

Chapter 4

Force Matching Approaches to Extend Density Functional Theory to Large Time and Length Scales



Rebecca K. Lindsey, Matthew P. Kroonblawd, Laurence E. Fried and Nir Goldman

Abstract We present methods for the creation of semi-empirical quantum approaches and reactive force fields through force matching to quantum simulation data for materials under reactive conditions. Our methodologies overcome the extreme computational cost of standard Kohn–Sham Density Functional Theory (DFT) by mapping DFT computed simulation data onto functional forms with linear dependence on their parameters. This allows for quick parameterization of our models by avoiding the nonlinear fitting bottlenecks associated with most molecular dynamics model development. We illustrate our approach with two different systems: (i) determination of density functional tight binding models for aqueous glycine dimerization, and (ii) determination of the Chebyshev Interactional Model for Efficient Simulation (ChIMES) reactive force field for metallic liquid carbon. In each case, we observe that our approach is easy to parametrize and yields a model that is orders of magnitude faster than DFT while largely retaining its accuracy. Overall, our methods have potential use for studying complex long time and length scale chemical reactivity at extreme conditions, where there is a significant need for computationally efficient atomistic simulations methods to aid in the interpretation and design of experiments.

R. K. Lindsey · M. P. Kroonblawd · L. E. Fried · N. Goldman (✉)
Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory,
Livermore, CA 94550, USA
e-mail: ngoldman@llnl.gov

R. K. Lindsey
e-mail: lindsey11@llnl.gov

M. P. Kroonblawd
e-mail: kroonblawd1@llnl.gov

L. E. Fried
e-mail: lfried@llnl.gov

N. Goldman
Department of Chemical Engineering, University of California, Davis, CA 95616, USA

© Springer Nature Switzerland AG 2019
N. Goldman (ed.), *Computational Approaches for Chemistry Under Extreme Conditions*, Challenges and Advances in Computational Chemistry and Physics 28,
https://doi.org/10.1007/978-3-030-05600-1_4

4.1 Introduction

Elucidating the chemistry of reactive materials under extreme conditions in the laboratory can require the investigation of a great number of permutations of different starting materials, thermodynamic conditions, and catalysts. The number of possible combinations can frequently be too numerous and costly to address with experimental trial and error alone. Experiments frequently rely on theoretical studies to elucidate measured data, including kinetic and spectroscopic properties. Nonetheless, in many cases, insufficient data exists for the equation of state and chemical reactivity of these materials under the extreme pressures attained during experiments [1]. For example, studies on carbon-rich materials under pressures up to 30 GPa (1 GPa = 10 kbar) suggest that slow chemical kinetics can extend beyond the timescales of nanosecond laser-driven compression experiments, even at temperatures of thousands of Kelvin [2]. Reported experimental temperatures can contain large uncertainties, making it difficult to adequately constrain the equation of state on the basis of experiment alone [3, 4]. Furthermore, low-resolution mesoscale simulations frequently require chemical kinetic input parameters that are challenging to obtain experimentally (e.g., due to difficulties obtaining atomic-scale resolution in experiments on bulk material synthesis) but can be readily computed via higher-resolution particle-based simulations [5–7].

Computer simulations such as molecular dynamics (MD) hold promise as an independent route to determining the equation of state and chemical reactivity during materials synthesis. Such studies can provide simple chemical pictures of ionized intermediates and reaction mechanisms, and can help identify atomic-scale properties that determine observed macroscopic kinetics (e.g., descriptors). These types of results can make experiments more tractable by aiding in their interpretation, and helping to narrow the number of different materials and reactive conditions to investigate.

Accurate modeling of the breaking and forming of bonds in condensed phases frequently requires the determination of many nuclear-body effects [8, 9], which are included in quantum theories such as Density Functional Theory (DFT). DFT remains one of the most popular and widely used modeling methods in condensed matter physics, computational chemistry, and materials science for predicting material properties and chemical reactivity. It has been shown to accurately reproduce the phase boundaries and thermal decomposition of many materials [10–13], particularly at extreme thermodynamic conditions such as planetary interiors [10, 11, 13], where long-range effects such as dispersion are less important. DFT-MD simulations, though, require immense computational effort per simulation time step and consequently are usually limited to picosecond time scales and nanometer system sizes. In contrast, many chemical events can occur over nanosecond timescales or longer [2, 11], and experiments can probe micron length scales or beyond [14–16].

Difficulty thus generally arises in determining models for chemical bonding that are both accurate and computationally efficient. Empirical models (e.g., [17]), where atomic forces are computed from parameterized potential energy surfaces, are

generally fit to properties of specific chemical reactants over a narrow range of thermodynamic conditions. These approaches are usually highly computationally efficient, but tend to contain numerous nonlinear parameters that can be time-consuming to optimize, and can also suffer from poor transferability. In contrast, semi-empirical quantum theories such as density functional tight binding with self-consistent charges (DFTB) retain some level of computational efficiency and tend to have a higher degree of transferability to different starting materials and reaction conditions. However, standard forms of these models are parameterized to gas phase quantum chemical calculations, and consequently can result in overbinding of reactants, e.g., reaction enthalpies that are up to 20 kcal/mol too high for simple hydrogenation reactions [18] and somewhat poor representation of condensed phase reactivity [19]. Thus, there exists a great need to explore novel methods for atomistic simulations of reactive materials that are computationally efficient while retaining the accuracy of higher order quantum methods.

In this chapter, we demonstrate how force matching can be used to determine both classical reactive MD and DFTB models through use of functional forms that are strictly linear in their parameters. Force matching maximizes the data set that can be obtained from DFT by fitting parameters of a potential energy function to each individual atomic force in an MD trajectory [20, 21], yielding a large quantity of data points for determining atomistic models (i.e., $3N$ data points per configuration, where N is the number of atoms in the system). It thus has potential to systematically determine MD models that yield highly accurate predictions of chemical reactivity for a given material or mixture and set of thermodynamic conditions. The use of linearly parametrized models allows for use of linear least-squares fitting, whereby optimum parameter values are rapidly determined in a single step. This removes the need for direct gradient minimization or iterative techniques (e.g., Levenberg–Marquardt) which can become trapped in a local minimum, or computationally intensive global energy minimum searches (e.g., simulated annealing). In this work, we discuss two application areas for this method: parameterization of a semi-empirical quantum model for glycine dimerization under ambient aqueous conditions, and generation of a reactive force field for molten carbon. Our intent is to provide a general overview of how these classes of MD models can be rapidly determined for reactive materials under a broad range of conditions.

4.2 Force Matching Overview

The force matching method (FM) was first developed by [20] for generation of an aluminum force field based on a repository of atomic configurations and corresponding *ab initio* forces [20]. This framework usually requires the generation of an MD training data set for given material at a specific set of state points, often through the use of DFT simulations. A force field functional form is then selected, e.g., $\phi(x; \{p\})$, which depends on interatomic distances x , and is subject to a set of parameters $\{p\}$.

Model parameters are then obtained by minimizing the following objective function:

$$\text{RMSE} = \sqrt{\frac{1}{3MN} \sum_{m=1}^M \sum_{i=1}^N \sum_{\alpha=1}^3 (F_{i_{m\alpha}, \text{DFT}} - F_{i_{m\alpha}, \phi\{p\}})^2}, \quad (4.1)$$

where RMSE is the root-mean-squared error, the subscript $i_{m\alpha}$ indicates the α th Cartesian component of the force on atom i in MD configuration m , and F_{DFT} are the forces to which model $\phi(x; \{p\})$ is to be fit.

As mentioned previously, generation of fitted parameters can be time consuming for complex target models due to the need for slow to converge nonlinear approaches such as Levenberg–Marquardt [22, 23] or global search methods such as simulated annealing [24]. If the target model is *linear* in fitted parameters, however, the FM optimization problem reduces to the following matrix equation: $\mathbf{M}\mathbf{p} = \mathbf{F}_{\text{train}}$. The elements of \mathbf{M} are given as $M_{rc} = \partial F_{r, \phi\{p\}} / \partial p_c$, where r represents the combined index over Cartesian components of the force on each atom in the training trajectory (i.e., matrix rows), and c the combined index over coefficients in the model (i.e., matrix columns), while p and F_{train} are the vector of target model parameters and vector of training forces, respectively. Standard methods such as Singular Value Decomposition [25], ridge [26, 27], or LASSO regression [28] can then be leveraged to solve for a unique set of model parameters in a matter of minutes. Well-developed procedures exist for determining the robustness of regression problems, such as cross-validation [29].

Though force matching was originally developed for generation of atomic potentials based on DFT data, we note that the approach is flexible and can be expanded to include other features including the system stress tensor (to tune model pressure) and energy. Furthermore, the force matching approach can be used to generate coarse-grained models from all-atom force fields (e.g., [30]) by mapping forces arising from several atoms to specific coarse grain interaction sites.

4.3 DFTB: Rapidly Tunable Models for High Throughput Quantum Molecular Dynamics

The density functional tight binding method (DFTB) [18, 31, 32] is a semi-empirical quantum simulation approach that yields a high degree of computational efficiency while potentially retaining the accuracy of the computationally intensive Kohn–Sham density functional theory (DFT). The formalism for DFTB with self-consistent charges (SCC) has been discussed in detail elsewhere [32–37, 40]. Briefly, the method assumes neutral, spherically symmetric charge densities on the atoms and expands the DFT Hamiltonian to second-order in charge fluctuations. The DFTB total energy is expressed as

$$E_{\text{DFTB}} = E_{\text{BS}} + E_{\text{Coul}} + E_{\text{Rep}}. \quad (4.2)$$

The first term, E_{BS} , corresponds to the band structure energy computed via sum over occupied electronic states from the approximate DFTB Hamiltonian. E_{BS} is usually computed from pre-tabulated Slater–Koster tables derived from DFT calculations with a minimal basis set, where both the electronic wave functions and electron density are subjected to separate confining or compression potentials. The compression potentials force the wavefunction/electron density to zero at relatively large distances from the nuclei, which has been shown to improve transferability of the Slater–Koster tabulations [31]. The precalculation of matrix elements in DFTB yields several orders of magnitude increase in computational efficiency over most DFT implementations, which project the electronic density onto a fine grid, and then use Fourier transforms to compute the Hartree potential and other terms. However, standard forms of DFTB still require calculation of eigenstates, which scales as $\mathcal{O}(N^3)$, where N is equal to the number of orbitals in the system. This generally remains the bottleneck of DFTB calculations, limiting the approach to system sizes of several hundred to thousands, e.g., far below the capabilities of classical MD codes.

The second term in (4.2), E_{Coul} , corresponds to a charge fluctuation term which is computed self-consistently. The third term, E_{Rep} (the repulsive energy), corresponds to ion–ion repulsions, as well as Hartree and exchange–correlation double counting terms. In practice, E_{Rep} is expressed as a short-ranged empirical function whose parameters are fit to reproduce DFT or experimental data, and can be either pairwise [13, 40] or contain multicenter interactions [37, 38]. This balance in DFTB of approximate quantum mechanics and short-ranged empiricism allows for the simultaneous interrogation of electronic states and ion dynamics over temporal scales that can approach those of many high temperature–pressure experiments (e.g., [39]). We note that the short-ranged nature of E_{Rep} generally allows for strong transferability of DFTB models (applicability to different materials and thermodynamic conditions), since the bulk of the interaction energy is computed through the quantum-mechanical E_{BS} and E_{Coul} terms.

Determination of an optimum E_{Rep} requires training data generation, which is a relatively straightforward endeavor for systems exhibiting rapid chemistry. Extracting uniformly spaced frames from an unbiased DFT-MD trajectory of a highly reactive system is likely to yield a training set with configurations corresponding to reactants, products, and various intermediates and transition states. In fact, this approach was shown to improve DFTB predictions for small species production during the rapid combustion of phenolic polymer resin [40]. However, in cases where chemistry is slow to either initiate or equilibrate, short unbiased DFT-MD simulations might not sample relevant intermediate or product configurations. Accelerated MD techniques such as umbrella sampling [41] or metadynamics [42] provide alternative approaches to training trajectory generation. In each of these methods, a biasing potential is applied to help escape local free energy minima and thus enhance sampling of configurational space. The biasing potential is typically applied to one or more “collective variables” that mathematically describe characteristic changes in system structure as the target process progresses. Optimal choice of collective variables is highly dependent on the situation and numerous kinds have been used

including bond separation vectors and/or coordination numbers [43], crystalline order parameters [44], and path variables [45] to name only a few.

Given these issues, force matching provides a systematic way to generate DFTB E_{Rep} parameters that can recover DFT-level accuracy for systems and states not considered in standard DFTB parameterizations. For this approach, the objective function is given as follows:

$$\text{RMSE} = \sqrt{\frac{1}{3MN} \sum_{m=1}^N \sum_{i=1}^M \sum_{\alpha=1}^3 [F_{i_{m\alpha}, \text{DFT}} - (F_{i_{m\alpha}, E_{\text{Rep}}\{p\}} + F_{i_{m\alpha}, \text{BS}} + F_{i_{m\alpha}, \text{Coul}})]^2}. \quad (4.3)$$

In this case, $F_{i_{m\alpha}, E_{\text{Rep}}\{p\}}$ (the forces arising from the repulsive energy term E_{Rep} subject to a set of model parameters $\{p\}$) are tuned to the values from the DFT training set with the DFTB forces from E_{BS} and E_{Coul} precomputed and subtracted out. We model the repulsive potential as a pairwise interaction between atom i of element e_i and atom j of element e_j that takes the form of the power series:

$$E_{\text{Rep}}^{e_i e_j}(r_{ij}) = \begin{cases} \sum_{n=2}^9 c_n^{e_i e_j} (r_c^{e_i e_j} - r_{ij})^n & r_{ij} \leq r_c^{e_i e_j}, \\ 0 & \text{otherwise.} \end{cases} \quad (4.4)$$

Here, r_{ij} is the interatomic separation distance, $r_c^{e_i e_j}$ is a radial cutoff distance, and $c_n^{e_i e_j}$ are parameters to be determined. Summing over all-atom pairs in the system gives the total repulsive energy E_{Rep} . Because (4.4) is linear in $c_n^{e_i e_j}$, the parameters can be determined from a simultaneous least-squares fit over all $e_i e_j$ pair types. Furthermore, selected $e_i e_j$ pair types can be excluded from the fit and substituted with some other interaction potential if desired, provided that corresponding contributions to the atomic forces are subtracted from the F_{DFT} values in (4.3).

4.3.1 Application to Glycine Dimerization: Computational Details

As an example of creating a DFTB for model of long timescale chemistry, we investigate the condensation reaction $2\text{Gly} \rightarrow \text{Gly-Gly} + \text{H}_2\text{O}$ in which two glycine molecules react to form diglycine and water. We begin by generating DFT training data via umbrella sampling along path collective variables, which are well-suited because the reactant and product species are known. Two path collective variables,

$$s(t) = \frac{\sum_k k \exp(-\lambda D[\mathbf{R}(t), \mathbf{R}_k])}{\sum_k \exp(-\lambda D[\mathbf{R}(t), \mathbf{R}_k])}, \quad (4.5)$$

and

$$z(t) = -\frac{1}{\lambda} \ln \left\{ \sum_k \exp(-\lambda D[\mathbf{R}(t), \mathbf{R}_k]) \right\}, \quad (4.6)$$

were used that respectively measure the progress along and distance from the path defined by reference configurations indexed by k at time t [45]. The variable $D[\mathbf{R}(t), \mathbf{R}_k]$ is the “distance” from the k th reference configuration and λ is a constant to be determined later. The path was defined with respect to two reference configurations, namely $k = 1$ for the two glycine reactant molecules and $k = 2$ for the single diglycine product molecule. We used the “distance” metric developed by Pietrucci and Saitta [46]:

$$D[\mathbf{R}(t), \mathbf{R}_k] = \sum_{i, e_i} [C_{i, e_i}(t) - C_{i, e_i}^k]^2, \quad (4.7)$$

where $C_{i, e_i}(t)$ is the number of atoms of type e_i coordinated to atom i at time t and C_{i, e_i}^k is the corresponding value in reference frame k . We tracked the coordination of the four C and two N atoms that form the diglycine backbone to all C, N, O, and H atoms in the system. The specific equations and parameters for computing $C_{i, e_i}(t)$ were taken from [46] and the reference C_{i, e_i}^k were obtained from short DFTB-MD simulations of neutral glycine and diglycine in explicit H₂O solvent. Following [46], we chose λ so that $\lambda D[\mathbf{R}_k, \mathbf{R}_{k+1}] \approx 2.3$, which for the present case was $\lambda = 0.70$. One particular advantage of using path collective variables to generate a DFT training set is that the resolution for sampling the reaction path can be controlled, which facilitates extracting important intermediate configurations and forces. The free energy surface $F(s, z)$ was computed as a function of s and z using umbrella sampling [41] and the weighted histogram analysis method (WHAM) [47, 48].

Simulations of glycine condensation were performed at $T = 300$ K for two glycine molecules solvated by 55 H₂O molecules, yielding a total of 185 atoms and a density of $\rho = 1$ g cm⁻³. Deuterium masses were substituted for hydrogen masses to allow for a longer time step. No empirical dispersion corrections were applied in our DFT-MD or DFTB-MD simulations of glycine. Umbrella sampling simulations were performed using the PLUMED 1.3 plugin [49, 50].

Car-Parrinello [51] DFT-MD simulations were performed using the Quantum Espresso *ab initio* simulation package [52, 53]. We closely followed the simulation protocol of [43] for simulations of glycine condensation. Three differences between [43] and the present work are the DFT software package, the choice of collective variables, and that the former used metadynamics [42] rather than umbrella sampling. Trajectories were integrated with a 0.145 fs (6.0 au) time step and electron mass of 700 au with a Nosé-Hoover thermostat [54, 55] applied to both the ionic and electronic degrees of freedom. Reducing the time step by a factor of four was necessary to conserve the Hamiltonian in some umbrella sampling simulations due to the strong bias potentials that were required to ensure adequate sampling of s and z space. The electronic structure was evaluated using the Perdew–Burke–Ernzerhof (PBE) functional [56] and ultrasoft pseudopotentials [57] with a planewave cutoff of 25 Ry.

Born–Oppenheimer DFTB–MD simulations were performed using LAMMPS [58, 59] with forces and stresses evaluated by the DFTB+ code [33, 60]. Isothermal–isochoric (NVT) trajectories were integrated with a 0.20 fs time step and a Nosé–Hoover–style thermostat [54, 55]. The electronic structure was computed with a minimal basis and with Fermi–Dirac thermal smearing with the electronic temperature set equal to the instantaneous ionic kinetic temperature at each time step. Electronic degrees of freedom were propagated using an extended Lagrangian approach [61–64] with four self-consistent charge cycles per time step. DFTB simulations were performed using the mio-1-1 parameter set (available at <http://www.dftb.org>), which is a typical off-the-shelf parameter set for CNHO systems. Selected mio-1-1 $E_{\text{Rep}}^{\alpha\beta}(r_{ij})$ potentials were replaced with fitted ones for the force-matched DFTB simulations, discussed below.

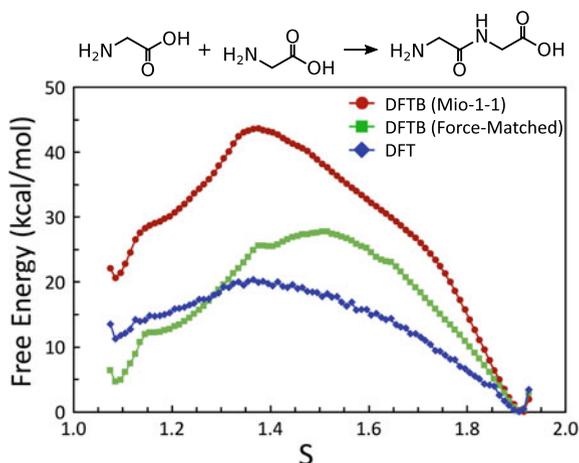
4.3.2 E_{Rep} Parameterization and Benchmarking

We begin with generation of the DFT training trajectory via umbrella sampling along “path progress” collective variable s , at fixed “path deviation” collective variable, z . Collective variables are constrained by the harmonic bias potential $V_{\text{Bias}}(s, z) = \frac{1}{2}K_s(s - s_0)^2 + \frac{1}{2}K_z(z - z_0)^2$, where z_0 is fixed at -0.10 , corresponding to the location of the two reference configurations and their associated local free energy minima, in z -space. A total of 37 umbrellas are utilized, evenly spaced between $1.050 \leq s \leq 1.950$, by increments of 0.025, where initial configurations were taken from 5 ps long unbiased simulations at $s = 1.950$. We note that the s coordinates for glycine and diglycine minima are approximately 1.1 and 1.9, respectively. Each of the 37 umbrella simulations were run for 2 ps, where the first picosecond was reserved for construction of the training trajectory, and the second half, for free energy calculation. The training trajectory was extracted by taking configurations at $t = 250$ and 500 fs from each umbrella, resulting in a total of 74 frames, and DFT forces were evaluated at the electronic ground state.

Because the ratio of C, N, O, and H atoms was highly uneven (4:2:59:120 for C:N:O:H, respectively), target training forces and design matrix elements $M_{r_i,c}$ for atom i were weighted by $1/N_{e_i}$ where the subscripts r_i and c of the design matrix represents the combined index over Cartesian values of the force on each atom and the combined index over model coefficients, respectively, and N_{e_i} is the number of atoms of type e_i . This reduces the contribution from interactions between the solvent H_2O molecules, which significantly outnumber the C–N, C–O, and N–H interactions that are more important in the condensation reaction. Repulsive potentials of the form in (4.4) were obtained for C–C, C–N, C–O, C–H, N–H, and O–H interactions using the weighted DFT training set with radial cutoffs $r_c^{e_i,e_j}$ set to mio-1-1 defaults. The N–N, N–O, O–O, and H–H repulsive potentials were taken from the mio-1-1 parameter set because those interactions were not well-sampled within their corresponding cutoffs.

In order to evaluate performance of the resulting model, simulations analogous to those described above (except using a 10 ps unbiased simulation and 20 ps umbrella

Fig. 4.1 Glycine condensation free energy surfaces $F(s, z = -0.10)$ predicted by DFTB with mio-1-1 and force-matched parameters, and DFT



simulations) were run for the resulting force-matched DFTB model, as well as for the standard mio-1-1 DFTB parameter set. A slice of the free energy surface $F(s, z = -0.10)$ was computed from histograms of s and z extracted from the last 10 ps of the force-matched and mio-1-1 DFTB and the last 1 ps of DFT using the two-dimensional WHAM equations (Fig. 4.1).

Beginning with results obtained from mio-1-1 DFTB calculations, we find the lowest free energy configuration in all three independent simulations is diglycine, so we chose those minima as the reference energy. We performed three independent sets of umbrella sampling simulations with this model as a means to assess the uncertainty in our calculations. The relative free energy differences from two unreacted, solvated glycine molecules were 12, 22, and 27 kcal mol⁻¹ across the three independent simulations, with maximum barrier heights of 41, 44, and 48 kcal mol⁻¹, respectively. The WHAM solution computed using histograms from all three independent runs combined yields a free energy of reaction of $\Delta F = 21$ kcal mol⁻¹ and an approximate free energy barrier of $\Delta F^{\text{barrier}} = 44$ kcal mol⁻¹ (for the back reaction). Determination of a free energy of activation (ΔF^{\ddagger}) and transition state would require a committer-style trajectory analysis [65], which is beyond the scope of the current study. We compute the uncertainties in ΔF and $\Delta F^{\text{barrier}}$ to be 8 and 3 kcal mol⁻¹, based on the standard deviation of these calculations.

Figure 4.1 also provides the free energy slices for the presently force-matched DFTB model as well as DFT. We note that DFT results should be viewed as an estimate, as the free energy calculations are not necessarily converged after one picosecond of averaging. We observe that the force-matched DFTB model predicts a significantly lower barrier height ($\Delta F^{\text{barrier}} = 28$ kcal mol⁻¹) than mio-1-1, which is much closer to the DFT prediction ($\Delta F^{\text{barrier}} = 20$ kcal mol⁻¹). It is unclear from this comparison whether ΔF is improved through force matching. The mio-1-1 and force-matched DFTB results respectively predict $\Delta F = 21$ and 5 kcal mol⁻¹, both of which are close to the DFT prediction $\Delta F = 11$ kcal mol⁻¹ within uncertainty. The topology of $F(s, z = -0.10)$ is qualitatively different between the DFTB and DFT

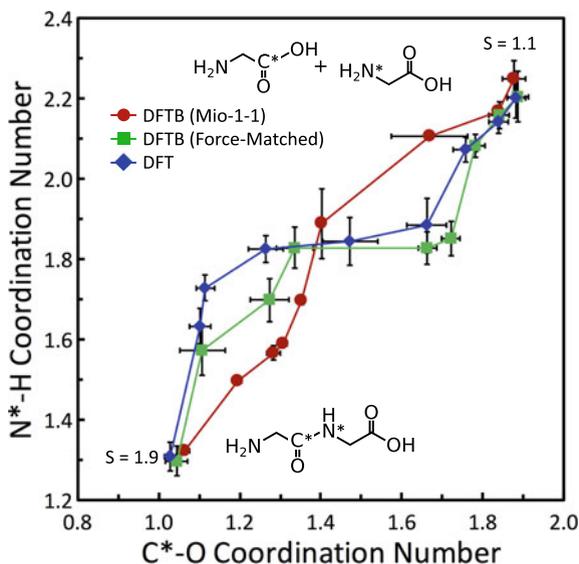


Fig. 4.2 Reaction paths expressed in terms of time-averaged coordination numbers for the dipeptide bond atoms C* and N* extracted from the umbrella sampling simulations used to compute the free energy surfaces in Fig. 4.1. Data points correspond to independent umbrella trajectories with $1.1 \leq s_0 \leq 1.9$ spaced in 0.1 increments. Uncertainties for the DFT and force-matched DFTB paths are standard deviations of the respective coordination numbers and uncertainties for the mio-1-1 DFTB path are standard errors computed over the paths predicted by three independent realizations

results, but the two DFTB predictions exhibit significant similarities aside from the numerical values for ΔF and $\Delta F^{\text{barrier}}$. One possibility is that topological features for $F(s, z)$ are largely determined through water solvation interactions, for which some second-order DFTB parameter sets are known to have significant errors [66].

All three models yield free energy surfaces that are in qualitative agreement with the DFT results for glycine condensation by Schreiner et al. [43] who predicted $\Delta F = 20 \text{ kcal mol}^{-1}$ and $\Delta F^{\text{barrier}} = 33 \text{ kcal mol}^{-1}$ at the same density and temperature. Quantitative differences between the Schreiner et al. and present (unconverged) DFT results are possibly due to the previously discussed differences in the free energy calculation protocols. Experiments [67, 68] and equation of state models [68, 69] suggest that glycine is slightly lower in free energy than diglycine by approximately 3 kcal mol^{-1} , which is in stark contrast to the mio-1-1 results and the results of Schreiner et al. The nominal equivalence in free energy for the reactants and products predicted by force-matched DFTB is generally more consistent with experiments.

To further assess the accuracy of our force-matched DFTB models for glycine condensation chemistry, we recast the reaction path in terms of the coordination between the atoms that form the dipeptide bond, namely the C* to O atoms and N* to H atoms with the same coordination number functions used to compute the matrix elements of (4.7). Reaction paths expressed in terms of these coordination numbers are shown in Fig. 4.2. We observe stark differences between mio-1-1 and DFT, with

agreement only at the end points. The mio-1-1 calculations yield a concerted mechanism with near simultaneous coordination of the O–H and H groups to C* and N*, resulting in a monotonically increasing linear form in its mapping. In contrast, DFT predicts a sequential process wherein H first coordinates to N* in diglycine followed by O–H coordinating to C*, resulting in the plateau shown in the middle of the plot. Force-matched DFTB also predicts a sequential process that is nearly identical to the DFT one within uncertainty. Similarities between the DFT and force-matched DFTB paths indicate that those two models likely have similar local minima in the regions of \mathbf{R} -space that map to these particular path-dependent coordinates. The computed uncertainties in the three reaction paths presented here indicate that care should be exercised when interpreting chemistry from subtleties in path coordinates, as they are not necessarily able to differentiate between chemically distinct configurations. Nevertheless, force-matched DFTB provides a straightforward means of accelerating predictive simulations of long timescale chemistry when combined with enhanced sampling methods such as Umbrella Sampling, as shown here.

4.3.3 *The Total Free Energy surface for Glycine Dimerization: Insights from Force-Matched DFTB*

The computational efficiency of DFTB allows for a more thorough exploration and convergence of long timescale chemical reactivity than is currently practical with DFT. The z -dependent free energy surface was determined for both mio-1-1 and force-matched DFTB over an interval that captures neutral reactant and product species. A total of 259 independent 20 ps umbrella sampling simulations were performed to generate each surface, which corresponds to a combined simulation time of 5.18 ns each. As before, the s_0 were set in 0.025 increments and the z_0 were evenly spaced in 0.05 increments in the interval $-0.15 \leq z_0 \leq 0.15$. Note that zwitterionic configurations map to $z > 0.2$ with the chosen parameters and reference configurations for (4.7). Free energy surfaces were computed following the same protocol as previously discussed and are shown in Fig. 4.3.

The total changes in free energy predicted by mio-1-1 and the force-matched model (vs. the free energy slices, discussed previously) are respectively $\Delta F = 20.7$ and $4.5 \text{ kcal mol}^{-1}$, which are both within uncertainty of the results for the surface section $F(s, z = -0.10)$ shown in Fig. 4.1. It is apparent that the $F(s, z = -0.10)$ section does not capture the minimum free energy barrier for either model. The mio-1-1 results predict the reaction barrier to be $40.5 \text{ kcal mol}^{-1}$, compared to the force-matched model minimum barrier is $16.7 \text{ kcal mol}^{-1}$. Comparing the two surfaces reveals different z -coordinates for the minimum barrier, with the location for mio-1-1 being at $z = 0.02$ and for the force-matched model at $z = 0.17$. The force-matched model minimum barrier is located much closer to the zwitterionic region of the surface ($z > 0.20$). We note that the variation in $F(s, z)$ with z is generally greater for the force-matched model than for mio-1-1. Significant z -dependency highlights

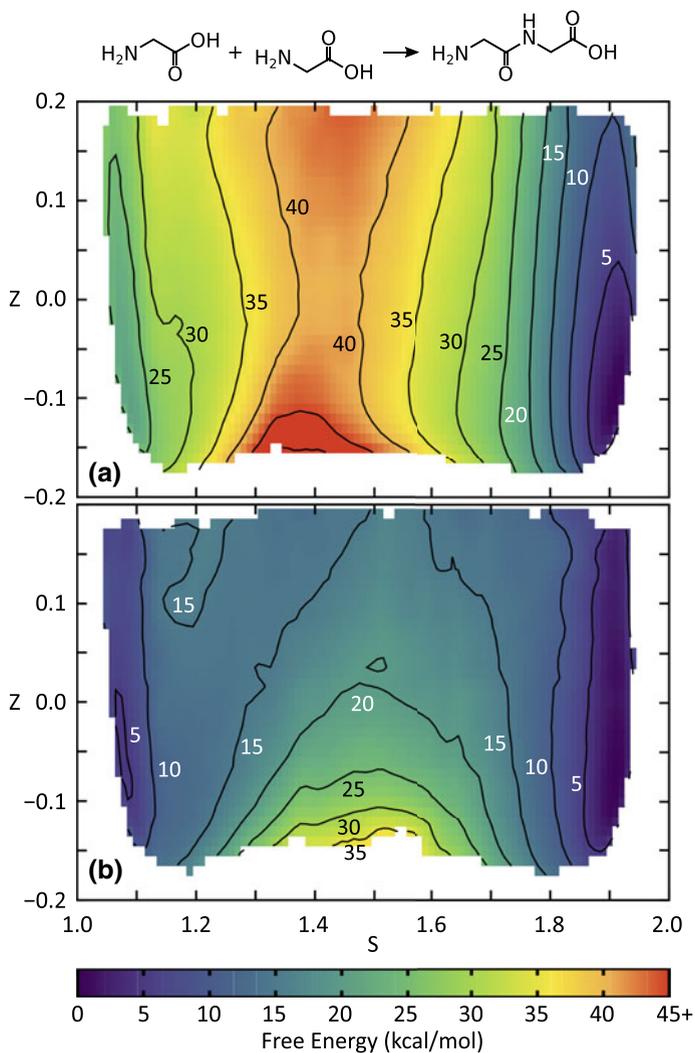


Fig. 4.3 Glycine condensation free energy surface $F(s, z)$ predicted by DFTB with **a** mio-1-1 and **b** force-matched repulsive potentials

the importance of extensively exploring the collective variable space when estimating the free energy and underscores the necessity for extending simulation times beyond those that can be practically reached with DFT.

4.3.4 DFTB Force Matching: Outlook

Accessing long timescales and performing large-scale ensemble studies beyond those that can be practically performed with DFT is critical for accurate determination of many condensed phase chemical reactions [40, 70]. We have shown that force-matched DFTB models hold promise as one avenue to extend quantum simulations to more realistic timescales while retaining much of the accuracy of DFT [40]. In particular, combining force-matched DFTB with accelerated sampling allows for the rapid generation of converged and accurate free energy surfaces for chemical reactions (discussed in further detail in recently published work [19]). This significantly expands the range of states that can be feasibly considered within the scope of a single study and allows for accurate quantum simulations that approach experimental time scales for reactivity. Our approach can help in the design and interpretation of laboratory synthesis studies, where there is a strong need for knowledge of the kinetics of chemical bond formation and breaking over extended periods of time.

4.4 ChIMES: Fast, Scalable Machine-Learned Reactive Force Fields

In this section, we discuss the problem of how to leverage force matching for ground-up construction of classical reactive MD models based on the Chebyshev Interaction Model for Efficient Simulation (ChIMES). ChIMES models are computationally efficient, fast to parameterize due to linear parameterization, and have the potential to maintain the accuracy of first principles methods across the large range of temperatures and pressures of interest for materials under extreme conditions. ChIMES does not include any explicit quantum-mechanical components, which affords significant gains in computational efficiency but can simultaneously complicate the development of transferable models relative to DFTB. Nonetheless, ChIMES can allow for large-scale million atom simulations where precise knowledge of electronic states might not be necessary for a specific research problem. We have applied ChIMES to water under ambient and high pressure–temperature conditions [71, 72], and metallic liquid carbon [73]. Here, we will focus on simple ChIMES force field development for metallic liquid carbon near the graphite–diamond–liquid triple point (e.g., ~ 15 GPa, 5000 K) as an illustrative example.

4.4.1 The ChIMES Force Field

There are numerous force fields designed to describe the physics governing chemical reactivity and metallic nature. For example, there are the ReaxFF [17], REBO [74], and COMB [75] potentials that leverage a reactive bond order description for bond

formation and breaking, the more general GAP [76, 77] and SNAP [78], methods which describe atomic environments through use of bispectrum components, and the EAM [79], and MEAM [80] embedded atom models. While highly successful in their intended applications, these models are not available for a large variety of systems or conditions, and thus frequently require reparameterization. Unfortunately, the complex nature of these reactive force fields combined with the large *ab initio* training data they frequently require can make their tailored parametrization for specific systems a challenging task.

ChIMES models are comprised of explicit n -body interatomic energy terms, which themselves, are constructed from linear combinations of Chebyshev polynomials. Resulting models are highly flexible and thus well suited for materials under extreme conditions. This allows for rapid parameterization of any number of new MD models through linear least-squares approaches like the singular value decomposition (SVD) [25].

To date, the ChIMES total energy of interaction has been described as the following sum of two- and three-body terms:

$$E_{\text{ChIMES}} = \sum_i \sum_{j>i}^N E_{ij} + \sum_i \sum_{j>i}^N \sum_{k>j}^N E_{ijk}. \quad (4.8)$$

Here, N is the total number of atoms in the system, E_{ij} is the pairwise interaction energy, and E_{ijk} is the energy between triplets of atoms. The two-body energy is expressed as follows:

$$E_{ij} = f_p^{e_i e_j}(r_{ij}) + f_c^{e_i e_j}(r_{ij}) \sum_{n=1}^{\ell_2} C_n^{e_i e_j} T_n(s_{ij}). \quad (4.9)$$

In this case, $T_n(s_{ij})$ represents a Chebyshev polynomial of order n , s_{ij} is the pair distance transformed over the interval $[-1, 1]$ (discussed below), and $C_n^{e_i e_j}$ is the corresponding coefficient for the interaction between atom types e_i and e_j . Permutational invariance of the polynomials is enforced for all interactions, e.g., $C_n^{e_i e_j} \equiv C_n^{e_j e_i}$. The term $f_c^{e_i e_j}(r_{ij})$ is a smooth cutoff function which is set to zero beyond a maximum distance defined for a given atom pair type, i.e., $f_c^{e_i e_j}(r_{ij}) = (1 - r_{ij}/r_{\text{max}}^{e_i e_j})^3$. In order to prevent sampling of r_{ij} distances below what is present in our DFT training set, we introduce a penalty function $f_p^{e_i e_j}(r_{ij})$, which we define as follows [71]:

$$f_p^{e_i e_j} = A_p (r_p^{e_i e_j})^3 \quad (4.10)$$

$$r_p^{e_i e_j} = \begin{cases} r_{\text{min}}^{e_i e_j} + d_p - r_{ij}, & r_{ij} - d_p < r_{\text{min}}^{e_i e_j} \\ 0, & \text{otherwise.} \end{cases} \quad (4.11)$$

The parameter A_p is a penalty function scaling factor and d_p is the penalty distance, which are preset to specific values. This allows for atoms to be “pushed” to larger

distances to avoid unphysical regions of the potential. We note that the penalty function was not sampled for any of the MD calculations presented here.

We map the interatomic distances over the interval of $[-1, 1]$ by first transforming the internuclear distance r_{ij} to the Morse variable, $x_{ij} = \exp(-r_{ij}/\lambda^{e_i e_j})$, where the λ parameter is the Morse variable range parameter [81–83], defined individually for each type of atomic pair interaction. The Morse variables lead to a natural decrease in the interaction strength as distance is increased. As a rule of thumb we set these values to correspond approximately to the nearest neighbor distance for each atom pair type from our fitting set, though in general we find that the results of our fit are relatively insensitive to these values. We then define the variable s_{ij} to be within the range $[-1, 1]$ through the operation $s_{ij} \equiv (x_{ij} - x_{\text{avg}}^{e_i e_j})/x_{\text{diff}}^{e_i e_j}$ where:

$$x_{\text{avg}}^{e_i e_j} = 0.5(x_{\text{max}}^{e_i e_j} + x_{\text{min}}^{e_i e_j}) \quad (4.12)$$

$$x_{\text{diff}}^{e_i e_j} = 0.5(x_{\text{max}}^{e_i e_j} - x_{\text{min}}^{e_i e_j}) \quad (4.13)$$

$$x_{\text{max}}^{e_i e_j} = \exp(-r_{e_i e_j}^{\text{min}}/\lambda^{e_i e_j}) \quad (4.14)$$

$$x_{\text{min}}^{e_i e_j} = \exp(-r_{e_i e_j}^{\text{max}}/\lambda^{e_i e_j}). \quad (4.15)$$

In this work, we will focus on development and application of a ChIMES model for two-body interactions, only, for the sake of simplicity. However, for completeness we include discussion of higher-body ChIMES terms as well. Similar to the two-body representation, the three-body energy is given as the product of Chebyshev polynomials for each of the three constituent atomic pairs:

$$E_{ijk} = f_c^{e_i e_j}(r_{ij}) f_c^{e_i e_k}(r_{ik}) f_c^{e_j e_k}(r_{jk}) \sum_{m=0}^{\mathcal{O}_3} \sum_{p=0}^{\mathcal{O}_3} \sum_{q=0}^{\mathcal{O}_3'} C_{mpq}^{e_i e_j e_k} T_m(s_{ij}) T_p(s_{ik}) T_q(s_{jk}). \quad (4.16)$$

The single sum given for the two-body energy is now replaced with a triple sum for the ij , ik , and jk polynomials, yielding a single permutationally invariant coefficient for each set of powers and atom types, $C_{mpq}^{e_i e_j e_k}$. The primed sum indicates that only terms for which two or more of the m , p , q indices are greater than zero are included, which guarantees that three bodies i , j , k enter into the expression. The expression for E_{ijk} also contains the f_c smoothly varying cutoff functions for each constituent pair distance. Penalty functions are not included for three-body interactions and are instead handled by the two-body E_{ChIMES} terms. We note that the ChIMES functional form can readily be extended to higher-body terms, e.g., four-body energies, by expanding (4.16) to include a multiplication of all $\binom{n}{2}$ pairwise interactions for that set (i.e., six total for four-body terms). Determination of greater than three-body ChIMES models is the subject of future work.

4.4.2 Application to Molten Carbon: Computational Details

In our efforts discussed here, ChIMES models can be viewed as “cooked to order,” where transferability is not considered an inherent feature of the force field. Instead, models are generated for a target region of state point space by force matching to one or more short DFT simulations (\mathcal{O} 5–10 ps) for the system of interest. Self-consistent refinement can be used for systems expected to undergo significant changes (i.e., as for formation of new phases) [73]. Here, we generate 2-body ChIMES models for molten carbon at 5000 K and 2.43 g/cm³. This system is particularly well-suited for demonstrative purposes as it is monoatomic and exhibits fast dynamics. Model sensitivity to user-specified parameters including polynomial order, outer cutoff, and Morse transformation variable will be discussed, and results will be compared against those from popular reactive carbon models LCBOP [84] and REBO [85].

We obtain a NVT -MD trajectory for 256 carbon atoms at our target state point via DFT-MD simulations utilizing the VASP software package [86–89]. The Perdew–Burke–Ernzerhof generalized gradient approximation functional (PBE) [90, 91] and projector-augmented wave pseudopotentials [92, 93] were used, along with a planewave cutoff of 1000 eV, a 0.5 fs time step, and a global Nose–Hoover thermostat [54, 55]. A total of 20 evenly spaced frames are selected from the resulting DFT trajectory, resulting in 250 fs between frames. Chebyshev polynomial coefficients are then obtained by minimizing the objective function shown in (4.1). We note that this step generally takes from a few seconds to several minutes to yield optimal parameter values.

Simulations using the presently developed ChIMES models are run with in-house developed MD software and a 0.25 fs time step, whereas LCBOP and REBO simulations are run with the LAMMPS suite [59] and a 0.5 fs time step. An atomic configuration from our DFT-MD training trajectory was used as the starting point for all simulations. All classical MD simulations are in the canonical ensemble at 5000 K and 2.43 g/cm³, and utilize a global Nose–Hoover thermostat.

4.4.3 ChIMES Sensitivity to User-Specified Parameters

A total of nine ChIMES models were fit, which varied by polynomial order $\mathcal{O}_{2B} = 4, 8, \text{ or } 12$, outer cutoff distance $r_{\max}^{\text{CC}} = 2.00, 3.15, 4.25, \text{ or } 5.00 \text{ \AA}$, and Morse variable $\lambda^{\text{CC}} = 1.25, 2.25, \text{ and } 3.25 \text{ \AA}$. The remaining model parameters, A_p^{CC} , d_p^{CC} and r_{\min}^{CC} were fixed at values of $10^8 \text{ kcal/mol} \cdot \text{\AA}^3$, 0.01 \AA , and 1.0 \AA , respectively, where the inner cutoff was set to the minimum distance observed in the training configurations, and A_p^{CC} and d_p^{CC} were selected to yield a penalty function strong enough to prevent the system from exploring distances smaller than $r_{\text{c,in}}^{\text{CC}}$, while otherwise having minimal influence on dynamics. Models have a total of $n_{\text{usrpar}} + \mathcal{O}_{2B}$ parameters, where n_{usrpar} is the number of user defined parameters, and for the

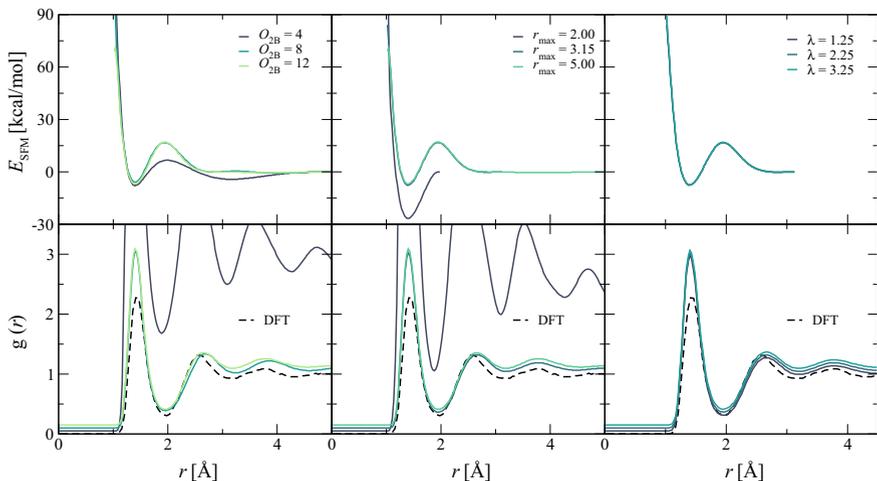


Fig. 4.4 Sensitivity of system structure to user-defined model parameters. Radial distribution functions have been offset by 0.05 \AA for clarity. E_{SFM} is the 2-body potential function derived from force matching

present work, encompasses the 6 parameters $\{\mathcal{O}_{2B}, r_{\text{min}}^{\text{CC}}, r_{\text{max}}^{\text{CC}}, \lambda^{\text{CC}}, A_{\text{p}}^{\text{CC}}, A_{\text{p}}^{\text{CC}}\}$. Thus, a model with $\mathcal{O}_{2B} = 10$ would have 16 parameters.

Figure 4.4 demonstrates how the potential energy surfaces (PESs) and subsequent radial distribution functions (RDFs) are altered as the Chebyshev polynomial order, outer cutoff distances, and Morse variable are increased. The results indicate that for both order and outer cutoff, some threshold value must be surpassed to obtain reasonable results, but beyond that value, the predicted system structures are indistinguishable from one another. Specifically, simulations using models with either $\mathcal{O} = 4$ or an $r_{\text{max}}^{\text{CC}}$ of 2.0 \AA solidify while all remaining simulations (using larger order and $r_{\text{max}}^{\text{CC}}$ values) yield RDFs that are in agreement with one another. Decreasing polynomial order limits model flexibility, and, as shown for the $\mathcal{O} = 4$ case, under-structuring of the PES (i.e., as compared to $\mathcal{O} = 8$ and 12 , between 1.5 and 2.25 \AA). The case with $r_{\text{max}}^{\text{CC}} = 2.0 \text{ \AA}$, on the other hand, fails because the cutoff distance forces the potential to zero where it would otherwise contain a maximum. Furthermore, the data suggests that a reasonable choice for selection of $r_{\text{max}}^{\text{CC}}$ would be the location of the second nonbonding minimum for a given $g(r_{\text{max}}^{\text{CC}})$. The data also indicate that the choice of λ^{CC} has little influence on resulting system structure. Nevertheless, a reasonable selection would be some characteristic “bonding” distance for the system, such as the location of the first peak in the pair RDF. An additional and pertinent user-specified aspect of the fit is the number of training frames utilized. For the present model problem, 20 frames were found to be sufficient for good fits, as including more did not yield significantly improved RMS errors or RDF predictions. In practice, one should consider the influence of training database size on a case-by-case basis, particularly

when increasing bodiedness of the ChIMES potential, and when targeting systems of multiple atom pair types and/or a variety of state points or phases.

4.4.4 Molten Carbon: Comparing ChIMES and Existing Reactive Models

Performance of the force-matched model with θ_{2B} , r_{\max} , and $\lambda = 12, 3.15,$ and 1.25 \AA , respectively [73] is compared to two existing reactive bond order carbon models, LCBOP and the 2002 parameterization of REBO [84, 85]. Both of these models contain numerous nonlinear parameters that are fit by iterative adjustment to either experimental or first-principles reference data. Notably, both models are fit to solid phase carbon data (i.e., graphite and diamond), and are designed to reproduce these corresponding material properties. Figure 4.5 presents the distribution of forces predicted from DFT, the present force-matched model, LCBOP, and REBO for a given set of configurations. Not surprisingly, the force-matched model yields significantly better agreement with DFT across the range of predicted values. Both LCBOP and REBO exhibit poor reproduction of the DFT force distribution; REBO overestimates the probability of observing low-lying forces, while LCBOP yields too high of a distribution of large magnitude forces.

The improved description of forces presented by the ChIMES model leads to more accurate system structure and dynamics, as is shown in Fig. 4.6. Both LCBOP and REBO predict overly ordered systems, with the first minimum in each RDF at nearly zero. REBO yields a better estimate of first peak height and recovers the DFT structure by the third peak, while LCBOP maintains solid-like structure over the

Fig. 4.5 Normalized distribution of forces acting on carbon at 5000 K and 2.43 g/cm^3

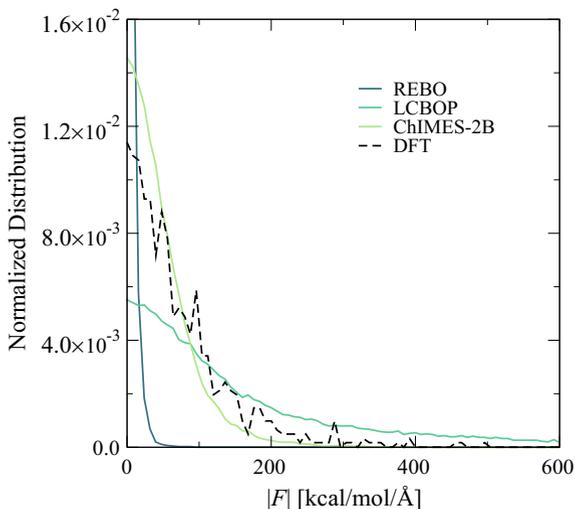
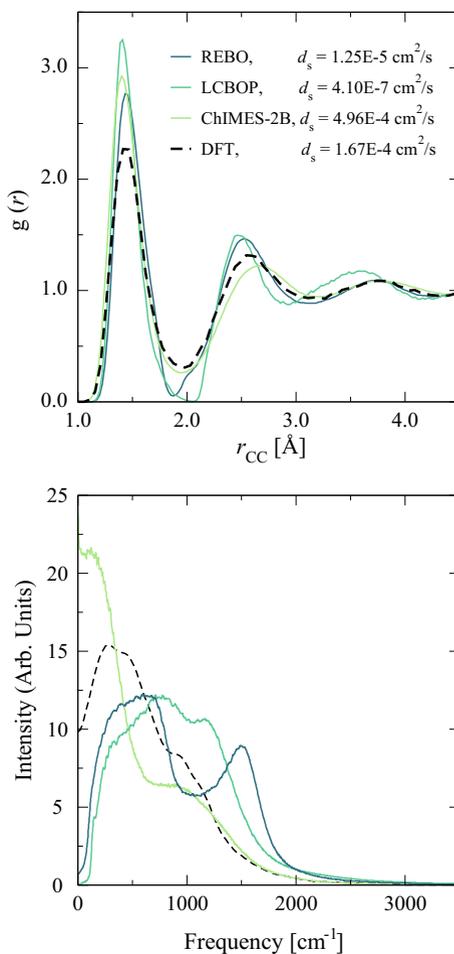


Fig. 4.6 Radial distribution functions (top) and vibrational power spectra (bottom) for carbon at 5000 K and 2.43 g/cm^3 . Self-diffusion constants are given in the legend for each model



range of distances presented. In contrast, the ChIMES model exhibits a reasonable description of the structure predicted by DFT over the entire range of distances. The overly-solid nature of the LCBOP and REBO systems is also reflected in the power spectra and self-diffusion constants. Both of these models yield peak intensities of nearly zero at low frequency, and accordingly, and predict diffusion constants that are several orders of magnitude smaller than the DFT result. Furthermore, the LCBOP and REBO power spectra do not capture the shape predicted by DFT, resulting in a broad distribution of high-intensity frequencies between 250 and 1500 cm^{-1} . In contrast, the ChIMES model captures the overall DFT power spectrum shape (a peak near 500 cm^{-1} and a shoulder near 1000 cm^{-1}) and exhibits a significantly improved diffusion constant, which is of the same order of magnitude as that from DFT. These results are particularly illustrative of the capability of ChIMES, given the exceedingly small training set and low-order functional form used here. We note, however, that

LCBOP and REBO are likely to be accurate over a wider range of conditions than the simple 2-body ChIMES model considered here. Better transferability with ChIMES can be obtained by adding many-body interactions and using more diverse training sets [73].

4.4.5 *ChIMES: Outlook*

Further ChIMES refinements of all of these properties have been determined through creation of three-body force fields, discussed in detail in recently published work [73]. Many-body effects can additionally be included in a number of ways such as explicit polarizability, or over-bonding terms (as in [71]). Our model problem discussed here demonstrates that the ChIMES force field and fitting framework provides a convenient approach for rapid generation of accurate DFT-informed reactive force fields. The utility of ChIMES extends far beyond the present application; in addition to use as a standalone molecular mechanics force field, three-body ChIMES models has been successfully leveraged as the repulsive term in DFTB for surface corrosion studies on plutonium [36].

4.5 Conclusions

Coupled advances in computing hardware and algorithms have enabled particle-based simulations to play an increasingly critical role in reactive materials research. Nevertheless, computer experiments based on quantum-mechanical methods remain too costly for many problems of interest. In this chapter, we have discussed two methods to force match models capable of extending the predictive power of quantum mechanics to larger scales. In the first, the repulsive term in the semi-empirical DFTB method was force-matched to DFT, to study glycine dimerization at low T and p conditions. Our model was found to yield a strong refinement of standard DFTB parameterizations, allowing for accurate determination of the diglycine formation free energy surface. In the second, force matching was used to construct a fully reactive ChIMES force field for molten metallic carbon, which is 10^5 times more efficient than DFT while retaining most of its accuracy. In general, our force-matched models are capable of serving as a high-fidelity proxy for obtaining DFT-quality forces on large time and length scales. There are any number of research areas that require atomic level knowledge of material reactivity, including studies of planetary interiors [94, 95], organic polymers subject to radiation and heat damage [96, 97], and the synthesis of materials with tailored properties [70]. The computationally efficient methods discussed in this chapter will have particular impact in these types of research areas, where there is traditionally a reliance on expensive DFT calculations for interpretation of imaging and spectroscopy experiments, and prediction of properties to guide materials synthesis.

Acknowledgements This work was partially supported by NASA Astrobiology: Exobiology and Evolutionary Biology Program Element NNH14ZDA001N. 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 16-LW-020 was funded by the Laboratory Directed Research and Development Program at LLNL with N.G. as principal investigator.

References

1. Barrios MA, Hicks DG, Boehly TR, Fratanduono DE, Eggert JH, Celliers PM, Collins GW, Meyerhofer DD (2010) *Phys Plasmas* 17:056307
2. Manaa MR, Reed EJ, Fried LE, Goldman N (2009) *JACS* 131(15):5483. <https://doi.org/10.1021/ja808196e>
3. Eggert J, Hicks D, Celliers P, Bradley D, McWilliams R, Jeanloz R, Miller J, Boehly T, Collins G (2010) *Nat Phys* 6(1):40
4. Goldman N, Reed EJ, Fried LE (2009) *J Chem Phys* 131(20):204103
5. Menikoff R, Sewell TD (2002) *Combust Theory Model* 6(1):103
6. Austin RA, Barton NR, Reaugh JE, Fried LE (2015) *J Appl Phys* 117(18):185902
7. Kroonblawd MP, Sewell TD (2016) *Propellants, Explos Pyrotech* 41(3):502
8. Goldman N, Fellers RS, Brown MG, Braly LB, Keoshian CJ, Leforestier C, Saykally RJ (2002) *J Chem Phys* 116(23):10148
9. Goldman N, Leforestier C, Saykally RJ (2005) *Philos Trans R Soc A* 363(1827):493
10. Goldman N, Reed EJ, Kuo IFW, Fried LE, Mundy CJ, Curioni A (2009) *J Chem Phys* 130:124517
11. Goldman N, Reed EJ, Fried LE, Kuo IFW, Maiti A (2010) *Nat Chem* 2:949
12. Cannella CB, Goldman N (2015) *J Phys Chem C* 119(37):21605
13. Goldman N, Fried LE (2012) *J Phys Chem C* 116(3):2198
14. Kenney JF, Kutcherov VA, Bendeliani NA, Alekseev VA (2002) *Proc Natl Acad Sci USA* 99(17):10976
15. Scott HP, Hemley RJ, Mao HK, Herschbach DR, Fried LE, Howard WM, Bastea S (2004) *Proc Natl Acad Sci USA* 101(39):14023
16. Kolesnikov A, Kutcherov VG, Goncharov AF (2009) *Nat Geosci* 2:566
17. Van Duin AC, Dasgupta S, Lorant F, Goddard WA (2001) *J Phys Chem A* 105(41):9396
18. Elstner M, Porezag D, Jungnickel G, Elsner J, Haugk M, Frauenheim T, Suhai S, Seifert G (1998) *Phys Rev B* 58:7260. <https://doi.org/10.1103/PhysRevB.58.7260>. <https://link.aps.org/doi/10.1103/PhysRevB.58.7260>
19. Kroonblawd MP, Pietrucci F, Saitta AM, Goldman N (2018) *J Chem Theory Comput* 14:2207
20. Ercolessi F, Adams JB (1994) *Europhys Lett* 26(8):583
21. Liu XY, Adams JB, Ercolessi F, Moriarty JA (1996) *Model Simul Mater Sci Eng* 4(3):293
22. Levenberg K (1944) *Quart Appl Math* 2:164
23. Marquardt D, *SIAM* (1963) *J Appl Math* 11:431
24. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) *Science* 220(4598):671
25. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) *Numerical recipes: the art of scientific computing*, 3rd edn. Cambridge University Press, Cambridge, UK
26. Tikhonov AN, Arsenin VY (1997) *Solutions of Ill-posed problems*, 3rd edn. Wiley, New York. Translated from Russian
27. Hoerl AE, Kennard RW (1970) *Technometrics* 12(1):55
28. Tibshirani R (1996) *J R Stat Soc Ser B (Methodological)* 267–288
29. Behler J (2015) *Int J Quantum Chem* 115:1032
30. Izvekov S, Rice BM (2014) *J Chem Phys* 140(10):104104

31. Porezag D, Frauenheim T, Köhler T, Seifert G, Kaschner R (1995) *Phys Rev B* 51:12947. <https://doi.org/10.1103/PhysRevB.51.12947>. <https://link.aps.org/doi/10.1103/PhysRevB.51.12947>
32. Koskinen P, Mäkinen V (2009) *Comput Mat Sci* 47(1):237. <https://doi.org/10.1016/j.commatsci.2009.07.013>. <http://www.sciencedirect.com/science/article/pii/S0927025609003036>
33. Aradi B, Hourahine B, Frauenheim T (2007) *J Phys Chem A* 111(26):5678. <https://doi.org/10.1021/jp070186p>. <http://dx.doi.org/10.1021/jp070186p>. PMID: 17567110
34. Gaus M, Cui Q, Elstner M (2011) *J Chem Theory Comput* 7:931
35. Goldman N (2015) *Chem Phys Lett* 622:128
36. Goldman N, Aradi B, Lindsey RK, Fried LE (2018) *J Chem Theory Comput* 14:2652
37. Goldman N, Srinivasan SG, Hamel S, Fried LE, Gaus M, Elstner M (2013) *J Phys Chem C* 117:7885
38. Srinivasan SG, Goldman N, Tamblyn I, Hamel S, Gaus M (2014) *J Phys Chem A* 118:5520
39. Armstrong MR, Zaug JM, Goldman N, Kuo IFW, Crowhurst JC, Howard WM, Carter JA, Kashgarian M, Chesser JM, Barbee TW, Bastea S (2013) *J Phys Chem A* 117:13051. <https://doi.org/10.1021/jp407595u>
40. Goldman N, Fried LE, Koziol L (2015) *J Chem Theory Comput* 11(10):4530. <https://doi.org/10.1021/acs.jctc.5b00742>. <http://dx.doi.org/10.1021/acs.jctc.5b00742>. PMID: 26574245
41. Torrie GM, Valleau JP (1974) *Chem Phys Lett* 28(4):578. [https://doi.org/10.1016/0009-2614\(74\)80109-0](https://doi.org/10.1016/0009-2614(74)80109-0). <http://www.sciencedirect.com/science/article/pii/0009261474801090>
42. Laio A, Parrinello M (2002) *Proc Nat Acad Sci* 99(20):12562. <https://doi.org/10.1073/pnas.202427399>
43. Schreiner E, Nair NN, Marx D (2009) *J Am Chem Soc* 131(38):13668. <https://doi.org/10.1021/ja9032742>. <http://dx.doi.org/10.1021/ja9032742> PMID:19725519
44. Khaliullin RZ, Eshet H, Kühne TD, Behler J, Parrinello M (2011) *Nat Mat* 10:693. <https://doi.org/10.1038/nmat3078>. <http://dx.doi.org/10.1038/nmat3078>
45. Branduardi D, Gervasio FL, Parrinello M (2007) *J Chem Phys* 126(5):054103. <https://doi.org/10.1063/1.2432340>. <http://dx.doi.org/10.1063/1.2432340>
46. Pietrucci F, Saitta AM (2015) *Proc Nat Acad Sci* 112(49):15030. <https://doi.org/10.1073/pnas.1512486112>. <http://www.pnas.org/content/112/49/15030.abstract>
47. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA (1992) *J Comput Chem* 13(8):1011. <https://doi.org/10.1002/jcc.540130812>. <http://dx.doi.org/10.1002/jcc.540130812>
48. Roux B (1995) *Comput Phys Commun* 91(1):275. [https://doi.org/10.1016/0010-4655\(95\)00053-I](https://doi.org/10.1016/0010-4655(95)00053-I). <http://www.sciencedirect.com/science/article/pii/001046559500053I>
49. Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, Donadio D, Marinelli F, Pietrucci F, Broglia RA, Parrinello M (2009) *Comput Phys Commun* 180(10):1961. <https://doi.org/10.1016/j.cpc.2009.05.011>. <http://www.sciencedirect.com/science/article/pii/S001046550900157X>
50. Plumed is available at <http://www.plumed.org/>
51. Car R, Parrinello M (1985) *Phys Rev Lett* 55:2471. <https://doi.org/10.1103/PhysRevLett.55.2471>. <https://link.aps.org/doi/10.1103/PhysRevLett.55.2471>
52. Giannozzi P, Baroni S, Bonini N, Calandra M, Car R, Cavazzoni C, Ceresoli D, Chiarotti GL, Cococcioni M, Dabo I, Corso AD, de Gironcoli S, Fabris S, Fratesi G, Gebauer R, Gerstmann U, Gougousis C, Kokalj A, Lazzeri M, Martin-Samos L, Marzari N, Mauri F, Mazzarello R, Paolini S, Pasquarello A, Paulatto L, Sbraccia C, Scandolo S, Sclauzero G, Seitsonen AP, Smogunov A, Umari P, Wentzcovitch RM (2009) *J Phys Condens Matter* 21(39), 395502 (2009). <http://stacks.iop.org/0953-8984/21/i=39/a=395502>
53. Quantum espresso is available at <http://www.quantum-espresso.org/>
54. Nosé S (1984) *J Chem Phys* 81(1):511. <https://doi.org/10.1063/1.447334>. <http://dx.doi.org/10.1063/1.447334>
55. Hoover WG (1985) *Phys Rev A* 31:1695. <https://doi.org/10.1103/PhysRevA.31.1695>. <https://link.aps.org/doi/10.1103/PhysRevA.31.1695>
56. Perdew JP, Burke K, Ernzerhof M (1996) *Phys Rev Lett* 77:3865. <https://doi.org/10.1103/PhysRevLett.77.3865>. <https://link.aps.org/doi/10.1103/PhysRevLett.77.3865>

57. Vanderbilt D (1990) *Phys Rev B* 41:7892. <https://doi.org/10.1103/PhysRevB.41.7892>. <https://link.aps.org/doi/10.1103/PhysRevB.41.7892>
58. Plimpton S (1995) *J Comput Phys* 117(1):1. <https://doi.org/10.1006/jcph.1995.1039>. <http://www.sciencedirect.com/science/article/pii/S002199918571039X>
59. Plimpton S (1995) *J Phys Chem C* 117(1):1
60. DFTB+ is available at <https://www.dftb-plus.info/>
61. Niklasson AMN, Tymczak CJ, Challacombe M (2006) *Phys Rev Lett* 97:123001. <https://doi.org/10.1103/PhysRevLett.97.123001>. <https://link.aps.org/doi/10.1103/PhysRevLett.97.123001>
62. Niklasson AMN (2008) *Phys Rev Lett* 100:123004. <https://doi.org/10.1103/PhysRevLett.100.123004>. <https://link.aps.org/doi/10.1103/PhysRevLett.100.123004>
63. Niklasson AMN, Steneteg P, Odell A, Bock N, Challacombe M, Tymczak CJ, Holmström E, Zheng G, Weber V (2009) *J Chem Phys* 130(21):214109. <https://doi.org/10.1063/1.3148075>. <http://dx.doi.org/10.1063/1.3148075>
64. Zheng G, Niklasson AMN, Karplus M (2011) *J Chem Phys* 135(4):044122. <https://doi.org/10.1063/1.3605303>. <http://dx.doi.org/10.1063/1.3605303>
65. Mullen RG, Shea JE, Peters B (2014) *J Chem Theory Comput* 10(2):659
66. Goyal P, Qian HJ, Irle S, Lu X, Roston D, Mori T, Elstner M, Cui Q (2014) *J Phys Chem B* 118(38):11007
67. Borsook H (1953) *Adv Protein Chem* 8:127. [https://doi.org/10.1016/S0065-3233\(08\)60092-3](https://doi.org/10.1016/S0065-3233(08)60092-3). <http://www.sciencedirect.com/science/article/pii/S0065323308600923>
68. Kitadaï N (2014) *J Mol Evol* 78(3):171. <https://doi.org/10.1007/s00239-014-9616-1>. <https://doi.org/10.1007/s00239-014-9616-1>
69. Shock EL (1992) *Geochim et Cosmochim Acta* 56(9):3481. [https://doi.org/10.1016/0016-7037\(92\)90392-V](https://doi.org/10.1016/0016-7037(92)90392-V). <http://www.sciencedirect.com/science/article/pii/001670379290392V>
70. Kroonblawd MP, Goldman N (2018) *Phys Rev B* 97(18):184106
71. Koziol L, Fried LE, Goldman N (2017) *J Chem Theory Comput* 13:135
72. Lindsey RK, Fried LE, and Goldman N (2019) *J Chem Theory Comput* 15:436
73. Lindsey RK, Fried LE, Goldman N (2017) *J Chem Theory Comput* 13:6222
74. Brenner DW (1990) *Phys Rev B* 42(15):9458
75. Yu J, Sinnott SB, Phillpot SR (2007) *Phys Rev B* 75(8):085311
76. Bartók AP, Payne MC, Kondor R, Csányi G (2010) *Phys Rev Lett* 104(13):136403
77. Bartók AP, Kondor R, Csányi G (2013) *Phys Rev B* 87(18):184115
78. Thompson AP, Swiler LP, Trott CR, Foiles SM, Tucker GJ (2015) *J Comput Phys* 285:316
79. Daw MS, Baskes MI (1984) *Phys Rev B* 29(12):6443
80. Baskes M (1992) *Phys Rev B* 46(5):2727
81. Wang Y, Shepler BC, Braams BJ, Bowman JM (2009) *J Chem Phys* 131(5):054511
82. Wang Y, Huang X, Shepler BC, Braams BJ, Bowman JM (2011) *J Chem Phys* 134(9):094509
83. Braams BJ, Bowman JM (2009) *Int Rev Phys Chem* 28(4):577
84. Los J, Fasolino A (2003) *Phys Rev B Condens Matter Mater Phys* 68(2):024107
85. Brenner DW, Shenderova OA, Harrison JA, Stuart SJ, Ni B, Sinnott SB (2002) *J Phys Condens Matter* 14(4):783
86. Kresse G, Hafner J (1993) *Phys Rev B Condens Matter Mater Phys* 47(1):558
87. Kresse G, Hafner J (1994) *Phys Rev B Condens Matter Mater Phys* 49(20):14251
88. Kresse G, Furthmüller J (1996) *Comput Mater Sci* 6(1):15
89. Kresse G, Furthmüller J (1996) *Phys Rev B Condens Matter Mater Phys* 54(16):11169
90. Perdew JP, Burke K, Ernzerhof M (1996) *Phys Rev Lett* 77(18):3865
91. Perdew JP, Burke K, Ernzerhof M (1997) *Phys Rev Lett* 78:1396
92. Blöchl PE (1994) *Phys Rev B Condens Matter Mater Phys* 50(24):17953
93. Kresse G, Joubert D (1999) *Phys Rev B Condens Matter Mater Phys* 59(3):1758
94. French M, Mattsson TR, Nettelmann N, Redmer R (2009) *Phys Rev B* 79:054107
95. Knudson MD, Desjarlais MP, Lemke RW, Mattsson TR, French M, Nettelmann N, Redmer R (2012) *Phys Rev Lett* 108:091102
96. Chenoweth K, Cheung S, van Duin ACT, Goddard WA, Kober EM (2005) *J Am Chem Soc* 127:7192
97. Melo LGA, Hitchcock AP, Susac D, Stumper J, Berejnov V (2018) *Phys Chem Chem Phys* 20:16625