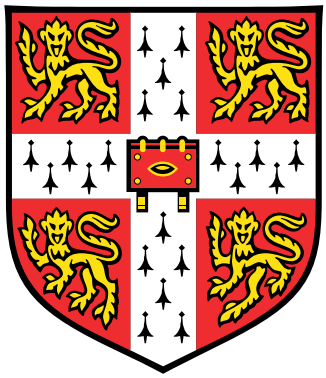
Locality of forces in molecular systems



Max David Veit
Churchill College

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of Master of Philosophy.

21st September 2015

Supervisor: Dr. Gábor Csányi

Declaration

This dissertation is substantially my own work and conforms to the University of Cambridge's guidelines on plagiarism. Where reference has been made to other research this is acknowledged in the text and bibliography. This dissertation does not exceed 15 000 words in length.

Acknowledgements

First, I would like to thank my supervisor, Gábor Csányi, for his guidance throughout the project; I would also like to thank him and all the other members of his research group for innumerable helpful discussions.

I gratefully acknowledge Shell Global Solutions International BV for funding my studies at the University of Cambridge. I also thank the computational chemistry team at the Shell Technology Centre in Bangalore, India, for useful discussions on hydrocarbon simulation.

Finally, I would like to thank the authors and contributors of the open software projects Python, IPython, SciPy, Matplotlib, and LaTeX, whose software allowed me to quickly and easily implement algorithms, test ideas, explore and visualize results, and communicate them in this dissertation.

Abstract

Locality is a basic requirement for most modern methods for modelling systems of thousands of atoms or more. Despite its widespread use, the assumption of locality remains largely unfounded in the underlying quantum mechanical description. This work presents an algorithm for quantifying the locality of forces in a general molecular system. The algorithm was tested on linear hydrocarbon systems; it confirmed the locality of the tight-binding model DFTB and quantified the extent to which chemical changes, such as bond conjugation and oxygen addition, introduce long-range effects within the more accurate DFT model. The results motivated the development of an intramolecular hydrocarbon potential using the GAP machine learning method; the new potential is a promising alternative to existing models used in hydrocarbon simulation.

Contents

Chapter 1

Introduction

The vast majority of methods for modelling large systems of atoms depend on the interactions between atoms to be local. For example, interatomic potentials used in chemistry generally write the Born-Oppenheimer potential energy of the system as a sum of terms between atoms directly connected via one, two, or three bonds. Long-ranged interactions are assumed to fall under well defined classes such as electrostatics and dispersion; these forces are accounted for separately. More recently, the assumption of locality has also seen increasing use in electronic structure methods that attempt to treat ever larger systems. Linear-scaling DFT scales linearly with the system size only because it assumes the electrons[1] are strongly localized [1].

The primary motivation for this work, however, is to further the development of machine learning methods for generating interatomic potentials [2, 3]. Like existing potentials, those generated by machine learning depend on the interactions between atoms to be local. Unlike most existing potentials, machine learning potentials address the issue of locality explicitly because they learn the energy as a flexible function of each atom's local environment. The sizes of these

[1] More precisely, the electron density matrix

environments are still limited because the cost of training and using a machine learning potential depends on the number of possible local environments, which itself increases steeply with the size of those environments. Nevertheless, since these methods treat the local environment size as a variable parameter, they are good tools for investigating locality in a systematic way.

A systematic study of locality is essential to justify the continued success of modelling methods for large systems. The assumption of locality remains unsupported by the Schrödinger equation governing all common materials. This equation describes a system's energy as an inseparable global property [4]. Although locality is expected to emerge in certain types of systems [1, 5] and it has been proven analytically under a restricted approximation of the underlying quantum mechanics [6], for the vast majority of existing modelling methods it remains an uncontrolled approximation. This work presents the first steps toward rigorously quantifying force locality in molecular systems and resolving this issue.

1.1 Theoretical background

In order to quantify the locality of interactions in a system, we first need to define the concept more rigorously. If the interaction between any given pair of atoms is local, that means we can neglect their interaction once the distance between the pair is sufficiently large. Concretely, this means we can decompose the total energy of the system into atomic contributions as follows:

$$E_{\text{tot}} = \sum_{i=1}^{N} \left(\epsilon(L_{r_C}(i)) + \delta(M \setminus L_{r_C}(i)) \right) \tag{1.1}$$

where E_{tot} is the Born-Oppenheimer potential energy of the system obtained by separating the electronic and nuclear components of the Schrödinger equation. The symbol M denotes the the set of positions, atomic numbers, and other relevant properties of all atoms in the molecule or system in question, while $L_{r_C}(i)$ denotes the **local neighbourhood** of atom i with radius r_C, which is the set of relevant properties of only those atoms located within a distance r_C of atom i. The radius r_C is called the **cutoff radius** of the neighbourhood[2]. The quantity ϵ is the **local energy**, the portion of that atom's energy contribution that depends only on its local neighbourhood. The error term δ is the remaining energy; this is the error incurred by ignoring other atoms outside the local neighbourhood. Often we can choose the local environment size to balance the computational cost of our modelling method against its accuracy, since both generally increase with the size of the local neighbourhoods.

The energy-based definition of locality above can be transformed to an equivalent version based on force, which is often a more accessible quantity than local energy. We arrive at this definition by taking the gradient of Equation (1.1), yielding

$$\begin{aligned}\mathbf{F}_k = -\nabla_k E_{\text{tot}} = \sum_{i=1}^{N} -\nabla_k \epsilon(L_{r_C}(i)) + \sum_{i=1}^{N} -\nabla_k \delta(M \setminus L_{r_C}(i)) = \\ = \sum_{i \in L_{r_C}(k)} \mathbf{f}_{ki} + \Delta_k(r_C)\end{aligned} \tag{1.2}$$

In other words, the force on a particular atom is the sum of forces on that atom from all other atoms in its local neighbourhood, plus an error term $\Delta(r_C)$ which depends on the size of the neighbourhood. It is this error term that we want to quantify and, eventually, control.

[2]Euclidean distance is not the only way of defining a local neighbourhood; in some cases, topological measures of distance like the number of bonds make more sense.

1.1.1 Locality in interatomic potentials

The above formulation of locality appears indirectly in the functional forms of most interatomic potentials. For example, take the **chemical forcefields**, which are interatomic potentials used in chemistry that write the total energy of a molecule as a sum of bond stretching, angle bending, torsional, improper torsion (general 4-body), and nonbonded (any atom pairs not included in the above) terms. For simplicity, consider a prototypical example of such a forcefield with only harmonic bond-stretching terms. This forcefield's energy expression is:

$$E_{\text{tot}} = \sum_{i,j \text{ bonded}} \frac{1}{2} k_{ij} (\mathbf{r}_j - \mathbf{r}_i)^2 \tag{1.3}$$

We can now write the local energy

$$\epsilon(\mathbf{r}_i, \{\mathbf{r}_j; j \in LB_1(i)\}) = \frac{1}{2} \sum_{j \in LB_1(i)} \frac{1}{2} k_{ij} (\mathbf{r}_j - \mathbf{r}_i)^2$$

that puts Equation (1.3) in the same form as Equation (1.1) with the error terms δ ignored; here, $LB_1(i)$ is the local neighbourhood consisting of all atoms within one bond of (i.e. bonded to) atom i and the additional factor of ½ corrects for double counting. This construction also works, with larger local neighbourhoods, for potentials that use bond-angle, torsion, and improper torsion terms.

Other types of potentials are more explicit in their use of local environments. Many-body potentials of the Tersoff or Brenner [7], embedded-atom [8], or machine learning [2, 9, 10] types use an explicit cutoff function to limit the range of the potential to only include each atom's local neighbourhood.

1.1.2 Nonlocal forces

The assumption of strictly local interactions is only justified in a small set of materials. The vast majority of systems are subject to various types of nonlocal interactions that are essential in determining the structure and properties of the material. These interactions are responsible, for example, for condensed phases [11]; they also play a large role in the packing and assembly of complex molecules such as proteins or DNA [12].

In order to model materials where nonlocal interactions are important, interatomic potentials include one or more extra terms inspired by the physics of the interactions between molecules (or parts of a large molecule). Although these interactions all ultimately derive from the electrostatic interactions between the electrons and nuclei of the system, they can broadly be classified into distinct physical effects. At long range, these effects are [11]:

Electrostatics

When two groups of atoms are both strongly polarized, i.e. they both have strongly asymmetric net charge distributions, the main interaction between the groups is the classical electrostatic interaction between their static, unperturbed charge distributions [11]. This interaction can be understood intuitively as the sum of electrostatic forces between the permanent charges, dipoles, and higher moments that arise in molecules containing species of differing electonegativities.

Obtaining the full charge distribution for a molecule in order to directly evaluate the classical electrostatic interaction energy is impractical for fast molecular modelling methods, as that would require a full quantum mechanical calculation for each geometry. Instead, these methods typically use approximations such as the distributed multipole expansion, where the charge distribution of a molecule

is approximated by a collection of point charges and higher moments assigned to a collection of sites throughout the molecule. For more detail on this method and how it assigns the electrostatic parameters, see Section 2.2.

Induction

Another type of interaction arises when the electron distribution of one group of atoms is distorted in the electric field of another. This interaction is more complicated to describe than the classical electrostatic energy because it requires a model for how an electron distribution responds to an applied field. Usually this is done by means of parameters called polarizabilities, which describe the induction of multipole moments at a site as a linear (although often frequency-dependent) reponse to the applied field [11].

Another reason the induction energy is difficult to describe is that it is a complex many-body effect; one polar molecule can polarize another molecule, but that molecule may in turn polarize a third molecule with its own electric field. This means the induction energy (in contrast to the electrostatic energy) cannot be expressed as a simple sum over pairwise interactions; in other words, it is non-additive. This complication, along with the relatively small role that the induction energy plays in many systems, means it is not explicitly included in most molecular modelling methods, although its effects may be included approximately in the electrostatic model [11].

Dispersion

By far the most common use of polarizabilities in molecular modelling methods is to describe dispersion interactions. When two widely separated groups of atoms both have neutral, unpolarized net charge distributions, this is the only force

between them.

Dispersion is an entirely quantum mechanical effect that results because the long-range correlation of electron motion in the two groups results in an overall lower-energy state [11]. The leading term of the dispersion energy decays as r^{-6} with intermolecular distance r. Higher-order terms also exist, including those that decay as r^{-8} and r^{-10}, but most potentials simply incorporate those effects into the sixth-power term.

Some potentials combine the dispersion interaction with another term that approximately describes the Pauli repulsion that occurs when atoms get close to each other. This approach has the advantage of removing the singularity of the inverse-power dispersion at short distances. The original and most well-known potential to use this approach is the Lennard-Jones potential, which includes a computationally efficient r^{12} repulsive wall. This potential is used in many of the most popular interatomic potentials in chemistry and physics. For instance, the AMBER [13], OPLS [14], and TraPPE [15] chemical forcefields use a Lennard-Jones potential plus Coulomb interactions between predefined partial charges for the nonbonded force terms. The related COMPASS forcefield [16] uses a similar form, except with r^{-9} repulsion instead of the traditional twelfth-power term. Even the AIREBO many-body potential for hydrocarbon simulation includes adaptively switched Lennard-Jones (12-6) potentials to represent intermolecular forces.

It was proposed as early as 1932 that an exponential repulsive wall might be more accurate [17, 11], although this 'exp-6' version[3] is still not as popular as the original Lennard-Jones potential. This functional form is used by the MM*n* forcefields, such as MM3 [19], and by the flexible Williams model of Tobias, Tu, and Klein [20].

[3]Sometimes called the Buckingham potential after the 1947 variant [18], although the true Buckingham potential also includes an eighth-power dispersion term

Although the sixth-power functional form of the dispersion energy (along with eigth- and tenth-power terms in more complete treatments) is widely agreed to be correct, the determination of the constant coefficient of r^{-6} is a subject of ongoing research. Usually the constant is determined using the same polarizabilities that parameterize the induction energy [12, 21, 22, 23]. Even among these methods, however, there are many different ways, at different levels of approximation and computational cost, of either defining the polarizabilities or of getting the coefficients from those polarizabilities. A new possiblity is to use machine learning to separately describe both the dispersion and short-range repulsion energies (see Section 4.4), although this avenue of research has not yet been pursued.

1.1.3 Quantum-mechanical models

Molecular modelling methods can be tuned to fit almost any reference data. The quality and predictive power of the method depend just as much on the quality of the fitting reference as on the method's functional form and fitting method. An early and popular reference is experimental data, which is accessible and often easy to translate into forcefield parameters. However, as more accurate first-principles modelling methods became available, it became viable to fit models to quantum-mechanical data. Such models are appealing because they are built from the bottom up, rather than constructed as empirical approximations, so their success is due to the success of the underlying physical principles.

To generate quantum-mechanical reference data, many methods are available. All of them attempt to solve the Schrödinger equation; they vary in the chosen compromise between speed and accuracy of approximation. Below are the methods most relevant to this work. For reference, the Schrödinger equation

for a material in the Born-Oppenheimer approximation is written:

$$\begin{aligned}\hat{\mathscr{H}}\Psi(\underline{\mathbf{r}};\underline{\mathbf{R}}) &= (\hat{T}_{\mathrm{el}} + \hat{V}_{\mathrm{nn}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{ext}})\Psi(\underline{\mathbf{r}};\underline{\mathbf{R}}) \\ &= E(\underline{\mathbf{R}})\Psi(\underline{\mathbf{r}};\underline{\mathbf{R}}) \end{aligned} \tag{1.4}$$

where $\underline{\mathbf{r}}$ is the set of all electronic coordinates and $\underline{\mathbf{R}}$ is the set of all nuclear coordinates. The operators in the Hamiltonian are, in order, the electronic kinetic energy, the repulsion between the nuclei, the attraction between the electrons and the nuclei, the repulsion between the electrons, and any external potential (due e.g. to an electric field).

Density functional theory

This method is based on a reformulation of the Schrödinger equation in terms of electron density. The original Schrödinger equation obeys a variational principle, where the expectation value of any wavefunction under the electronic Hamiltonian is always greater than or equal to the ground state energy, equality being reached only with the ground-state wavefunction. A similar principle applies when the electron density $n(\mathbf{r})$ is used in place of the wavefunction: The energy functional

$$\begin{aligned} E[n(\mathbf{r})] &= T_0[n(\mathbf{r})] \\ &\quad + \int n(\mathbf{r}) \left(V_{\mathrm{ext}}(\mathbf{r}) + \int \frac{1}{2} \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \, \mathrm{d}^3\mathbf{r}' \right) \mathrm{d}^3\mathbf{r} \\ &\quad + E_{XC}[n(\mathbf{r})] \end{aligned} \tag{1.5}$$

has a unique minimum at the ground-state density $n_0(\mathbf{r})$ [4]. In the above, T_0 gives the kinetic energy of a fictitious system of non-interacting electrons with

the density $n(\mathbf{r})$, the external potential V_{ext} now includes the repulsion between the nuclei as well as the attraction of the electrons to the nuclei, and the exchange-correlation energy E_{XC} must account for the effects of indistinguishable electrons and electron correlation, both of which the first two terms ignore. If the exchange-correlation functional is perfect, optimizing this energy expression with respect to electron density gives the electronic energy and density of the real system's ground state. This formalism is known as density functional theory, or DFT. The most common way to optimize the functional (1.5) is to write it as a superposition of non-interacting electrons, each acting under an effective potential, and then find the eigenstates (known as the Kohn-Sham orbitals) of that potential.

In practice, it is difficult to approximate the true exchange-correlation functional accurately. Existing approximations use either the local electron density, its gradient, or a nonlocal exchange term based on the Hartree-Fock method (see Section 1.1.3) and are fitted either to simple model systems or to existing calculations on large samples of molecules [4]. Approximations of this type generally capture intramolecular interactions and other short-range forces accurately. This accuracy, combined with its low computational cost relative to wavefunction-based methods, make DFT probably the most successful and widely used method to date for modelling medium-sized molecules and solid-state systems [4].

None of the standard functionals, however, adequately account for long-range intermolecular interactions. They are especially bad at capturing dispersion forces [22, 23]. There are functionals that correct the energy to account for dispersion [12]; however, these usually use simple r^{-6} functional forms that could just as well be explicitly added in to an existing interatomic potential.

Tight-binding approximation

Even though DFT offers an excellent compromise between speed and accuracy for a wide range of systems, it is still too slow for many common tasks. For example, running hundreds of thousands of steps of molecular dynamics to sample a DFT energy landscape is still intractable for large biological or solid-state systems containing thousands of atoms or more [24, 25, 26]. For such applications, methods are available that are faster and more approximate than DFT but are still founded in quantum mechanics, and are thus more accurate and transferable than empirical interatomic potentials.

One such method is based upon an expansion of the DFT energy functional near some reference density $n_{\text{ref}}(\mathbf{r})$. For a system with no spin polarization, the second-order expansion in some perturbation δn about the reference is [24, 25]:

$$\begin{aligned} E[n_{\text{ref}} + \delta n] \approx & \sum_i n_i \langle \psi_i(\mathbf{r}) | \hat{\mathscr{H}}_{\text{ref}} | \psi_i(\mathbf{r}) \rangle + \\ & E_{XC}[n_{\text{ref}}(\mathbf{r})] + V_{\text{nn}} - \int V_{XC}[n_{\text{ref}}(\mathbf{r})] n_{\text{ref}}(\mathbf{r}) \, \mathrm{d}^3\mathbf{r} + \\ & - \frac{1}{2} \iint \frac{n_{\text{ref}}(\mathbf{r}) n_{\text{ref}}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}^3\mathbf{r} \, \mathrm{d}^3\mathbf{r}' + \\ & \frac{1}{2} \iint \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta^2 E_{XC}}{\delta n(\mathbf{r}) \delta n(\mathbf{r}')} \bigg|_{n_{\text{ref}}(\mathbf{r})} \right) \delta n(\mathbf{r}) \delta n(\mathbf{r}') \, \mathrm{d}^3\mathbf{r} \, \mathrm{d}^3\mathbf{r}' \end{aligned} \tag{1.6}$$

with the reference Hamiltonian

$$\hat{\mathscr{H}}_{\text{ref}} = -\frac{1}{2}\nabla^2 + \hat{V}_{\text{ext}} + \frac{1}{2} \int \frac{n_{\text{ref}}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}^3\mathbf{r}' + \hat{V}_{XC}[n_{\text{ref}}(\mathbf{r})] \tag{1.7}$$

The ψ_i are the Kohn-Sham eigenstates of the reference system. The exchange-correlation energy also appears above as a potential $V_{XC}[n(\mathbf{r})] = \frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})}$ and the internuclear repulsion V_{nn} has again been written separately from the external potential. In this expansion, only terms in the zeroth and second order of the

density fluctuation appear; the first-order terms cancel.

If the above expansion is truncated to zeroth order taking only the first three lines of Equation 1.6, the electronic Hamiltonian no longer depends on the density itself and no self-consistent iteration is necessary. To solve the equation, the Kohn-Sham orbitals of the reference density are first expanded in an atom-centred basis set, whose approximate overlap and expectation values are precomputed and tabulated. The Schrödinger equation then reduces to a single matrix equation that uses precomputed values. This process is characteristic of a tight-binding method: No self-consistent iterations or explicit orbitals are needed to compute an energy. This specific formulation is called density functional based tight binding (DFTB) because it is parameterized using DFT calculations rather than empirical data. The precomputed values, lack of self-consistent iteration, and minimal basis make this method much less expensive than DFT on a comparable system [26].

A limitation of the zero-order method is that it cannot account for long-range forces. This method, along with other tight-binding models, has been shown to exhibit strict force locality in most insulating systems [6], meaning it can only account for short-range bonding interactions. The DFTB model can be modified to include long-range interactions by taking the Hamiltonian expansion to second order in the density fluctuation. This version, known as charge–self-consistent DFTB (SCC-DFTB), involves a self-consistent assignment of partial charges to each of the atoms. Although this modification makes the method more accurate and better able to account for electrostatic interactions, it is also more expensive than the zero-order expansion.

In this work, however, DFTB was only used to sample the molecular configuration space efficiently; more accurate long-range forces were simply computed

using DFT. Since the sampling method had to be efficient rather than accurate, it was done using only the zero-order DFTB method.

Quantum chemistry

On the other end of the continuum of speed versus accuracy lie the quantum chemistry methods. These methods are based on the Hartree-Fock procedure, which computes the many-electron wavefunction of a system in an iterative, self-consistent way. This procedure exactly accounts for exchange by antisymmetrizing the many-particle wavefunction. Modern, post–Hartree-Fock quantum chemistry methods also account for correlation by considering states with excited electrons. These methods include many-body perturbation theory, more commonly known as Møller-Plesset theory, as well as configuration-interaction methods, which optimize the energy in a basis of excited wavefunctions [27].

The main downside of these methods is their computational cost. Even on small molecules they are typically many times slower than standard DFT methods. Their cost also scales unfavourably with the number of electrons in the system, typically $\mathcal{O}(N^5)$ or worse [27, 28].

The main strength of these methods is their ability to account for both exchange and correlation effects with no uncontrolled approximations. For an accurate, systematic description of intermolecular forces, quantum chemistry is usually the best choice.

Chapter 2

Methodology

2.1 Locality algorithm

The method used to quantify locality is inspired by Equation 1.2. If the local neighbourhood of the atom labelled k is fixed, the local contribution to the force $\sum_{i \in L_{r_C}(k)} \mathbf{f}_{ki}$ is constant. We can therefore obtain an estimate of the remaining force $\Delta_k(r_C)$ by generating a sample of molecular geometries, all with the same local neighbourhood of atom k, and computing the standard deviation of the force $\mathbf{F}_k$ on the central atom over the sample. The local contribution drops out and we are left with an estimate of the error incurred by ignoring forces outside the local neighbourhood.

The quality of this estimate depends on the quality of the sample of geometries used to compute the standard deviation. Ideally the sample would be representative of of configurations in all the systems we want to simulate. For example, if we are only interested in locality of saturated hydrocarbon systems, the nonlocal sample only needs to include good coverage of saturated hydrocarbons. If we want to simulate unsaturated hydrocarbons, on the other hand, we should ideally include both saturated and unsaturated geometries in the non-

local sample. Section 2.1.2 gives more detail on how to generate chemically varied samples.

The above procedure tells us about the locality of only one geometry of the local neighbourhood. Usually, however, we are interested in a more general and approximate definition of locality: we want to know the *average* dependence of the force on an atom on changes outside its local environment. To do this, we need to compute force variances across a sample of local environment geometries and average them together. If we do this with many different local environment sizes, i.e. many different radii r_C, we obtain a profile of how the average locality of the force on a given atom depends on the size of its local environment.

2.1.1 Algorithmic details

The samples of molecular configurations used to compute the force variances in this work were generated by running molecular dynamics with the atoms inside the local environment constrained in place. The forces used to evolve the dynamics were computed using zero-order DFTB (see Section 1.1.3) via the implementation in the QUIP library [29]; the method used the standard '`mio`' parameter set[1]. The dynamics were evolved using the standard velocity Verlet algorithm with a timestep of 0.1 fs to capture the fast motion of hydrogen atoms. No thermostat was used, i.e. the dynamics sampled the NVE ensemble.

The dynamics were run either until a set number of steps was reached (4000 unless otherwise noted) or one of the mobile atoms entered the local neighbourhood, i.e. came within r_C of the central atom. Had atoms been allowed to move within the space of the local neighbourhood, then the neighbourhood would no longer have been constant throughout the trajectory and the computed variance

[1] Parameters available at `http://www.dftb.org/parameters/download/`

would no longer have represented the real force error. The only way to preserve the accuracy of the sample in such situations without introducing artificial forces was to terminate the trajectory, saving all timesteps up until that point. This necessity also means that the saved trajectories had widely varying lengths, which presented a challenge for data storage and later analysis.

While DFTB provides forces accurate enough to generate acceptable thermal samples of molecular configurations, it cannot capture long-range interactions such as electrostatics or dispersion. To obtain more accurate locality data that included some of these interactions, DFT calculations were done on a subset of the DFTB trajectories. The trajectories were subsampled at an interval of 20 fs and the resulting geometries were fed to the MOLPRO code [30, 31, 32, 33]. The DFT calculations used the minimal 6-31G* basis set and the local B-LYP exchange-correlation functional. These parameters were chosen for speed rather than accuracy due to the large number of calculations that were required; the accuracy should still be much better – especially qualitatively – than that of zero-order DFTB. Any calculations that did not converge were ignored and excluded from the final analysis.

In order to obtain a good sample of local environments to compute an average force variance, the starting structure of the test system was assigned random thermal velocities corresponding to some fixed initial temperature (4000 K unless otherwise noted) and equilibriated using molecular dynamics. The dynamics were run using DFTB and with the same parameters as the constrained systems described earlier, only in this case all of the atoms were free to move and the dynamics were only run for 1000 steps (100 fs). This equilibriation was repeated several times to obtain a number N_{inst} of instances of the system with different atomic positions and velocities.

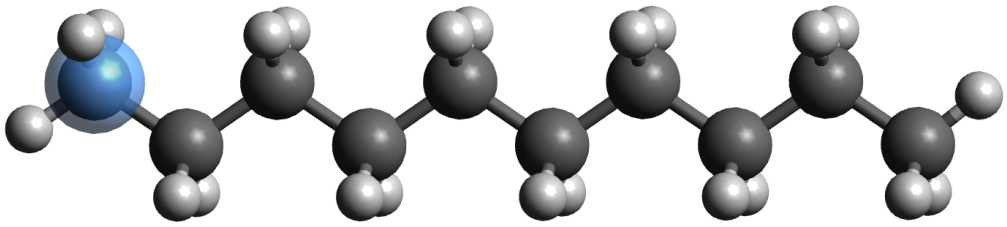
Each instance was used as the starting configuration of a separate locality computation wherein local environments of different sizes were frozen, the rest of the system was mutated and equilibriated as described in Section 2.1.2, dynamics were run on those mutations as described above, and variances of those samples were computed. The list of radii $\underline{r_C}$ used to determine a local environment was the same across all instances to ensure consistency. The result of this procedure was a list of N_{inst} lists of variances, one list per dynamics instance, each consisting of N_{env} variances, one per local environment radius. For each local environment radius, all variances corresponding to that size were averaged together, weighting by the number of configurations that were used to compute each individual variance. Finally, the square roots of the averaged variances were taken to give the force error as a function of local environment radius.

All of the dynamics, DFT calculations, and data analysis was controlled and automated using a Python code, interfaced to QUIP, as described in Appendix A.

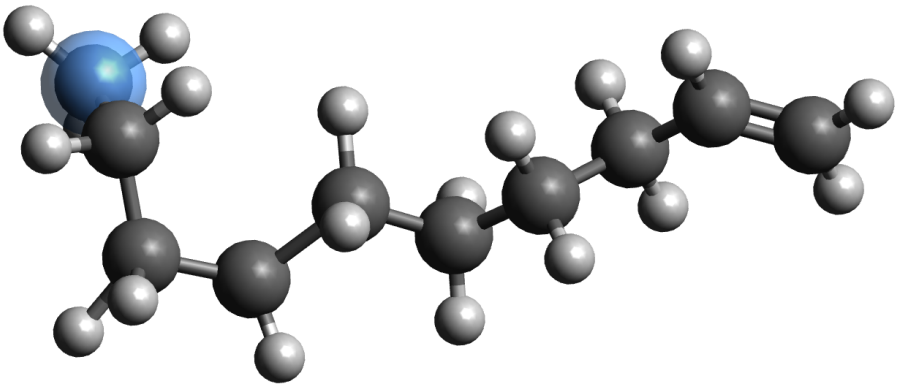
2.1.2 Mutations

One might also be interested in locality as a function not only of the motions of atoms outside a local environment, but also of changes in the number and type of atoms outside that environment. To carry out such locality computations, the capability was implemented to mutate molecules in chemically sensible ways. The two types of mutations were the conjugation of a carbon-carbon bond, i.e. the removal of one hydrogen atom on each end of the bond, and the addition of oxygen by replacing two hydrogen atoms on the same carbon atom. These mutations are illustrated in Figure 2.1.

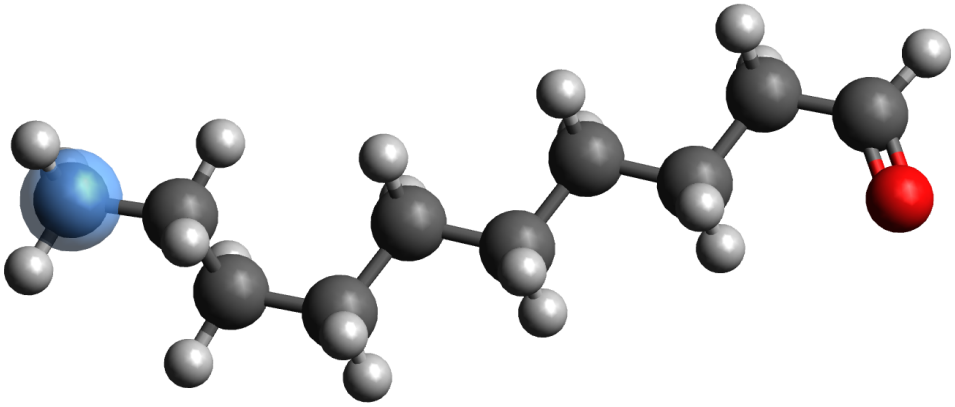
These mutations were done on each dynamics instance after the end of the initial equilibriation trajectory. Several copies of the equilibriated structure were



(a) decane



(b) decane with a conjugated bond



(c) decane with oxygen

Figure 2.1: The molecular structures of the test system, decane, along with two mutated examples. The mutated structures were first equilibriated and relaxed with DFTB *forces. The centres of the local environments used in the algorithm are highlighted in blue.*

made and the two mutations were applied once to one copy and twice to another copy; only the parts of the structure outside the local neighbourhood were mutated. Together with the unmutated structure, this process resulted in five chemically different structures with identical local neighbourhoods. Each structure was first relaxed using the BFGS linesearch optimization method until the maximum of the forces fell below 0.05 eV/Å; this step dealt with any unfavourable geometries resulting from the programmatic geometry changes. After the relaxation, the atomic velocities were rescaled to match the system's temperature before the mutation step, at the end of the instance's initial equilibriation. This rescaling ensured that all mutations of an instance were run at the same temperature, namely, the temperature that the system would have had if it had not been mutated and relaxed. The dynamics were then evolved, as described in Section 2.1.1, for each of the mutated structures independently. Finally, since the trajectories for each of the structures had identical local environments, they were concatenated (after removing 20 fs of equilibriation from the beginning of each trajectory) and variances were computed across the concatenated lists. The variances were then averaged across instances as described in Section 2.1.1. This procedure allowed locality to be computed with respect to chemical changes as well as physical ones.

2.1.3 Test systems

All of the calculations done in this work used a single linear alkane (saturated hydrocarbon chain) with a length in the range of 10–24 carbon atoms as the starting structure; one example (decane, 10 carbons) is shown in Figure 2.1a. These are ideal systems for testing the algorithm because the number of atoms in a single molecule scales only linearly with the maximum diameter of the system,

and hence with the range of local neighbourhood radii that can be explored.

2.2 Electrostatics

The locality algorithm can also be used to test whether existing models of long-range interactions can be used to make the force more local by subtracting the predicted nonlocal components. This application is explored here with the simplest nonlocal interaction to model, classical electrostatics. Although the physical basis of the interaction is simple, the practical computation of the interaction energy presents difficulties that are addressed with a large variety of techniques and approximations.

2.2.1 Parameterizations

The most complete, correct way to evaluate the electrostatic interaction energy is to obtain the electron distributions of the molecules or groups in question, then evaluate the classical Coulomb interaction energy of the distributions just as DFT and quantum chemistry codes do. Fast molecular modelling methods, on the other hand, do not have the luxury of accessing the full electron distribution for each geometry they encounter, so they must use some other approximation to predict the electrostatic energy.

The most common approximation is to assign electrostatic parameters such as charges, dipoles, and higher moments to a molecule or to various sites within a molecule and assume that these moments do not change during the simulation [11]. Such parameterizations are chosen to mimic the net charge distribution of the molecule or group in some way, for example by reproducing its electrostatic field at long range.

One method that is relevant especially in the context of small molecules is to assign each molecule a single series of multipole moments centred on that molecule, effectively expanding its electrostatic potential in spherical harmonics. This expansion has the disadvantage that it may diverge when the molecules get close to each other. For this reason, it is much more common to replace the single expansion centred on the molecule with multiple expansions centred on sites distributed throughout the molecule.

The latter method, termed a distributed multipole expansion [11], has much more freedom in the choice of sites and assignment of moments than the molecule-centred expansion. The simplest choice of sites is to assign a site to each atom in the molecule. Although it may be advantageous to use additional sites, the atom-site convention is the simplest and most practical one for most molecules. For this reason, it is the convention adopted by most molecular modelling methods as well as in this work.

2.2.2 Derivation of electrostatic moments

Even when the sites are restricted to lie on the atoms in a molecule, there is still considerable freedom in deciding how to assign charges and higher moments to the sites. The distributed multipole expansion does not specify any single optimal assignment of these parameters, in part because there is no clear target for what should be optimized. Accordingly, there are numerous established methods for computing these parameters for a given molecular geometry. Some try to give insight into the behaviour of atoms as part of a molecule, others try to match the molecule's electrostatic potential in the far field, while still others try to find a compromise. Below is a selection of some of the most popular methods.

The earliest example of a multipole moment assignment method is the Mul-

liken population analysis [34]. This method is simple to implement; it takes the output of electronic structure methods performed with atom-centred basis sets (atomic orbitals) and assigns charge to an atom by finding the population of all orbitals centred on that atom. Initially this method was developed to assist chemical intuition and decide how ionic or covalent a bond is. In reproducing the electrostatic forces on a molecule, however, Mulliken moments perform poorly [11].

More recent methods either use a partitioning of the electron density of a molecule or fit the electrostatic potential generated by that density. The former class of methods are usually known as atoms-in-molecules (AIM) methods; the latter are known as electrostatic potential fitting (ESP) methods. The AIM methods include Bader's method [35], which segments the charge distribution along zero-flux surfaces (minima in the electron density), and the Voronoi deformation density [36], which uses a Voronoi diagram with the nuclei as centres to divide the electron density. Other AIM methods decompose the charge density into overlapping regions; the Hirshfeld partitioning, for example, works by imagining the system as a superposition of non-interacting free atoms, then partitioning the real electron density at each point in proportion to the free-atom contributions [37]. To obtain the electrostatic moments for an atom, these methods expand the charge density assigned to an atom into a series of multipoles. Such methods can also be used to assign an effective volume to each atom, which can be used to derive polarizabilities for the induction and dispersion energies. In particular, the Hirshfeld definition has inspired several methods for computing effective atomic polarizabilities [21, 22, 23].

Electrostatic potential fitting methods generally use the computed charge distribution to evaluate the potential at a grid of points outside the molecule; they then use a simple least-squares fit to find charges that reproduce the potential

at those points [38]. A recent method [39] combines this approach with the AIM formalism to produce an all-purpose compromise between near-field and far-field electrostatic potential accuracy.

While all of the above methods can provide distributed multipole expansions to arbitrary order, most molecular modelling methods adopt the simpler convention of using only partial charges, in effect truncating the expansion at the zeroth order. Simply ignoring higher moments can lead to a large error [11], so many methods, such as electrostatic potential fitting, constrain the expansion so as to obtain the best approximation with only point charges. Another technique that can improve the accuracy of point-charge models is to use additional sites; however, this approach is usually only practical for small molecules such as water.

With the wide array of electrostatic moment assignment methods available, it is difficult – and often arbitrary – to choose one for a particular application. The ideal assignment for the creation of a local interatomic potential is the one that accounts for all long-range electrostatic interactions so that the remainder of the total force is strictly local. However, it is difficult to find this forcefield directly, so for now an approximation must be chosen. For the purpose of this work, the best compromise between accuracy and computational efficiency was offered by the Bader AIM charges, implemented in a compact and efficient code[2] [40, 41, 42]. The input to this code was an electron density file generated by MOLPRO using its default rectilinear grid with points spaced by about 0.25 Å. This spacing was relatively large and led to inaccuracies in the total electron count, but the parameter set was deemed good enough for a first trial of the electrostatic force field correction. In any case, the choice of the electrostatic assignment method and parameters is subject to change in future work on this problem.

[2]Code available at `http://theory.cm.utexas.edu/henkelman/code/bader/`

2.2.3 Use of electrostatic force fields

If a molecular modelling method uses an electrostatic force field derived using one of these methods, it must combine the electrostatics with the local force derived from each atom's local energy. Simply adding the forces is not a good approach because multipole-based electrostatic force fields all have singularities when two sites approach each other too closely. In the full quantum-mechanical description these singularities are resolved because the interpenetration of the charge distributions at short range reduces the electrostatic force between the two sites. Molecular modelling methods can simulate this effect by using functions that damp the electrostatic interaction at short range [11].

A more common approach, however, is simply to turn off the interaction between sites that are too close. Bond-based potentials such as AMBER ignore electrostatic forces between atoms that are within three bonds of each other (atoms separated by exactly three bonds are sometimes assigned a scaled-down interaction) under the assumption that those forces are already included in the bond, angle, or torsion terms [13]. Similarly, methods based on explicit cutoff radii can turn off electrostatic forces between atoms separated by less than the cutoff, effectively separating the energy expression into disjoint local and nonlocal terms. For machine learning methods, this means that any short-range electrostatic damping effects can be incorporated into the learning procedure as part of the local energy; the electrostatic model then only needs to accurately reproduce the far-field electrostatic potential.

Until now we assumed that the electrostatic moments do not change during a simulation. This means that the moments are precomputed or assigned with reference to some ideal, equilibrium geometry, or are averaged over some thermal sample of molecular configurations. To obtain the most accurate de-

scription of electrostatic forces, however, we need to allow the moments to vary with the geometry of the molecule. This task is an ideal application for machine learning. If the moments on an atom depend only on the local neighbourhood of that atom, then a sample of geometries with quantum-mechanically derived electrostatic moments can be used to train a machine learning model to reproduce those moments at any other geometry of interest. This method has been applied successfully to ionic systems [43], where the redistribution of charge is a major contributor to the energy of the system, and in multicomponent solid systems [44]. It would be interesting to apply this idea to molecular systems as well, as soon as a good method for computing the long-range electrostatic forces in these systems can be decided.

Chapter 3

Locality results

The locality algorithm provides an estimate of the variability of the force on a given atom when the atomic environment outside its local neighbourhood changes. This quantity, observed as a function of local neighbourhood size, gives insight into the physics of the interactions in a system. Not only is this information interesting in itself, but it can also be used as a tool to assist in the design of new interatomic potentials. Below are the results of this algorithm on a selection of hydrocarbon chains used as test systems.

3.1 Forces from DFTB

First, we can investigate the locality of the DFTB forces in the system. Zero-order DFTB provides forces good enough to evolve dynamics and explore the molecule's configuration space, but we do not expect much quantitative accuracy compared to a full quantum-mechanical model. Indeed, tight-binding models have been shown to always give strictly local (exponentially decaying) forces in systems

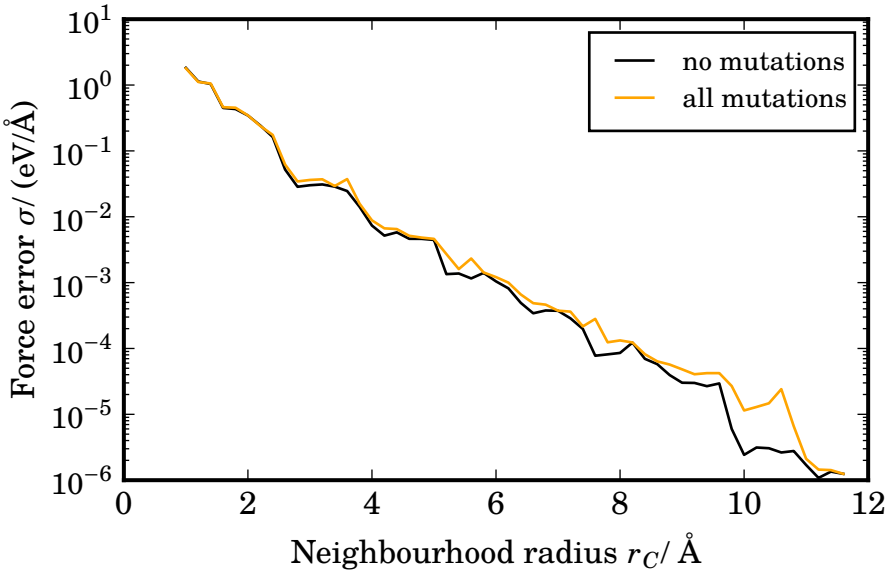


Figure 3.1: Errors of the DFTB forces on decane. The centre is at one end of the molecule; the results comprise $N_{inst} = 20$ dynamics instances. The black line shows force errors with no mutations, the orange line shows them with conjugation and oxygen mutations. The force axis is logarithmic.

that behave as insulators[1] [6], even including systems where we know nonlocal forces to be present, e.g. molecules with polar, interacting parts.

The locality algorithm bears out this prediction, as shown in Figure 3.1 for the example of a 10-atom hydrocarbon chain (*n*-decane). The locality centre is the carbon atom at one end of the chain; the distance between the carbon atoms at each end of the straightened chain – in effect, the length of the molecule – is 11.3 Å. The plot shows the standard deviations of the force on the central atom as a function of local neighbourhood radius. The standard deviations were first averaged over all dynamics instances, weighting by the square root of the number of points used to compute each data point, then the Cartesian components of the force errors were averaged together.

The trend in the plot is linear, indicating an exponential decay with distance

[1] More precisely, the strict locality holds only if the Coulomb interaction between different parts of the molecule is screened, which is what is expected in insulating systems.

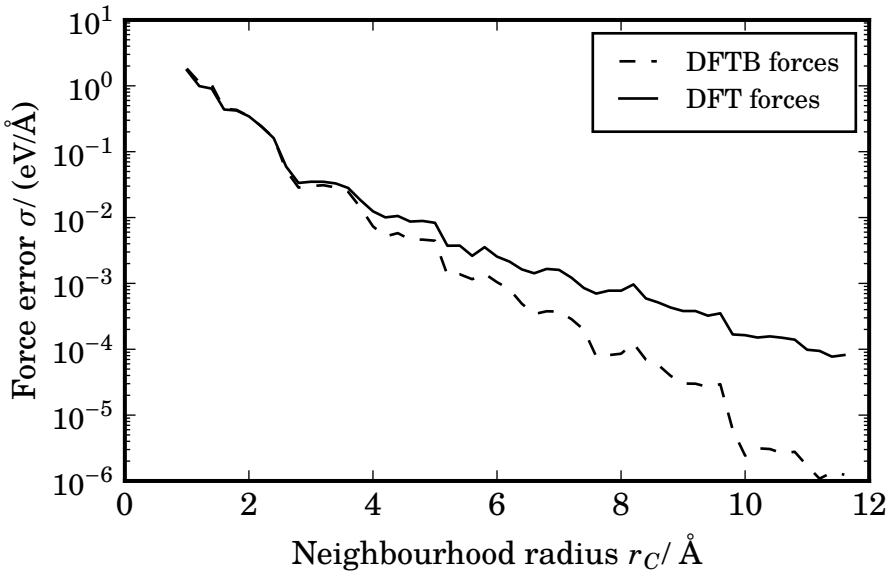


Figure 3.2: Comparision of the DFTB versus the DFT force errors on decane

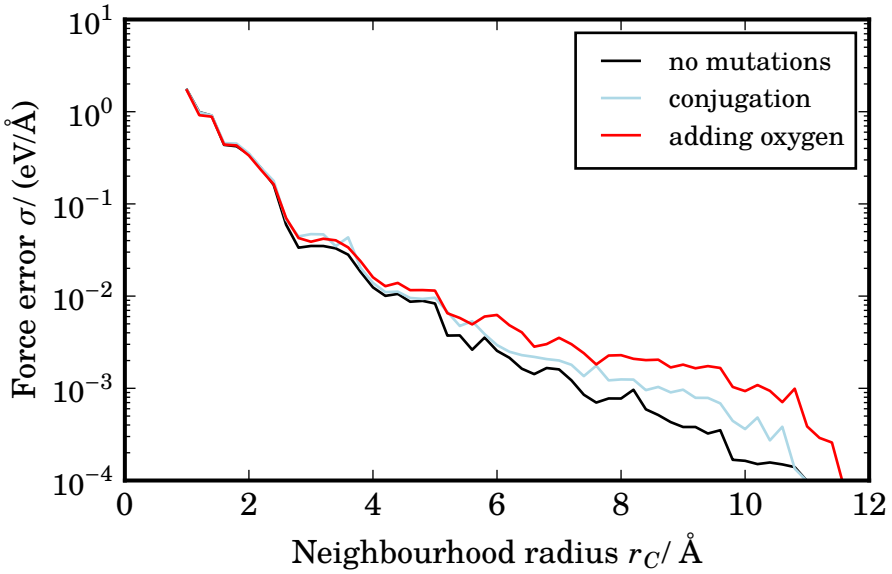


Figure 3.3: Force errors on decane with two types of mutations

of the forces in the molecule. This trend is not significantly affected by the inclusion of conjugation and oxygen mutations. Since both types of mutation result in closed-shell insulating systems, this retention of the exponential decay is what was expected.

3.2 Forces from DFT

For a more realistic description of the interactions in the molecule we need to use DFT. As described in Section 2.1.1, DFT calculations were done on geometries sampled from the DFTB trajectories to obtain a more accurate set of forces.

Even without including mutations, we see in Figure 3.2 that DFT predicts the forces to be less local than DFTB does. After about 4 Å, or two carbon-carbon bonds, the DFT force errors begin to deviate strongly from the DFTB errors.

In contrast to the zero-order DFTB model, the DFT forces are affected by mutations of the molecule. Figure 3.3 shows the force variance including conjugation and oxygen mutations separately. The forces are most strongly affected by the inclusion of oxygen atoms, although the reduction in locality is still not as large as expected. This could be because the plot is based on forces on a carbon atom. Perhaps a larger effect could be seen if the locality of the forces on an oxygen atom were computed instead (using an already mutated molecule as the starting point), since oxygen atoms participate more strongly in the electrostatic forces that reduce locality.

3.2.1 Size dependence

To determine whether the locality results are affected by the finite size of the molecule, the locality algorithm was also run on hydrocarbon chains of different

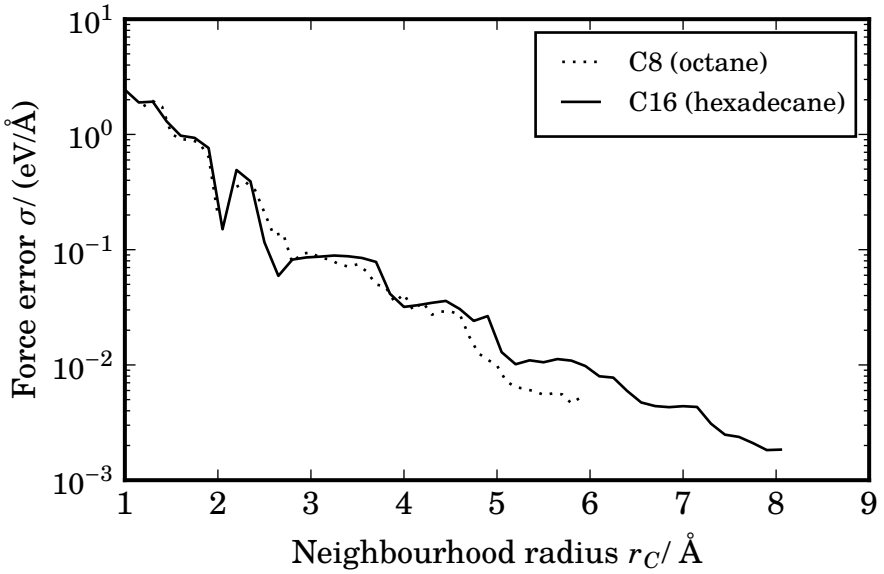


Figure 3.4: Comparison of the DFT *force locality of hydrocarbon chains of two different lengths. The local neighbourhood centres are now near the centres of the molecules, i.e. the fourth carbon for C8 and the eighth for C16. No mutations were done on the systems; each of* $N_{inst} = 50$ *instances was equilibriated for* 60 fs *without prior relaxation and run for an additional* 400 fs.

sizes, once on octane (8 carbons) and once on hexadecane (16 carbons). Figure 3.4 shows the results, which show no evidence of finite-size effects in this type of system.

3.3 Electrostatics

Returning to the decane system, the electrostatic field generated by Bader's AIM partial charges was additionally computed and subtracted from the DFT forces. The charges were computed at each geometry where a DFT calculation had been done; electrostatic forces were ignored between atoms separated by less than 5 Å. The resulting forces should ideally be better localized than the original forces, but the actual results were unsatisfactory. As Figure 3.5 shows, this correction only increased the force errors, instead of systematically decreasing them as the

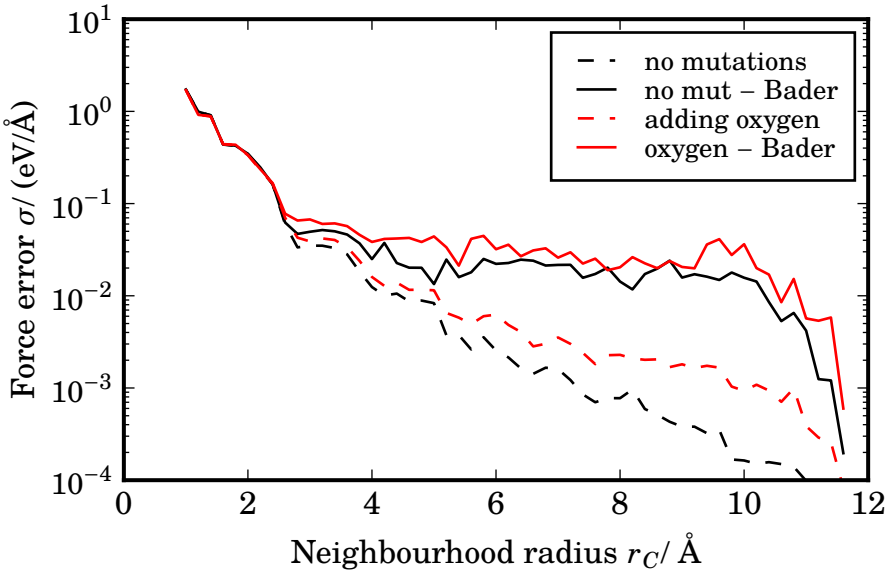


Figure 3.5: Force errors with electrostatic forces due to Bader charges subtracted between atom pairs more than than 5 Å *apart*

correction was intended to do. The adverse effect is particularly prominent when considering only unmutated systems. Unmutated decane is a nonpolar molecule and long-range forces due to its static charge distribution should be negligible, but the Bader charges still appear to generate a strong electrostatic field. Plain Bader charges thus provide an inadequate description of the electrostatic field in hydrocarbons since they seem to always overpredict the forces, so in future work a better electrostatic model will be necessary.

Chapter 4

Hydrocarbon GAP

4.1 Motivation

The linear alkanes have long been studied as a classical example of local systems. Their local behaviour is evidenced by their electronic structure; for example, the total electronic energy is essentially linear in the chain length [1] and the electron density matrix decays exponentially with spatial separation [45].

Force locality is a stronger condition than electron localization, though, and the results in Section 3.2 offer direct numerical evidence that the force itself is well localized. Not only does this evidence justify the creation of local potentials for hydrocarbons, it also provides valuable information that we can use to optimize a local potential for accuracy or for computational efficiency. For example, we can use the locality information to select a cutoff radius at which the average force error falls below a certain target. Additionally, once we select a cutoff, we can use the locality information to find the minimum force error that we can theoretically achieve. Both uses of the locality data will be explored below by fitting a Gaussian approximation potential (GAP) to model saturated and unsaturated linear hydrocarbons.

This new GAP could have applications beyond simply exploring the locality of hydrocarbon systems. Although many potentials are available to model the properties of hydrocarbons, most of them fail under conditions far away from those for which they were parameterized. These failures are especially apparent when computing quantities, such as transport properties, that are very sensitive to the details of the intramolecular force model. One such property is the shear viscosity of liquid hydrocarbons, which becomes increasingly difficult to predict accurately as either the pressure or the chain length is increased [46, 47]. The GAP method tested here offers a systematic way of overcoming those failures by generating accurate potentials for hydrocarbons that can be made to work even under extreme temperature or pressure conditions.

4.1.1 Existing hydrocarbon potentials

Systems made chiefly of carbon and hydrogen are central subjects of research in many different fields. They play important roles in chemical engineering, specifically in the study of petrochemicals; in biophysics, specifically in the study of lipids and membranes; as well as in materials science, for example in the study of diamond and of carbon nanostructures. The demand for good carbon-hydrogen potentials from these areas of research has resulted in many specialized models.

One of the earliest models optimized specifically for hydrocarbons is the optimized potentials for liquid simulation (OPLS) [14], which is a chemical forcefield with fixed bonds and angles, but with the torsional and nonbonded (Lennard-Jones 12-6 plus Coulomb) potentials fitted to a combination of experimental data and existing potentials computed for various liquid hydrocarbons. The OPLS is a united-atom model, meaning the hydrogen atoms are implicitly represented by the carbon centres to which they are attached, but an all-atom version [48] was

later developed. A forcefield with similar goals to OPLS is the flexible Williams model of Tobias, Tu, and Klein [20], which includes bond stretching and angle terms and replaces the Lennard-Jones plus Coulomb nonbonded interactions with an exp-6 potential. The TraPPE united-atom potential [15], on the other hand, was developed with emphasis on transferability and accuracy of hydrocarbon phase equilibria.

Potentials using functional forms other than the traditional chemical bond-angle-nonbonded form have also been applied successfully to carbon-hydrogen systems, with the added benefit that they can describe not only molecular systems, but condensed materials and nanostructures as well. One such potential is the modified embedded-atom method (MEAM) [8]; while the EAM was originally developed for use on metals, the angular terms added in the modified version allow it to describe covalently bonded networks as well. The MEAM has even been extended to molecular systems such as saturated hydrocarbons [49]. Another popular class of potentials for carbon-hydrogen systems includes those based on the Tersoff, and later Brenner, many-body potential [7]. A more recent potential of this type is the adaptive intermolecular reactive empirical bond order potential (ARIEBO) [50], which includes an adaptively switched Lennard-Jones 12-6 potential for intermolecular interactions. Becuase this potential is a flexible many-body potential that can model chemical reactions and because it performs well for alkane liquids, it was chosen as a point of comparison for the newly fitted GAP models.

The AIREBO potential was evaluated using the LAMMPS molecular dynamics code[1] [51]; the cutoff of the Lennard-Jones potentials was 5.0σ for each nonbonded pair, where σ is the distance where the L-J potential crosses zero. The AIREBO parameters were those from the original 2000 paper, [50].

[1] `http://lammps.sandia.gov`

4.2 Methodology

Gaussian approximation potentials (GAP s) are a systematic way of generating interatomic potentials from quantum-mechanical data using machine learning. A GAP takes a set of geometries, along with the corresponding quantum-mechanical total energies and forces (together called the **training set** or **training data**), and interpolates those quantities to new geometries using a nonparametric regression method based on Gaussian processes. Like other interatomic potentials, a GAP expresses the total energy of a system as a sum of atomic contributions. Unlike most interatomic potentials, a GAP directly interpolates these local energies in the space of local atomic neighbourhoods extracted from the training geometries. One difficulty in performing this interpolation is that local energies are usually not available from DFT or quantum chemistry training data. The GAP method works around this limitation by expressing the available quantities (forces and total energies) as linear operators on the local energies that it is trying to learn [2].

The method of Gaussian processes used to perform the interpolation is founded in Bayesian statistics [52]. It fits a function to the available data while resisting overfitting; that is, it produces a fit that ignores the pattern of random noise specific to the training set. The robustness of the method depends on a sensible choice of Bayesian priors for the fit parameters. In atomic systems, the length and energy scales of the Born-Oppenheimer potential energy surface are known well enough (within an order of magnitude) in most situations that the priors can be confidently chosen to produce a robust fit.

4.2.1 Descriptors

In order for a potential to compute the energy of a local atomic neighbourhood, it needs to transform that neighbourhood to obtain the quantities that appear in its energy expression. These transformed quantities should respect the symmetries of the local energy, i.e. they should not change if the entire neighbourhood is translated or rotated in space, or if the labels of atoms of the same species are permuted. These transformed quantities are called **descriptors** of the local neighbourhood, especially in the context of machine learning potentials. However, most other types of local potentials make use of them as well. Chemical forcefields, for instance, transform the neighbourhood into a set of bond lengths, angles, torsions, and other terms.

Machine learning potentials like GAP, on the other hand, have a variety of descriptor types they can choose from. While the set of atomic positions can be used directly, they are not good descriptors because they lack the symmetries of the local energy. Fortunately, many other descriptors are available that do respect these symmetries. The best-performing of these is called SOAP; it has been used successfully to create potentials for solid-state systems such as silicon and tungsten [9, 53]. A similar descriptor is the 'symmetry functions' approach developed for neural network potentials [10], although these are not nearly as good as SOAP at providing unique representations of atomic environments [9]. One downside of descriptors in this family is that they are more expensive to compute, being based on expansions of the local neighbour density, than the simpler chemical descriptors. As always, this cost must be considered alongside the greater accuracy achievable using these descriptors.

4.2.2 Training sets

For the machine learning potential to work well on a variety of sizes of hydrocarbons, its training data should include local neighbourhoods both at the ends of a chain as well as those in the middle, far away from either end. In order to provide the potential with a good variety of configurations, the molecule *n*-tetracosane (a 24-carbon linear hydrocarbon chain) was chosen for the training configurations.

The sample of training geometries was generated in much the same way as the trajectories for the locality algorithm in Section 2.1.1. The difference in this case is that none of the atoms were frozen; for each dynamics instance, only one DFTB trajectory was run with 500 equilibriation steps (50 fs) discarded and 19200 additional steps (1920 fs) recorded. Each trajectory was subsampled at an interval of 20 fs and DFT calculations were done on the samples, just as described in Section 2.1.1. The DFT results were then simply concatenated across all available instances to form the training set.

In order to assess the effects of temperature on the performance of the potential, this procedure was done twice, once with the instances initialized to a temperature of 3000 K (the system equilibriated to about 1500 K; 59 total instances were run) and once with an initial temperature of 500 K (equilibriating to about 350 K; 60 total instances were run). The latter temperature is closer to the ambient temperature conditions for which most existing potentials have been parameterized. The resulting training sets consisted of 5659 geometries with DFT energies at the hot temperature and 5726 such geometries at the cold temperature.

In addition, two more training sets for unsaturated hydrocarbons, consisting of 50 instances (4842 geometries) at the hot temperature and 50 instances (4792 geometries) at the cold temperature, were created by changing two bonds on the

tetracosane molecule to double bonds, thereby conjugating them. The resulting molecule is called 6,14-tetracosene. The sampling procedure just described was also applied to this system to obtain two training sets with unsaturated local neighbourhood geometries.

4.2.3 Implementation and parameters

The GAPs in this work were trained and evaluated using the original implementation[2] within QUIP [29]. The code implements Gaussian process regression as well as a variety of descriptors. More details on the process of computing descriptors and training a Gaussian process, the parameters involved in these computations, and the way that GAP works around local energies not being available, are described in [54] and [9]. The relevant parameters used in this work are summarized below; the others were left at their defaults.

The code was used to generate three different GAP models for each training set. Each was trained with a prior noise value of 0.05 eV per atom on the energies and 0.2 eV/Å on the forces.

The first of the models used the SOAP descriptor on nearest neighbours only. This was done by generating two different sets of descriptors, which formed the basis of two separate Gaussian processes whose results were later added together to obtain the total energy. The first descriptor set was only computed for local environments of carbon atoms, so it had a cutoff of 2.5 Å. The second descriptor set was only computed for local environments of hydrogen atoms, so it had a cutoff of 1.8 Å. Each cutoff was smoothed with a cosine curve of width 0.5 Å to avoid discontinuities in the descriptors and their derivatives. Both SOAP models were limited to $n_{\max} = 10$ radial basis functions and angular momentum

[2]Code available at `http://www.libatoms.org`; GAP version was 1432414728.

functions of up to $l_{\max} = 8$. The SOAP descriptors were used with a covariance function of the form $k(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) = (\mathbf{p}^{(i)} \cdot \mathbf{p}^{(j)})^{\zeta} + \delta_{ij}\sigma_e$, where the $\mathbf{p}$ are the normalized vectors of SOAP descriptor coefficients, σ_e is the prior error on the local energies, and the sensitivity parameter ζ was set to 4.0. The width of the atomic Gaussians used to compute the SOAP atomic density was left at the default of 0.5 Å.

The other two GAP models used two-body and three-body chemical descriptors, again using only nearest neighbours. These models described the energy of individual bonded-atom pairs or triplets, so the total energy was expressed as a sum over each pair or triplet in the system. The two-body descriptor was calculated separately for C-C, C-H, and H-H bonds, again using separate Gaussian processes. Two atoms were considered bonded if they were within the cutoff of 1.8 Å (again smoothed over a width of 0.5 Å). The covariance function in this case was a squared exponential, where the difference between two bond distances was squared, scaled, and its negative exponential taken. The scaling factor was computed using automatic relevance determination (ARD), in which the scaling is determined by the range of the descriptor across the training set. For the two-body descriptor, the differences were scaled by $\theta_{\text{rel}} = 0.2$ times the descriptor range.

The three-body GAP, on the other hand, incorporated the two-body descriptors just described, in addition to three-body descriptors on H-C-H, C-C-C, and H-C-C bonded atom triplets with (cosine-smoothed) bond distance cutoffs of 1.6 Å, 2.0 Å, and 2.0 Å, respectively. The three-body descriptors also used a squared exponential kernel with ARD, this time with a scaling factor of $\theta_{\text{rel}} = 1.0$ times the descriptor range. The three-body descriptors themeselves are not single distances; rather, each triplet was processed into three numbers: the sum of the two bond

distances to the central C atom, the squared difference of those distances, and the distance between the two outer atoms.

Gaussian process regression formally requires the building and inversion of a square covariance matrix with one row for each training example, an operation which scales as the cube of the number of examples. Moreover, training examples taken from dynamics simulations are usually highly correlated and thus redundant, meaning it would be an enormous waste of computational power to use all the local neighbourhoods in the training set. This problem is solved by a procedure called **sparsification**, which randomly selects representative configurations from the local neighbourhoods in the training set and expresses the remaining configurations as linear combinations of the representatives [2, 53]. Since the computational cost is governed by the number of sparse points rather than the full training set, this procedure considerably reduces the computational cost of evaluating a GAP, while largely preserving the coverage of configuration space present in the full training set.

The sparsification procedure used in this work was also implemented in the GAP code. The code randomly selected a preset number of descriptors from the training data, where each bond distance, each set of three-body distances, and each set of SOAP coefficients was counted as a descriptor. For the GAP models trained here, 100 sparse points were selected from the two-body descriptors, while 1000 sparse points were selected from the three-body and SOAP descriptors.

4.3 Results

4.3.1 Saturated hydrocarbons

The performance of a potential can be assessed by comparing the forces and energies it produces against some reference, in this case the DFT results used to train the potentials. Because the sparsification procedure ensures that only a small subset of the training data enters into the definition of the potential, the performance of the GAP models on their own training sets is a good measure of how well the potential will generalize to data not in the training set. This property, in combination with the Bayesian regularization built into the GAP, also means no separate validation set is necessary to guard against overfitting of the potential, since the validation data is essentially already contained in the training set.

Figure 4.1 shows how the predicted total energies of each of the geometries of the test system compare against the reference for each of several GAP models trained using different descriptors. The figure also shows how their performance compares to that of the AIREBO potential. The quantitative performance of the potentials is summarized in Table 4.1, which gives the root-mean-square (RMS) errors of the energies predicted by the models.

Evidently, the two-body GAP model failed to capture most of the variation in the training data. The RMS error of this potential is as large as the variation in the training set itself. This result is unsurprising, as the two-body nearest-neighbour GAP effectively only includes bond stretching terms, which are known to usually contribute little to the total energy of a molecule under ambient conditions (indeed, some hydrocarbon potentials, including OPLS [14], simply freeze bond lengths and angles). The three-body GAP model, on the other hand, is much

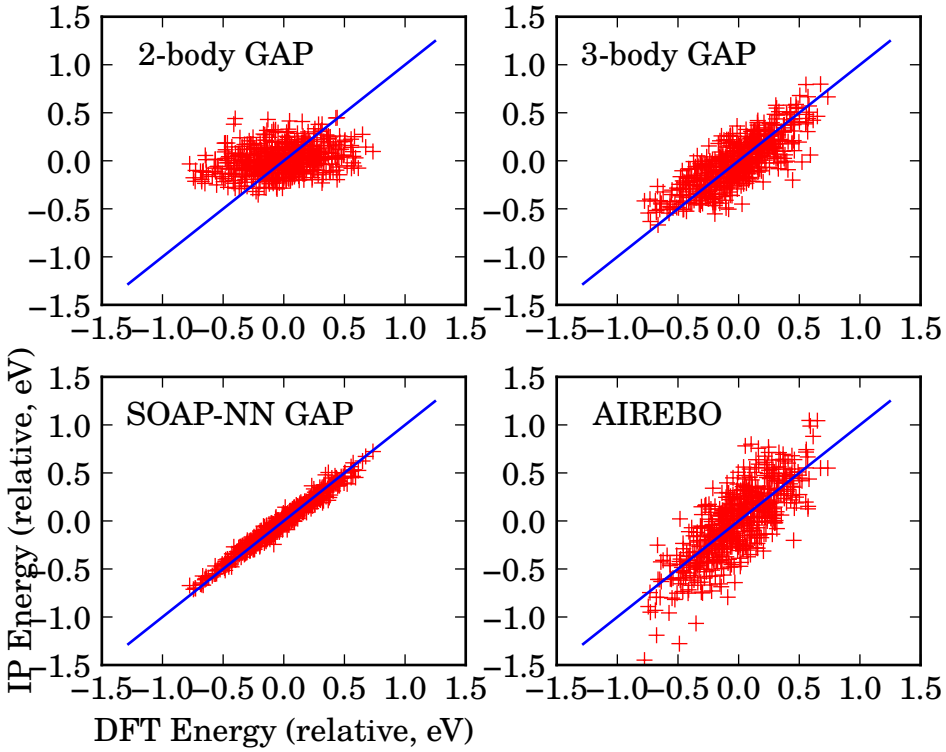


Figure 4.1: A comparison of GAP *models and the* AIREBO *potential for cold saturated hydrocarbons. The data points plotted here have been subsampled by a factor of 10 for legibility. Each set of energies has also had its mean subtracted so that it is centred about zero. The blue lines represent the target, $y = x$.*

	Energy error / (meV per atom)				
Test set	Prior (DFT)	2-body	3-body	SOAP-NN	AIREBO
Cold saturated	3.77	3.80	2.30	0.638	3.29
Cold unsaturated	5.52	5.72	3.36	1.52	52.7

Table 4.1: Errors of the GAP *models and of* AIREBO *on the two cold test sets. The prior error is just the standard deviation of the total energies in the test set. The other errors are the root-mean-square (*RMS*) differences between the predicted and the* DFT *total energies, where each set first had its mean subtracted.*

more successful. Its performance is comparable to that of the AIREBO model, even though AIREBO is a true many-body potential and should have more predictive capability than a model that includes only two- and three-body terms.

Finally, the model based on SOAP performs the best out of all of these. To compare its performance with the locality results in Chapter 3, the RMS error was computed on the SOAP-predicted forces as well. The resulting 37.9 meV/Å is much lower than the DFT force error at 2.5 Å from Figure 3.2. This surprisingly low error is probably because the cold training set on which this potential was evaluated contains a much smaller range of configurations than the ones accessible to the locality algorithm, which was run with an initial temperature of 4000 K. For a better comparison, the SOAP-NN model was trained and evaluated on the hot saturated test set (initial temperature 3000 K) as well. The resulting force error was 124 meV/Å, much closer to the order of magnitude (0.1 eV/Å) suggested by Figure 3.2, which indicates that the SOAP-NN potential approaches the limit of accuracy available to a local saturated hydrocarbon potential with a cutoff of 2.5 Å.

4.3.2 Unsaturated hydrocarbons

The potentials tested above were also trained on the unsaturated test sets. Since, as Figure 3.3 shows, the locality of the DFT forces within 2.5 Å is not significantly affected by conjugation mutations, a GAP trained on unsaturated hydrocarbons should be able to achieve similar performance to the one for saturated hydrocarbons. Even the AIREBO potential is formulated, with the inclusion of bond order parameters, so that it should be able to model unsaturated hydrocarbons as well.

The results in Figure 4.2 are the predicted versus reference energies for the cold unsaturated test set. The RMS errors on this test set are also shown in

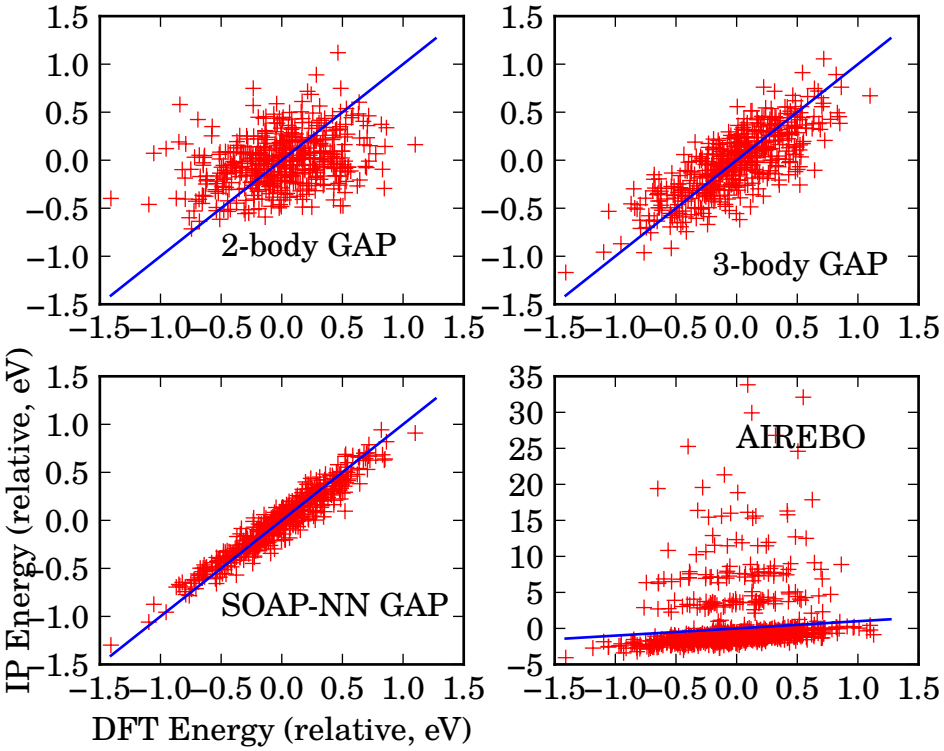


Figure 4.2: A comparison of potentials on cold unsaturated hydrocarbons. The AIREBO *plot was only subsampled by a factor of 5, the others by a factor of 10.*

Table 4.1. The two- and three-body GAP models each captured about the same proportion of variation in the training set, while the SOAP model fared proportionally worse. The most surprising change, however, is the degradation in performance of the AIREBO potential. As Figure 4.2 shows, AIREBO segments the training data into different classes, each one shifted by a different amount from the reference DFT energies. No pattern has yet been found in the training data that could explain this segmentation.

4.3.3 Range dependence

All of the GAP models trained so far are *intra*molecular potentials due to their short ranges, which only extend to nearest neighbours. A complete description of hydrocarbons must also account for the long-range *inter*molecular forces, most importantly dispersion, that help determine the properties of hydrocarbons in the bulk.

A naïve approach to capturing these forces would be to simply increase the size of the local neighbourhoods used to train the potentials. However, the computational cost and overall difficulty of fitting a GAP explodes with an increase in the range of interactions considered in the fit, since increasing the radius of the local environments combinatorially increases the number of configurations that can fit in that environment. This difficulty can be circumvented in a number of ways, for example by explicity parameterizing the nonlocal interactions as most existing forcefields do.

Another way to work around the combinatorial explosion is to use simple descriptors of the local environment with more reasonable scaling. The simplest of these is the two-body distance descriptor. Since it only consists of one number for each pair of atoms within a certain distance, it is effectively one-dimensional, so increasing the range of the descriptor only linearly increases the size of the space that the GAP has to describe. Increasing the range does increase the number of atom pairs in any given system, but the sparsification step keeps the corresponding rise in computational cost under control.

To see how well a two-body GAP model would perform with a longer range, two additional models were trained with ranges (cosine-smoothed two-body pair cutoffs) of 3.8 Å and 4.9 Å, corresponding roughly to second-nearest-neighbour (hereafter referred to as '2NN') and third-nearest-neighbour ('3NN') interac-

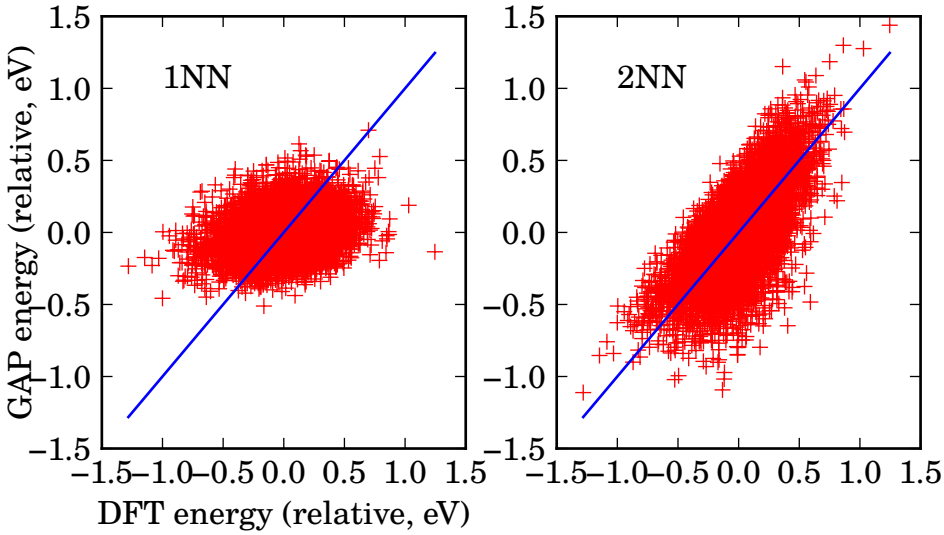


Figure 4.3: A comparison of two-body GAP *models with ranges of* 1.8 Å *(left) and* 3.8 Å *(right). Neither plot was subsampled.*

Model	Energy error / (meV per atom)	Force error / (meV/Å)
1NN	3.80	405
2NN	3.29	188
3NN	3.48	195
AIREBO	3.29	518

Table 4.2: Errors of two-body GAP *models, with various ranges, on the cold saturated test set. The original nearest-neighbour two-body model ('1NN') and* AIREBO *are included for comparison.*

tions. The models used a squared exponential ARD kernel with $\theta_{\text{rel}} = 0.2$ and 100 sparse points as before; they were trained on the cold saturated hydrocarbons only.

The qualitative improvement of the long-ranged potentials over the nearest-neighbour version is shown in Figure 4.3. The alignment of the 2NN energies with the DFT energies shows that this model is at least capturing some of the variation of the data.

The quantitative improvement of the long-range potentials, on the other hand, is less clear. Table 4.2 lists both force and energy errors of the two-body GAP mod-

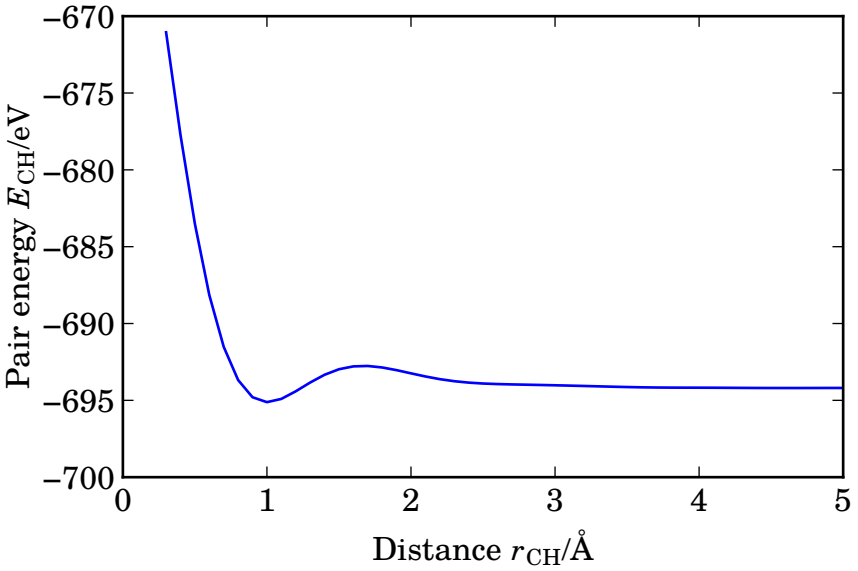


Figure 4.4: The carbon-hydrogen pair potential predicted by the two-body 3NN GAP model

els. It shows that the RMS energy errors were not reduced by much compared to that of the nearest-neighbour potential. The forces, on the other hand, improved significantly. Incidentally, both long-range potentials have comparable energy errors and lower force errors than the AIREBO model, although the three-body and SOAP models in Section 4.3.1 still offer a better qualitative improvement over AIREBO than these long-range models.

Two-body potentials such as the GAP models trained here offer the additional benefit that they can be visualized easily. The carbon-hydrogen pair potential from the 3NN GAP model is show in Figure 4.4. The potential does exhibit bonding behaviour, with a minimum at approximately the right location (the C-H bond length in alkanes is usually taken to be about 1.1 Å [16, 29]). However, it appears to be repulsive at medium range, possibly as a result of exchange repulsion between different parts of the molecule. At long range, there is no attractive force. This is to be expected from a potential derived from DFT data, since most DFT functionals totally lack long-range dispersion [12, 23].

4.4 Further development

Although the intramolecular GAP models for hydrocarbons show promise, much work remains to describe intermolecular forces accurately and so obtain a useful model for bulk hydrocarbons. These interactions can be captured by two main strategies: The one used by most existing potentials is to parameterize the interactions and find parameters either by fitting to some reference, such as experimental data, or by deriving quantities from electronic structure calculations (see Section 1.1.2). The other strategy is to use a nonparametric method such as GAP to fit the interactions directly to some reference.

The parametric method is the easiest short-term strategy, as the parameters and functional forms used in existing forcefields can simply be incorporated into an existing intramolecular GAP model. The GAP could be used, for instance, to fit the difference between the intramolecular force field and the DFT data, using a relatively short range in order to ignore the DFT description of intermolecular forces. This method could encounter problems, however, if the optimal parameters of the intermolecular force field are very sensitive to the intramolecular force field used. If that is the case, the intermolecular force field may need to be reparameterized and fit in conjunction with the intramolecular GAP model.

The nonparametric method, in contrast, may be better at describing intermolecular forces due to such a potential's more general functional form. However, the quality of a nonparametric potential still depends strongly on the quality of the data used to fit it. Since standard DFT is notoriously bad at capturing dispersion [12, 22, 23], a higher-quality reference is needed. One could use a DFT functional with dispersion corrections, but since those corrections are usually done with fixed, parametric functional forms of the same type used in interatomic potentials, there is no point in fitting a nonparametric model to such calculations.

The most sensible reference for nonparametric fitting of intermolecular forces is thus an adequate level of quantum chemistry. This means reference data will be difficult and expensive to generate. Preliminary work on the methane dimer indicates that calculations of at least the coupled-cluster level will be necessary to describe dispersion in alkanes. Such calculations are still limited in practice to small systems (a few tens of atoms, or perhaps hundreds with the aid of large supercomputers[3], for single calculations, although new linear scaling algorithms are beginning to make larger systems feasible [28]) so the reference data will be limited to small model systems such as dimers or small-molecule clusters.

An additional challenge is to find descriptors of the molecular system that capture the information necessary to describe long-range interactions without suffering from the steep increase in complexity with increasing interaction range that affects local machine learning potentials. The two- and three-body potentials described earlier are good candidates, as they have inherently low dimensionalities regardless of the interaction range. They should work especially well if the long-range interactions can easily be decomposed into two- and three-body terms, as is the case for dispersion and electrostatics in many molecular systems [11]. Another possible descriptor is SOAP with the width of the Gaussians that represent the atoms set very large, in effect changing the length scale of the information that the descriptor captures. Such a descriptor could be useful for capturing long-ranged many-body effects.

In summary, while the parametric fitting method takes advantage of the physics of the dispersion and electrostatic interactions involved to arrive at a compact, inexpensive force model, the nonparametric fitting method trades this efficiency for generality, allowing it to capture what the approximations in the parametric method may have left out. The parametric, physics-aware strategy

[3]See, for example, `http://www.nwchem-sw.org/index.php/Benchmarks`

has been used successfully in nearly all interatomic potentials, so it may well be accurate enough for the class of problems the new GAP is designed to address. In the long term, however, a nonparametric fit of the intermolecular forces may turn out to be a better, more flexible, and more accurate strategy.

Chapter 5

Conclusions

The locality algorithm in this work was developed to provide a rigorous, quantitative estimate of the range of the forces in a system. This algorithm can help us not only understand the physics of a molecule, but since it works with the force directly, it can help us see why existing interatomic potentials work and how we can design better potentials in the future. The algorithm was tested on linear hydrocarbon systems with various chemical mutations. It was found that the tight-binding model predicted strictly localized forces as expected, while the DFT forces were much less local (by about two orders of magnitude at a distance of 10 Å). The DFT forces were even less local when conjugation and especially oxygenation mutations were considered, with oxygenation worsening the error by about one order of magnitude at 10 Å. The effects of these mutations agree broadly with chemical intuition about the types of forces in a molecule.

In the future, the algorithm could be used on many other types of systems. For instance, it might be applied to the free motion of separate molecules outside the local neighbourhoods in order to probe truly *inter*molecular forces. Another potentially fruitful study would be to mutate the molecules before freezing the local environments so as to investigate the forces on different types of atoms.

Rather than carrying out such extended studies of the locality algorithm itself, this work turned to an application of the results already available. These results were used to inform the design of a family of intramolecular potentials, based on GAP, for linear hydrocarbons. The new potentials reproduced the DFT training data as good as or better than an existing potential, AIREBO, and approached the maximum accuracy available with local nearest-neighbour potentials. If they are combined with a suitable intermolecular force model, these new potentials offer the promise of tackling problems in hydrocarbon simulation that were previously hindered by the lack of sufficiently accurate potentials. If this approach is successful, it would enable the creation of potentials for a wide variety of related organic compounds with unprecedented accuracy.

Appendix A

Technical Notes

The code mentioned in Section 2.1.1 was written in Python version 2.7.9[1] using components from the SciPy ecosystem[2], including the libraries NumPy 1.9.2, SciPy 0.15.1, the Matplotlib [55] plotting package (version 1.4.3), and IPython [56] version 3.2.0 for interactive work and data analysis. The code interfaced to QUIP [29] (Git revision `5b152d5` with minor modifications) using the included `quippy` wrapper, which also included an interface to the MOLPRO (version 2012.1) computational chemistry package. The initial molecular configurations for the dynamics and the molecular graphics in this document were produced using Avogadro [57].

This project also made use of the GAP code distributed for use with QUIP; more details are in Section 4.2.3. The analysis and plotting of the GAP models also made use of IPython and Matplotlib.

[1] `http://python.org`
[2] `http://scipy.org/`

Bibliography

[1] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999), URL http://link.aps.org/doi/10.1103/RevModPhys.71.1085.

[2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010), URL http://link.aps.org/doi/10.1103/PhysRevLett.104.136403.

[3] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007), URL http://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

[4] R. M. Martin, *Electronic structure : basic theory and practical methods* (Cambridge University Press, Cambridge, UK, 2008), 1st ed., ISBN 9780521534406.

[5] S. Ismail-Beigi and T. A. Arias, Phys. Rev. Lett. **82**, 2127 (1999), URL http://link.aps.org/doi/10.1103/PhysRevLett.82.2127.

[6] H. Chen and C. Ortner (2015), arXiv:1505.05541, URL http://arxiv.org/abs/1505.05541.

[7] D. W. Brenner, Phys. Rev. B **42**, 9458 (1990), URL http://link.aps.org/doi/10.1103/PhysRevB.42.9458.

[8] M. I. Baskes, Phys. Rev. B **46**, 2727 (1992), URL http://link.aps.org/doi/10.1103/PhysRevB.46.2727.

[9] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013), URL http://link.aps.org/doi/10.1103/PhysRevB.87.184115.

[10] J. Behler, J. Chem. Phys. **134**, 074106 (2011), URL http://scitation.aip.org/content/aip/journal/jcp/134/7/10.1063/1.3553717.

[11] A. J. Stone, *The theory of intermolecular forces* (Oxford University Press, Oxford, 2013), 2nd ed., ISBN 9780199672394.

[12] S. Grimme, J. Comput. Chem. **27**, 1787 (2006), URL http://dx.doi.org/10.1002/jcc.20495.

[13] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995), URL http://dx.doi.org/10.1021/ja00124a002.

[14] W. L. Jorgensen, J. D. Madura, and C. J. Swenson, J. Am. Chem. Soc. **106**, 6638 (1984), URL http://dx.doi.org/10.1021/ja00334a030.

[15] M. G. Martin and J. I. Siepmann, J. Phys. Chem. B **102**, 2569 (1998), URL http://pubs.acs.org/doi/abs/10.1021/jp972543%2B.

[16] H. Sun, J. Phys. Chem. B **102**, 7338 (1998), URL http://dx.doi.org/10.1021/jp980939v.

[17] M. Born and J. E. Mayer, Z. Phys. **75**, 1 (1932), URL http://link.springer.com/10.1007/BF01340511.

[18] R. A. Buckingham and J. Corner, Proc. R. Soc. Lond. A **189**, 118 (1947), URL http://rspa.royalsocietypublishing.org/content/189/1016/118.

[19] N. L. Allinger, Y. H. Yuh, and J. H. Lii, J. Am. Chem. Soc. **111**, 8551 (1989), URL http://dx.doi.org/10.1021/ja00205a001.

[20] D. J. Tobias, K. Tu, and M. L. Klein, J. Chim. Phys. **94**, 1482 (1997), URL http://cat.inist.fr/?aModele=afficheN&cpsidt=2819547.

[21] A. D. Becke and E. R. Johnson, J. Chem. Phys. **124**, 14104 (2006), URL http://dx.doi.org/10.1063/1.2139668.

[22] A. Olasz, K. Vanommeslaeghe, A. Krishtal, T. Veszprémi, C. Van Alsenoy, and P. Geerlings, J. Chem. Phys. **127**, 224105 (2007), URL http://dx.doi.org/10.1063/1.2805391.

[23] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009), URL http://link.aps.org/doi/10.1103/PhysRevLett.102.073005.

[24] T. Frauenheim, G. Seifert, M. Elstner, T. Niehaus, C. Köhler, M. Amkreutz, M. Sternberg, Z. Hajnal, A. D. Carlo, and S. Suhai, J. Phys. Condens. Matter **14**, 3015 (2002), URL http://stacks.iop.org/0953-8984/14/i=11/a=313.

[25] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998), URL http://link.aps.org/doi/10.1103/PhysRevB.58.7260.

[26] M. Elstner, Theor. Chem. Acc. **116**, 316 (2006), URL http://dx.doi.org/10.1007/s00214-005-0066-0.

[27] T. Veszprémi and M. Fehér, *Quantum chemistry : fundamentals to applications* (Kluwer Academic/Plenum, Dordrecht, 1999), ISBN 0306461641.

[28] C. Riplinger and F. Neese, J. Chem. Phys. **138**, 034106 (2013), URL http://scitation.aip.org/content/aip/journal/jcp/138/3/10.1063/1.4773581.

[29] A. P. Bartók, G. Csányi, A. Nichol, L. Mones, W. Szlachta, J. Kermode, A. De Vita, N. Bernstein, and L. Pastewka, *libAtoms+QUIP* (2015), URL http://libatoms.org.

[30] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, WIREs Comput Mol Sci **2**, 242 (2012).

[31] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, et al., *MOLPRO, version 2012.1, a package of ab initio programs* (2012).

[32] M. Schütz, R. Lindh, and H.-J. Werner, Mol. Phys. **96**, 719 (1999), URL http://dx.doi.org/10.1080/00268979909483008.

[33] R. Lindh, Theor. Chim. Acta **85**, 423 (1993), URL http://dx.doi.org/10.1007/BF01112982.

[34] R. S. Mulliken, J. Chem. Phys. **23**, 1833 (1955), URL http://scitation.aip.org/content/aip/journal/jcp/23/10/10.1063/1.1740588.

[35] R. F. W. Bader, *Atoms in molecules : a quantum theory* (Clarendon, Oxford, 1990), ISBN 0198551681.

[36] C. Fonseca Guerra, J.-W. Handgraaf, E. J. Baerends, and F. M. Bickelhaupt, J. Comput. Chem. **25**, 189 (2004), URL http://dx.doi.org/10.1002/jcc.10351.

[37] F. L. Hirshfeld, Theor. Chim. Acta **44**, 129 (1977), URL http://dx.doi.org/10.1007/BF00549096.

[38] U. C. Singh and P. A. Kollman, J. Comput. Chem. **5**, 129 (1984), URL http://doi.wiley.com/10.1002/jcc.540050204.

[39] T. A. Manz and D. S. Sholl, J. Chem. Theory Comput. **8**, 2844 (2012), URL http://dx.doi.org/10.1021/ct3002199.

[40] G. Henkelman, A. Arnaldsson, and H. Jónsson, Comput. Mater. Sci. **36**, 354 (2006), URL http://www.sciencedirect.com/science/article/pii/S0927025605001849.

[41] E. Sanville, S. D. Kenny, R. Smith, and G. Henkelman, J. Comput. Chem. **28**, 899 (2007), URL http://dx.doi.org/10.1002/jcc.20575.

[42] W. Tang, E. Sanville, and G. Henkelman, J. Phys. Condens. Matter **21**, 84204 (2009), URL http://stacks.iop.org/0953-8984/21/i=8/a=084204.

[43] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, Phys. Rev. B **92**, 045131 (2015), URL http://link.aps.org/doi/10.1103/PhysRevB.92.045131.

[44] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011), URL http://link.aps.org/doi/10.1103/PhysRevB.83.153101.

[45] P. E. Maslen, C. Ochsenfeld, C. A. White, M. S. Lee, and M. Head-Gordon, J. Phys. Chem. A **102**, 2215 (1998), URL http://dx.doi.org/10.1021/jp972919j.

[46] R. S. Payal, S. Balasubramanian, I. Rudra, K. Tandon, I. Mahlke, D. Doyle, and R. Cracknell, Mol. Simul. **38**, 1234 (2012), URL http://dx.doi.org/10.1080/08927022.2012.702423.

[47] W. Allen and R. L. Rowley, J. Chem. Phys. **106**, 10273 (1997), URL http://scitation.aip.org/content/aip/journal/jcp/106/24/10.1063/1.474052.

[48] G. Kaminski, E. M. Duffy, T. Matsui, and W. L. Jorgensen, J. Phys. Chem. **98**, 13077 (1994), URL http://dx.doi.org/10.1021/j100100a043.

[49] S. Nouranian, M. A. Tschopp, S. R. Gwaltney, M. I. Baskes, and M. F. Horstemeyer, Phys. Chem. Chem. Phys. **16**, 6233 (2014), ISSN 1463-9084, URL http://pubs.rsc.org/EN/content/articlehtml/2014/cp/c4cp00027g.

[50] S. J. Stuart, A. B. Tutein, and J. A. Harrison, J. Chem. Phys. **112**, 6472 (2000), URL http://scitation.aip.org/content/aip/journal/jcp/112/14/10.1063/1.481208.

[51] S. Plimpton, J. Comput. Phys. **117**, 1 (1995), ISSN 00219991, URL http://www.sciencedirect.com/science/article/pii/S002199918571039X.

[52] D. J. C. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, 2003), ISBN 9780521642989.

[53] W. J. Szlachta, A. P. Bartók, and G. Csányi, Phys. Rev. B **90**, 104108 (2014), URL http://link.aps.org/doi/10.1103/PhysRevB.90.104108.

[54] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015), URL http://doi.wiley.com/10.1002/qua.24927.

[55] J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007).

[56] F. Pérez and B. E. Granger, Comput. Sci. Eng. **9**, 21 (2007), URL http://ipython.org.

[57] *Avogadro: an open-source molecular builder and visualization tool. Version 1.1.1*, URL http://avogadro.openmolecules.net/.