Genetic Programming for Multi-Timescale Modeling

Kumara Sastry,1 D. D. Johnson,1 David E. Goldberg,2 Pascal Bellon1
1Department of Materials Science & Engineering, and the Fredrick Seitz Materials Research Laboratory.
2Department of General Engineering
University of Illinois Urbana-Champaign, Urbana IL.
(Dated: March 22, 2022)

A bottleneck for multi-timescale dynamics is the computation of the potential energy surface (PES). We explore the use of genetic programming (GP) to symbolically regress a mapping of the saddle-point barriers from only a few calculated points via molecular dynamics, thereby avoiding explicit calculation of all the barriers. The GP-regressed barrier function enables use of kinetic Monte Carlo (KMC) to simulate real-time kinetics (seconds to hours) using realistic interactions. To illustrate, we apply a GP regression to vacancy-assisted migration on a surface of a binary alloy and predict the diffusion barriers within 0.1–1% error using 3% (or less) of the barriers, and discuss the significant reduction in CPU time.

PACS numbers: 02.60.-x, 02.70.Wz, 31.50.-x, 68.35.Fx

Molecular dynamics (MD) is extensively used for kinetic modeling of materials. Yet MD methods are limited to nanoseconds of real time, and hence fail to model directly many processes. Recently several approaches were proposed for multiscale bridging methods such as temperature-accelerated dynamics (TAD)2 provide significant acceleration of MD but they still fall 3–6 orders of magnitude short of real processing times. These methods assume that transition-state theory applies, and concentrate only on infrequent events. An alternative approach to bridge timescales1 uses kinetic Monte Carlo (KMC)12 combined with MD by constructing an a priori list of events (i.e., “look-up table”). The table look-up KMC yields several orders of magnitude increase in simulated time over MD (e.g., see 11). The table of events is commonly comprised of atomic jumps, but collective motions (or off-lattice jumps), e.g., see 7, may need to be added but may not be known a priori. Additionally, tabulating barrier energies from a list of events is a serious limitation. For example, multicomponent alloys have an impossibly large set of barriers, due to configurational dependence, making their tabulation impractical, especially from first-principles. An alternative approach is calculating energies “on-the-fly”13, but it too has serious time limitation (see Fig. 1). Recent developments and limitations of KMC methods are given, e.g., in 13.

Symbolically-Regressed Table KMC (sr-KMC): To avoid the need or expense of explicit calculation of all activation barriers—frequent or infrequent—and thereby facilitate an effective hybridization of MD and KMC for multiscale dynamics modeling, we utilize genetic programming (GP)—a genetic algorithm that evolves computer programs—to symbolically regress the PES (in our case saddle-point barriers only) from a limited set of calculated points on the PES. An accurate GP-regressed PES extends the KMC paradigm, as suggested in Fig. 11 to machine learn the “look-up table” and get an in-line barrier function for increasing number of active configurations (or complexity) and providing simulation over experimentally relevant time frames, which may not be possible from standard table look-up or on-the-fly KMC. Interfacing GP with TAD-MD and/or pattern-recognition methods will further extend its applicability, e.g., by finding system-specific mechanisms. Of course, sr-KMC benefits from any advances in KMC methods. In addition, GP-based symbolic regression holds promise in other multiscaling areas, e.g., regressing constitutive rules and chemical reaction pathways, which we are studying. Also, as we exemplify, standard basis-set regression are generally not competitive to GP for fixed accuracy due to the difficulty in choosing appropriate basis functions.

To demonstrate, we discuss GP and its application to a non-trivial case of vacancy-assisted migration on (100) surface of phase-separating Cu₇₆Co₁₄₋ₓ. Although there are millions of configurations, only the environmental atoms locally around vacancy and migrating atom significantly influence the barrier energies. We refer to these as the active configurations. The results show that GP predicts barriers within 0.1–1% error using calculated barriers of less than 3% of the total active configura-
olutions. For alloys, this technique can be combined with a local cluster expansion technique \cite{14} that reduces the explicit barrier calculations to $\sim 0.3\%$ of the active configurations. Our initial results hold promise to enable the use of KMC (even with realistic potentials) for increased problem complexity with a scale-up of simulation time.

Genetic programming \cite{15} is a genetic algorithm that evolves computer programs. The program is represented by a tree consisting of functions in the internal nodes and terminals in the leaf nodes (Fig. 2a). Here we use the function set $F = \{+,-,*,/,*\exp,\sin\}$ and the terminal set $T = \{\vec{x},\mathcal{R}\}$, where $\vec{x}$ is a vector representing the active alloy configuration, and $\mathcal{R}$ is an ephemeral random constant $\mathbb{R}$. Since we use GP for predicting the barriers, a tree represents a PES-prediction function that takes a configuration and ephemeral constants as inputs and returns the barrier for that configuration as output.

A tree’s quality is given by its fitness $f$. For this, we calculate the barriers $\{\Delta E_{\text{calc}}(\vec{x}_1), \ldots, \Delta E_{\text{calc}}(\vec{x}_M)\}$ for $M$ random configurations $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_M\}$. These configurations are used as inputs to the tree and the barriers $\{\Delta E_{\text{pred}}(\vec{x}_1), \ldots, \Delta E_{\text{pred}}(\vec{x}_M)\}$, are predicted. The fitness is then computed as a weighted average of the absolute error between the predicted and calculated barriers:

$$f = \frac{1}{M} \sum_{i=1}^{M} w_i |\Delta E_{\text{pred}}(\vec{x}_i) - \Delta E_{\text{calc}}(\vec{x}_i)|$$

with $w_i = |\Delta E_{\text{calc}}|^{-1}$, which gives preference to accurately predicting lower energy (most significant) events.

Unlike traditional search methods, GP uses a population of candidate solutions (PES prediction functions) that are initially created using the ramped half-and-half method \cite{17}. Once the population is initialized and evaluated, the following genetic operators are repeatedly applied till one or more convergence criteria are satisfied:

- **Selection**: allocates more copies to solutions with better fitness values. We use an $s$-wise tournament selection \cite{16}, where $s$ candidate solutions are randomly chosen and pitted against each other in a tournament. A solution with the best fitness wins.

- **Recombination** combines bits and pieces of two solutions to create new, hopefully better, solutions. We use subtree crossover \cite{16}, where a crossover point for each solution is randomly chosen and subtrees below the point are swapped to create two new solutions, see Fig. 2.

- **Mutation** locally but randomly modifies a solution. We use two mutation techniques (see Fig. 2): Subtree mutation, where a subtree is randomly replaced with another randomly created subtree, and point mutation where a node is randomly modified.

**Case Study**: We consider the prediction of diffusion barriers for vacancy-assisted migration on (100) surface of phase-separating Cu$_2$Co$_{1-x}$. The system consists of five layers with 100 to 625 atoms in each layer (see Fig. 3). The bottom three layers are held fixed to their bulk bond distances, while the top layers are either held fixed (as a test) or fully relaxed via MD. We consider only first and second nearest-neighbor (n.n.) jumps, along with 1st (as a test) and 2nd n.n. environmental atoms in the active configuration, as shown in Fig. 3. This system already exhibits large complexity and is still small enough so that table look-up and GP-regressed KMC can be implemented and directly compared. Table II gives the number of active configurations when 1st and 2nd n.n. environments are considered for a binary alloy.

We model the atomic interactions with a simple Morse potential \cite{17} and a tight-binding potential with second-moment approximation (TB-SMA) \cite{18}. To validate interactions, we model vacancy-assisted migration on (100)-surface of Cu and consider only first n.n. jumps. The predicted barrier for n.n. vacancy jumps with fully relaxed lattice in Cu is 0.39 eV for Morse and 0.45 eV.
for TB-SMA, agreeing with $0.42 \pm 0.08$ ($0.47 \pm 0.05$) from 
*ab initio* (EAM) calculations.

We now consider the barrier regression via GP for vacancy-assisted migration on (100)-surface of Cu$_x$Co$_{1-x}$. The input to the barrier regression (i.e., prediction) function, $\tilde{x} = \{x_j\}$ is a binary-encoded vector sequence, where $x_j = 0$ (1) represents a Cu (Co) atom. For simplicity, we begin by considering only seven 1st n.n. environmental atoms yielding 128 active configurations. About 20, i.e., 16%, different active configurations are randomly chosen and their barriers are computed using the conjugate-gradient method and are used in the GP fitness function, see Eq. 1. The barriers predicted by the conjugate-gradient method and are used in the GP performance is independent of the configuration set database of barriers and has no knowledge (nor the need) of the underlying potential used. We also find that the GP performance does not depend on the potentials used, e.g., Fig. 5 shows results for both Morse, and non-additive and non-linear tight-binding potentials. The regression only requires a database of barriers and has no knowledge (nor the need) of the underlying potential used. We also find that the GP performance is independent of the configuration set used in calculating the fitness function, the order in which they are used, and the labeling scheme used to convert the configuration into a vector of inputs. Differences in activation-energy scale on the PES prediction via GP is also negligible. That is, even though the barriers for the

| Table I: Number of active configurations for 1st and 2nd n.n. jumps, and for 1st and 2nd n.n. active atoms |
|-------------------------------------------------------------------------------------------------|
| 1st n.n. active configurations | 128 | 128 |
| 2nd n.n. active configurations | 2048 | 8192 |
| Total configurations | $\gg 2^{100}$ | $\gg 2^{100}$ |

FIG. 4: Activation energies (in eV) predicted by regression. GP (circles) and a quadratic polynomial (crosses) are compared to the calculated (Morse) barriers for 1st n.n. jumps on (100)-surface of Cu$_{0.5}$Co$_{0.5}$ for relaxed lattices. As a simple test, only first n.n. environments are considered in the active configuration. The line is a guide for the eye.

FIG. 5: (Upper) Calculated vs. GP-predicted barriers (in eV) for 2nd n.n. jumps on relaxed (100)-surface of Cu$_{0.5}$Co$_{0.5}$ active configurations up to 2nd n.n. for Morse and TB-SMA. (Lower) GP predicts the barriers with 0.1% (1%) error for most- (less-) significant events with $\Delta E < 4.8$ eV ($\Delta E > 4.8$ eV) from only 3% of active configurations.
1st and 2nd n.n. jumps differ by an order of magnitude, GP predicts the barriers with similar accuracy. Moreover, for more complex, cooperative effects, such as island diffusion via surface dislocations, sr-KMC could be interfaced with pattern-recognition methods (see [25] for long-range fields).

Time enhancements by coupling a GP-regressed barriers with KMC (or sr-KMC) are simple to estimate. For our example, with ~33 times fewer calculated barriers GP symbolically regresses an in-line barrier function—rather than the complete look-up table—and thus, sr-KMC provides a direct CPU savings of ~100 over table look-up methods. Additionally, each time step of sr-KMC requires only $10^{-3}$ CPU-seconds for an in-line function evaluation, as opposed to on-the-fly KMC which require seconds (empirical potentials) to hours (quantum methods), thus providing a gain of $10^{4}$–$10^{7}$ CPU-seconds. For our example, one relaxed barrier calculation takes ~10 secs (~1800 secs) for Morse (TB-SMA). One important question, especially for bulk diffusion, is how the gain from sr-KMC scales with system complexity. While we cannot fully answer this question yet, in the present study it is remarkable and promising that the fraction of explicit barrier calculations required by sr-KMC decreases as the number of active configurations increases.

To summarize, potential energy surface prediction using symbolic regression via genetic programming (GP) holds promise as an efficient tool for multi-scaling in dynamics. The GP based KMC approach avoids the need or expense of calculating the entire potential-energy surface, is highly accurate, and leads to significant scale-up in simulation time and a reduction in CPU time. We have shown on a non-trivial example of vacancy-assisted migration on a surface of Cu$_x$Co$_{1-x}$ that GP predicts all barriers with 0.1–1% error from calculations for only 3% of active configurations, allowing seconds of simulation time. For alloy problems, the number of direct barrier calculations can further be reduced by over an order of magnitude by hybridizing GP with cluster expansion methods [14]. We emphasize that the GP is non-trivially regressing a function and its coefficients that approximates the potential-energy surface, and its efficacy over standard basis-set regression is clear. Moreover, GP approach is not problem specific and requires little modification, if any (say by choice of operators and functions), to address increasingly complex cases, and as suggested, is potentially useful in other multiscaling areas.

This work was supported by the NSF at Materials Computation Center (ITR grant DMR-99-76550), CPSD (ITR grant DMR-0121695), AFOSR (grant F49620-00-0163), Computational Science & Engineering fellowship and the Dept. of Energy through the Fredrick Seitz MRL (grant DEFG02-91ER45439) at UIUC. We gratefully acknowledge discussions with R.S. Averback, Y. Ashkezary, and J.-M. Roussel.

[1] A. F. Voter, Phys. Rev. Lett. 78, 3908 (1997).
[2] A. F. Voter, Phys. Rev. B 57, R13985 (1998).
[3] A. F. Voter, F. Montalenti, and T. C. Germann, Annu. Rev. Mater. Res. 32, 321 (2002).
[4] J. Jacobsen, B. H. Cooper, and J. P. Sethna, Phys. Rev. B 58, 15847 (1998).
[5] T. Diaz De La Rubia et al., J. Comp.-Aided Mat. Design 5, 243 (1998).
[6] G. Henkelman and H. Jónsson, J. Chem. Phys. 111, 7010 (1999), ibid 113, 9978 (2000); ibid 115, 9657 (2001).
[7] M. R. Sørensen and A. F. Voter, J. Chem. Phys. 112, 9599 (2000).
[8] W. Cai, M. H. Kalos, M. de Koning, and V. V. Bulatov, Phys. Rev. E 66, 046703 (2002).
[9] M. M. Steiner and P.-A. Genilloud, Phys. Rev. B 57, 10236 (1998).
[10] G. T. Barkema and N. Mousseau, Phys. Rev. Lett. 77, 4358 (1996).
[11] G. T. Barkema and N. Mousseau, Comput. Mat. Sc. 20, 285 (2001).
[12] K. Binder, ed., Monte Carlo methods in statistical physics (Springer, Berlin, 1986).
[13] J. L. Bocquet, Defect and Diffusion Forum 203–205, 81 (2002).
[14] A. Van der Ven and G. Ceder, Phys. Rev. B 64, 184307 (2001).
[15] J. R. Koza, Genetic programming: On the programming of computers by means of natural selection (MIT Press, Cambridge, MA, 1992).
[16] D. E. Goldberg, B. Korb, and K. Deb, Complex Systems 3, 493 (1989).
[17] L. Girifalco and V. Weizer, Phys. Rev. 114, 687 (1959).
[18] N. Levanov et al., Phys. Rev. B 56, 15847 (1997).
[19] G. Boisvert and L. Lewis, Phys. Rev. B 65, 184307 (2002).
[20] L. Girifalco and V. Weizer, Phys. Rev. Lett. 74, 2760 (1995).
[21] Y. L. Bouar and F. Soisson, Phys. Rev. B 57, 15847 (1998).
[22] N. Levanov et al., Phys. Rev. B 56, 15847 (1997).
[23] The average relative error for $N'_{\text{dfs}}$ configurations within the desired energy range is given by

$$\varepsilon_{\text{rel}} = \frac{100}{N'_{\text{dfs}}} \sum_{i=1}^{N'_{\text{dfs}}} \left| \frac{\Delta E_{\text{pred}}(\vec{x}_i) - \Delta E_{\text{calc}}(\vec{x}_i)}{\Delta E_{\text{calc}}(\vec{x}_i)} \right| ,$$

[24] Talat Rahman, private communication
[25] For long-range fields (e.g., elastic fields from coherent interfaces, such as multilayers or precipitates), a description based solely on local configurations may have to be extended, say, with phase field methods.
Calculated activation energy (eV)

GP-predicted $\Delta E$ (eV)

Percentage of active configurations to fit GP

Avg. rel. error (%)

Low-Energy events: $\Delta E < 4.8\text{eV}$

- Rigid lattice
- Relaxed lattice