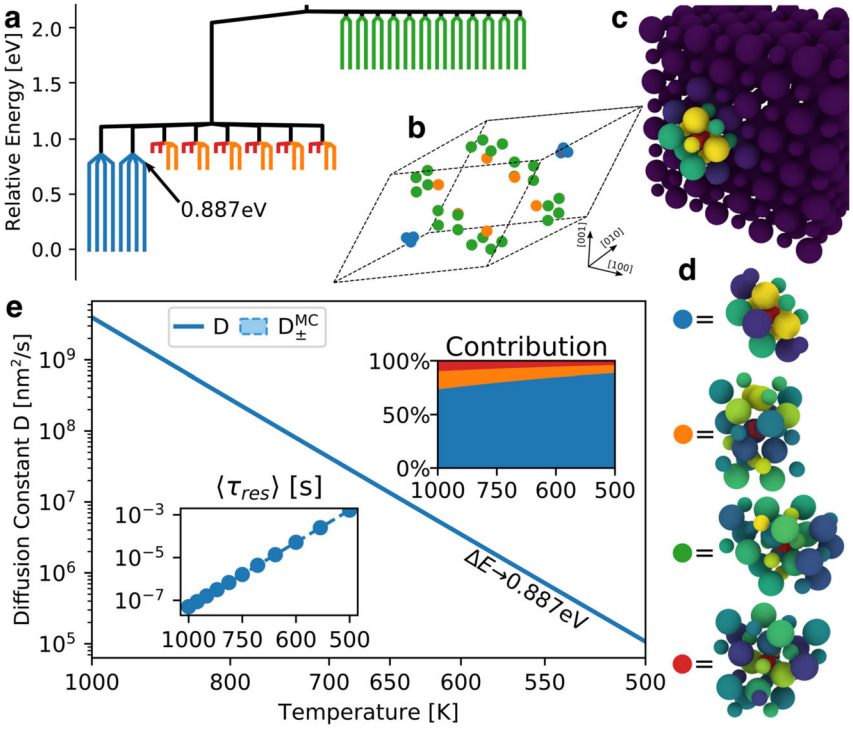

Fig. 2 Sampling of Hexa-interstitial in MgO. a Connectivity graph[23] between states irreducible under translation. The very low diffusion barriers (<0.1 eV) observed in previous work[7,8] can be seen. **b** Nonequivalent positions in the primitive unit cell for the four irreducible states. **c** Cross-section of the lowest energy state. Atoms colored by centrosymmetry[51]. **d** Atomic structure of the four lowest energy irreducible states. **e** Single eigenvalue of the diffusion tensor at various temperatures. Error bounds, calculated through the Monte Carlo procedure detailed in the main text, are not visible on the presented scale. Insets show $\langle \tau_{\text{res}} \rangle$ at various temperatures and the contribution to the overall diffusivity from self transitions within each state. The fast diffusing states have increasing contribution at high temperatures as they are occupied more frequently.

DISCUSSION

We have presented a fully automatable and efficient method to evaluate the transport tensors resulting from the arbitrarily complex diffusion processes of crystal defects, with well-defined convergence criteria based on quantitative measures of sampling uncertainty combined with a Monte Carlo procedure to sample the admissible diffusion tensors consistent with sampling uncertainty and a detailed balance. The method was demonstrated on a surface trimer in tungsten and a hex-interstitial in magnesium oxide. By effectively eliminating user input beyond the seeding of some initial state(s), the presented approach demonstrates sufficient computational and critically *end-user* efficiencies to extend the phenomenological reach of high throughput computations to point defect kinetics. Future work will exploit these efficiencies to analyze defect transport over a wide range of material systems, and the influence of breaking detailed balance through external driving forces, giving nonzero limits for the drift vector $\boldsymbol{\mu}$.

METHODS

Massively parallel sampling routine

Through the use of `NEB` calculations[47] and transition, state theory[48]-`TAMMBER` constructs a transition matrix $\mathbf{K}$ of rank equal to the number of discovered states, giving a continuous-time Markov chain $\dot{\mathbf{P}}(t) = [\mathbf{K} - \mathbf{K}^{tot}]\mathbf{P}(t)$. Exploiting the known Poissonian distribution of exit times from a suitably thermalized basin[24], a Bayesian likelihood (and thus posterior distribution) for the $\{\mathrm{k}_i^u\}$ was derived using parallel trajectory data obtained through a modified temperature accelerated dynamics method[18]. Continuous self-optimization was achieved by calculating the derivative of each k_i^u with respect to additional computational work, allowing the degree of temperature acceleration to be statewide optimized and the massively parallel sampling effort distributed across states to differentially maximize a key measure of model validity, the expected residence time before absorption[28], Eq. (1). Typically, a simulation starts with one state then the rank of $\mathbf{P}$, $\mathbf{K}$, and $\mathbf{K}^{tot}$ increases as states are discovered. If $\mathbf{P}_0$ is fixed to be unity for the initial state and zero otherwise, i.e., $[\mathbf{P}_0]_j = \delta_{ij}$ for initial state i, $\langle \tau_{\text{res}} \rangle$ is monotonically increasing with sampling effort, a key consequence of the estimation procedure for $\{\mathrm{k}_i^u\}$. An implementation of this method is available as an open-source code at https://github.com/tomswinburne/tammber, whose near-ideal parallel efficiency has been demonstrated on massively parallel resources employing 1000 to more than 80,000 cores.

Isomorphically compressed representation

Isomorphic configurations are identified in two ways—the connectivity graph for every state is duplicated and reindexed into McKay's pseudo unique canonical order[35], which is identical for isomorphic states. Alternatively, the `VF2` graph matching routine[49] is applied to a single state to find self-isomorphisms. To determine a given reindexing $\mathcal{R}_{ij}$ between two isomorphic states i, j we first find the mappings $\mathcal{R}_{ic}$ and $\mathcal{R}_{jc}$ to the McKay canonical order then obtain $\mathcal{R}_{ij} = \mathcal{R}_{ic}\mathcal{R}_{cj}^{-1}$.

As isomorphisms will be with respect to the simulation supercell, not the host crystal structure, the relevant point group is $\overline{\mathcal{G}} = \mathcal{W} \cap \mathcal{G} \in \mathcal{O}_h$, the intersection of the point group of $\mathcal{G}$ and the supercell point group $\mathcal{W}$, where $\mathcal{W}$ is a subgroup of the cubic group $\mathcal{O}_h$. Unlike an arbitrary element of $\mathcal{G}$ or $\mathcal{W}$, any element of $\overline{\mathcal{G}}$ is guaranteed to leave both the perfect lattice and supercell unchanged up to a translation and reindexing. We then iterate through all members of $\overline{\mathcal{G}}$, applying the point transform, aligning the first indexed atom then checking for minimum image coincidence atom-by-atom, rejecting the candidate member of $\overline{\mathcal{G}}$ at the first failure.

DATA AVAILABILITY

The datasets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The `TAMMBER` source code, installation guide, and cloud-hosted analysis tools are freely available at https://github.com/tomswinburne/tammber.

Received: 20 February 2020; Accepted: 13 November 2020;
Published online: 11 December 2020

REFERENCES

1. Fu, C.-C., Dalla Torre, J., Willaime, F., Bocquet, J.-L. & Barbu, A. Multiscale modelling of defect kinetics in irradiated iron. *Nat. Mater.* **4**, 68–74 (2005).
2. Osetsky, Y. N., Beland, L. K., Barashev, A. V. & Zhang, Y. On the existence and origin of sluggish diffusion in chemically disordered concentrated alloys. *Curr. Opin. Solid State Mater. Sci.* **22**, 65–74 (2018).
3. Sefta, F., Hammond, K. D., Juslin, N. & Wirth, B. D. Tungsten surface evolution by helium bubble nucleation, growth and rupture. *Nucl. Fusion* **53**, 073015 (2013).
4. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **1**, 37 (2009).
5. Böller, B., Durner, K. M. & Wintterlin, J. The active sites of a working Fischer–Tropsch catalyst revealed by operando scanning tunnelling microscopy. *Nat. Catal.* **2**, 1027–1034 (2019).
6. Sorensen, M. R., Mishin, Y. & Voter, A. F. Diffusion mechanisms in Cu grain boundaries. *Phys. Rev. B* **62**, 3658–3673 (2000).
7. Uberuaga, B. et al. Structure and mobility of defects formed from collision cascades in MgO. *Phys. Rev. Lett.* **92**, 115505 (2004).
8. Uberuaga, B. et al. Exploring long-time response to radiation damage in MgO. *Nucl. Instrum. Methods Phys. Res. Sect. B* **228**, 260–273 (2005).
9. Uberuaga, B., Hoagland, R., Voter, A. & Valone, S. Direct transformation of vacancy voids to stacking fault tetrahedra. *Phys. Rev. Lett.* **99**, 135501 (2007).
10. Perez, D., Uberuaga, B. P., Shim, Y., Amar, J. G. & Voter, A. F. Accelerated molecular dynamics methods: introduction and recent developments. *Annu. Rep. Comput. Chem.* **5**, 79–98 (2009).
11. Béland, L. K., Brommer, P., El-Mellouhi, F., Joly, J.-F. & Mousseau, N. Kinetic activation-relaxation technique. *Phys. Rev. E* **84**, 046704 (2011).
12. Martínez, E. & Uberuaga, B. P. Mobility and coalescence of stacking fault tetrahedra in Cu. *Sci. Rep.* **5**, 1–5 (2015).
13. Dezerald, L., Proville, L., Ventelon, L., Willaime, F. & Rodney, D. First-principles prediction of kink-pair activation enthalpy on screw dislocations in bcc transition metals: V, nb, ta, mo, w, and fe. *Phys. Rev. B* **91**, 094105 (2015).
14. Swinburne, T. D. et al. Fast, vacancy-free climb of prismatic dislocation loops in bcc metals. *Sci. Rep.* **6**, 30596–30604 (2016).
15. Alexander, R. et al. Ab initio scaling laws for the formation energy of nanosized interstitial defect clusters in iron, tungsten, and vanadium. *Phys. Rev. B* **94**, 024103 (2016).
16. Voter, A. F. Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908 (1997).
17. Voter, A. F. Parallel replica method for dynamics of infrequent events. *Phys. Rev. B* **57**, R13985 (1998).
18. Sorensen, M. & Voter, A. Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **112**, 9599–9606 (2000).
19. Chatterjee, A. & Bhattacharya, S. Uncertainty in a Markov state model with missing states and rates: application to a room temperature kinetic model obtained using high temperature molecular dynamics. *J. Chem. Phys.* **143**, 114109 (2015).
20. Chill, S. T. & Henkelman, G. Molecular dynamics saddle search adaptive kinetic Monte Carlo. *J. Chem. Phys.* **140**, 214110 (2014).
21. Henkelman, G. & Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **111**, 7010–7022 (1999).
22. Wales, D. J. Discrete path sampling. *Mol. Phys.* **100**, 3285–3305 (2002).
23. Wales, D. J. *Energy Landscapes* (Cambridge, 2003).
24. Le Bris, C., Lelievre, T., Luskin, M. & Perez, D. A mathematical formalization of the parallel replica dynamics. *Monte Carlo Methods Appl.* **18**, 119–146 (2012).
25. Lelièvre, T. *Mathematical Foundations of Accelerated Molecular Dynamics Methods*, 1–32 (Springer International Publishing, Cham, 2018). https://doi.org/10.1007/978-3-319-42913-7_27-1.
26. Henkelman, G. Atomistic simulations of activated processes in materials. *Annu. Rev. Mater. Res.* **47**, 199–216 (2017).
27. Aristoff, D., Chill, S. & Simpson, G. Analysis of estimators for adaptive kinetic Monte Carlo. *Commun. Appl. Math. Comput. Sci.* **11**, 171–186 (2016).
28. Swinburne, T. D. & Perez, D. Self-optimized construction of transition rate matrices from accelerated atomistic simulations with Bayesian uncertainty quantification. *Phys. Rev. Mater.* **2**, 053802 (2018).
29. Bartók, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
30. Marinica, M. Towards better efficiency of interatomic linear machine learning potentials. *Comput. Mater. Sci.* **166**, 200–209 (2019).
31. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
32. Boulougouris, G. C. & Frenkel, D. Monte Carlo sampling of a Markov web. *J. Chem. Theory Comput.* **1**, 389–393 (2005).
33. Bhoutekar, A., Ghosh, S., Bhattacharya, S. & Chatterjee, A. A new class of enhanced kinetic sampling methods for building Markov state models. *J. Chem. Phys.* **147**, 152702 (2017).
34. Perez, D., Cubuk, E. D., Waterland, A., Kaxiras, E. & Voter, A. F. Long-time dynamics through parallel trajectory splicing. *J. Chem. Theory Comput.* **12**, 18–28 (2015).
35. McKay, B. D. & Piperno, A. Practical graph isomorphism, II. *J. Symb. Comput.* **60**, 94–112 (2014).
36. Scott, W. R. *Group Theory* (Courier Corporation, 2012).
37. Bortz, A. B., Kalos, M. H. & Lebowitz, J. L. A new algorithm for Monte Carlo simulation of Ising spin systems. *J. Comput. Phys.* **17**, 10–18 (1975).
38. Trinkle, D. R. Variational principle for mass transport. *Phys. Rev. Lett.* **121**, 235901 (2018).
39. Landman, U. & Shlesinger, M. F. Stochastic theory of multistate diffusion in perfect and defective systems. I. Mathematical formalism. *Phys. Rev. B* **19**, 6207 (1979).
40. We note that the calculated transition rates presently employ harmonic transition state theory, which can be in error at high temperatures [50]. However, as the present method is generally applicable for any employed rate theory we do not consider such errors.
41. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2012).
42. Marinica, M.-C., Willaime, F. & Crocombette, J.-P. Irradiation-induced formation of nanocrystallites with $c15$ laves phase structure in bcc iron. *Phys. Rev. Lett.* **108**, 025501 (2012).
43. Yang, L., Wirth, B., Perez, D. & Voter, A. F. Mobility of tungsten clusters on tungsten surfaces. *Nucl. Instrum. Methods Phys. Res. Sect. B* **453**, 61–66 (2019).
44. Marinica, M. C. et al. Interatomic potentials for modelling radiation defects and dislocations in tungsten. *J. Phys.* **25**, 395502 (2013).
45. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
46. Juslin, N. & Wirth, B. D. Interatomic potentials for simulation of the bubble formation in W. *J. Nucl. Mater.* **432**, 61–66 (2013).
47. Henkelman, G., Uberuaga, B. P. & Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
48. Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **62**, 251–342 (1990).
49. Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. An improved algorithm for matching large graphs. in *Proc. 3rd IAPR-TC15 workshop on graph-based representations in pattern recognition*, 149–159 (2001).
50. Swinburne, T. D. & Marinica, M.-C. Unsupervised calculation of free energy barriers in large crystalline systems. *Phys. Rev. Lett.* **120**, 135503 (2018).
51. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO—the open visualization tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2010).

ACKNOWLEDGEMENTS

T.D.S. gratefully recognizes support from the Agence Nationale de Recherche, via the MEMOPAS project ANR-19-CE46-0006-1. This work was granted access to the HPC resources of IDRIS under the allocation AP010910718 and A0070910965 attributed by GENCI. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training program 2019–2020 under grant agreement No. 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. Work at Los Alamos National Laboratory was supported by the U.S. Department of Energy, Office of Nuclear Energy and Office of Science, Office of Advanced Scientific Computing Research through the Scientific Discovery through Advanced Computing (SciDAC) project on Fission Gas Behavior. Los Alamos National Laboratory is operated by Triad National Security LLC, for the National Nuclear Security Administration of the U.S. DOE under contract No. 89233218CNA0000001.

AUTHOR CONTRIBUTIONS

T.D.S. and D.P. designed the research program. T.D.S. derived the theoretical results, ran the simulations, and designed the Monte Carlo algorithms. T.D.S. and D.P. and wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00463-8.

Correspondence and requests for materials should be addressed to T.D.S.

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.