Uncertainty Driven Active Learning of Coarse Grained Free Energy Models: Supplemental Information

Blake R. Duschatko*

John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138, USA

Jonathan Vandermause

John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138, USA and
Department of Physics, Harvard University, Cambridge, MA 02138, USA

Nicola Molinari and Boris Kozinsky†

John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138, USA and
Robert Bosch LLC, Watertown, MA 02472, USA

I. SUPPLEMENTARY METHODS

A. OPLS-AA Parameters

In order to generate force field parameters for both pentane and octane, the LigParGen [1, 2] server is used. Interactions between different species are defined using a geometric mixing rule for the parameters. The pentane system, consisting of 70 molecules, and the octane system consisting of 40 molecules, are placed around the edges of a box to begin. The fire minimization scheme implemented in LAMMPS [3] is first performed, with a time step of 0.1 fs, with an energy and force tolerance of 1e-7 and 1e-9, respectively. Random velocities are then drawn for all atoms for a Boltzmann velocity distribution with temperature 62.5K.

For the pentane liquid, using a timestep of 0.05 fs, 500,000 timesteps are run in the isothermal-isobaric ensemble from a temperature of 62.5K to 250K, and from a pressure of 0 to 1 atmosphere. The damping parameters for the thermostate and barostat are set to 100 and 1,000 times the timestep, respectively.

Another isothermal-isobaric ensemble segment is performed, this time with a timestep of 0.5 fs for 800,000 steps at constant temperature and pressure. Finally, 200,000 steps are made with a 1 fs timestep at a constant temperature of 250K. The resulting equilibrated structure is used as the starting point for all subsequent simulations, including constrained dynamics, initial on-the-fly frames, and the starting point for all mapped simulations.

For the octane liquid, the first isothermal-isobaric stage is performed with a 0.01 fs timestep for 1,000,000 steps, from a temperature of 25K to 300K, and from a pressure of 1 to 151 atmospheres. The damping parameters are the same as for pentane. In the second isothermal-isobaric stage, a 0.5 fs timestep is used over 800,000 steps, from a temperature 300K to 250K, and from 151 to 1 atmosphere of pressure. A third isothermal-isobaric stage is run with a 0.5 fs timestep for 200,000 steps at constant temperature and pressure. Finally, an isothermal stage with a timestep of 1 fs is run for 200,000 steps at constant temperature.

[*] bduschatko@g.harvard.edu

[†] bkoz@g.harvard.edu

B. OPLS-UA Parameters

To compare our coarse grained models against a common baseline for n-alkane systems, we run a simulation of the same pentane system with the parameter set of the OPLS United Atom approach [4, 5]. In particular, the potential energy takes the form

$$\begin{aligned} E(r^N) &= \sum_{\text{bonds}} K_b (r - r_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\ &+ \sum_{\text{dihedrals}} \sum_{k=1}^{4} \frac{V_k}{2} (1 + (-1)^{k+1} \cos(k\phi)) \\ &+ \sum_{i \neq j} 4\varepsilon \left(\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right) \end{aligned} \tag{1}$$

The groupings in the OPLS-UA formalism chosen here are electrically neutral and give no Coulomb contributions. The system is broken into two species, the CH_3 ground and the CH_2 group. The longest values reported in [5] are for the butane groups, and we apply these parameters (see Supplementary Tables I-IV) to the pentane system.

C. ASE On-the-Fly Parameters

The on-the-fly molecular dynamics loop is run with ASE [6]. The NPT ensemble is used for the Nose-Hoover thermostat implementation, but with the barostat turned off and kept at constant volume. A timestep of 1 fs is used, with a damping parameter for the thermostat equal to 100 times the timestep. On-the-fly trajectories are performed for 100,000 steps within ASE's molecular dynamics engine prior to being mapped and run in LAMMPS.

D. Constrained Dynamics Parameters

After reconstructing the all-atom representation for them coarse grained representation (as well as when collecting fixed training and test data for which no reconstruction is necessary), the free degrees of freedom are allowed to evolve at a constant 250K temperature. For the on-the-fly runs, we use a 0.5 fs timestep during constrained dynamics and find that averaging forces over 10,000 frames sampled 200 steps apart is sufficient for mean-force convergence (see Supplementary Figure 2). This trend can be verified both by looking at the standard error of the mean of a selection of force components, as well as the marginal log likelihood of an SGP trained on a limited amount of data as a function of mean-force sampling. Once sufficient sampling is reached, the model noise becomes dominant over the data noise, and further constrained dynamics will not benefit the models performance.

E. Mapped Sparse Gaussian Process Simulations

Once the trained SGP's are mapped, pentane and octane models are equilibrated in LAMMPS with the learned force field for 250,000 steps and a timestep of 1 fs at constant temperature. The damping parameter of the Nose-Hoover thremostat is set to 100 times the timestep. Production runs to acquire radial distribution functions are run for 400,000 steps, sampling frames every 400 steps.

We demonstrate also the efficiency that is gained over all-atom simulations using ML CG models. While we have used classical force fields in this work as our baseline all-atom reference, this need not be the case. Because classical force fields are generally quite fast, and not nearly as accurate as ML models, comparing the Gaussian process models to them would not be a fair comparison. Instead, we demonstrate the efficiency of the single and two species CG models with respect to an all-atom two species Gaussian process model of the same system. We show this for the pentane systems, but the octane models follow the same trend as their densities are highly similar. In Supplementary Figure 1, we can see clearly the substantial gain in computational speed with coarse graining. In particular, the single species models which have shown to be quite accurate are faster by a factor of 30-40.

F. Hyperparameter Selection

As in Ref. [7], we maximize the log marginal likelihood of the sparse Gaussian processes in order to choose optimal hyperparameters. We optimize the likelihood with respect to the kernel power, ξ, basis expansion parameters n and ℓ, as well as the environment cutoff radius for a pentane model (see Supplementary Figure 3). For consistency of comparison, we use the optimized parameters for the two species model for all single species models. The octane models use the same parameters. While the cutoff radius is optimized by the likelihood at 4.2 Å, we find better overall performance of each given model when using a cutoff of 4.5 Å.

II. SUPPLEMENTARY DISCUSSION

A. Structure of n-alkane Chains

In the main text we explore the behavior of an ensemble of SGP models in the presence of nearby local free energy minima. Supplementary Figure 4 motivates the study of pentane by examining the end to end chain distance distribution of increasingly long n-alkane chains.

B. Model Stability with Larger Timesteps

The benefits of coarse graining are two fold. First and foremost, the reduction of the number of degrees of freedom tracked throughout the simulation reduces the computational cost of the simulation on a per step basis. In addition, removing fast degrees of freedom allows for greater timesteps to be used to enhance sampling efficiency. Here we show the stability of our models with respect to larger integration timesteps. We perform the analysis for the pentane single-species liquid system, but the same conclusions hold for octane where the densities and interactions are highly similar.

All-atom simulations of hydrocarbons would typically use timesteps between 0.5-1 fs. The results presented in the main text for the coarse grained models are all obtained with a 1 fs integration step in order to ensure fair comparison. Supplementary Figure 5 shows that in fact, for timestep sizes between 2-5 fs, the sampling of the pentane models with CG models remains stable, and a larger integration step could be used if desired. On the other hand, we find that above 2 fs timesteps for the all atom model, simulations no longer remain stable. Further, this argument is strengthened by considering the energy drift in NVE simulations, shown in Supplementary Figure 6. Here, the 2 fs timestep all-atom simulation already begins to show appreciable energy drift, while the CG models do not.

C. Transferability with Non-active Learning

In the case of direct models (i.e. those trained on the system they are intended to be deployed on) for hydrocarbon liquids, the variety of sparse environments sampled in any given simulation frame reduces the need for active learning in general. With a sufficient number of training frames, a diverse set of sparse environments can efficiently be sampled.

However, in the context of building transferable models from existing data sets, it is essential that care is taken with regards to what new information is being added.

The goal is to minimize the overall amount of new data to be added. By matching the number of training frames and sparse environments that are added to the existing training sets via random sampling versus what was deemed optimal via active learning, there is relatively little new data added about the new system compared to direct models. It is therefore crucial that the data added be maximally informative to reduce unnecessary noise in the data set.

This is particularly evident in Supplementary Figures 7 and 8. Here, we show the intramolecular distributions for single and two species models averaged over 40 model instances. In the case of single species, a single additional training frame is added to reflect a similar amount of data added by the active learning scheme. For two species models, we instead add 15 frames of additional data. In all cases, 50 sparse environments are added per training frame.

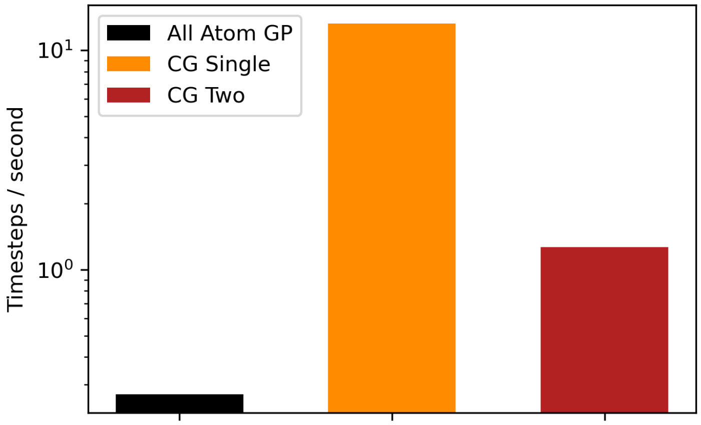
While the single-species ML CG model produces distributions that are overall quite good, we note the spurious bond length and bond angle peaks picked up by the non-active learning model. The models learned with the non-active approach are substantially worse in the case of two species models, where the higher sensitivity and expressivity of the descriptors can easily lead to model deterioration with insufficient training sets. Here, all distributions fail to even qualitatively reproduce the all-atom distributions, underlining a stark deficiency of random sampling for additional data compared to uncertainty-aware active learning.

D. Interaction Correlations in Octane

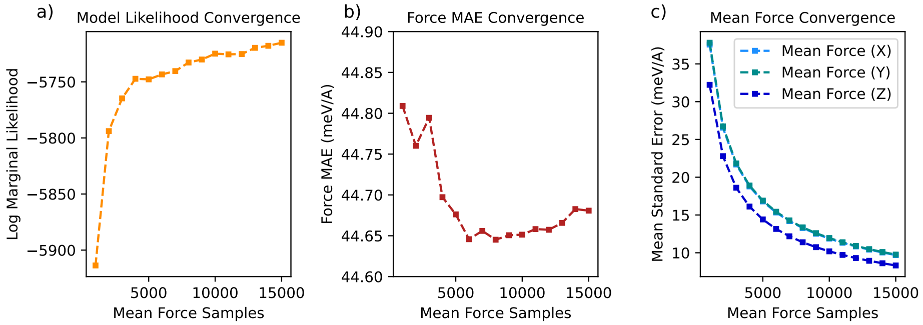
We further inspect the structural correlations of the octane system, characterized by the set of dihedral angle values along the backbone. Similarly to pentane, the bond length and angle distributions are unimodal (as seen in Supplementary Figures 7 and 8), and we do not consider them in the correlations. However, there are many possible unique sets of dihedral states. In particular, in Supplementary Figure 9 we show the relative sampling of all unique sets of five dihedral angles along the octane backbone, categorized by whether they are trans or gauche. Similar to the non-bonded interactions explored in Figure 5 of the main text as well as Supplementary Figures 7 and 8), there is a noticeable discrepancy of the unadapted two species model relative to the all-atom ground truth compared to the single species case. This discrepancy is alleviated upon adaptation of the two models. In particular, both single and two species models reproduce structural distribution functions with much higher fidelity when octane data is added to the training set. However, we note difficulties of both the direct and adapted models at accurately capturing the correlations of some structures.

Unlike the pentane system, the entire octane molecule does not lie within the cutoff radius of any given atomic environment. As such, it is more challenging for the octane model to identify and learn the interaction of dihedral pairs. Approximations of the PMF are known to, in some cases, fall short in this regard. For example, Rudzinski et. al. identify this issue with a similar system of hexane for a variety of CG mappings [8]. Addressing the ways in which the approximations made in CG models fail remains of great importance for future work. Such approximations include, but are not limited to, the choice of cutoff radius, basis set sizes, and body-order of interactions. Regardless, we emphasize that the transferability argument is not compromised by this complication, as it is clear in the data that the adapted models approach the performance of the direct model.

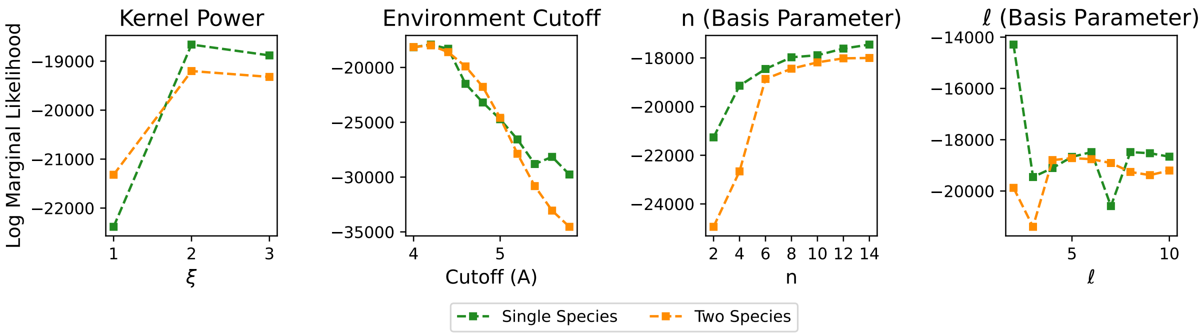
III. SUPPLEMENTARY FIGURES



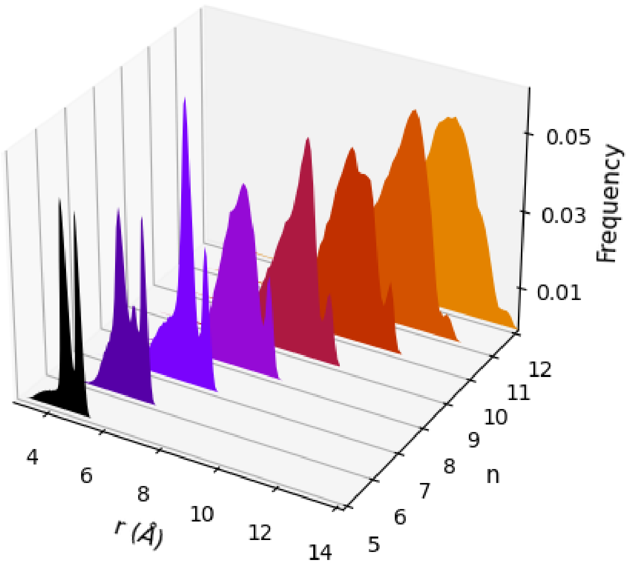
Supplementary Figure 1. The number of timesteps per second is shown for single species CG, two species CG, and two species all-atom models. These computations are all performed with a single CPU on an Intel Icelake node. A custom compilation of LAMMPS using the flare pair style was used to run mapped GP models.



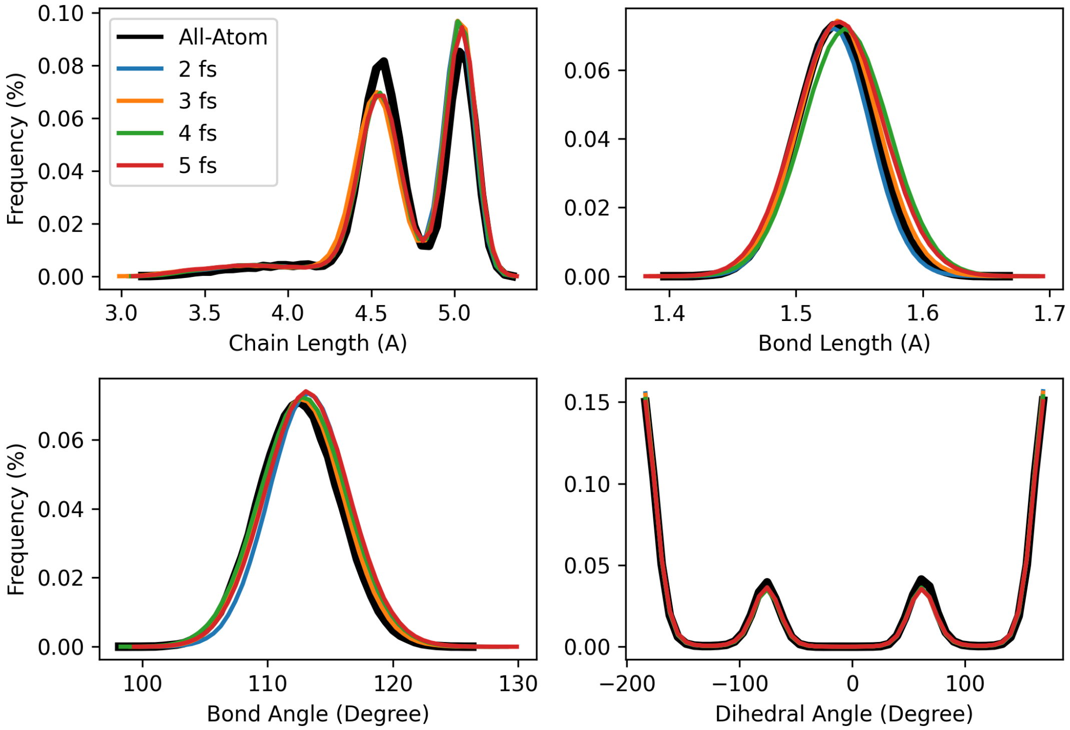
Supplementary Figure 2. The marginal likelihood (a), mean absolute force error on a test set (b) and the average standard error of the mean of PMF derivatives, i.e. forces, (c) are shown as a function of constrained dynamics sampling. The x-axis indicates the number of uncorrelated frames in which data was averaged over to provide force labels to the corresponding model.



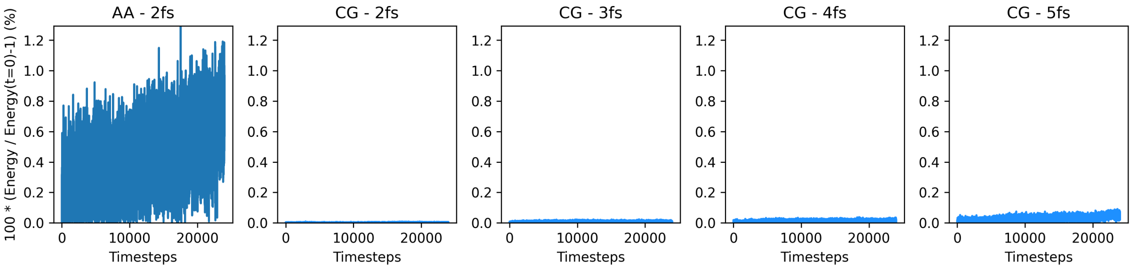
Supplementary Figure 3. The four hyperparameters for the sparse Gaussian processes are optimized by maximizing the log marginal likelihood. For each parameter sweep, the values are fixed at $\xi = 2$, $r_{\rm cut} = 4.5$ Å, $n = 5$ and $\ell = 12$ for those not being varied.



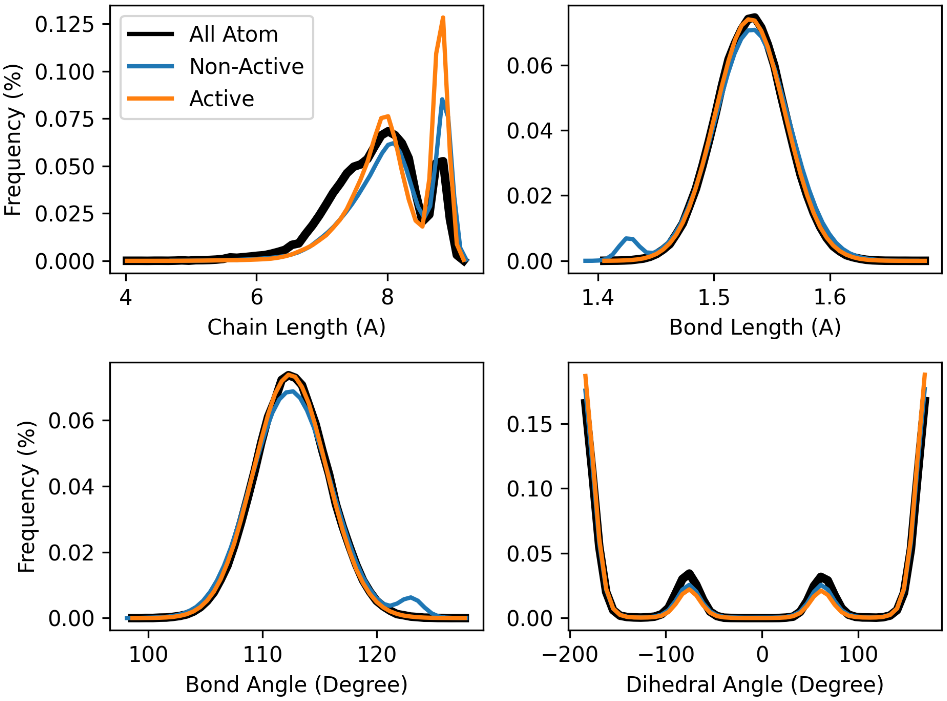
Supplementary Figure 4. The end-to-end chain distance distributions for n-alkanes from pentane to dodecane. At longer chain distances, the bimodal distribution begins to disappear. This is the result of many more dihedral energy minima along the chain.



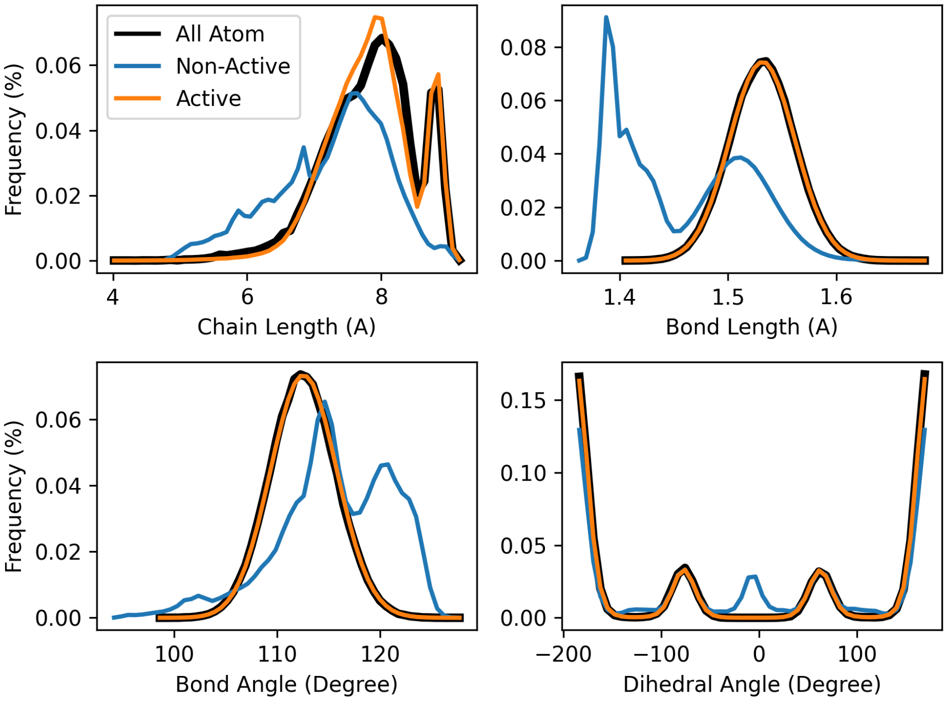
Supplementary Figure 5. Distributions are shown for single-species Gaussian process models of pentane using various time steps. All models are the same as those in the main text, with the mapped simulations using different integration steps.



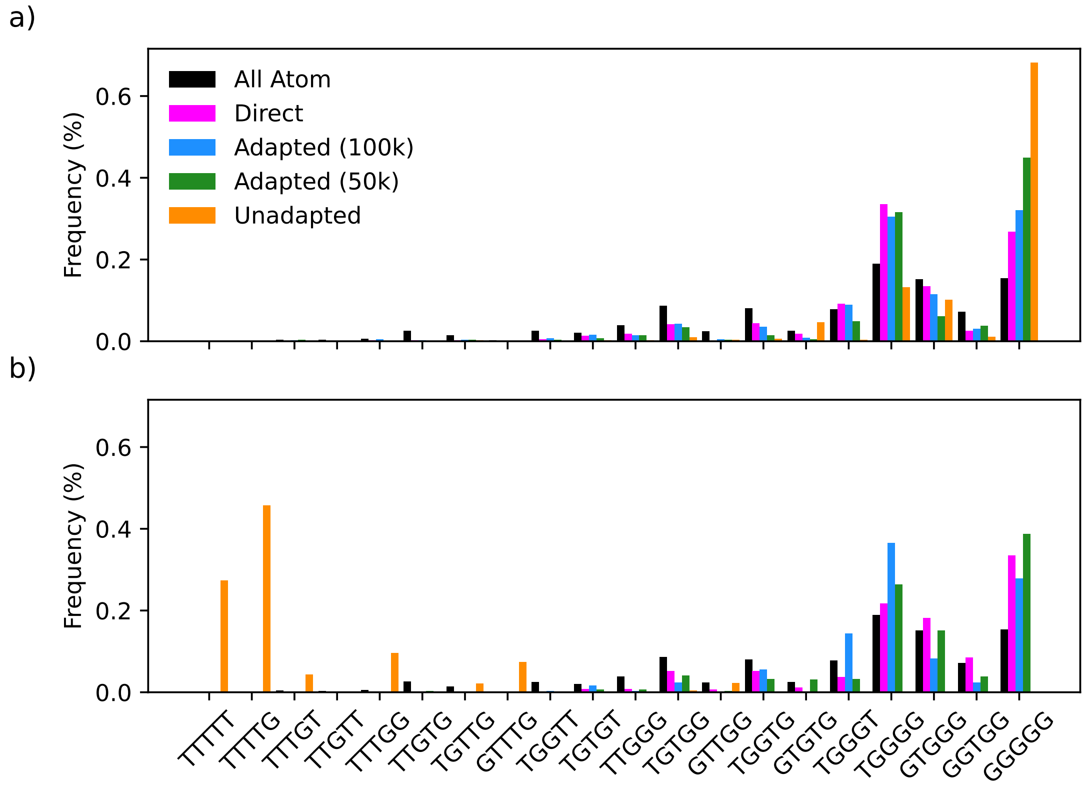
Supplementary Figure 6. Energy drift for simulations performed in the NVE ensemble. All models were run for 1.2 million time steps. The all-atom model used here is the OPLS model for efficiency, rather than the Gaussian process.



Supplementary Figure 7. The end-to-end chain length, bond length, bond angle, and dihedral angle distributions are shown for averages of 40 single species models with and without active learning, compared to the all-atom baseline, in the top left, top right, bottom left, and bottom right plots, respectively.



Supplementary Figure 8. The end-to-end chain length, bond length, bond angle, and dihedral angle distributions are shown for averages of 40 two species models with and without active learning, compared to the all-atom baseline, in the top left, top right, bottom left, and bottom right plots, respectively.



Supplementary Figure 9. Shown is the relative sampling of the unique sets of dihedral states for octane. Each molecules has five dihedrals, which we categorize as being either trans (T) or gauche (G). Due to the symmetry of the molecule, some states are redundant and not shown. a) the relative sampling for single species GP models with direct, adapted (100k), adapted (50k) and unadapted models shown. b) the relative sampling of dihedral states for the same model types but with a two-species representation.

IV. SUPPLEMENTARY TABLES

Atom Type	α (CH$_\alpha$)	σ (Å)	ε ($\frac{\text{kcal}}{\text{mol}}$)
1	3	3.905	0.175
2	2	3.905	0.118

Supplementary Table I. The OPLS United Atom vander-waals parameters used to simulate pentane.

Bond Type	Atom 1	Atom 2	K_b ($\frac{\text{kcal}}{\text{mol}}$)	r_0 (Å)
1	2	1	260.0	1.526
2	2	2	260.0	1.526

Supplementary Table II. The OPLS United Atom bond parameters used to simulate pentane.

Angle Type	Atom 1	Atom 2	Atom 3	k_θ ($\frac{\text{kcal}}{\text{mol}}$)	θ_0 (Degrees)
1	1	2	2	63.0	112.4
2	2	2	2	63.0	112.4

Supplementary Table III. The OPLS United Atom bond angle parameters used to simulate pentane.

Dihedral Type	Atom 1	Atom 2	Atom 3	Atom 4	V_1 ($\frac{\text{kcal}}{\text{mol}}$)	V_2 ($\frac{\text{kcal}}{\text{mol}}$)	V_3 ($\frac{\text{kcal}}{\text{mol}}$)	V_4 ($\frac{\text{kcal}}{\text{mol}}$)
1	1	2	2	2	0.0	0.0	2.0	0.0

Supplementary Table IV. The OPLS United Atom dihedral angle parameters used to simulate pentane.

V. SUPPLEMENTARY REFERENCES

[1] Dodda, L., Vilseck, J. Z., Rives-Tirado, J. & Jorgensen, W. L. 1.14*cm1a-lbcc: Localized bond-charge corrected cm1a charges for condensed-phase simulations. *J. Phys. Chem. B* **121**, 3864–3870 (2017).

[2] Dodda, L. S., Cabeza de Vaca, I., Rives-Tirado, J. & Jorgensen, W. L. Ligpargen web server: An automatic opls-aa parameter generator for organic ligands. *Nucleic Acids Research* **45**, W331–336 (2017).

[3] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).

[4] Beveridge, D. L. & Jorgensen, W. L. The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *Annu. Rev. Biophys. Bioeng.* **110**, 18 (1988).

[5] Racker, J. A. *et al.* Tinker 8: Software tools for molecular design. *J. Chem. Theory Comput.* **14**, 5273–5289 (2018).

[6] Larsen, A. H. *et al.* The atomic simulation environment - a python library for working with atoms. *J. Condens. Matter Phys.* **29** (2017).

[7] Vandermause, J., Xie, Y., Lim, J. S., Owen, C. J. & Kozinsky, B. Active learning of reactive bayesian force fields applied to heterogeneous catalysis dynamics of h/pt. *Nat. Comm.* **13** (2022).

[8] Rudzinski, J. F. & Noid, W. G. Investigation of coarse-grained mappings via an iterative generalized yvon-born-green method. *J. Phys. Chem. B* **118**, 8295–8312 (2014).