Active Learning for Robust, High-Complexity Reactive Atomistic Simulations

Rebecca K. Lindsey,[1, a)] Laurence E. Fried,[1] Nir Goldman,[1, 2] and Sorin Bastea[1]

[1)] *Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

[2)] *Department of Chemical Engineering, University of California, Davis, California 95616, United States*

(Dated: 15 July 2020)

Machine learned reactive force fields based on polynomial expansions have been shown to be highly effective for describing simulations involving reactive materials. Nevertheless, the highly flexible nature of these models can give rise to a large number of candidate parameters for complicated systems. In these cases, reliable parameterization requires a well-formed training set, which can be difficult to achieve through standard iterative fitting methods. Here we present an active learning approach based on cluster analysis and Shannon information theory to enable semi-automated generation of informative training sets and robust machine learned force fields. Use of this tool is demonstrated for development of a model based on linear combinations of Chebyshev polynomials explicitly describing up to four-body interactions, for a chemically and structurally diverse system of C/O under extreme conditions. We show that this flexible training repository management approach enables development of models exhibiting excellent agreement with Kohn–Sham density functional theory (DFT) in terms of structure, dynamics, and speciation.

[a)] To whom correspondence should be addressed.lindsey11@llnl.gov

I. INTRODUCTION

Machine Learning (ML) has gained significant traction in the force field development community[1–3], largely due to the afforded versatility and potential to significantly decrease the human effort required to generate high fidelity, complex models. ML methods are well suited for problems demanding first principles-level accuracy in conjunction with the computational efficiency of force field-based methods. This is particularly true for materials under extreme conditions (e.g. 1000s of K and 10s to 100s of GPa) which can be highly reactive and exhibit phase separation over several ns, leading to chemical and structural heterogeneity on several-nm scales (e.g. in the case of reaction-driven carbon condensation in shock-compressed materials[4–6]). In these cases, machine learned interatomic potentials can offer an ideal balance between predictive power and computational efficiency, allowing simulations to more closely approach experimental time and length scales than possible with quantum simulations alone.

Machine-learning can be leveraged in a diversity of manners for model development, ranging from the model and/or mathematical representation[7,8] to the scheme used for training[9–14] and data generation[15–19]. Though powerful, each of these applications have significant associated challenges. For example, the literature contains several success stories surrounding neural network- and polynomial expansion-based interatomic interaction potentials, but these models can be enormous (e.g. comprised of thousands to tens of thousands of parameters), and thus highly susceptible to overfitting. Generating quality training data for these cases is difficult because the immense amount of training data needed to prevent overfitting makes human inspection challenging (i.e., determining what the data "looks like" is not a trivial matter). Moreover, these data are often obtained from molecular dynamics (MD) trajectories, which can yield correlated configurations on computationally accessible timescales.

Active learning (AL) has become a popular means of managing these challenging training problems. Typically, this concept centers around the manner in which a training repository is updated and maintained and can be accomplished through a variety of methods[20]. In general, these approaches involve an iterative framework (i.e. initial models are fit to DFT-MD data and subsequent refinements are made via single-point DFT calculations on frames generated through MD simulations with the i^{th} model and subsequent addition to the training

repository, from which the $i{+}1^{th}$ model is generated), but also include a scheme for deciding which new configurations should be added to the central training repository. The first such method applied to force field development was based on the concept of "committee-driven" decisions, where configurations yielding large disagreement among multiple models fit to distinct subsets of a training set are added to the central repository[15,18,19]. Other successful methods include "distance" based decisions, where configurations exhibiting a great enough distance from previous configurations in fingerprint space are added to the central repository[16], as well as approaches arising from design of experiments[17], etc[21]. The success and efficiency of each of these methods is strongly linked to the nature of the target model and its implementation. For example, committee-based active learning is best used with models of high computational efficiency but relatively slow non-linear fitting as the final "model" requires evaluating each run-time simulation frame with the entire "committee."

Here, we are concerned with developing an active learning approach well suited for parameterically linear machine learned models (e.g. SNAP[12], ChIMES[22], Shapeev's moment tensor potentials[23], etc.), for which the parameter solution step is rapid compared to parametrically nonlinear models. We present an alternative active learning approach leveraging cluster analysis in conjunction with Shannon information theory,[24] and demonstrate its application to one such model with explicit two-, three-, and four-body interactions (i.e. approximately 4000 parameters and thus highly susceptible to overfitting) for a C/O system under reactive conditions. The following sections provide the model functional form and description of the proposed AL approach. Resulting models are benchmarked against DFT in terms of predicted small molecule chemistry, pressure, diffusion coefficients, atomic forces, as well as molecular and overall system energetics.

In this section, we provide an overview of ChIMES (i.e. the testbed machine-learned force field for this work), revisit a fitting problem for which the necessary model complexity precluded generating a ChIMES force field through the standard iterative approach, and present the active learning framework developed to overcome this challenge.

A. The ChIMES Force Field

The recently developed machine-learned Chebyshev Interaction Model for Efficient Simulation (ChIMES) provides an excellent testbed for the proposed active learning framework.

ChIMES models are comprised of linear combinations of Chebyshev polynomials explicitly describing many-body interactions, where the high degree of flexibility afforded by the basis makes ChIMES highly suitable for problems in chemistry and capable of "quantum accuracy." As a consequence of this polynomial basis, ChIMES models are completely linear in fitted coefficients and thus rapidly parameterizable, computationally efficient and scale linearly with system size. As will be described in greater detail below, ChIMES force fields are fit to forces (and optionally energies and stresses) arising from Kohn–Sham density functional theory (DFT) simulations of 2-20 ps and have typically leveraged 2+3-body interaction terms and an iterative refinement scheme, where frames from molecular dynamics (MD) simulations with the i^{th} ChIMES force field are occasionally sent back to DFT for single point calculation and combined with the training repository, from which an $i+1^{\text{th}}$ ChIMES model is generated; the cycle is repeated until desired model performance is achieved. This iterative approach has worked well for molten carbon[22], ambient water[25], and dissociative carbon monoxide[26], where in all cases species were small and chemistry was rapid when present. However, upon application to systems in which larger and more complex species form, shortcomings arising from use of a 2+3-body ChIMES many-body truncation have been identified[26]. Though the ChIMES equations can be easily extended to include higher-bodied interactions, the resulting increase in model complexity necessitates intelligent and automated model development tools.

The generalized ChIMES potential energy equation is given by:

$$E_{n_\text{B}} = \sum_{i_1}^{n_\text{a}} {}^1E_{i_1} + \sum_{i_1>i_2}^{n_\text{a}} {}^2E_{i_1 i_2} + \sum_{i_1>i_2>i_3}^{n_\text{a}} {}^3E_{i_1 i_2 i_3} + \cdots + \sum_{i_1>i_2\ldots\ i_{n_\text{B}-1}>i_{n_\text{B}}}^{n_\text{a}} {}^{n_\text{B}}E_{i_1 i_2 \ldots i_{n_\text{B}}}, \qquad (1)$$

where E_{n_B} is the total ChIMES system energy, n_B is the maximum bodiedness, ${}^nE_{i_1 i_2 \ldots\ i_n}$ is the n-body ChIMES energy for a given set of n atoms with indices $\boldsymbol{i} = \{i_1, i_2, \ \ldots\ , i_n\}$, and n_a is the total number of atoms in the system.

In the ChIMES framework, single-body energies are constant values and n-body energies are constructed from the product of polynomials of transformed atom pair distances. Thus, a 2-body interaction would involve a single pair, ij, while a three-body interaction would involve 3 pairs, ij, ik, jk, a 4-body interaction would involve $\binom{4}{2}$ pairs, and so on. Taking as an example a 3-body interaction, we define the following: $\boldsymbol{A} = \{i, j, k\}$ is the index over atoms within an interaction cluster, with the corresponding set of pairs given by $\boldsymbol{P} = \{ij, ik, jk\}$, their element pair types by $\boldsymbol{E} = \{e_i e_j, e_i e_k, e_j e_k\}$, and the polyno-

mial orders for each pair given by $\boldsymbol{O} = \{\alpha, \beta, \gamma\}$. Analogous conventions are used for a 4-body interaction. $\boldsymbol{A} = \{i, j, k, l\}$, $\boldsymbol{P} = \{ij, ik, il, jk, jl, kl\}$, element pair types are $\boldsymbol{E} = \{e_i e_j, e_i e_k, e_i e_l, e_j e_k, e_j e_l, e_k e_l\}$, and the polynomial orders for each pair are given by $\boldsymbol{O} = \{\alpha, \beta, \gamma, \delta, \epsilon, \zeta\}$. Two mapping functions are used to relate pair indices $\boldsymbol{P}$ to the three aforementioned pair properties: $m_1 = \boldsymbol{P} \rightarrow \boldsymbol{E}$, and $m_2 = \boldsymbol{P} \rightarrow \boldsymbol{O}$. The index y refers to a particular component of $\boldsymbol{P}$, defining an interaction pair.

Using these definitions, we write the generalized ChIMES energy for a cluster of n atoms as:

$$ {}^{n}E_{\boldsymbol{A}} = \prod_{y \in \boldsymbol{P}} f_{\mathrm{s}}^{m_1(y)}\left(r_y\right) \times \sum_{\boldsymbol{O}}^{\mathcal{O}_n^*} c_{\boldsymbol{O}}^{\boldsymbol{E}} \prod_{y \in \boldsymbol{P}} T_{m_2(y)}\left(s_y^{m_1(y)}\right). \tag{2} $$

As is given above, the $\sum_{\boldsymbol{O}}$ notation indicates a multiple sum for which there are $\binom{n}{2}$ distinct indices, $\mathcal{O}_n^*$ is the maximum polynomial order for an n body interactions, and the asterisk indicates a sufficient number of non-zero terms exist that the graph formed by the edges of interacting atoms connects all n atoms, which guarantees a true n-body interaction. $T_{m_2(y)}\left(s_y^{m_1(y)}\right)$ is a Chebyshev polynomial of order $m_2(y)$ that depends on pair distance $s_y^{m_1(y)}$ for pair y of atom types $m_1(y)$ that has been transformed from r_y, to ensure it exists in the $[-1, 1]$ domain over which Chebyshev polynomials are defined, and $f_{\mathrm{s}}^{m_1(y)}\left(r_y\right)$ is a cutoff function that ensures smooth behavior at the outer cutoff. For the special case of a two-body interaction one has:

$$ {}^{2}E_{ij} = f_{\mathrm{p}}^{e_i e_j}\left(r_{ij}\right) + f_{\mathrm{s}}^{e_i e_j}\left(r_{ij}\right) \sum_{\alpha=1}^{\mathcal{O}_{2\mathrm{B}}} c_{\alpha}^{e_i e_j} T_{\alpha}\left(s_{ij}^{e_i e_j}\right), \tag{3} $$

where f_p is a short-ranged repulsive interaction described later. Higher body interactions follow the form of Eq. 2. We have for a three-body interaction:

$$ {}^{3}E_{ijk} = f_{\mathrm{s}}^{e_i e_j}\left(r_{ij}\right) f_{\mathrm{s}}^{e_i e_k}\left(r_{ik}\right) f_{\mathrm{s}}^{e_j e_k}\left(r_{jk}\right) \sum_{\alpha=0}^{\mathcal{O}_{3\mathrm{B}}} \sum_{\beta=0}^{*} \sum_{\gamma=0} c_{\alpha,\beta,\gamma}^{\boldsymbol{E}} T_{\alpha}\left(s_{ij}^{e_i e_j}\right) T_{\beta}\left(s_{ik}^{e_i e_k}\right) T_{\gamma}\left(s_{jk}^{e_j e_k}\right), \tag{4} $$

and for a four-body interaction:

$$ \begin{aligned} {}^{4}E_{ijkl} = {} & f_{\mathrm{s}}^{e_i e_j}\left(r_{ij}\right) f_{\mathrm{s}}^{e_i e_k}\left(r_{ik}\right) f_{\mathrm{s}}^{e_i e_l}\left(r_{il}\right) f_{\mathrm{s}}^{e_j e_k}\left(r_{jk}\right) f_{\mathrm{s}}^{e_j e_l}\left(r_{jl}\right) f_{\mathrm{s}}^{e_k e_l}\left(r_{kl}\right) \\ & \sum_{\alpha=0}^{} \sum_{\beta=0}^{} \sum_{\gamma=0}^{\mathcal{O}_{4\mathrm{B}}} \sum_{\delta=0}^{*} \sum_{\epsilon=0}^{} \sum_{\zeta=0}^{} c_{\alpha,\beta,\gamma,\delta,\epsilon,\zeta}^{\boldsymbol{E}} T_{\alpha}\left(s_{ij}^{e_i e_j}\right) T_{\beta}\left(s_{ik}^{e_i e_k}\right) T_{\gamma}\left(s_{il}^{e_i e_l}\right) T_{\delta}\left(s_{jk}^{e_j e_k}\right) T_{\epsilon}\left(s_{jl}^{e_j e_l}\right) T_{\zeta}\left(s_{kl}^{e_k e_l}\right) \end{aligned} \tag{5} $$

Transformed pair distances $s_y^{m_1(y)}$ are obtained via[8]:

$$x_y^{m_1(y)} = \exp(-r_y/\lambda^{m_1(y)}) \tag{6}$$

$$x_{\text{avg}}^{m_1(y)} = 0.5(x_{\text{c,out}}^{m_1(y)} + x_{\text{c,in}}^{m_1(y)}) \tag{7}$$

$$x_{\text{diff}}^{m_1(y)} = 0.5|x_{\text{c,out}}^{m_1(y)} - x_{\text{c,in}}^{m_1(y)}| \tag{8}$$

$$s_y^{m_1(y)} = (x_y^{m_1(y)} - x_{\text{avg}}^{m_1(y)})/x_{\text{diff}}^{m_1(y)} \tag{9}$$

where $s_y^{m_1(y)}$ is the pair distance and $\lambda^{m_1(y)}$ can be considered a characteristic bonding distance, typically set to the location of the first peak in the DFT radial distribution function for the $m_1(y)$ atom pair type, and $r_{\text{c,in}}^{m_1(y)}/r_{\text{c,out}}^{m_1(y)}$ are the corresponding inner/outer cutoff radii. We note that Eq. 6 enforces a natural decrease in interaction strength as distance is increased, and increasing $\lambda^{m_1(y)}$ has the effect of decreasing the rate of interaction decay as $r_{\text{c,out}}^{m_1(y)}$ is approached, in a Morse-like[8] fashion.

Finally, $f_{\text{s}}^{m_1(y)}\left(r_y\right)$ in Eq. 2 ensures the potential goes smoothly to zero at the outer cutoff, $r_{\text{c,out}}^{m_1(y)}$. In departure from earlier ChIMES work where $f_{\text{s}}^{m_1(y)}\left(r_y\right)$ took on a cubic form, we employ a Tersoff style cutoff function[27] in the present work:

$$f_{\text{s}}^{m_1(y)}\left(r_y\right) = \begin{cases} 0, & \text{if } r_y > r_{\text{c,out}}^{m_1(y)} \\ 1, & \text{if } r_y < d_{\text{t}} \\ \frac{1}{2} + \frac{1}{2}\sin\left(\pi\left[\frac{r_y - d_{\text{t}}}{r_{\text{c,out}}^{m_1(y)} - d_{\text{t}}}\right] + \frac{\pi}{2}\right), & \text{otherwise} \end{cases} \tag{10}$$

where the threshold distance is given by $d_{\text{t}} = r_{\text{c,out}}^{m_1(y)}\left(1 - f_{\text{O}}\right)$, and f_{O} is a value in [0,1] taken to be 0.5, here. This new form exhibits a smooth step, allowing ${}^{n}E_{\boldsymbol{A}}$ to remain unmodified by the smoothing function for all $r_{m_1(y)} < d_{\text{t}}$; this is particularly useful for many-body interactions of large n, where the product of $\binom{n}{2}$ $f_{\text{s}}^{m_1(y)}\left(r_y\right)$ factors is used, and can otherwise severely reduce ${}^{n}E_{\boldsymbol{A}}$ contributions to the total energy for $n > 2$.

To prevent spurious close contact, a penalty term is added to each two-body energy ${}^{2}E$ in the system:

$$f_{\text{p}}^{m_1(y)}\left(r_y\right) = \begin{cases} A_{\text{p}}^{m_1(y)}\left(r_{\text{c,in}}^{m_1(y)} + d_{\text{p}}^{m_1(y)} - r_y\right)^3, & \text{if } r_y < r_{\text{c,in}}^{m_1(y)} + d_{\text{p}} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $A_{\mathrm{p}}^{m_1(y)}$ is the penalty prefactor and $d_{\mathrm{p}}^{m_1(y)}$ is the penalty initiation distance set to 10^5 kcal/(mol Å^3) and 0.01 Å here, respectively.

Permutational invariance of the energy is explicitly enforced. In particular, we require ${}^nE_A = {}^nE_{\Pi A}$, where Π is a permutation operator acting on the n atoms in the cluster. This leads to equality conditions among the coefficients

$$c_{\boldsymbol{O}}^{\boldsymbol{E}} = c_{\Pi\boldsymbol{O}}^{\Pi\boldsymbol{E}}. \tag{12}$$

As an example, for a 4-body interaction we have

$$c_{\alpha\beta\gamma\delta\epsilon\zeta}^{e_ie_j,e_ie_k,e_ie_l,e_je_k,e_je_l,e_ke_l} = c_{\alpha\delta\epsilon\beta\gamma\zeta}^{e_je_i,e_je_k,e_je_l,e_ie_k,e_ie_l,e_ke_l} \tag{13}$$

which is derived by permuting atoms i and j. In our implementation of ChIMES, permutational invariance is enforced by treating permutationally related coefficients as the same unique fitting variable.

In this work, we will consider the following objective function, which contains terms for per-atom forces and per-system-configuration energies, though we note additional terms for the system stress tensor can also easily be included[14,22,25,26]:

$$F_{\mathrm{obj}} = \frac{1}{n_{\mathrm{f}}\left(3n_{\mathrm{a}}+1\right)} \sum_{i=1}^{n_{\mathrm{f}}} \left[\sum_{j=1}^{n_{\mathrm{a}}} \sum_{k=1}^{3} w_{\mathrm{F}_{ijk}}^2 \left(\Delta F_{ijk}\right)^2 + w_{\mathrm{E}_i}^2 \left(\Delta E_i\right)^2 \right], \tag{14}$$

where $\Delta X = X^{\mathrm{DFT}} - X^{\mathrm{ChIMES}\{c\}}$. F_{obj} and $\{c\}$ are the weighted root-mean-squared error and model coefficients, respectively. The number of frames and atoms are given by n_{f} and n_{a}, respectively, and the factor of 1 in the denominator arises from inclusion of a single per-configuration energy, E_i. F_{ijk} indicates the k^{th} Cartesian component of the force on atom j in configuration i. Units of kcal mol^{-1} Å^{-1} and kcal mol^{-1} and weights of 1.0 and 5.0 were used for forces and energies, respectively (i.e. $w_{\mathrm{F}_{ijk}}$ and w_{E_i}). The superscripts "ChIMES" and "DFT" indicate forces/energies predicted from the present force-matched model and the DFT molecular dynamics (DFT-MD) training trajectory, respectively.

Since ChIMES is entirely linear in its fitted parameters, the model optimization problem can be recast as the following over-determined matrix equation:

$$\boldsymbol{wMc} = \boldsymbol{wX}_{\mathrm{DFT}}, \tag{15}$$

where $\boldsymbol{X}_{\mathrm{DFT}}$ is the vector of F_{ijk}^{DFT} and E^{DFT} values, $\boldsymbol{w}$ is a diagonal matrix of weights to be applied to the elements of $\boldsymbol{X}_{\mathrm{DFT}}$ and rows of $\boldsymbol{M}$, and the elements of design matrix $\boldsymbol{M}$

are given by:

$$M_{ab} = \frac{\partial X_{a,\mathrm{ChIMES}\{c\}}}{\partial c_b}. \tag{16}$$

In the above, a represents a combined index over force and energy components, and b is the index over permutationally invariant model parameters.

To avoid overfitting in determining c, we find that a regularization method must be used for most problems. In the present work, the least absolute shrinkage and selection operator (LASSO)[28] method is used to regularize c. The LASSO method minimizes the objective function:

$$F_{\mathrm{LASSO}} = F_{\mathrm{obj}} + 2\lambda \sum_{i=1}^{n_p} |c_i|, \tag{17}$$

where n_p is the total number of unique fitting parameters. The parameter λ penalizes large magnitude coefficients and reduces the likelihood of overfitting. Moreover, the absolute values (L1 norm) used in the LASSO objective function have the effect of setting certain coefficient values c_i to 0, which can lead to substantial gains in model efficiency.

We use a locally written code that implements the Least Angle Regression (LARS) algorithm[29,30] for LASSO. The LARS algorithm proceeds in stages where variables are added or removed one at a time. Each stage in the LARS algorithm is a solution of LASSO optimization for a value of λ larger than the requested value. When the requested value is reached, the algorithm terminates. We find that the LARS algorithm has better convergence properties than direct optimization of F_{LASSO}, and additionally allows the analysis of solutions for each iteration. Our code distributes the $\boldsymbol{M}$ matrix between computing nodes, allowing for solution of large (e.g. $>$ 1 TB) problems, and uses rank 1 Cholesky decomposition updates to solve linear equations that arise in LARS.

The standard ChIMES iterative fitting approach is given below. We note that the success of this method hinges upon the notion that the relatively inaccurate models produced during early iterations are more likely to reach unsampled regions of physicochemical space and can thus be considered a means of rare-event sampling. As briefly mentioned above, this approach has been successful for systems of non-reactive small molecules, or those exhibiting rapid chemistry. To generate these models, a simple iterative refinement framework was employed, where (i) training trajectories were obtained from short DFT-MD trajectories at the state points of interest for the system, (ii) a model was obtained by minimizing the objective function (i.e. Eq. 14), (iii) a MD trajectory was launched using this i^{th} force

field, (iv) configurations from this simulation were periodically sent to DFT for single point calculation to be merged with the existing repository, and (v) steps ii through iv were repeated until the model exhibited the target level of accuracy.

B. Shortcomings of the Standard Parameterization Approach

The ChIMES model development scheme described above, (i.e. iterative fitting), has been found insufficient when at least one of the following two conditions are met: the system undergoes relatively slow chemistry, with species lifetimes exceeding 1 ps (i.e. where kinetics are limited by relatively large reaction barriers), or the system contains structurally complex species, e.g. oligomers, heterocycles, fused rings, etc. where greater-than 3-body interactions (e.g. intramolecular torsion) have been shown[26] to play a significant role. As an example of this, consider a system of 50/50% C/O at 2400 K at 1.79 g cm^{-3}. As shown in Fig. 1a and 1c, DFT simulations predict that both criteria are met for the above system. Previously, a ChIMES model was developed with the intention of transferability over T = 9350, 6500, and 2400 K, and ρ = 2.56, 2.5, and 1.79 g cm^{-3}[26]. This model was fit through the standard iterative approach using 2+3-body interactions and was found to work well between the 9350 and 6500 K state points, but failed at 2400 K. As indicated in Fig. 1b and 1c, one of two structural themes emerged during ChIMES MD simulations at 2400 K using models derived from successive iterative iterations; either exclusively small linear species, or unphysical ring-like structures featuring highly coordinated oxygen. These two outcomes arose from an insufficiently complex ChIMES interaction model; to a model containing only 2- and 3-body interactions, the coordinated oxygen structure is reasonable, containing bond and angle distances reminiscent of those found in carbon dioxide, ethylene oxide, dimethyl ether, etc. However, when iteratively added to the training repository, DFT assigns a high energy, resulting in subsequent models that bias away for cyclic structures. As a consequence, predicted chemistry is far off from the DFT-computed vales.

The above issues can be resolved through addition of 4-body terms, which in the language of molecular mechanics force fields would allow for description of bonds (2-body), angles (3-body) and dihedrals/impropers (4-body). However, doing so substantially increases the number of parameters considered in the fitting process. For example, in a system with two atom types, polynomial orders of $\mathcal{O}_{2B}/\mathcal{O}_{3B}/\mathcal{O}_{4B}$ = 12/7/0 yields a maximum of 806

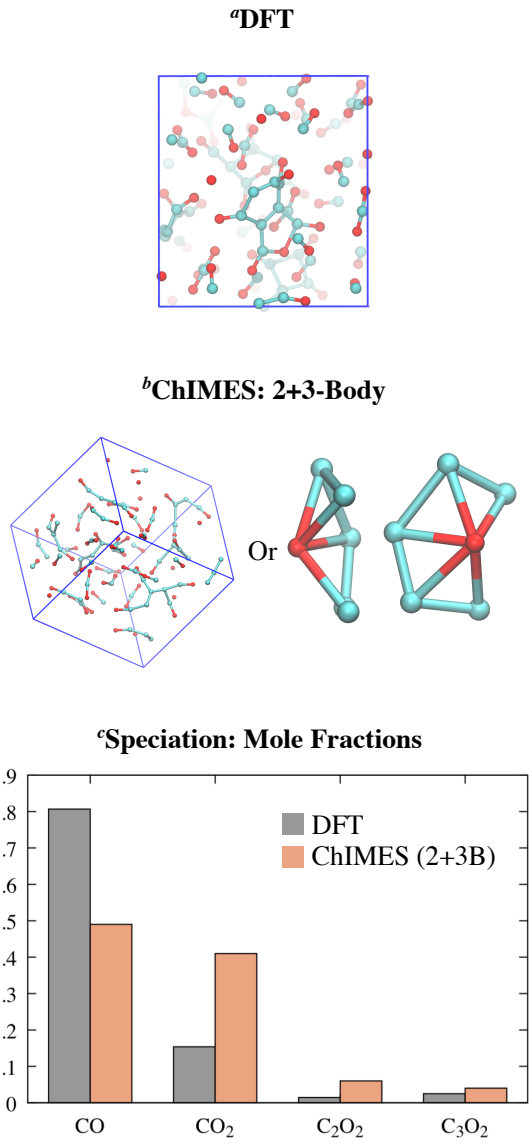

FIG. 1. Sample system configuration predicted by (a) DFT and (b) successive, iteratively generated $2+3-$body ChIMES force fields[26] for carbon monoxide at 2400 K and 1.79 g/cm^3; (c) comparison of corresponding mole fractions predicted by DFT and ChIMES.

parameters, whereas $\mathcal{O}_{2B}/\mathcal{O}_{3B}/\mathcal{O}_{4B} = 12/7/3$ (i.e. the polynomial orders used in the present work), yields a maximum of 3978 parameters. Increasing model complexity gives rise to an additional set of problems; beyond the concomitant increase in risk of over-fitting, a far greater number of iterative cycles are required to generate a converged force field, which is prohibitive for reactive carbon-containing systems (i.e. that exhibit significantly higher chemical complexity than those of only O, H, and/or N). In the following sections we describe

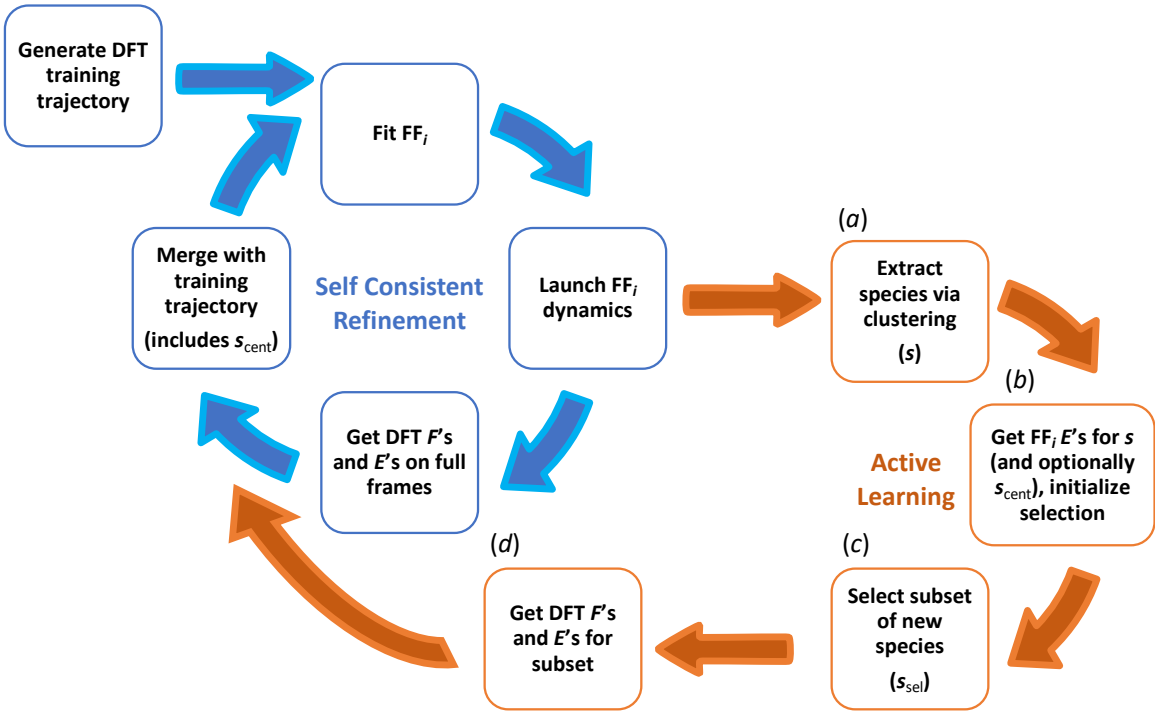

FIG. 2. The ChIMES active learning framework.

strategies to overcome these challenges, through development of an active learning framework combining cluster analysis and Shannon information theory.

C. Active Learning Framework Overview

As described in section I A, development of our active learning framework began with examination of our standard fitting approach. ChIMES training repositories contain both DFT- and iteratively-obtained frames from MD trajectories, comprised of $3n_{\mathrm{a}}$ coordinates and forces, and a single overall system energy. The convoluted nature of this fitting problem becomes apparent if one considers fitting a $\mathcal{O}_{2\mathrm{B}}/\mathcal{O}_{3\mathrm{B}}/\mathcal{O}_{4\mathrm{B}} = 12/7/3$ force field for a C/O system to only energies from DFT; doing so requires assignment of 3978 parameters such that the energies for each dimer, trimer, and tetramer sum to the single value for each frame.

The total energy in a given frame arises from the sum of bonded and non-bonded interactions, inter- and intra-molecular interactions, etc. Thus, by decomposing a given frame into a collection of individual atom clusters (i.e. nominal molecules), computing the corresponding DFT forces and energies, and adding them to the training repositories, resulting

fits contain greater information on how to assign the 10s to 1000s of of parameters giving rise to different 2-, 3-, and 4-body energetic contributions. This concept is similar to generating training configurations from a direct cluster expansion, but can be more effective for molecular systems because resulting configurations are relevant to both the model and the target physicochemical space. Note that the ChIMES model space can allow formation of unphysical many-body cluster configurations (e.g. 3 atoms forming an equilateral triangle with all distances equal to the corresponding 3-body inner cutoff). Thus, it is important that the training repository contain configurations of this nature to inform the fit of their unfavorability and prevent their spurious appearance during MD simulations. However, this approach is still highly inefficient for several reasons: (1) successive MD frames are highly correlated (e.g. due to the limited exploration of physicochemical space in short-time MD simulations at a fixed temperature(s)), (2) the species contained in each frame can be very similar (e.g. distributed about some chemical and/or conformational minimum energy), and (3) the computational cost of evaluating the possible tens of thousands of species via DFT would be prohibitive from a practical standpoint. In order to increase the efficiency of our fits, we aim to increase the information contained in the training repository while simultaneously maintaining a minimal size, by developing a method for selecting subsets of possible species. As will be discussed in greater detail in following sections, we do so by defining a "feature" for each species, simply taken to be $E_{\mathrm{ChIMES},i}$, the energy per species atom as computed by the i^{th} ChIMES force field; using this construct, the information contained in any given subset of species is maximized when the corresponding probability distribution of $E_{\mathrm{ChIMES},i}$ values is flattened. These concepts can be combined with the standard ChIMES iterative fitting procedure to form the active learning cycle shown in Fig. 2. We note the distinction between active learning ("AL"), represented by the orange components of the cycle in Fig. 2, and an active learning cycle ("ALC"), which involves *both* iterative refinement and active learning, i.e. the blue *and* orange components in Fig. 2.

D. Cluster-Based Species Identification

Fig. 3 provides a pictorial representation of the active learning (i.e. orange) portion of the overall fitting scheme given in Fig. 2. The first step of this process is extraction of all possible molecular species from the MD trajectory, i.e. step "a" in both Figs. 2 and 3. In this section,

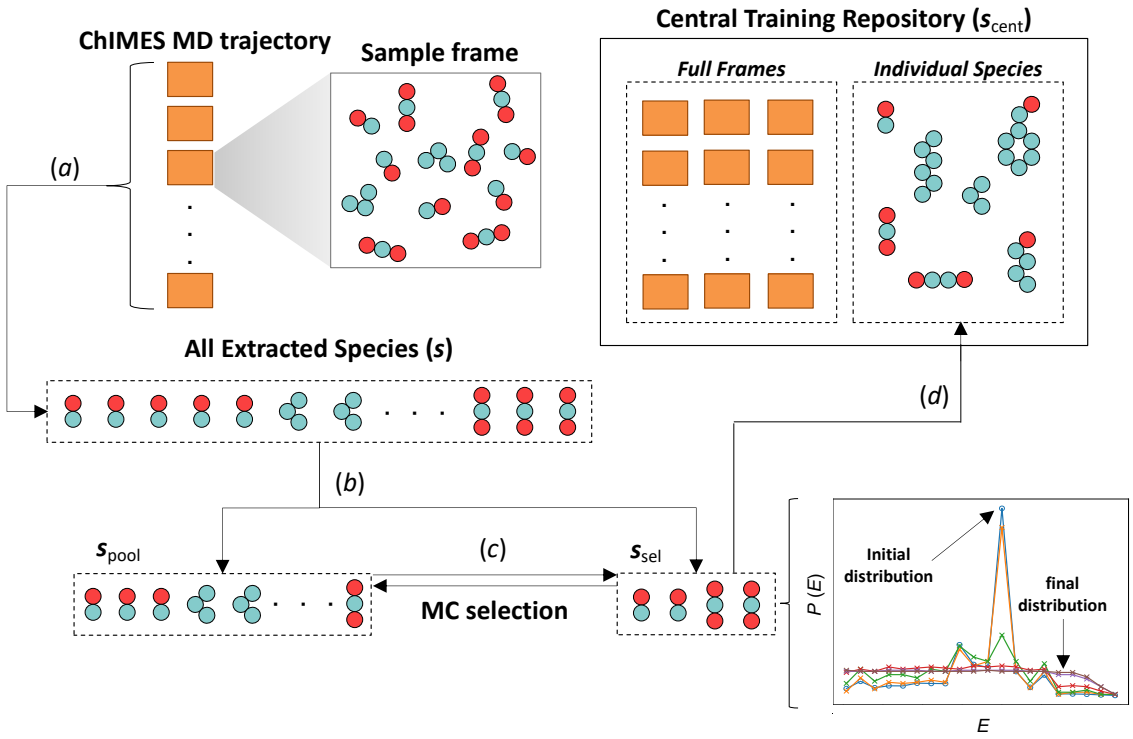

FIG. 3. Schematic representation of the active learning (i.e. orange) portion of an active learning cycle (i.e Fig. 2). Arrow a indicates species extraction via clustering and per-cluster energy assignment by the i^{th} ChIMES force field. Arrow b indicates randomized initialization of the cluster selection process by splitting all extracted species into an initial selection subset and candidate pool. The c arrows indicate exchange of species between the pool and selection subset via MC (i.e. flattening the histogram of selected cluster energies), and arrow d indicates single point DFT evaluation for each selected cluster and subsequent addition to the ever-growing central repository.

we describe the simple clustering approach through which this is achieved. Note that, for the remainder of this work, "cluster" is used to mean a collection of atoms presumed to be bonded or in the process of bonding, and can simply be considered an umbrella term for molecule, intermediate, and/or transition state species. High accuracy condensed phase chemistry is a central ChIMES goal and requires both molecules and "transition"/intermediate species be well described - capturing the former is critical for recovering reaction energy minima and thus speciation, while the latter influences reaction barriers, and thus predicted lifetimes. To ensure both types of species are identified we use a simple double-pass clustering approach.

The first pass is used to identify nominal molecules, where a relatively short-ranged distance-based criterion is identified for each possible pair type; atoms are then considered

part of the same cluster if their distance to any other cluster member are within this "tight" criteria. The second pass identifies nominal transition-state species, and uses looser distance criteria; typically, this pass results in a smaller number clusters with larger overall sizes. Because species identified in the second pass may be identical to those in the first pass (i.e. in the case of well-separated molecules), we take the final set, $\boldsymbol{s}$ of identified species as those unique across both passes, noting that in our algorithm, two clusters are considered identical only if they are obtained from the same frame with the same atomic coordinates and indices. In this work, respective values of 1.9, 1.8, and 1.7 Å are used as the "tight" criteria for CC, CO, and OO pairs, respectively (i.e. location of the first minima in each pair radial distribution function), while "loose" criteria were taken as $\approx 117\%$ these values, (i.e. corresponding roughly to the midpoint between the first minima and second maxima). Species were extracted from 250 evenly-spaced frames from the 6 ps MD simulations.

E. Monte Carlo for Cluster Subset Selection

Once the set of candidate clusters have been extracted from the MD trajectory as described above, the next task is identification of a much smaller subset of those clusters which are maximally informative to the fit (i.e. steps b and c in Figs. 2 and 3). This batch mode pool-based process[20] is critical from an efficiency standpoint as it drastically reduces the number of single-point DFT calculations required during addition to the training repository (i.e. step d); moreover, in selecting this subset by maximizing information, fewer active learning cycles (i.e. Fig. 2) will be required to achieve a converged result.

To achieve this, we look to statistical mechanics (SM) and Shannon information theory (IT). IT provides a measure of the information entropy (I) contained in a given dataset according to $I = -\sum p_i \ln p_i$, where p_i is the probability to observe the i^{th} value; the information contained in the data set is said to increase with increasing I. Note that the definition of I differs from the statistical mechanical definition of entropy by only a factor of k_B, where in SM, p_i is the probability to observe the i^{th} energy state; in both cases, information (entropy) is maximized when the distribution of p_i is uniform. In the present work we are concerned with obtaining a subset of all extracted clusters that will maximally inform our fits. According to SM, a subset constructed via *random selection* from a larger set will exhibit the same energetic distribution (e.g. Boltzmann). Randomly generated clusters

avoid this problem; however, generation of such a set would be computationally inefficient as our training data needs only to contain clusters of relevance to the model domain. Thus, we have implemented an intermediate approach which combines the SM and IE definitions of entropy to establish that the down-selected subset of fixed size n_{sel} possessing the maximum possible information in our feature space (i.e. $E_{\text{ChIMES},i}$, the energy per cluster atom predicted by the i^{th} ChIMES force field), yield a uniform probability distribution. Because our goal is selection of a finite number of species from a set, our probability distributions are comprised of bins with finite width, where high complexity species within a given bin can be structurally or conformationally isoenergetic. Thus, we devise a simple Monte Carlo (MC) approach for cluster selection; by using a statistical method (i.e. rather than selection of clusters from each bin in a single pass), we ensure a degree of randomness across selected species contained in each distribution bin.

Prior to the start of the Monte Carlo (MC) selection process, clusters extracted from the trajectory of interest are stored in a set ($\boldsymbol{s}$), of size n, as shown in Fig. 3. An energy ($E_{\text{cluster}}^{\text{ALC}-X}$) is assigned to each of the clusters in $\boldsymbol{s}$ with the X^{th} ChIMES model. The MC process aims to select a subset ($\boldsymbol{s}_{\text{sel}}$) of n_{sel} clusters from $\boldsymbol{s}$ which yields a flat distribution in energy. To initiate the process, n_{sel} random clusters are moved from $\boldsymbol{s}$ to $\boldsymbol{s}_{\text{sel}}$ and the remainder are put into a pool ("$\boldsymbol{s}_{\text{pool}}$") containing $n_{\text{pool}} = n - n_{\text{sel}}$ clusters. A histogram ($\boldsymbol{h}$) of cluster energies in $\boldsymbol{s}_{\text{sel}}$ is then constructed, defined over $[\min(\boldsymbol{s}), \max(\boldsymbol{s})]$, with n_{bins} bins. We note that later sections will explore alternative definitions of $\boldsymbol{h}$ and the effect of changing n_{bins}. Fig. 3 shows that, as expected, this initial histogram is sharply peaked (i.e. the orange line), and closely resembles the histogram obtained by considering all possible clusters (i.e. the blue line). In the remainder of this section we discuss how this histogram shape, indicating the the $\boldsymbol{s}_{\text{sel}}$ contains many similar clusters, can be flattened to evolve $\boldsymbol{s}_{\text{sel}}$ into a maximally informative subset of species for subsequent addition to our training repository (i.e. step d).

The MC selection process proceeds as follows: (1) a random cluster (i_{old}) having associated energy $E_{i_{\text{old}}}^{\text{ALC}-X}$ is selected from $\boldsymbol{s}_{\text{sel}}$ and its probability is taken as $p_{i,\text{old}} = \boldsymbol{h}(E_{i_{\text{old}}}^{\text{ALC}-X})/n_{\text{bins}}$ (2) a random cluster (i_{new}) having associated energy $E_{i_{\text{new}}}^{\text{ALC}-X}$ is selected from $\boldsymbol{s}_{\text{pool}}$ and its probability is taken as $p_{i,\text{new}} = \boldsymbol{h}(E_{i_{\text{new}}}^{\text{ALC}-X})/n_{\text{bins}}$ (3) if $1.0 + \frac{p_{i,\text{new}} - p_{i,\text{old}}}{2} > \text{rand}[0,1]$ the attempted move is rejected; otherwise the move is accepted (i.e. i_{old} is moved to $\boldsymbol{s}_{\text{pool}}$, i_{new} is moved to $\boldsymbol{s}_{\text{sel}}$, and $\boldsymbol{h}$ is recomputed). rand$[a,b]$ is defined to be a uniformly distributed

random number between a and b. We note that many MC acceptance criteria could be used here, and that ours is simply intended to bias toward selections yielding a flat probability distribution (i.e. equivalent histogram bins). In particular, the present criteria only accepts moves for which $p_{i,\mathrm{new}} < p_{i,\mathrm{new}}$, forcing the MC process to converge (i.e. stop accepting moves) once the histogram is completely flattened. This process is repeated for the total requested number of MC cycles (n_{cyc}), where one cycle is equal to the number of requested Monte Carlo steps divided by n, allowing each cluster in $\boldsymbol{s}$ to be considered for selection at least n_{cyc} times, in principle. Typically, n_{cyc} is set to $n_{\mathrm{bins}}/10$.

The Monte Carlo process is rapid, contributing only ≈ 5 minutes to the overall time required for any given ALC. Users must specify the following options to execute the above MC algorithm: n_{sel}, n_{cyc}, the number of histogram bins (n_{bins}), and the specific definition of $\boldsymbol{h}$ utilized. The two former options are set to 400 and 2 in this work unless otherwise stated. We note that choice of n_{sel} was based on identification of four small molecules present in the DFT-MD simulation with mole fractions of at least 0.02, i.e. CO, CO_2, C_2O_2, and C_3O_2. In principle, $n_{\mathrm{sel}} = 400$ allows for 100 of each aforementioned species to appear in the final $\boldsymbol{s}_{\mathrm{sel}}$ set; however in practice, a diversity of species is observed. We note that optimal values of n_{cyc} can be rapidly identified by tracking evolution of Shannon information (which increases as the histogram flattens) during the MC process and will depend on the target system and we calculate information via trapezoidal numerical integration of the normalized histograms (i.e. rather than by summation) to ensure calculated information does not significantly differ for calculations of the same distribution using different numbers of histogram bins. The discussion section will present studies on influence of the latter two user-specified options, n_{bins} and $\boldsymbol{h}$ definition, as they can significantly impact ALC efficiency.

F. Computational details

The initial training trajectory comprised full frames arising from short (< 10 ps) spin-restricted DFT-MD simulations of dissociative carbon monoxide at (9350, 2.56), (6500, 2.5), and (2400, 1.79) (K, g cm^{-3}). We note that previous studies have shown choice of spin restriction minimally impacts condensed-phase and isolated molecule forces for the present system[26]. Simulations at 9350 K were comprised of 64 atoms, while simulations at 6500 and 2400 K contained 128 each. We note that DFT-MD simulations at each temperature are

initialized as molecular CO, but decompose into various species as the simulation progresses. All DFT data was generated with VASP[31–34], where a planewave cutoff of 700 eV, Fermi smearing with width equivalent to the ion temperature, the Perdew-Burke-Ernzerhof generalized gradient approximation functional[35,36] (PBE), projector-augmented wave pseudopotentials[37,38] (PAW), and the DFT-D2 method for description of dispersion interactions[39]. DFT-MD simulations utilized a 0.5 fs time-step and a Nose-Hoover thermostat[40,41]. The initial training trajectory contained 80 evenly spaced full frames from the 2400 and 6500 K state points, and 160 from the 9350 state point, for a total of 320 initial frames. Simulations at higher temperature were included because they are likely to sample a broader region of configuration space as well as close contacts, which informs the repulsive portion of the potential and ultimately reduces the number of required iterative cycles. All other simulations (i.e. during active learning) are conducted at 2400 K and 1.79 g cm^{-3}, which we emphasize are the target thermodynamic conditions for benchmarking the actively-learned models.

Recall that an active learning cycle involves both a simple iterative and an active learning component (i.e the blue and orange portions of Fig. 2). During active learning cycles, ChIMES simulations contained 128 atoms and were run at 2400 K and 1.79 g cm^{-3} for 6 ps with our locally developed ChIMES-MD software using a global Hoover thermostat with periodic boundary conditions. A 0.1 fs timestep is used during the ALC process since early models generally learn from incomplete training repositories and tend to exhibit rapidly varying potential energy surfaces. As will be discussed in later sections, models generated in successive cycles yield far smoother potential energy surfaces, enabling use of a larger timestep during production simulations. The present 2400 K/1.79 g cm^{-3} DFT simulations yield significantly slower kinetics than is observed at the two higher T/p state points, which serves to exacerbate uncertainty in predicted chemistry arising from small system sizes and the effects of initial conditions. Because multiple independent DFT-MD simulations are too time consuming and block averaging is dubious for short (i.e. non-equilibrated) simulations, we instead estimate errors in examined properties by computing standard deviation across eight independent ChIMES-MD simulations.

In addition to cluster selection, the iterative component of any given ALC involves selection of a handful of full frames from ChIMES-MD for single-point DFT calculation and subsequent addition to the central training repository, for more complete exploration of training phase space. For this process, we select 20 frames evenly spaced over the duration

of the simulation, rather than using a special technique to identify optimal frames. We take this approach for two reasons: (i) to ensure the entire simulation progression is represented in the training repository, and (ii) because the AL portion of each ALC is performed on much more frequently sampled frames (i.e. 250), and thus should identify important events. In addition to the 20 frames discussed above, any frames for which pair distances $r_{\mathrm{c,in}}^{m_1(y)} < r_y < r_{\mathrm{c,in}}^{m_1(y)} + d_{\mathrm{p}}$ are sampled are added to the repository and considered during the ALC process to inform the fit in the typically undersampled region near the model's inner cutoffs.

II. ACTIVE LEARNING RESULTS AND DISCUSSION

In the following sections we present application of the AL approach discussed above to development of ChIMES models explicitly describing 2-, 3- and 4-body interactions for a C/O system under reactive conditions. As described in sections I B and I C, coupling between model and physicochemical complexity make this a particularly challenging problem and thus well suited for validating the present AL scheme. Three versions of our AL approach are described, which vary by how the histogram $\boldsymbol{h}$ used during MC selection is constructed. In the protocol described in section I E and Fig. 3, (i.e. where $\boldsymbol{h}$ is constructed considering only clusters in $\boldsymbol{s}$) there is no memory of clusters selected in previous active learning cycles (i.e. those in the central repository $\boldsymbol{s}_{\mathrm{cent}}$). In addition to this "no-memory" mode, we explore cases where there is "partial memory" (i.e. only some $\boldsymbol{s}_{\mathrm{cent}}$ clusters are remembered when constructing $\boldsymbol{h}$), and full memory (all $\boldsymbol{s}_{\mathrm{cent}}$ clusters are remembered when constructing $\boldsymbol{h}$).

A. No-Memory Active Learning

We begin with investigation of the most basic active learning approach. The first task is generation of the ALC-0 force field, trained to the 320 frames extracted from the 2400, 6500, and 9350 K DFT trajectory, each of which contain $3n_{\mathrm{a}}$ coordinates and forces, and a single frame energy. Following identification, all clusters are extracted from 250 evenly spaced frames spanning each DFT-MD trajectory (i.e. forming $\boldsymbol{s}$), using the clustering approach described in section I D. Energies for these clusters, $E_i^{\mathrm{ALC-0}}$ are then computed with the ALC-0 force field and a subset of these species, $\boldsymbol{s}_{\mathrm{sel}}$ is selected through the Monte Carlo

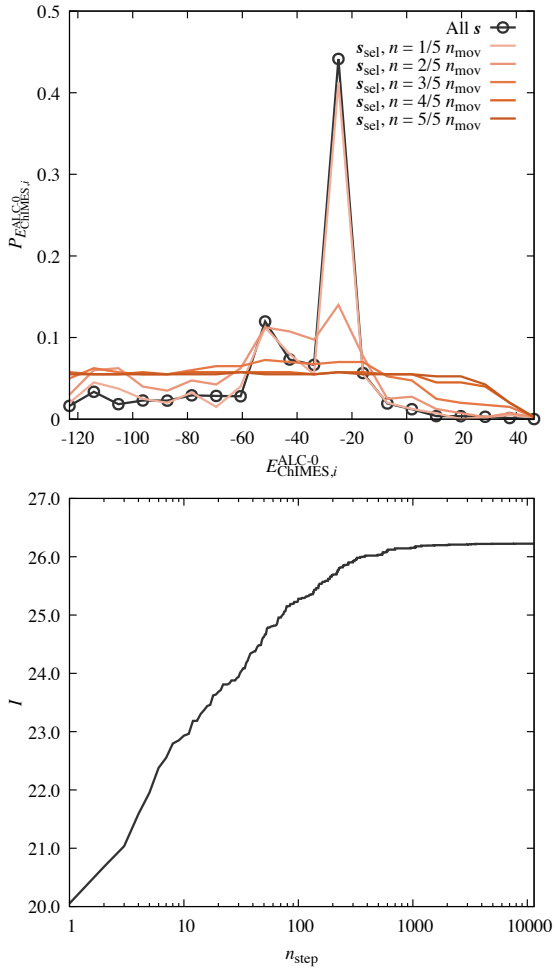

FIG. 4. Evolution of normalized $\boldsymbol{h}$ (i.e. probability) (top) and Shannon information (bottom) with Monte Carlo step for no-memory ALC-0.

(MC) approach presented in section I E (i.e. step a in Fig. 3). For the MC procedure (i.e. steps b and c in Fig. 3), $n_{\rm cyc}$ and $n_{\rm bins}$ are set to 2 and 20, respectively for identification of $n_{\rm sel} = 400$ species. We take $\boldsymbol{h} = \left[\min\left(\boldsymbol{s}\right), \max\left(\boldsymbol{s}\right)\right]$, which gives a histogram over values in the current $\boldsymbol{s}_{\rm sel}$) as the "no-memory" definition of $\boldsymbol{h}$. In this scheme, each active learning cycle selects clusters "from scratch," with no knowledge of clusters selected and added to the central training repository, in previous ALCs.

Focusing on the top plot of Fig. 4, the black data show the normalized histogram over all $E_i^{\rm ALC-0}$ values for clusters in $\boldsymbol{s}$ (i.e. all clusters extracted from the initial training trajectory). The histogram spans approximately -125 to 45 kcal mol^{-1} and rather than exhibiting a Boltzmann or normal distribution, is sharply peaked at -25 kcal mol^{-1} and has a tail

extending to positive $E_i^{\mathrm{ALC}-0}$ values. The remaining lines in the top plot of Fig. 4 show how the distribution of $E_i^{\mathrm{ALC}-0}$ in s_{sel} change as the MC selection process continues. Initially, these distributions are similar to that for s, but rapid flattening is observed with successive MC steps. We note that the sloping feature at positive $E_{\mathrm{ChIMES},i}^{\mathrm{ALC}-0}$ of the final distribution is due to an insufficient number of high energy configurations to yield $n_{\mathrm{sel}}/n_{\mathrm{bins}}$ configurations in each bin. The bottom plot of Fig. 4 provides evolution of Shannon Information, I, (i.e. a measure of histogram flatness) as the MC selection process progresses and also indicates rapid convergence during the selection process. Specifically, I values are converged within ≈ 5000 MC steps (i.e. ≈ 1 cycle) and result in 30% increase in information relative to the initial value. The final 400 selected species are then sent to DFT for single point force/energy calculation and added to the central training repository (i.e. step d in Fig. 3), completing ALC-0. Successive ALCs progress by generating new ChIMES models fit to the updated training repository, launching ChIMES-MD simulations with the resulting models, repeating the MC selection, and updating the central repository with both selected clusters and 20 evenly-space full frames from the ChIMES-MD simulation. These steps are repeated until target model accuracy is achieved.

Here, we consider eight active learning cycles (i.e. up to ALC-7). Each ALC requires approximately 6 hours of walltime on Intel Xeon E5-2695 hardware, where tasks are either serial on a single processor, or in parallel using 4 36-core nodes. Tasks used the following number of walltime hours: two for model generation (parallel), molecular dynamics (parallel), and post-processing (serial), one for cluster extraction (serial), $E_i^{\mathrm{ALC}-X}$ calculations (parallel), and $\boldsymbol{s}_{\mathrm{sel}}$ generation (serial), and 3 hours for single point DFT calculations (parallel), though we note that these timings do not necessarily represent an optimized process. Fig. 5 shows evolution of normalized $\boldsymbol{h}$ for ALCs 3 and 6, from which several notable features emerge. Distributions exhibit a similar overall shape to ALC-0, with a sharp peak (shifted to -30 kcal mol^{-1}) and a tail at positive $E_i^{\mathrm{ALC}-X}$. We find the ranges of observed $E_i^{\mathrm{ALC}-X}$ to be similar between the two ALCs, but substantially smaller than that observed in ALC-0; here, values are found to fall between ≈ -40 to -10 kcal mol^{-1}. We note that $\boldsymbol{s}$ for ALC-1 includes clusters extracted from DFT at 6500 and 9350 K in addition to 2400 K. This is in contrast to configurations extracted in subsequent ALCs that are harvested from ChIMES-MD simulations (i.e. which are *only* the target 2400 K state point); as a consequence, the range of $E_i^{\mathrm{ALC}-X}$ considered in ALC-1 is expected to be greater than in successive ALCs.

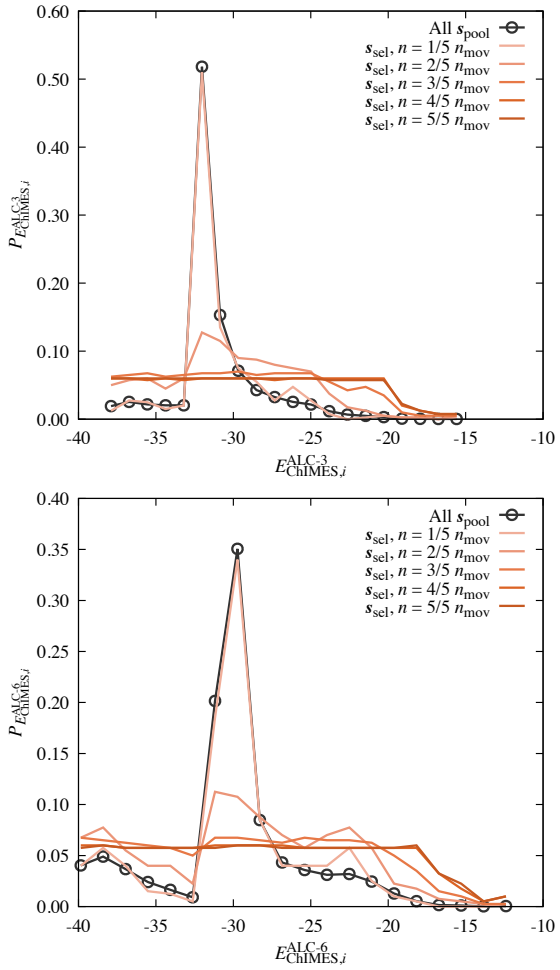

FIG. 5. Evolution of $\boldsymbol{h}$ with Monte Carlo step for no-memory ALC-3 (top) and ALC-6 (bottom).

Nevertheless, similarity in the range of $E_i^{\mathrm{ALC}-X}$ values between ALC-3 and ALC-6 speaks to convergence in ChIMES force fields generated at late ALC and resulting ChIMES-MD simulations.

For a more direct means of comparing performance of ChIMES models arising from successive ALCs, we consider speciation of small molecules CO, CO_2, C_2O_2, and C_3O_2, in terms of mole fractions (x_i), and corresponding lifetimes (t_i). For this analysis, a molecule was considered persistent if all constituent bonds remained within a specified cutoff distance, for a specified time. The distance criteria for each pair were set to the location of the first minimum in the radial pair distribution function (i.e. 1.9, 1.8, and 1.7 Å for CC, CO, and OO pairs, respectively), while the lifetime cutoff was set to 50 fs, allowing one-to-two bond vibrations. Overall, the results, given in Fig. 6 indicate a tendency for values

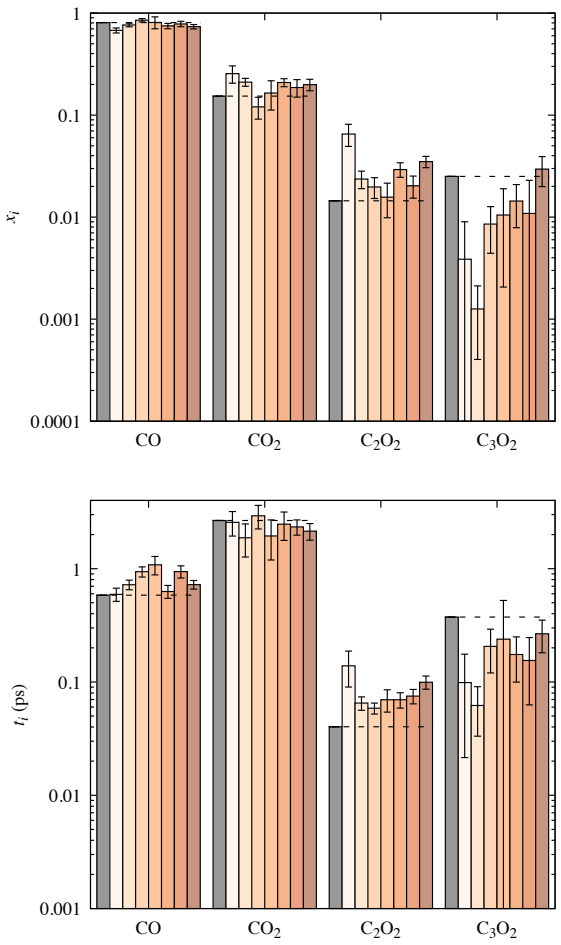

FIG. 6. Evolution of predicted mole fractions (x_i) and lifetimes (t_i) as a function of no-memory ALC, for carbon monoxide at 2400 K and 1.79 g/cm^3. DFT results are given in gray while ALCs 1–7 are given in successively darker orange bars, with standard deviations given as error bars. Horizontal dashed lines serve as a guide to the eye and give the DFT value for each species.

to converge to the DFT result at late ALC, with error bars generally decreasing for rarer species such as C_2O_2 and C_3O_2. We note that, at early in early ALCs, simulations may become unstable and terminate prematurely, contributing to large error bars in predicted values. In terms of species lifetime, significant fluctuations about the DFT value for CO_2 are found in conjunction with relatively large error bars, which could indicate a large variance in physically reasonable values.

Per-atom force and system energy recovery were also investigated for the 2400 K ChIMES-MD trajectory. As shown in Fig. 7, excellent force recovery is observed, suggesting the

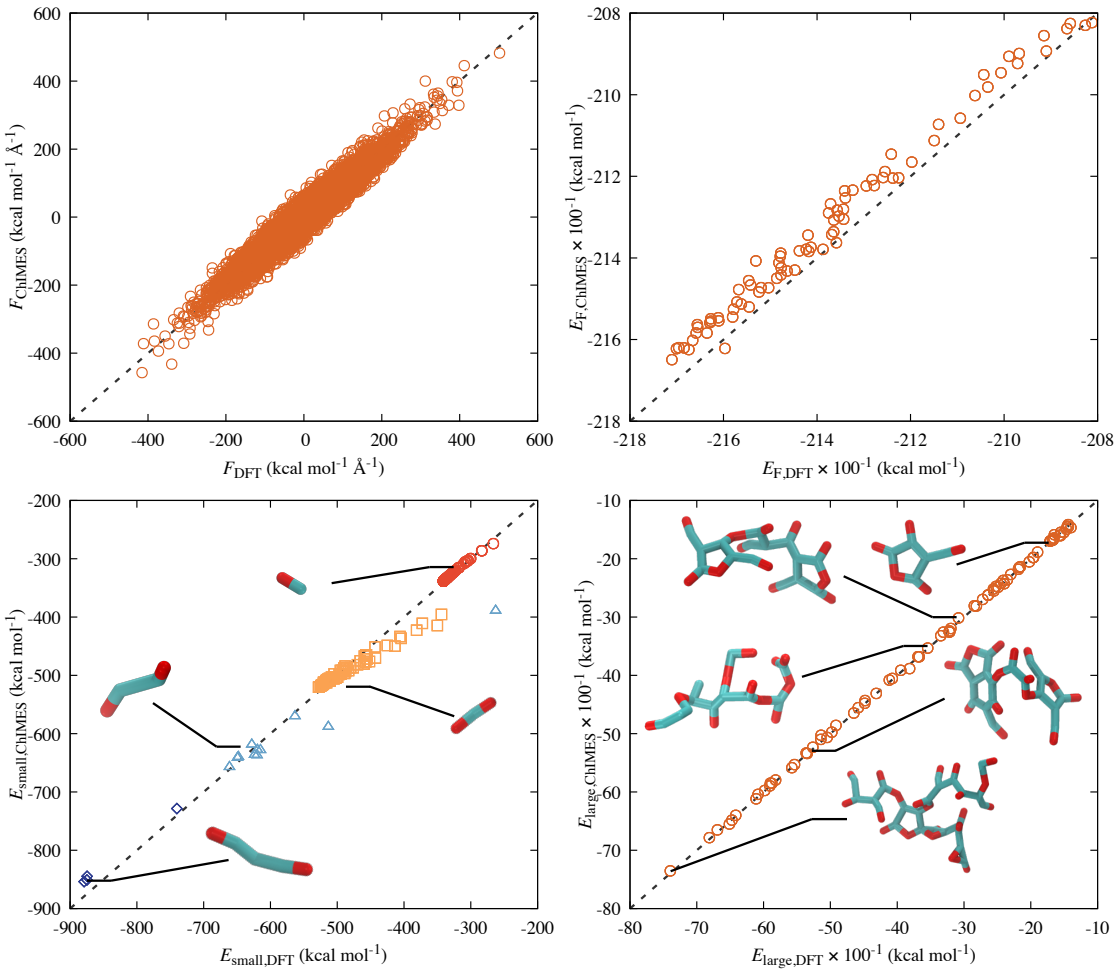

FIG. 7. Comparison of per-atom forces (top left), system energies (top right), small molecule energies (bottom left) and large molecule energies (bottom right) predicted by DFT and the ChIMES force field arising from no-memory ALC-7, for carbon monoxide at 2400 K and 1.79 g/cm^3. Note that energies for full frames and large species have been divided by 100 for clarity.

model should yield good reproduction of DFT structure. The largest deviations are found for large magnitude force, which generally corresponds to unusual and less favorable structures, which are useful for fitting purposes but generally not indicative of model performance under the target conditions. Overall, we find a reduced root-mean-squared error in the forces (i.e. $\mathrm{RMSE}_{\mathrm{r},F} = \mathrm{RMSE}_F/\langle|F|\rangle$) of 0.345, consistent with previous ChIMES models[22,25,26]. Fig. 7 also provides a comparison between overall frame energies predicted by DFT and ChIMES. The data are found to be highly linear with an R^2 and slope of 0.993 and 0.969, respectively, however, the data are slightly offset from the $x = y$ line, with an intercept

at -5.978 kcal mol^{-1} 100^{-1}. This behavior is attributed to inclusion of frame energies corresponding to simulations at 6500 and 9350, each of which used distinct electronic temperatures and thus strictly correspond to different electronic potential energy surfaces; in fact, when frame energies are plotted for all three state points, we find data for 6500 and 9350 K to be slightly below and above $x = y$, respectively. Nevertheless, we note that this intercept improves substantially upon that of the previous 2+3-body parameterization[26] (i.e. ≈ -57 kcal mol^{-1} 100^{-1}), by an order of magnitude. Rather than modifying the current study to account for this effect, we simply note that it can be mitigated by excluding highest temperature state points or re-computing forces and energies using a smaller thermal smearing parameter. Overall, these data exhibit a reduced RMSE of $\text{RMSE}_{\text{r},E_\text{T}} = 0.003$, i.e. an order of magnitude less than in previous work[26], suggesting the present model should still yield high accuracy condensed phase energetics.

In systems such as the present, which exhibit a diversity of chemical species with varied complexity, it is helpful to further decompose energetic contributions, thus the final two plots of Fig. 7 provide comparisons between DFT and ChIMES energetics for small species (i.e. CO, CO_2, C_2O_2, and C_2O_3), and species for which $10 \leq n_\text{C} + n_\text{O} \geq 50$; the former plot speaks to the predicted concentrations and lifetimes of species that seed formation of larger structures, and the latter to molecular conformations in complex species. We note that these species correspond to the final $\boldsymbol{s}_\text{sel}$ from ALC-0 and includes configurations from all three temperatures identified as nominal molecules *and* nominal intermediates (i.e. both high- and low-energy configurations). Focusing first on the small molecules, we find reasonable agreement, with a corresponding reduced RMSE of $\text{RMSE}_{\text{r},E_\text{sml}} = 0.036$. The largest discrepancies between DFT and ChIMES are found for higher-energy configurations involving CO_2, i.e. pseudo-intermediate state species which inform energetic maxima in reaction coordinate space, and thereby play an important role in predicting species lifetimes. Thus, it is somewhat unsurprising that the error bars for t_{CO_2} are among the largest observed for lifetimes given in Fig. 6. Possible explanations for this include insufficiently high 2- or 3-body order, or inconsistent DFT energetics for CO_2 arising from the three considered state points (i.e. due to use of different smearing parameters). We find excellent recovery of large species energetics, with $\text{RMSE}_{\text{r},E_\text{lrg}} = 0.009$, suggesting the present model provides a high-accuracy description of conformations in complex species. These RMSE_r values are also listed in table I along with pressures, diffusion coefficients, and minimized CO and CO_2

TABLE I. Pressures, diffusion coefficients, root-mean-squared errors in forces and energetics, and geometries in CO and CO_2 predicted by DFT and each examined AL approach.

	DFT	No-Memory	Full-Memory	Partial-Memory	
n_{bins}	–	20	20	20	40
P (GPa)	9	10.8_3	8.9_8	9.6_1	11.4_4
d_{s}^{O} (10^{-8} m^2/s)	1.8	1.7_2	1.4_2	1.5_2	1.4_4
d_{s}^{C} (10^{-8} m^2/s)	1.5	1.4_2	1.4_2	1.3_2	1.4_4
$\text{RMSE}_{\text{r},F}$	–	0.345	0.356	0.338	0.331
$\text{RMSE}_{\text{r},E_{\text{T}}}$	–	0.003	0.004	0.003	0.003
$\text{RMSE}_{\text{r},E_{\text{sml}}}$	–	0.036	0.047	0.033	0.035
$\text{RMSE}_{\text{r},E_{\text{lrg}}}$	–	0.009	0.010	0.009	0.008
$l_{\text{eq,CO}}^{\text{C-O}}$ (Å)	1.15	1.14	1.14	1.14	1.14
$l_{\text{eq,CO}_2}^{\text{C-O}}$ (Å)	1.18	1.18	1.17	1.17	1.17
$\theta_{\text{eq,CO}_2}^{\text{O-C-O}}$ (deg.)	180	180	180	180	180

structures predicted by DFT and the present ChIMES model. Noting that stress tensors were not included in the present fits, we find a 20% over-prediction in pressure, but otherwise, all other metrics are in excellent agreement with DFT.

B. Full-Memory Active Learning

Though the above "no-memory active" learning approach improved substantially upon early parameterization efforts for CO at 2400 K and 1.79 g cm^{-3} (e.g. by removing the propensity for the ChIMES-MD simulations to form hyper-coordinated structures), there is still room for refinement, most notably in recovery of pressure and small molecule energetics.

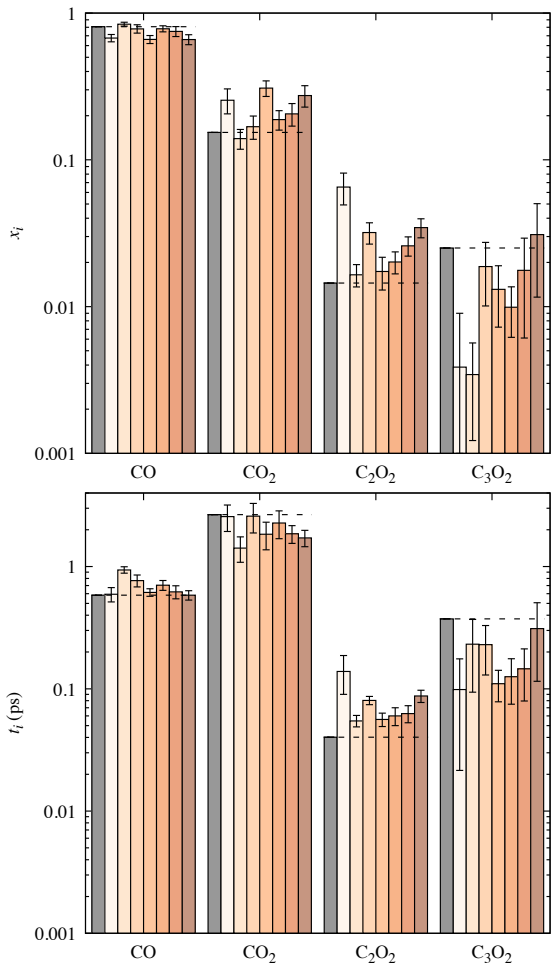

FIG. 8. Evolution of predicted mole fractions (x_i) and lifetimes (t_i) as a function of full-memory ALC, for carbon monoxide at 2400 K and 1.79 g/cm^3. DFT results are given in Gray while ALCs 1–7 are given in successively darker orange bars, with standard deviations given as error bars. Horizontal dashed lines serve as a guide to the eye and give the DFT value for each species.

In this section, we consider an active learning framework with "full-memory" of the central repository where we aim to flatten a histogram of *both* clusters selected in previous ALCs (i.e. $\boldsymbol{s}_{\text{cent}}$) and current selections from $\boldsymbol{s}$, $\boldsymbol{s}_{\text{sel}}$. This is achieved by constructing $\boldsymbol{h}$ from *both* the set of clusters in the central repository ($\boldsymbol{s}_{\text{cent}}$) and those in $\boldsymbol{s}_{\text{sel}}$; $\boldsymbol{h}$ is defined over the minimum and maximum values among both $\boldsymbol{s}_{\text{cent}}$ and $\boldsymbol{s}$. In principle, this approach should improve model performance by preventing redundancy among species selected in successive ALCs. As a practical point, we note that this approach requires re-calculation of $E_i^{\text{ALC}-X}$ for $\boldsymbol{s}_{\text{cent}}$ species each ALC, which for the present system, adds no more than 15 minutes

to the overall time required to complete an ALC. Fig. 8 provides the mole fractions and lifetimes predicted from full-memory ALCs 1–7. In general, average mole fractions, $\langle x_i \rangle$ predicted by the no-memory ALC-7 model are closer to DFT than those from full-memory ALC-7, however both methods produce values within error of one another. In contrast, ALC-7 full-memory average lifetimes, $\langle t_i \rangle$, are generally closer to DFT, with 3 out of 4 values within error between the two methods. In both methods, $\langle x_i \rangle$ and $\langle t_i \rangle$ values across ALCs 1–7 appear converged to the same extent. Moving to table I, we find pressure in better agreement with DFT while diffusion coefficients and predicted CO/CO_2 structures are close to the no-memory values.

It is not entirely surprising that full-memory active learning yields such marginal improvements over the no-memory model. In the former framework, early ALC ChIMES-MD simulations can give rise to unphysical structures to which DFT typically assigns high energy; moreover, ALC-1 includes clusters from 6500 and 9350 K, which can exhibit substantially higher energies than those sampled at 2400 K. As more ALCs are performed, resulting ChIMES models begin yielding more reasonable configurations and assigning energies more consistent with those arising from DFT. As a consequence, early ALCs set the upper bound histogram value causing late ALCs to cluster about relatively low histogram values and giving rise to a long tail at high histogram values. The practical implication of this is that only a fraction of the 20 available bins are allocated to histogram "active space" at late ALC.

C. Partial-Memory Active Learning

The pitfalls of full-memory active learning can readily be overcome setting an energetic cutoff for central repository configurations "remembered" during the sub-selection process, effectively imposing bounds on the possible range of $\boldsymbol{h}$. In the present work, this is achieved by constraining the domain of $\boldsymbol{h}$ to the minimum and maximum $E_i^{\mathrm{ALC}-X}$ values among *only* $\boldsymbol{s}$ and populating it only with $\boldsymbol{s}_{\mathrm{cent}}$ and $\boldsymbol{s}_{\mathrm{sel}}$ values that fall within this domain; any $\boldsymbol{s}_{\mathrm{cent}}$ values outside of that range are ignored during the MC selection process; henceforth we refer to this approach as "partial-memory" active learning. Note that this is in contrast to the full-memory approach, which constructed $\boldsymbol{h}$ from *all* clusters in $\boldsymbol{s}_{\mathrm{cent}}$ *and* those in $\boldsymbol{s}_{\mathrm{sel}}$. Mole fractions and corresponding lifetimes predicted by ChIMES force fields developed with partial-memory ALC are given in Fig. 9. Comparing with the no-memory model, we find

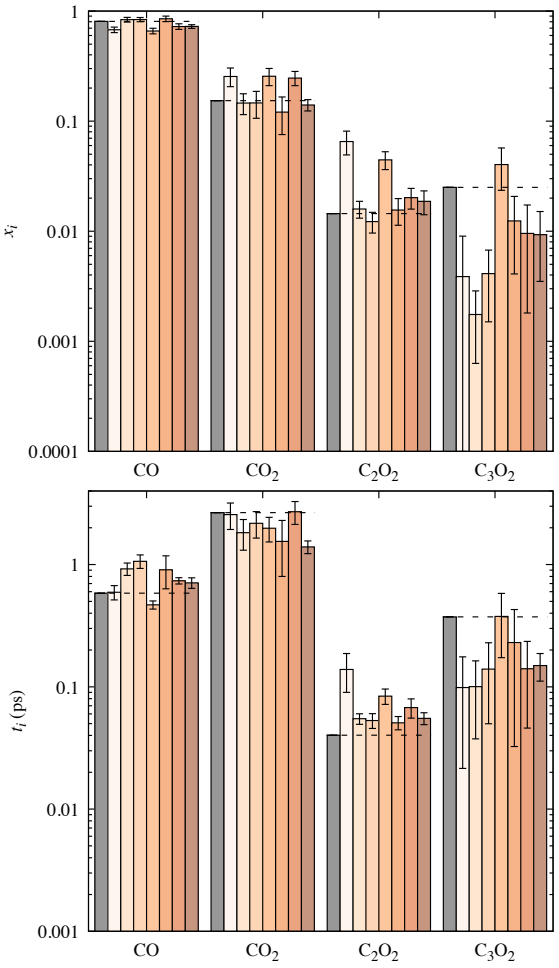

FIG. 9. Evolution of predicted mole fractions (x_i) and lifetimes (t_i) as a function of partial-memory ALC using 20 bins, for carbon monoxide at 2400 K and 1.79 g/cm^3. DFT results are given in Gray while ALCs 1–7 are given in successively darker orange bars, with standard deviations given as error bars. Horizontal dashed lines serve as a guide to the eye and give the DFT value for each species.

ALC-7 $\langle x_i \rangle$ for the present model are generally closer to DFT. Values for no-memory $\langle t_i \rangle$ are closer to those predicted by DFT for all but C_2O_2, however all partial-memory predictions for these species are within error of the full-memory results. Table I also shows that the both models predict pressure and diffusion coefficients within error of one another, and identical predictions for CO and CO_2 geometries. Additionally, table I indicates that, along with no-memory, the partial-memory model exhibits the lowest value of $\mathrm{RMSE}_{\mathrm{r},E_\mathrm{T}}$, and has the lowest value for all other RMSE_r considered. Ultimately, the partial-memory model is found

to provide the best overall performance, though differences between the various models are small.

It stands to reason that the present active learning framework should yield improved results when $\boldsymbol{s}_{\text{hist}}$ is constructed with finer resolution. Thus, we explore the effect of doubling n_{bins}, i.e. from 20 to 40. To ensure the MC selection process is converged, we also double n_{cyc}, i.e. from 2 to 4. Fig. 10 gives the mole fractions and corresponding lifetimes predicted from this 40-bin partial-memory active learning process. Most notably, the present active learning approach yields speciation which is more obviously converged than that of the previous methods, suggesting enhanced efficiency. Moreover, this indicates the likelihood of serendipitous predictions are decreased in the 40-bin case. Comparing to ALC-7 speciation from the 20-bin-partial-memory approach, 40-bin $\langle t_i \rangle$ are closer than DFT for all but C_2O_2; 40-bin $\langle t_i \rangle$ values are closer to DFT for only half the species, though they are within error of the 40-bin prediction for the remaining two. Table I shows the 40-bin-partial-memory model yields the worst pressure prediction relative to DFT, but diffusion coefficients and CO and CO_2 geometries are in agreement with the other three approaches within error. Furthermore, $\text{RMSE}_{\text{r},E_{\text{T}}}$ and $\text{RMSE}_{\text{r},F}$ are both the smallest of all considered models. Computed $\text{RMSE}_{\text{r},E_{\text{sml}}}$ and $\text{RMSE}_{\text{r},E_{\text{lrg}}}$ values for the 40-bin-partial memory model are not directly comparable with the other three models, since the final ALC-0 $\boldsymbol{s}_{\text{sel}}$ contains different species (i.e. because they were selected using different n_{bin} values). Nevertheless, we find $\text{RMSE}_{\text{r},E_{\text{sml}}}$ is the second smallest and $\text{RMSE}_{\text{r},E_{\text{lrg}}}$ the smallest of the four active learning methods considered. We note that additional simulations of the same size and length were run for the 40-bin-partial memory ALC-7 model using an increased time-step (i.e. from 0.1 to 0.5 fs), for which resulting predictions were consistent with the 0.1 fs results presented here. As a final comparison between each active learning approach, Fig. 11 provides the radial pair distribution functions (RDFs) predicted for the 2400 K system by the ALC-7 model from each approach. As with all other explored validation metrics, the RDFs are all similar to one another and DFT, though a slight over-structuring is observed in the second CC 20-bin-partial-memory peak ($r \approx 2.5$), and the first OO peak predicted by the 20-bin-partial-memory and the no-memory models.

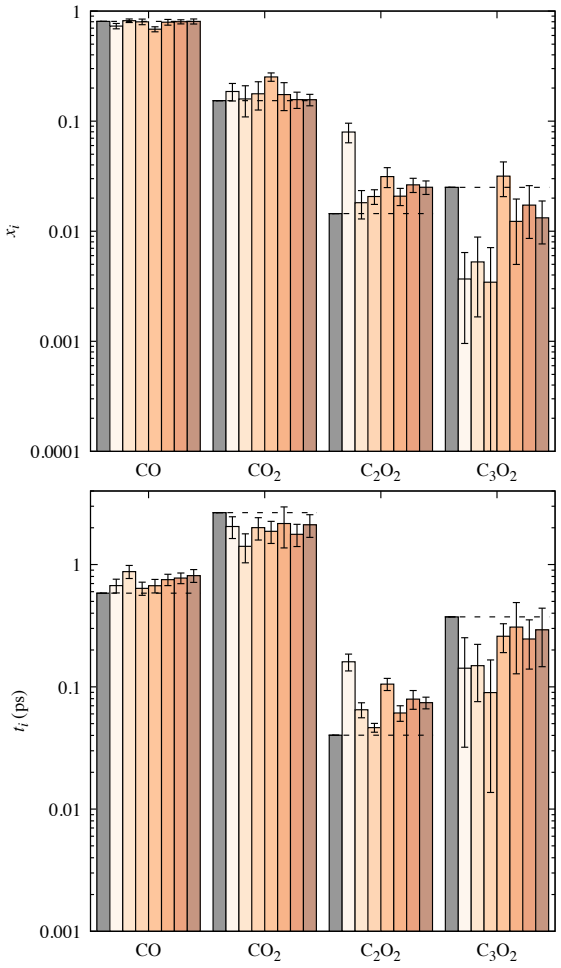

FIG. 10. Evolution of predicted mole fractions (x_i) and lifetimes (t_i) as a function of partial-memory ALC using 40 bins, for carbon monoxide at 2400 K and 1.79 g/cm^3. DFT results are given in Gray while ALCs 1–7 are given in successively darker orange bars, with standard deviations given as error bars. Horizontal dashed lines serve as a guide to the eye and give the DFT value for each species.

III. CONCLUSIONS

Active learning provides an automated and flexible means of achieving training repository completeness, a primary factor determining accuracy and robustness of high complexity machine-learned force fields. In this paper, we have demonstrated design of an AL approach for semi-automated development of high-fidelity reactive ChIMES models by sampling only from relevant configurations found in a condensed phase. Moreover, this approach is broadly

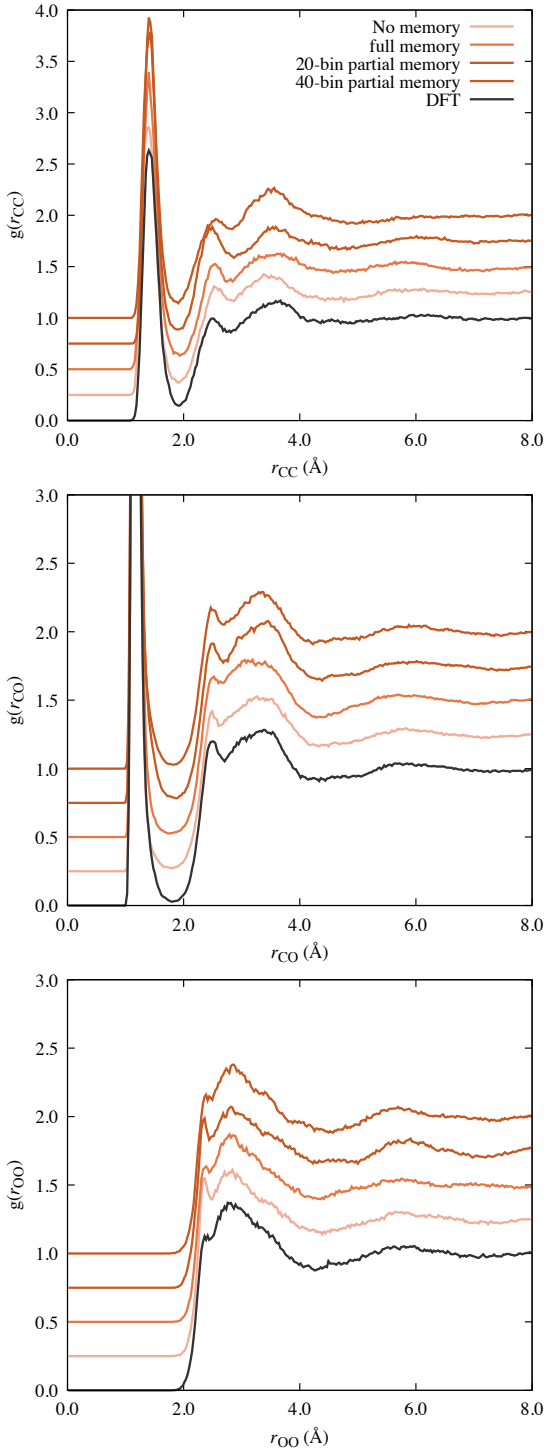

FIG. 11. Radial pair distribution functions for C/O at 2400 K and 1.79 g cm^{-3}. Curves have been vertically offset by increments of 0.25 for clarity.

applicable and well-suited for any parameterically linear high-complexity model. At a high

level, this approach involves identification, extraction and selection of potentially important species by way of clustering, energetics and a criterion inspired by Shannon information theory.. The result is an evolving central training repository that enables deconvolution of inter- and intra-molecular contributions to DFT forces and energies, allowing for an improved description of structure, dynamics, and speciation.

To demonstrate the suitability of this AL framework, we have applied it to development of a high-complexity (i.e approximately 4000 parameter) explicitly many-bodied machine learned force field for C/O systems under reactive conditions and shown resulting models exhibit excellent agreement with DFT. Model development through a partial memory active learning process with 40 bins was found to yield convergent behavior by 8 active learning cycles, and predicted structure, dynamics, and speciation in excellent agreement with DFT. We note that the present AL approach has an added benefit; instabilities are often encountered during initial ChIMES runs on large systems due to an enhanced probability of sampling new regions of phase space with increased system size. Using the present AL approach, these unstable configurations can be dealt with on a cluster-scale, rather than overall-system scale, drastically reducing the computational requirements for additional ALCs. This work represents a significant step forward in ML model development methodology by substantially enhancing automation and reproducibility. Furthermore, this approach is highly flexible; current efforts are focused on extending this fitting framework to enable transferability through parallel learning at multiple state points and further refining this process by addition of structural criteria during MC selection, as well as a configuration filter based on the expected model change method[20].

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The project 17-ERD-011 was funded by the Laboratory Directed Research and Development Program at LLNL with S.B. as principal investigator. LLNL-JRNL-812206.

The data that supports the findings of this study are available within the article.

REFERENCES

[1]C. M. Handley and P. L. Popelier, "Potential energy surfaces fitted by artificial neural networks," The Journal of Physical Chemistry A **114**, 3371–3383 (2010).

[2]J. Behler, "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations," Physical Chemistry Chemical Physics **13**, 17930–17955 (2011).

[3]J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," Angewandte Chemie International Edition **56**, 12828–12840 (2017).

[4]N. R. Greiner, D. Phillips, J. Johnson, and F. Volk, "Diamonds in detonation soot," Nature **333**, 440 (1988).

[5]V. N. Mochalin, O. Shenderova, D. Ho, and Y. Gogotsi, "The properties and applications of nanodiamonds," Nature nanotechnology **7**, 11 (2012).

[6]M. Bagge-Hansen, L. Lauderbach, R. Hodgin, S. Bastea, L. Fried, A. Jones, T. van Buuren, D. Hansen, J. Benterou, T. May, C aand Graber, B. J. Jensen, and T. M. Willey, "Measurement of carbon condensates using small-angle x-ray scattering during detonation of the high explosive hexanitrostilbene," Journal of Applied Physics **117**, 245902 (2015).

[7]M. J. Jordan, K. C. Thompson, and M. A. Collins, "The utility of higher order derivatives in constructing molecular potential energy surfaces by interpolation," The Journal of chemical physics **103**, 9669–9675 (1995).

[8]B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," International Reviews in Physical Chemistry **28**, 577–606 (2009).

[9]R. M. Balabin and E. I. Lomakina, "Support vector machine regression (ls-svm)–an alternative to artificial neural networks (anns) for the analysis of quantum chemistry data?" Physical Chemistry Chemical Physics **13**, 11710–11718 (2011).

[10]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Physical review letters **108**, 058301 (2012).

[11]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Physical review letters **104**, 136403 (2010).

[12]A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral

neighbor analysis method for automated generation of quantum-accurate interatomic potentials," Journal of Computational Physics **285**, 316 (2015).

[13]P. L. Popelier, "Qctff: on the construction of a novel protein force field," International Journal of Quantum Chemistry **115**, 1005–1011 (2015).

[14]L. Koziol, L. E. Fried, and N. Goldman, "Using force-matching to determine reactive force fields for bulk water under extreme thermodynamic conditions," J. Chem. Theory Comput. **13**, 135 (2017).

[15]S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, "Bayesian ensemble approach to error estimation of interatomic potentials," Physical review letters **93**, 165501 (2004).

[16]V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate ab initio molecular dynamics," International Journal of Quantum Chemistry **115**, 1074–1083 (2015).

[17]E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials," Computational Materials Science **140**, 171–180 (2017).

[18]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," The Journal of chemical physics **148**, 241733 (2018).

[19]L. Zhang, D.-Y. Lin, H. Wang, R. Car, and E. Weinan, "Active learning of uniformly accurate interatomic potentials for materials simulation," Physical Review Materials **3**, 023804 (2019).

[20]B. Settles, "Active learning literature survey," Tech. Rep. (University of Wisconsin-Madison Department of Computer Sciences, 2009).

[21]T. D. Loeffler, T. K. Patra, H. Chan, M. Cherukara, and S. K. Sankaranarayanan, "Active learning the potential energy landscape for water clusters from sparse training data," The Journal of Physical Chemistry C **124**, 4907–4916 (2020).

[22]R. K. Lindsey, L. E. Fried, and N. Goldman, "Chimes: A force matched potential with explicit three-body interactions for molten carbon," J. Chem. Theory Comput. **13**, 6222–6229 (2017).

[23]A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Modeling & Simulation **14**, 1153–1173 (2016).

[24]C. E. Shannon, "A mathematical theory of communication," Bell system technical journal **27**, 379–423 (1948).

[25]R. K. Lindsey, L. E. Fried, and N. Goldman, "Application of the chimes force field to

nonreactive molecular systems: Water at ambient conditions," J. Chem. Theory Comput. **15**, 436–447 (2019).

[26]R. Lindsey, N. Goldman, L. E. Fried, and S. Bastea, "Development of the chimes force field for reactive molecular systems: Carbon monoxide at extreme conditions," (2019).

[27]J. Tersoff, "Modeling solid-state chemistry: Interatomic potentials for multicomponent systems," Physical Review B **39**, 5566 (1989).

[28]R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Statist. Soc. B **58**, 267–288 (1996).

[29]B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Ann. Stat. **32**, 407–499 (2004).

[30]J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," J. Stat. Softw. **33**, 1 (2010).

[31]G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," Phys. Rev. B: Condens. Matter Mater. Phys. **47**, 558 (1993).

[32]G. Kresse and J. Hafner, "Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium," Phys. Rev. B: Condens. Matter Mater. Phys. **49**, 14251 (1994).

[33]G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15–50 (1996).

[34]G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," Phys. Rev. B: Condens. Matter Mater. Phys. **54**, 11169 (1996).

[35]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).

[36]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple [erratum to phys. rev. lett. 77, 3865 (1996)]," Phys. Rev. Lett. **78**, 1396–1396 (1997).

[37]P. E. Blöchl, "Projector augmented-wave method," Phys. Rev. B: Condens. Matter Mater. Phys. **50**, 17953 (1994).

[38]G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B: Condens. Matter Mater. Phys. **59**, 1758 (1999).

[39]S. Grimme, "Semiempirical gga-type density functional constructed with a long-range dis-

persion correction," J. Comput. Chem. **27**, 1787–1799 (2006).

[40]S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," J. Chem. Phys. **81**, 511–519 (1984).

[41]W. G. Hoover, "Canonical dynamics: equilibrium phase-space distributions," Phys. Rev. A: At., Mol., Opt. Phys. **31**, 1695 (1985).