Quantum Chemical Studies of Four-center Iron CO_2 Reduction Electrocatalyst and Multifaceted Development of a Small Molecule Force Field

By

Hyesu Jang
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Lee-Ping Wang, Chair

Alexei Stuchebrukhov

Davide Donadio

Committee in Charge

2021

Table of Contents

Abstract

The development of carbon dioxide (CO_2) reduction electrocatalysts is an intensively studied area in the development of CO_2 capture, utilization, and storage strategies. In this work, density functional theory (DFT) and *ab initio* molecular dynamics (AIMD) are employed to study redox properties and the pathway of a side reaction of CO_2 reduction electrocatalyst $[Fe_4N(CO)_{12}]^-$. The material of this chapter was published as an article titled "Quantum chemical studies of redox properties and conformational changes of a four-center iron CO_2 reduction electrocatalyst" in Chemical Science (2018). I also present the multifaceted development of a molecular mechanics force field, an important piece of molecular mechanics simulations that is widely used for molecular structure and property prediction. An open-source python package for restrained electrostatic potential (RESP), *respyte* is implemented for developing improved electrostatic potentials (ESP) based charge models for a better description of electrostatic interactions in force fields. I describe further improvements in Open Force Field small molecule force field after the first optimized small molecule force field from Open Force Field Initiative, which includes modification of parameter set for improved performance in certain chemical spaces, and a more careful design of quantum mechanics (QM) training data used to re-optimize bonded parameters.

Acknowledgements

I would like to dedicate this work to my family — especially my parents, Heebok Yoo (유희복) and Jaehwa Jang (장재화); and my beloved aunt, Heewon Yoo (유희원) — and my friends who have believed in me even when I didn't believe in myself throughout my Ph.D. journey.

To Lee-Ping Wang, my Ph.D. adviser, I appreciate your academic guidance and advice over the course of my Ph.D. studies. I truly enjoyed being your student for the past five years and you've been an ideal academic advisor for me.

I am so thankful for the current and former members of the Wang group. I want to thank Yudong Qiu for being a good mentor and my academic role model, Nanhao Chen and Marshall Hutchings for helping me a lot especially when I first joined the group, Lisa Oh for not just being a good lab mate, but a true friend.

Chapter 2 is a reprint of the published work with a small modification: Hyesu Jang, Yudong Qiu, Marshall E. Hutchings, Minh Nguyen, Louise A. Berben, Lee-Ping Wang, "Quantum chemical studies of redox properties and conformational changes of a four-center iron CO_2 reduction electrocatalyst", Chem. Sci., 2018, 9, 2645-2654. I acknowledge financial support from the ACS Petroleum Research Fund, award number 58158-DNI6. I am also thankful to my co-authors for all their help.

For chapter 3, I acknowledge Christopher Bayly for his scientific advice and his initial work on the implementation of RESP in Python. I am also thankful to Paul Nerenberg and his group members, especially Charles Metzler-Winslow, for the fun collaboration and valuable feedback on my work.

For chapter 4, I acknowledge Open Force Field Consortium for giving me a chance to be a part of the wonderful collaboration and the financial support. I am so thankful

to Yudong Qiu, Christopher Bayly, David Mobley, Jeffrey Wagner, Jessica Maat, and all current and former members of the consortium.

Above all, I thank God for everything he has given me.

1 Introduction

1.1 Grand Challenges in Chemistry[1]

There are several grand challenges in chemistry that the current and the next generation of chemists should work on to make the world a better place. One of the challenges is mitigating the atmospheric concentration of carbon dioxide (CO_2). Global warming has been accelerated as the greenhouse gases including CO_2 have accumulated in Earth's atmosphere. Since the largest source of anthropogenic CO_2 emission is from fossil fuel combustion, much research has been conducted to decrease the dependence on fossil fuels by transitioning to renewable energy sources. However, the transition is not feasible yet due to the high capital cost of renewable energy technologies and lack of energy storage technologies. For solar energy, for example, improvement in silicon purification methods, reduction of production cost of silicon photovoltaic module, and minimization of the environmental impact of the production process should be achieved to make it competitive. Another approach to actively reducing atmospheric CO_2 concentration is developing technologies for effective CO_2 capture, utilization, and storage (CCUS). CO_2 utilization techniques can provide a possibility to produce useful chemical products from CO_2, and electrochemical reduction of CO_2 is one of the techniques that have been studied intensively. For designing catalysts that can efficiently reduce CO_2 to specific desired products, understanding their catalytic reaction mechanisms is essential, which involves a combined experimental and theoretical effort.

In addition to protecting ecological health, chemistry plays an important role in the development of new therapeutics for protecting human health. Another grand challenge associated with this aspect is a fast and accurate prediction of structures and properties of molecules not yet synthesized. The crystal structure of a drug de-

termines its properties, such as stability, solubility, and rate of dissolution. Therefore, an accurate prediction of crystal structures of drug molecules is crucial in the pharmaceutical industry. while the prediction over small rigid molecules has been rapidly improved in recent years, it is still considered extremely challenging to accurately predict more complex molecules. Accurate evaluation of the binding affinity of the drug is also crucial for computer-aid drug discovery. While there are a number of different techniques for affinity evaluation have been developed, they all have their specific applications and limitations.[2] Also, protein structure prediction is a crucial challenge in drug developments, by taking into account the fact that there is a huge gap between the number of known protein sequences and the number of corresponding structures revealed. Molecular structure and property prediction largely rely on molecular mechanics. Therefore, improving the accuracy of molecular mechanics simulations will lead to a more reliable prediction of molecular structure and properties of interest.

1.2 Overview of Computational Chemistry Methods[3]

Theoretical chemistry is the subfield of chemistry where mathematical methods and fundamental laws of physics are applied to study the structure and dynamics of chemical systems. Given a chemical system of interest, theoretical chemistry can attempt to compute: (1) The geometrical arrangements of the nuclei corresponding to stable molecules; (2) the molecular properties such as energies, dipole and quadrupole moment, dipole polarizability, NMR spin-spin coupling constants; (3) the rate of transformation of one stable molecule to another; (4) the time evolution of molecular structures and properties, etc.

The majority of the systems of interest consist of more than two electrons. The

fundamental issue is that only one or two-body systems have exact solutions. Therefore, numerical methods should be applied to solve many-body problems, which require a great number of mathematical operations. The development and widespread of electronic computers enabled the treatment of the many-body systems with high speed and high accuracy, with the advent of a new field in chemistry, computational chemistry.

Decades of research in theoretical and computational chemistry have produced a diverse library of methods for modeling the potential energy surface of a molecule. Two principal categories of such methods are electronic structure method (quantum mechanical method) and Force Field method (molecular mechanics method). Both are models of the molecular energy and properties as a function of the nuclear coordinates. The main differences between the two methods are how they describe a chemical system and interactions between particles. While electronic structure methods choose atomic nuclei and electrons as fundamental units, the force field method chooses atom as a building block, i.e., electrons are not treated explicitly, instead, their effects are taken into account implicitly.

1.2.1 Electronic Structure Methods (Quantum Mechanics)

In electronic structure methods, the electronic structure (the wave function of the electrons and its corresponding energy) in a chemical system is determined by solving the Schrodinger equation associated with the electronic molecular Hamiltonian (H_{el}), which is a sum of kinetic and potential energies of the nuclei and electrons:

$$H_{el} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{A,i}^{M} \frac{Z_A}{r_{Ai}} + \sum_{i<j} \frac{1}{r_{ij}} \tag{1}$$

where Z_A is atomic number of nuclear A, r_{Ai} is a distance between electron i and nuclear A. Here the atomic system is used, i.e. ($\hbar = e = m_e = 1$). Since there is no exact solution to the electronic Schrodinger equation for many-electron systems, we have to rely on approximations and numerical methods for the description of many-electron systems.

One of the most basic approximate electronic structure methods is Hartree-Fock (HF) method. It applies the variational method, which states that the energy of any approximate wave function is an upper bound to the exact energy. It chooses a trial wave function with parameters and finds the best trial wave function that gives the lowest energy possible, by minimizing the expectation value of Hamiltonian (the energy of a wave function is the expectation value of the Hamiltonian operator, divided by the norm of the wave function.) with respect to the parameters. Due to the Pauli exclusion principle, which states two or more electrons cannot be in the same quantum state within a quantum system at the same time, the electronic wave function must be antisymmetric with respect to the exchange of any two-electron coordinates. And in quantum mechanics, the simplest antisymmetric wave function satisfying the Pauli principle is expressed as Slater determinants. For the general case of N electrons and N spin orbitals, a Slater determinant is given as

$$\Psi(x_1, x_2, ..., x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_i(x_1) & \chi_j(x_1) & ... & \chi_k(x_1) \\ \chi_i(x_2) & \chi_j(x_2) & ... & \chi_k(x_2) \\ ... & ... & & ... \\ \chi_i(x_N) & \chi_j(x_N) & ... & \chi_k(x_N) \end{vmatrix} \tag{2}$$

The approximation in HF that the exact wave function can be approximated by a single Slater determinant, introduces the neglect of electron correlation effect, which

is electron-electron interaction in the electronic structure of a quantum system. (It only accounts for the electron-electron interactions in an average fashion.) Therefore, post-HF methods such as configuration interaction (CI)[4], coupled-cluster (CC)[5], and Moller-Plesset(MP) perturbation theory[6] have been developed to include the electron correlation effect.

CI utilizes a linear combination of configuration state functions, which are linear combinations of Slater determinants, to describe the wave function. The first term of the CI wavefunction is the HF ground-state and the higher terms are some electronically excited states, which are to account for the electron correlation effects. If all possibilities of excited configurations are included, termed Full-CI, the numerically exact solution to the electronic Schrodinger equation can be achieved, which is in practice impossible to compute. MP perturbation theory and CC treat electron correlation by using Rayleigh-Schrodinger perturbation theory to second, third, or fourth-order (MP2, MP3, MP4) and by constructing a wavefunction using an exponential cluster operator, respectively. The details of the methods are intentionally omitted because they are out of scope for the present work.

Another widely used electronic structure method is density functional theory (DFT). While post-HF methods give a systematic approach to the exact solutions, they are expensive and not tractable for certain systems. DFT is known to give a good combination of accuracy and computational cost. The method is based on the Hohenberg-Kohn theorem, whose first theorem states that the ground-state electron density contains the complete ground state properties of a many-electron system. Electron density is a much simpler quantity than wave function because while wave function has 3N spatial variables for an N-electron system and its complexity increases exponentially, the electron density is a function in 3d space and has constant complexity with the number of electrons. Thus, if we could extract the ground state

properties from the ground state density, we could potentially have a faster route to solving quantum mechanics (QM) problems. The second H-K theorem is about "how we get the energy from the density", which states that given an approximate density that originates from an antisymmetric N-electron wave function the energy given by this density is an upper bound to the exact ground state energy. The problem here is the exact functional connecting electron density and the ground state energy is not known. Therefore, the fundamental goal of DFT method development has been designing functionals connecting the electron density with the energy.

Early attempts to design DFT models tried to formulate all the energy terms as a functional of the electron density alone, but the orbital-free models had the main problem which is the poor representation of the kinetic energy. To solve the issue in the orbital-free models, Kohn-Sham (KS) theory[7] introduced orbitals. The idea in the KS theory is to split the kinetic energy term into two parts, kinetic energy of non-interacting electrons and a small correction term. While the electrons are interacting in reality, the kinetic energy of non-interacting electrons provides 99 % of the exact kinetic energy. Therefore, the errors from inaccuracies in the functionals are much smaller in KS DFT models compared to orbital-free DFT models, thus KS theory has been the basis of modern DFT.

While the KS-DFT method is comparable to HF in terms of its numerical methods and computational cost, it gives much more reliable results. Therefore, it has become an important tool in many areas of chemistry and material science. The main downsides of the DFT method are the absence of a systematic way to improve the results towards the exact solution and the inaccuracy in describing intermolecular interactions, such as dispersion interactions.

There is a hierarchy of approximations to the exchange-correlation functional in DFT, often referred to as Jacob's ladder. The lowest rung of Jacob's ladder is the

local density approximation (LDA)[8]. In LDA, it is assumed that the local density can be treated as a homogeneous electron gas, and under the assumption, the exchange-correlation energy is only dependent on the electron density. The second rung is the Generalized gradient approximation (GGA), where the exchange-correlation energy is expressed not only by the density but also by its gradient.[9–12] With the additional information, GGAs give a significant improvement over LDA, and popular GGA functionals include BLYP[13,14] and PBE[15]. The third rung is meta-GGA, where higher-order derivatives of the electron density or orbital kinetic energy density are included in the exchange-correlation energy as a variable.[16–21] Fourth rung is hybrid-GGA, which mixes some exact HF exchange with a GGA.[22–25] The inclusion of a suitable fraction of exact HF exchange is found to improve the accuracy of calculated results, becoming a standard feature of DFT method development. Examples of hybrid-GGAs include PBE0[25] and B3LYP[22], one of the most popular DFT functionals. The fifth rung is generalized random phase approximation (RPA), which includes virtual orbitals for a better description of dispersion interactions.[26,27] Accuracy and computational cost tend to increase up the ladder.

1.2.2 Force Field Methods (Molecular Mechanics)

In force field methods, the atom is used as a fundamental unit for describing a chemical system, i.e., electrons are not treated explicitly in the methods, and molecules are described by a 'ball and spring' model, where atoms are represented as spheres with varying size and pairwise interactions, and bonds are springs with varying stiffness.

Solving the electronic Schrodinger equation to calculate electronic energy for a given nuclear configuration is very demanding and complex. For most molecular dynamics applications, it is not practical to directly use the electronic structure method to evaluate the potential energy at every time step. To reduce the computational cost

and complexity, a force field approximates the electronic energy as a parametric function of the nuclear coordinates. Many different functional forms of force field exist for modeling the potential energy surface. But in general, force fields are expressed as a sum of bonded and non-bonded interactions.

$$E_{\mathrm{FF}} = E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{dihedral}} + E_{\mathrm{vdW}} + E_{\mathrm{electrostatic}} \tag{3}$$

where the first three terms describe bonded interactions, representing bond stretching, angle bending, torsional interactions, respectively, and the last two terms describe non-bonded interactions, representing van der Waals (vdW) interaction and electrostatic interaction, respectively.

For bond stretching interaction, the properties of the chemical bond are defined in terms of the two bonded atom types. An *atom type* represents an atom of a particular element within its local chemical environment, for example, an *sp3* carbon. Given two bonded atoms, the force field provides parameters for the bonded energy based on the bonded atom types. In the simplest model, the bond stretching interaction is described as a simple harmonic motion.

$$E_{\mathrm{bond}}(r^{\mathrm{AB}}) = k^{\mathrm{AB}}(r^{\mathrm{AB}} - r_0^{\mathrm{AB}})^2 \tag{4}$$

where r_0^{AB} is the equilibrium bond length and k^{AB} is a force constant, a measure of the stiffness of the bond. Note that most force fields are not capable of describing bond dissociation.

The angle bending energy describes the energy change associated with bending an angle formed by three connected atoms A-B-C. Similar to the stretching interactions, it is usually given by a Taylor series around an equilibrium bond angle and truncated

at the second-order in its simplest form.

$$E_{\text{angle}}(\theta^{\text{ABC}}) = k^{\text{ABC}}(\theta^{\text{ABC}} - \theta_0^{\text{ABC}})^2 \tag{5}$$

where θ_0^{ABC} is the equilibrium bond angle and k^{ABC} is a force constant, a measure of the stiffness of the angle. As in the case of bonding, the parameters are defined in terms of the three bonded atom types involved in the interaction.

The torsional energy is the energy required for rotation around a B-C bond in four sequentially connected atoms A-B-C-D. Unlike the stretching and the bending energies, the torsional energy is not expanded as a Taylor series because the torsional angle can significantly deviate from the equilibrium angle. Instead, it is usually written as a Fourier series.

$$E_{\text{dihedral}} = \sum_{n=1} V_n\big(1 + \cos(n\theta - \text{phase}_n)\big) \tag{6}$$

where θ is dihedral angle, n is periodicity, V_n is a rotational barrier, phase_n is a phase shift. Because the torsional energy is intermediate between the bonding and non-bonding regimes, it is often the most heavily parameterized component of the whole force field.

Non-bonded terms consist of vdW interactions and electrostatic interactions. The vdW energy describes the non-polar part of the interactions between atoms that are not directly bonded. One of the widely used models is Lennard-Jones (LJ) potential, where repulsive interaction is proportional to r^{-12} and attractive interaction is proportional to r^{-6}:

$$E_{LJ}(r^{AB}) = \epsilon\left[\left(\frac{r_0}{r^{AB}}\right)^{12} - 2\left(\frac{r_0}{r^{AB}}\right)^{6}\right] \tag{7}$$

where r^{AB} is the distance between atom A and atom B, ϵ is the depth of the potential wall, r_0 is the distance where the potential energy is zero.

The electrostatic energy describes the Coulombic interaction between atoms with partial charges. Molecular charge distribution can be approximated by several different charge models and the simplest approximation is the simple point charge model, where the partial charges are assigned to each atom center. In the simple point charge model, the electrostatic interaction between atom A and atom B is given by:

$$E_{el}(r^{AB}) = \frac{q^A q^B}{\epsilon r^{AB}} \tag{8}$$

where ϵ is a dielectric constant and q^A and q^B are atomic partial charges of atom A and atom B respectively.

The main advantage of force field methods over quantum mechanics methods is the greatly reduced computational cost, which enables simulations of large systems, including biomolecular systems, with relatively long timescales (on the order of microseconds). On the other hand, the performance of the current generation of widely used force fields in predicting thermodynamic properties shows that there is still scope for improvement in the quality of parameters and the choice of functional forms.

1.3 Overview of my Ph.D. Research

As an effort in developing CO_2 reduction electrocatalyst, first half of my Ph.D. work focused on theoretical study of four-centered iron CO_2 reduction electrocatalysts $[Fe_4N(CO)_{12}]^-$. The complex is first found by Muetterties and coworkers[28,29], and in its resting state, it is found to be able to act as a selective electrocatalyst for CO_2 reduction to formate in aqueous solution.[30] After it undergoes two reduction events, slow CO dissociation from the cluster is observed experimentally, making the cluster

inactive. While understanding the mechanism of the side reaction is important to the overall strategy for designing a more robust molecular catalyst, the mechanism was not yet uncovered. To uncover the mechanism, we employed DFT and *ab initio* molecular dynamics (AIMD) to estimate the first two redox potentials of the complex and explore the pathway of its side reaction involving CO dissociation.[31]

The second half of my Ph.D. work focused on the multifaceted development of a small molecule force field. While the molecular structure and property prediction largely rely on molecular mechanics, the accuracy of molecular mechanics simulations is critically dependent on the quality of the molecular mechanics force field. Therefore, improvement in the molecular mechanics force field will lead to more reliable prediction of molecular structure and properties of interest.

Most molecular mechanics force fields use point charges to approximate the charge distribution around atoms in a system. Restrained electrostatic potential (RESP), where partial charges are fitted against quantum chemical electrostatic potentials with restraints, is the most widely used ESP-based charge model for determining atomic partial charges.[32] However, several not-yet-resolved challenges remain, including significant dependence of the fitted charges on many heuristic choices, indicating that there is still room for improvement in the method. As a first step to develop the next generation of the ESP-based charge model, we implemented an open-source Python package for RESP, *respyte*, a tool capable of exactly reproducing the original implementation and easily extensible for developing improved ESP-based charge models.

Developing a high-quality general force field for the application to the wide range of small organic molecules is challenging due to the vastness of the chemical space that must be covered. While Current efforts in improving force fields mostly involve in-house or commercial closed-source efforts, Open Force Field Consortium aims to

develop and release an automated, open infrastructure and open datasets for producing and benchmarking force fields, so that anyone can access and contribute to the force field development. In our previous work, we showed our initial progress toward the goals, the development of the Open Force Field toolkit, an open-source software package for the development and application of force fields, and the first application of the infrastructure to create a small molecule force field, OpenFF1.0.0, code-named Parsley.[33] The present work describes the further improvements in Parsley that have been made, which include modifications and additions of parameter definitions for improved performance in certain chemical spaces, and a more careful design of QM reference data used to refit bonded parameters of Parsley.

2 Quantum Chemical Studies of Redox Properties and Conformational Changes of a Four-center Iron CO_2 Reduction Electrocatalyst[31]

2.1 Introduction

Development of economically viable technologies for reducing CO_2 concentration in Earth's atmosphere is one of the global environmental problems that we must solve in the near future. One of the major research fields in modern chemistry is to develop CO_2 capture, utilization and storage strategies. Electrochemical CO_2 reduction has been studied as one of CO_2 utilization techniques, which can give us the possibility to produce useful products from CO_2.

The discovery of CO_2 reduction electrocatalysts represents a significant advance in CO_2 utilization.[34] Certain metallic electrodes have been reported to have a catalytic activity for carbon dioxide reduction; Hori reported the formation of hydrocarbons and alcohols in electrochemical reduction of carbon dioxide at copper electrodes in aqueous solution and discussed the reaction mechanism in 1989.[35] In recent years, several metal and metal dichalcogenide nanostructured catalysts with high surface area have been proposed as promising candidates for electrocatalysts for the CO_2 reduction.[36–41] In addition to the heterogeneous catalysts, a number of molecular catalysts have also been investigated for CO_2 reduction and reviewed in several papers.[42–44]

In 2011, Rail and Berben has found that an Earth-abundant metal complex, first described by Muetterties and coworkers[28,29] and denoted as $[Fe_4N(CO)_{12}]^-$ or $\mathbf{1}^-$ in its resting state, can act as a selective electrocatalyst for CO_2 reduction to formate in aqueous solution.[30] The preference of the catalyst for hydrogen evolution vs. CO_2 reduction can be adjusted by tuning the strength of the acid used as a proton donor.

An isoelectronic compound, $[Fe_4N(CO)_{12}]^{2-}$, was found to be a catalyst for hydrogen evolution only.[45] In more recent work, Taheri and Berben further characterized the CO_2 reduction mechanism and proposed the reduced hydride H-**1**$^-$ as a key reaction intermediate.[46] The hydricity, or hydride donor-ability of H-**1**$^-$ was proposed as a thermodynamic predictor of selectivity for hydrogen evolution or CO_2 reduction; a free energy window was proposed to explain the activity of **1**$^-$ for CO_2 reduction as opposed to its isoelectronic analogues.

$[Fe_4N(CO)_{12}]^-$ is experimentally known to undergo two reduction events. When **1**$^-$ is electrochemically reduced to **1**$^{3-}$, slow CO dissociation from the cluster is observed, resulting in the $[Fe_4N(CO)_{11}]^{3-}$ or **2**$^{3-}$ species; the catalytic activity of this species is unknown but presumed to be inactive. In a companion experimental work, the X-ray crystal structure of **2**$^{3-}$ is reported.[47] Simulations that uncover mechanisms of side-reactions are important to the overall strategy for designing molecular catalysts which are resistant to them. In this respect, this article describes the redox properties and CO dissociation pathway of this complex using computational quantum chemistry to complement the experimental findings and provide atomic-resolution insights.

The use of density functional theory (DFT) to study the electronic properties of metal carbonyl clusters has precedent in the literature. In particular, Schaefer and coworkers have produced a series of studies on the structures and metal-metal bonding of Iron carbonyls and their derivatives.[48–67] Several other groups also have carried out DFT studies for geometry optimization and vibrational frequency analysis of iron carbonyl complexes.[68–71] Presumably, the strong fields from the CO ligands promote a low-spin and single-reference electronic state, making DFT a qualitatively appropriate method for studying these otherwise daunting multi-center inorganic clusters. Likewise, the application of DFT and solvent models for calculating redox potentials is well established.[72–74] On the other hand, we are not aware of any theoretical stud-

ies that have investigated the redox properties and reactivity of **1**; the significant metal-metal bonding and variation of charge states in this cluster may pose significant challenges for the density functional approximation and solvent model. For this reason, it is vital to compare calculated observables with experimental data where available.

In this theoretical study, we characterize the structures and energetics of the series of redox states: $\mathbf{1}^{0}$, $\mathbf{1}^{-}$, $\mathbf{1}^{2-}$ and $\mathbf{1}^{3-}$, and provide mechanistic insight into the CO dissociation side-reaction: $\mathbf{1}^{3-} \rightarrow \mathbf{2}^{3-} + \mathrm{CO}$. Our calculations of the one-electron reduction potentials show close agreement with the experimentally measured values and provide some evidence that the BP86 density functional approximation[13,75] performs more accurately for this system than the hybrid B3LYP functional.[22] The dissociation pathway was found using high-temperature AIMD and relaxed to the minimum energy path to calculate the activation barrier.[76] The calculations predict a structure of $\mathbf{2}^{3-}$ in remarkable agreement with the X-ray crystal structure that was determined concurrently,[47] lending further confidence to the level of theory used in this study. We also compare the CO dissociation barrier height to the analogous reaction after only one reduction event: $\mathbf{1}^{2-} \rightarrow \mathbf{2}^{2-} + \mathrm{CO}$ and show that dissociation from this electronic state is energetically uphill, though the activation free energy of CO dissociation is similar in both states. Our usage of DFT approximations is checked using natural orbital occupation numbers from multireference complete active space self consistent field (CASSCF) calculations at key geometries.[77,78]

2.2 Computational Methods and Results

2.2.1 Redox Potential Calculations

We evaluated the relative free energies between the redox intermediates $\mathbf{1}^{0}$, $\mathbf{1}^{-}$, $\mathbf{1}^{2-}$ and $\mathbf{1}^{3-}$ using unrestricted Kohn-Sham DFT[79] with the implicit solvent environment, conductor-like screening model (COSMO)[80] for comparison to experimentally determined redox potentials. These calculations were carried out using the TeraChem software, which uses graphics processing units to accelerate the computation of the Coulomb and exchange matrices,[81–83] effective core potentials (ECPs)[84,85] and solvent response[86] that appear in the SCF calculation. A recently developed geometry optimization method using translation-rotation internal coordinates was employed to accelerate the energy minimization calculations.[87]

$$G = G^{\text{solv}} + H_{\text{SCF}} + \text{ZPE} + H_{\text{tr,rot,vib}} - TS_{\text{tr,rot,vib}} \tag{9}$$

$$\Delta G = G_{\text{red.}} - G_{\text{ox.}} = -FE^{0} \tag{10}$$

Geometry optimization was used to derive the self-consistent field (SCF) electronic energy together with the solvation free energy. Vibrational frequency calculations were used to derive the zero point energy and Gibbs free energy within the harmonic approximation. To calculate the relative redox potential, we took the differences of the free energies of redox pairs and subtracted the absolute potential of the reference electrode, which is 4.67 V for the saturated calomel electrode (SCE). This value is based on the absolute potential of the NHE which was determined by Reiss and Heller to be 4.43 V,[88] though this quantity is difficult to measure and values in the range of 4.2–4.7 V have been reported in the literature.[89]

We tested the dependence of results on the choice of DFT approximation by per-

forming calculations using three functionals: the BP86 gradient-corrected semilocal functional,[13,75] the B3LYP hybrid functional,[22] and the PBE0 hybrid functional.[25] Previous studies have noted that BP86 may perform more reliably than B3LYP in the study of the compounds in this paper.[90,91] We also investigated whether adding diffuse basis functions affects the calculation results, because previous gas-phase DFT suggest that diffuse basis functions are needed for the description of anions.[92–94] For light elements (H, C, N, O) we used the def2-TZVP triple-valence Gaussian basis set[95,96] with f and higher angular momentum functions removed, denoted as def2-TZVP(-f). For the iron atoms we used either the LANL08 or LANL08+ basis set / ECP combination,[97] which differ by the addition of a diffuse d angular momentum function in the latter. We further tested the effects of adding a minimal set of diffuse functions on light elements.[98] The combined basis sets are called def2-TZVP(-f)-LTZ, def2-TZVP(-f)-LTZ+, and ma-def2-TZVP(-f)-LTZ+ respectively. Finally, since the cyclic voltammetry experiments to measure the redox potentials were carried out in MeCN/H_2O (95:5) solvent, we also conducted the calculations employing the dielectric constants of water (78.4) and MeCN (37.5). From Table 1 and Table 2, we concluded that the system has minor dependences on the choice of basis set and solvent while the functional dependence is significant. The BP86 functional gave closer agreement with experiment (RMSE $<$ 0.2 V) than the B3LYP and PBE0 hybrid functionals (RMSE $>$ 0.4 V); the improved agreement is not due to shifting the absolute electrode potential, because the BP86 RMSE is still much lower than the other two functionals with the average gap subtracted out. Overall, the combination of the BP86 functional, the def2-TZVP(-f)-LTZ+ basis, and the dielectric constant of water yields good agreement with experimental data with a root-mean-squared error (RMSE) of 0.15 V. The relatively high accuracy of BP86 (compared to hybrid functionals) for this system is consistent with previously published DFT studies of 3d transition metal containing

complexes.[90] To check the possible higher spin multiplicities for each state of the catalyst, we also calculated energies of higher spin multiplicities for each redox state (triplet and quintet for even-electron systems, quartet and sextet for odd-electron systems) and found that increasing the spin multiplicity significantly increases the total energy by over 10 kcal mol^{-1}. From these findings we conclude that higher spin multiplicities do not participate in the redox chemistry and reaction pathways in this paper.

	def2-TZVP(-f)_LTZ+/ water (V)			
	$1^{0/1-}$	$1^{1-/2-}$	$1^{2-/3-}$	RMSE
exp	>0.2	-1.23	-1.60	
B3LYP	-0.13	-1.37	-2.07	0.42
PBE0	-0.26	-1.42	-2.19	0.50
BP86	**0.37**	**-1.16**	**-1.54**	**0.15**

Table 1: **Functional dependence of the redox potential calculation and comparison with the experimentally determined redox potentials.** The BP86 results are shown in bold as they are judged to be most reliable from the present data and literature precedent.

	BP86/ acetonitrile (V)			
	$1^{0/1-}$	$1^{1-/2-}$	$1^{2-/3-}$	RMSE
exp	>0.2	-1.23	-1.6	
def2-TZVP(-f)_LTZ	0.41	-1.30	-1.62	0.14
def2-TZVP(-f)_LTZ+	0.34	-1.22	-1.64	0.13
ma-def2-TZVP(-f)_LTZ+	0.39	-1.27	-1.55	0.15

Table 2: **Addition of diffuse functions and solvent dependences of the redox potential calculation.** The line shown in bold is identical to the bolded line in Table 1.

2.2.2 Computational Discovery of Dissociation Pathways by *ab initio* Molecular Dynamics

We used AIMD to explore the chemical and structural rearrangements of $Fe_4N(CO)_{12}$ in its different electronic states. In the Born-Oppenheimer MD (BOMD) framework, the motion of atoms is simulated by applying the nuclear gradients of the energy as classical forces to the atoms, then accelerating the atoms along the force vectors using Newton's second law. In order for the simulations to broadly sample the chemical space and discover many reaction pathways while keeping the computational cost affordable, accelerated sampling techniques are needed to cross over potential barriers more rapidly.[99–104] In this study, we simply ran unbiased AIMD at elevated temperatures to accelerate the sampling. A velocity Verlet integrator was used with a time step of 1.0 fs. A Langevin thermostat was used with the equilibrium temperature set to 1000 K and a collision frequency of 1.0 ps^{-1}. Several simulations were started from the energy-minimized structures of $\mathbf{1}^{2-}$ and $\mathbf{1}^{3-}$, as well as the protonated isoelectronic species, i.e. H-$\mathbf{1}^{1-}$ and H-$\mathbf{1}^{2-}$. These simulations used the B3LYP functional and a hybrid basis set combining 6-31G*[105] for light elements and the LANL2DZ basis set / ECP for Fe[106], abbreviated as 6-31G*-LDZ.

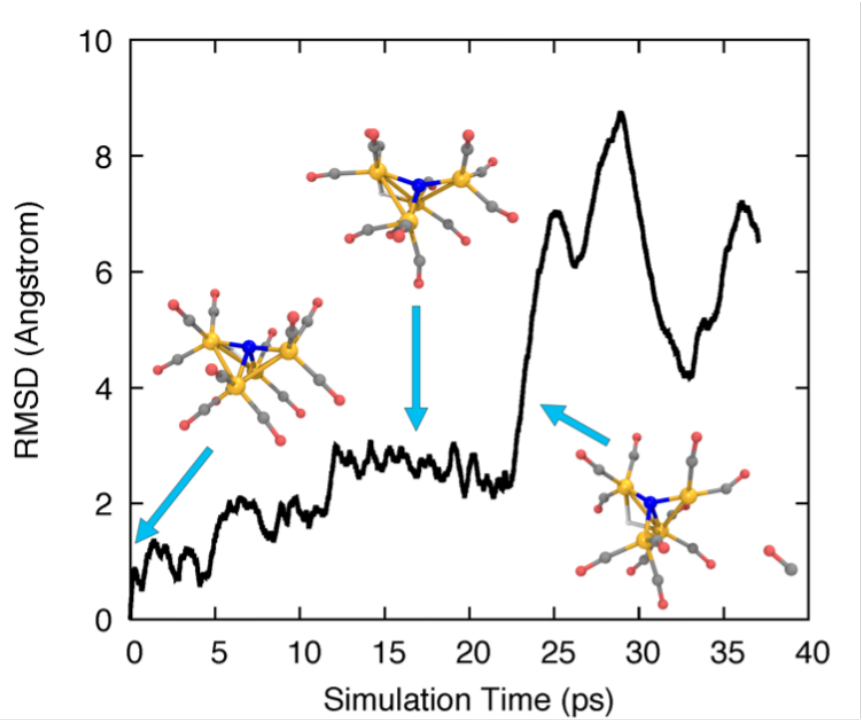


Figure 1: **RMSD time series to the initial optimized structure for a high-temperature AIMD simulation of H-$\mathbf{1}^{1-}$.** Several trajectory snapshots are shown, along with blue arrows indicating their corresponding simulation time. Fe, orange; C, grey; N, blue; O, red; H, white.

The AIMD trajectories at elevated temperatures features highly fluxional behavior of the CO ligands. Figure 1 shows the all-atom root-mean-square deviation (RMSD) of the trajectory frames to the initial structure and several trajectory snapshots of the simulation of H-$\mathbf{1}^{1-}$. The RMSD rapidly reaches 1 Åafter 1000 simulation steps (1 ps) and increases steadily over the course of $\sim$ 15 ps to almost 3 Åas larger geometric rearrangements took place. The conformational changes include concerted rotation of multiple CO groups bonded to the same iron atom (analogous to torsion about a single bond), as well as the exchange of CO ligands on different iron atoms. At frame 22500, we observed a significant increase of RMSD to $>$ 6 Åwhere a CO ligand dissociated from the cluster. The distance between the dissociated CO and the catalyst molecule continued to increase until the simulation was terminated at frame 37000.

2.2.3 Characterization of Optimized Structures

The AIMD simulation explores the potential landscape very broadly, but a closer examination of the optimized structures and barriers is needed to assess the feasibility of the discovered pathways at experimental conditions. We focused on trajectory frames numbered 22000–22390 where the CO is observed to dissociate from the complex and optimized a total of 40 trajectory frames evenly spaced by 10 frames (i.e. spanning 400 simulation time steps). The proton was deleted from the trajectory frames prior to optimization. The charge and spin multiplicity of the twice-reduced state were set to –3 and 1 respectively, prompted by CASSCF (8,8) calculations which indicated that the lowest-energy state is a closed shell singlet.

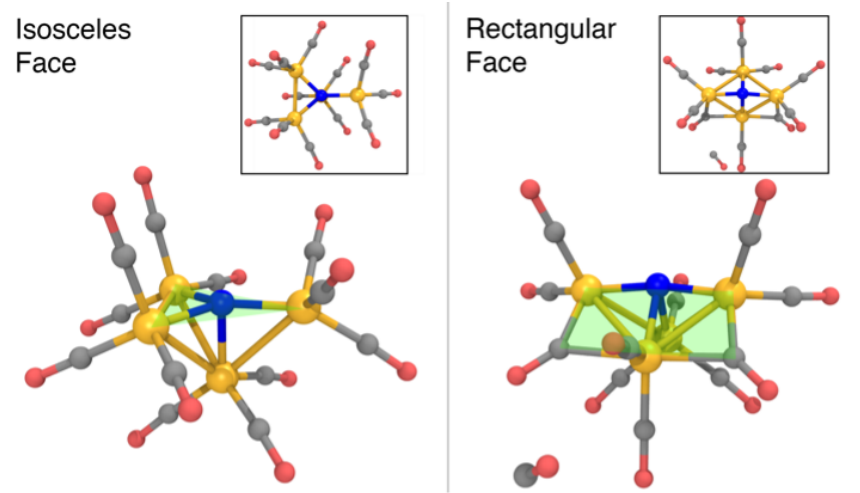


Figure 2: **Optimized structures of 1^{3-} (left) and 2^{3-}+CO (right) at the B3LYP/6-31G*-LANL2DZ level of theory, before and after CO dissociation.** The structures are characterized by a nearly planar isosceles triangular face (left) and a rectangular face (right) that contain the nitrogen atom. Fe, orange; C, grey; N, blue; O, red; H, white.

Figure 2 summarizes the main results when the cluster is optimized in the –3 charge, singlet electronic state. The lowest energy structure (Figure 2, left) is close to Cs-symmetric with a single mirror plane; the Fe atoms surround the central N in an

isosceles trigonal pyramidal arrangement. The central N is nearly in the plane made by three iron atoms, with three in-plane Fe-N-Fe angles of 137, 137 and 84 degrees respectively, summing up to 358 degrees. The other three Fe-N-Fe angles are between 85 and 90 degrees. Each Fe atom has three CO ligands with a tight distribution of Fe-C distances ranging from 1.74–1.77 Å; the *ab initio* bond order (BO) indices computed using Mayer's method[107] range from 1.05 to 1.25, indicating single bond order.

The lowest-energy structure with a dissociated CO ligand (Figure 2, right) features two CO ligands bridging a pair of Fe atoms. The three Fe atoms, central nitrogen and two bridging carbons form nearly a planar rectangle, with Fe-N-Fe and C-Fe-C angles of 168 and 174 degrees respectively. The cluster is also nearly Cs-symmetric with a single mirror plane. Moreover, the bridging COs have significantly larger Fe-C distances of 1.83 Å(left and right edges of rectangle) and 2.08 Å(bottom edge). The increased lengths of the Fe-C bonds along the bottom edge of the rectangle suggest that they possess a different electronic character; indeed these two bonds have *ab initio* bond orders of 0.55, which are almost exactly half of the others. Our interpretation is that the C-Fe-C is a three-center two-electron bond, which compensates for the two σ-electrons that are lost in the dissociation process. To support this interpretation, Figure 3 shows a doubly occupied CASSCF (8,8) optimized molecular orbital that shows significant electron delocalization across the C-Fe-C bond; this is the only orbital we observed that possesses bonding character for these atoms. A comparison of the optimized structure with the experimentally determined X-ray crystal structure[47] revealed an excellent agreement of 0.13 Å, lending confidence to the accuracy of the theoretical methods used; the calculations were performed without knowledge of the crystal structure, and the comparison was only performed later. The experimental crystal structure also contains three Na^+ counterions that further stabilize the $\mathbf{2}^{3-}$

structure; these were not included in the present calculations.

To assess the possibility that **2** may be a catalyst for CO_2 reduction, we computed redox potentials of the $\mathbf{2}^{0}/\mathbf{2}^{-}$, $\mathbf{2}^{-}/\mathbf{2}^{2-}$ and $\mathbf{2}^{2-}/\mathbf{2}^{3-}$ couples in analogy to **1**. Our computed potentials are +0.30, –0.45, and –1.05 V vs. SCE, respectively. Because all of these potentials are more positive than the applied potential for electrocatalysis, we do not think these species are participating redox intermediates in the main CO_2 reduction reaction.

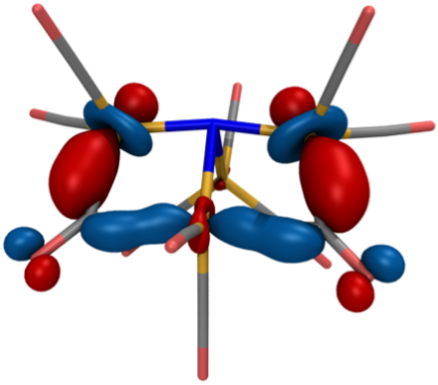


Figure 3: Optimized, doubly-occupied molecular orbital of $\mathbf{2}^{3-}$ at the CASSCF(8,8)/6-31G*-LANL2DZ level of theory, indicating a delocalized bond that connects the three Fe centers and two bridging C atoms in the foreground. The orbital is plotted with an isosurface value of 0.07.

2.2.4 Calculation of Barrier Heights of CO Dissociation

The AIMD simulation that discovered the dissociation pathway is a good starting point for estimating the activation barrier separating the initial and final states. An initial reaction pathway is obtained by concatenating the MD trajectory frames with the output frames from the geometry optimization. From these structures, an "initial chain" of 21 equally spaced frames is selected. Because the initial chain may contain kinks that interfere with the convergence of reaction path optimization methods, we performed an initial smoothing by minimizing an elastic band energy function that

depends solely on internal coordinate displacements. The resulting "smoothed chain" is free of kinks and has a shorter arc-length than the initial chain, and is input into a nudged elastic band (NEB) calculation. The NEB uses a climbing-image approach to ensure the highest-energy structure is as close as possible to the true transition state.

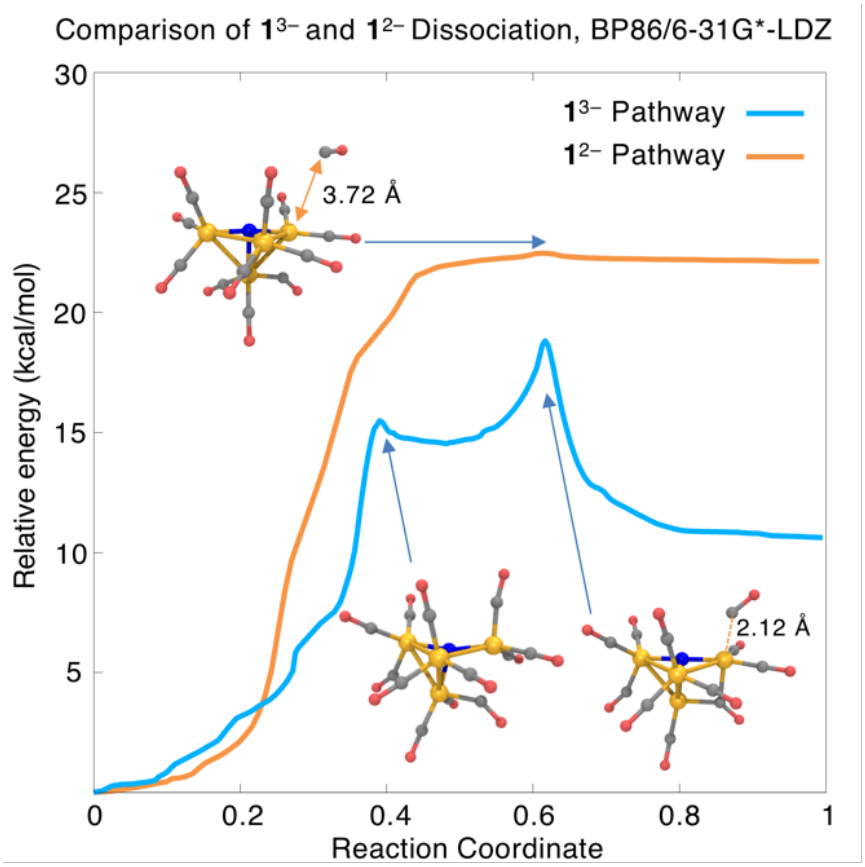


Figure 4: Comparison of relative energies along CO dissociation coordinate from $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$ calculated using BP86/6-31G*-LDZ. The Fe–C distance for the dissociating CO ligand is highlighted. Fe, orange; C, grey; N, blue; O, red; H, white.

	$\mathbf{1}^{3-} \rightarrow \mathbf{2}^{3-}$ + CO		$\mathbf{1}^{2-} \rightarrow \mathbf{2}^{2-}$ + CO	
Structure	ΔE	ΔG	ΔE	ΔG
Optimized IRC at BP86/6-31G*-LDZ				
Initial	0.0	0.0	0.0	0.0
TS	18.8	17.4	22.4	18.1
Final	10.6	7.2	22.1	16.3
Separated	13.3	1.4	25.2	12.8
Optimized IRC at M06-L/6-31G*-LDZ				
Initial	0.0	0.0	0.0	0.0
TS	20.1	18.3	–	–
Final	11.0	7.6	–	–
Separated	15.9	3.6	22.9	12.9
Optimized IRC at B3LYP/6-31G*-LDZ				
Initial	0.0	0.0	0.0	0.0
TS	25.3	23.1	25.2	16.5
Final	12.2	7.8	24.9	14.7
Separated	14.8	3.3	27.2	11.5
Optimized IRC at M06/6-31G*-LDZ				
Initial	0.0	0.0	0.0	0.0
TS	23.5	22.5	–	–
Final	11.1	9.9	–	–
Separated	15.6	4.2	21.9	10.0

Table 3: **Relative energies and free energies (via harmonic approximation) for CO dissociation from $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$.** Each group of four rows refers to calculations performed using a different DFT approximation. In the fourth row of each group, energies and free energies are calculated as a sum of the separated species in the product. All energies are reported in kcal mol^{-1}.

The blue curve in Figure 4 shows the total energy for CO dissociation from $\mathbf{1}^{3-}$ along the BP86/6-31G*-LDZ optimized reaction coordinate. The first part of the path involves a torsional motion of six CO ligands, allowing the two highlighted Fe-C

distances to come into closer contact. An intermediate is found with relative energy of ΔE_1 = +14.5 kcal mol^{-1} and activation barrier of E_{a1} = +15.5 kcal mol^{-1}; the structure contains an additional Fe-C bond (distance = 2.09 Å; BO = 0.63). The second transition state has energy E_a = +18.8 kcal mol^{-1} ($\Delta G^{\ddagger}$ = 17.4 kcal mol^{-1}) and the CO ligand beginning to dissociate from the cluster; this is followed by a relatively flat energy basin where the two newly formed Fe-C bonds (the three-center bond) become equal in length. The final CO-dissociated structure gives a reaction energy ΔE = +10.6 kcal mol^{-1} (ΔG = +7.2 kcal mol^{-1}). We also computed the reaction energy by treating the products as completely separate species and obtainedΔE_{sep} = 13.3 kcal mol^{-1} (ΔG_{sep} = +1.4 kcal mol^{-1}). The higher value of ΔE_{sep} is attributed to dissociating intramolecular interactions and the lower value of ΔG_{sep} to the translational and rotational entropy of separated dissociation products. The slightly uphill ΔG and moderate $\Delta G^{\ddagger}$ values indicate this mechanism may be operative for forming the experimentally observed $\mathbf{2}^{3-}$ species.

We also investigated CO dissociation from the $\mathbf{1}^{2-}$ electronic state; because dissociation is not observed from $\mathbf{1}^{2-}$ in the experiment, we presume that the calculated thermodynamic and/or kinetic parameters should be less favourable compared to $\mathbf{1}^{3-}$. In searching for the reaction energies and activation barriers for the $\mathbf{1}^{2-}$ state, we proceeded from the same initial structures from the AIMD trajectory; the charge and spin multiplicity were set to –2 and 2 respectively. Our BP86 calculations found an uphill and nearly barrierless dissociation pathway (orange curve in Figure 4) with ΔE = 22.1 kcal mol^{-1} and E_a = 22.5 kcal mol^{-1} (ΔG = +16.6 kcal mol^{-1}; $\Delta G^{\ddagger}$ = 18.1 kcal mol^{-1}). The reaction energy calculated using separated species as the products is ΔE = 25.2 kcal mol^{-1} (ΔG = 12.8 kcal mol^{-1}).

Comparison of the dissociation pathways from $\mathbf{1}^{3-}$ vs. $\mathbf{1}^{2-}$ gives reaction free energies of ΔG = +7.2 vs. +16.3 kcal mol^{-1}; with separated product species, ΔG_{sep}

= +1.4 vs. +12.8 kcal mol^{-1}. These values indicate that CO dissociation from $\mathbf{1}^{2-}$ is thermodynamically less favourable than from $\mathbf{1}^{3-}$, consistent with the experimental findings. On the other hand, though the energy barrier for $\mathbf{1}^{3-}$ is lower than for $\mathbf{1}^{2-}$ (ΔE = 18.8 vs. 22.4 kcal mol^{-1}), the calculated activation free energies are nearly the same ($\Delta G^{\ddagger}$ = 17.4 vs. 18.1 kcal mol^{-1}). Comparison of the overall shape of the dissociation curve shows some other important differences; whereas the $\mathbf{1}^{3-}$ pathway has two clearly defined barriers and an intermediate, the $\mathbf{1}^{2-}$ pathway is nearly barrierless which indicates tunnelling effects may play a significant role in determining the reaction rate.[108] In summary, CO dissociation from $\mathbf{1}^{2-}$ is found to be thermodynamically less favourable, but more detailed reaction rate and free energy calculations may be needed to accurately compare the kinetics of these two pathways.

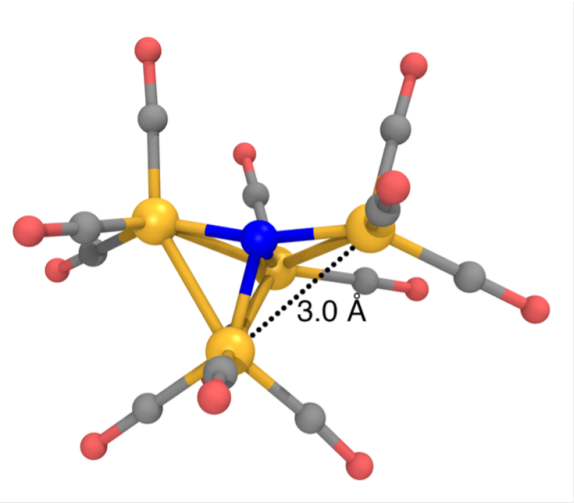


Figure 5: **Optimized structure of $\mathbf{1}^{2-}$ at the BP86/6-31G*-LANL2DZ level of theory, characterized by a "crooked butterfly" structure with a single elongated Fe-Fe distance of 3.0 Å.** Fe, orange; C, grey; N, blue; O, red; H, white.

2.2.5 Validation of Electronic Structure Method

The veracity of our predictions regarding CO dissociation rests upon the choice of method. In this section we provide some justifications for our use of DFT in general,

and the BP86/6-31G*-LDZ level of theory in particular. Our comparison tests include four DFT approximations (BP86, B3LYP, the meta-GGA functional M06-L[109] and the hybrid meta-GGA M06[110]). Whereas the former two functionals contain minimal empiricism, the latter two functionals contain 30+ parameters fitted to databases of diverse molecular properties. Optimized IRCs from $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$ were computed using all four functionals in the 6-31G*-LDZ basis (Table 3). We also tested for basis set effects in the BP86 and B3LYP calculations by comparing a smaller double-zeta basis 6-31G*-LDZ (6-31G* for main group, LANL2DZ for Fe) and a larger triple-zeta basis TZVP-LTZ (TZVP[111,112] for main group, LANL2TZ for Fe). ΔE and E_a in the large basis set were estimated by taking differences of single-point energies along the small-basis-optimized pathway following the IRCMax approach.[113] Our results for comparing BP86 vs. B3LYP and basis set effects in the 13– dissociation pathway are shown in Supplementary Table S1 and Supplementary Figure S1.

In all of our results, we found that increasing the basis set size has a relatively small effect. In BP86/TZVP-LTZ calculations of CO dissociation from $\mathbf{1}^{3-}$, ΔE is essentially unchanged from the 6-31G*-LDZ result (10.6 kcal mol^{-1}); E_a is slightly lower at 18.3 kcal mol^{-1}. For the $\mathbf{1}^{2-}$ pathway, BP86/TZVP-LTZ predicts a slightly higher value for ΔE = 22.7 kcal mol^{-1}, and there is no energy maximum on the pathway; this is perhaps not surprising given the nearly barrierless dissociation curve. In B3LYP/TZVP-LTZ calculations, the ΔE and E_a values changed by < 1 kcal mol^{-1} from the corresponding B3LYP/6-31G*-LDZ values. The choice of DFT functional has a more significant impact. B3LYP/6-31G*-LDZ predicts ΔE = 12.2 kcal mol^{-1} and E_a = 25.3 kcal mol^{-1} for CO dissociation from $\mathbf{1}^{3-}$; notably, E_a is 6 kcal mol^{-1} higher than in BP86. Despite differences in the barrier height, the structures along the $\mathbf{1}^{3-}$ IRCs are highly similar for both functionals, as evidenced by B3LYP single-point calculations along the BP86 optimized pathway and *vice versa* (Supplementary

Table S1).

The most significant DFT functional dependence is seen in the $\mathbf{1}^{2-}$ dissociation pathway. For the reactant ($\mathbf{1}^{2-}$) structure B3LYP predicts a pyramidal structure with an isosceles triangular base, almost identical to the structure of $\mathbf{1}^{3-}$ in Figure 2, left. On the other hand, BP86 predicts a "crooked butterfly" structure (Figure 5) that is closer to the $\mathbf{1}^{-}$ resting state; the largest Fe-N-Fe angle is 165 degrees, and one of the Fe-Fe distances is elongated to 3.01 Å(the others Fe-Fe distances are between 2.55–2.65 Å). These structures are only stable on the potential surfaces of their respective functionals, as a BP86 optimization started from the B3LYP-optimized structure leads to the BP86 minimum and vice versa. Clearly a more objective measure is needed to determine which DFT approximation is more appropriate for this system.

The differences in BP86 vs. B3LYP in the $\mathbf{1}^{2-}$ state originates from the electronic character of the ground state Kohn-Sham (KS) wavefunction. We computed the expectation value of the squared total spin operator $\langle S^2 \rangle$ to measure any deviations of the KS wavefunction from a pure doublet (Supplementary Figure S2). Along the BP86 pathway, the $\langle S^2 \rangle$ value of the BP86 KS wavefunction is stable around 0.77, close to the value of 0.75 for a pure doublet; on the other hand, the B3LYP wavefunction has higher $\langle S^2 \rangle$ values ranging from 0.84 to 1.08, indicating a higher degree of spin contamination. The spin contamination is even greater along the B3LYP IRC, where the B3LYP wavefunction has $\langle S^2 \rangle$ close to 2.0 at the dissociated state. BP86 also predicts $\langle S^2 \rangle$ values around 1.6–1.7 for these structures, indicating a broken-symmetry KS wavefunction containing more than one unpaired electron.

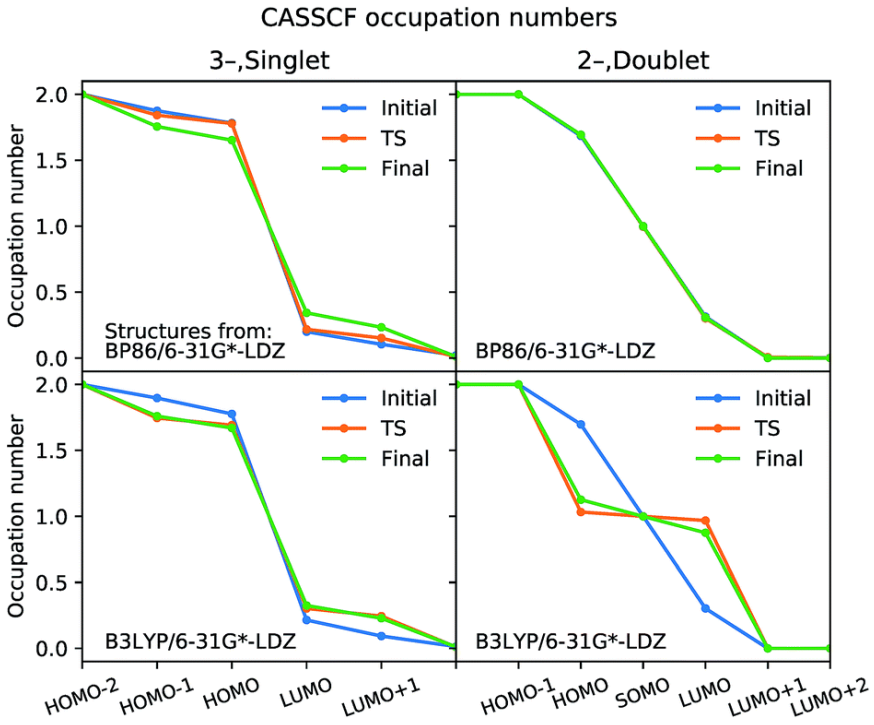


Figure 6: **Natural orbital occupation numbers calculated from CASSCF.** The input geometries are from IRCs of CO dissociation from $\mathbf{1}^{3-}$ (left) and $\mathbf{1}^{2-}$ (right) optimized using BP86 (top) and B3LYP (bottom). Active spaces of (4,6) and (3,6) were used for all structures from $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$ respectively.

The significant spin contamination along the B3LYP IRC for $\mathbf{1}^{2-}$ points to a multireference ground state that is not well-described by a KS determinant. To investigate this further, we carried out single-point CASSCF calculations at the initial, TS, and final geometries along the CO dissociation pathway for both the $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$ IRCs calculated using BP86 and B3LYP. These calculations employ the same 6-31G*-LDZ basis set as the DFT calculations, and active spaces of (4,6) and (3,6) were used for all states from $\mathbf{1}^{3-}$ and $\mathbf{1}^{2-}$ state pathways respectively. These calculations were carried out in the ORCA software package.[114,115]

The optimized CASSCF molecular orbitals are very close to the natural orbitals that diagonalize the density matrix; the eigenvalues are within 10^{-4} of the diagonal elements, and off-diagonal elements are all $< 10^{-4}$. The natural orbital occupation

numbers for initial, TS, and final structures optimized using B3LYP and BP86 are plotted in Figure 6; the more the occupation numbers deviate from 2.0 and 0.0 (for occupied and virtual orbitals), the greater the multireference character. Our analysis for $\mathbf{1}^{3-}$ shows that the natural orbitals at the "frontier" have occupation numbers in the range of 1.9–1.7 and 0.1–0.2. The variation in these values are small when comparing the initial, TS, and final structures, indicating there is no qualitative change in the electronic character along the reaction pathway. Moreover, none of the natural orbitals have occupation numbers near 1.0, which is a hallmark of wavefunctions that display strong multireference character; this is the case for diradicals and homolytic dissociation of N_2.[77]

For CO dissociation from $\mathbf{1}^{2-}$, the CASSCF calculations using BP86-optimized structures show a similar pattern to $\mathbf{1}^{3-}$, except a singly occupied molecular orbital is present. On the other hand, a major change in the electronic character is seen for the B3LYP-optimized structures. The TS and final structures have occupation numbers close to 1.0 in three orbitals, indicating strong ground-state multireference character; this result agrees with the spin contamination observed in DFT wavefunctions for the same structures. When comparing the BP86 and B3LYP functionals, only the BP86-optimized structures have CASSCF ground states with consistent electronic character; we thus conclude that BP86 gives the more reliable result overall.

We also calculated reaction energies and activation energies of the reactions using the M06 and M06-L functionals to confirm the accuracy of BP86 for this system (Table 3). These calculations were performed in Q-Chem 5.0. We could not find a TS structure for CO dissociation from $\mathbf{1}^{2-}$ using these functionals, again possibly owing to the nearly barrierless dissociation curve. The M06-L results are in close agreement with BP86, which is reasonable given that both functionals contain no Hartree-Fock (HF) exchange; spin contamination along the BP86-optimized $\mathbf{1}^{2-}$ dissociation path-

way is low, with $\langle S^2 \rangle$ = 0.79–0.80. The M06 results are closer to B3LYP, perhaps because both functionals contain a similar amount of HF exchange (28 % vs. 20 %). M06 also shows similar amounts of spin contamination to B3LYP along both the BP86-optimized and B3LYP-optimized $\mathbf{1}^{2-}$ dissociation pathways.

2.2.6 Calculated Vibrational Analyses

Infrared (IR) absorption spectra provide a meaningful connection between theory and experiment; a harmonic vibrational analysis calculation provides a series of frequencies and intensities that may be converted to a simulated spectrum by applying artificial broadening to each absorption peak. The results of two frequency calculations are shown in Figure 7, where we compare the IR absorption peaks of $\mathbf{1}^{2-}$ and $\mathbf{2}^{3-}$, the presumed initial and final states of CO dissociation. The approximate spectra cannot accurately reproduce the widths of the experimental peaks, and only the shifts in the peak positions, or the appearance of new peaks, is meaningful. The most notable feature in the spectrum of $\mathbf{2}^{3-}$ is a new peak that appears in a region red-shifted from the main CO-stretching band by about 150 cm^{-1}. The vibrational mode of this peak corresponds to CO-stretching of the bridging CO ligands. The reduced frequency indicates a slightly lower force constant in the CO bond of the bridging ligands that donate more electron density to the Fe centers. This red-shifted stretching peak may be used as a vibrational signature of CO dissociation.

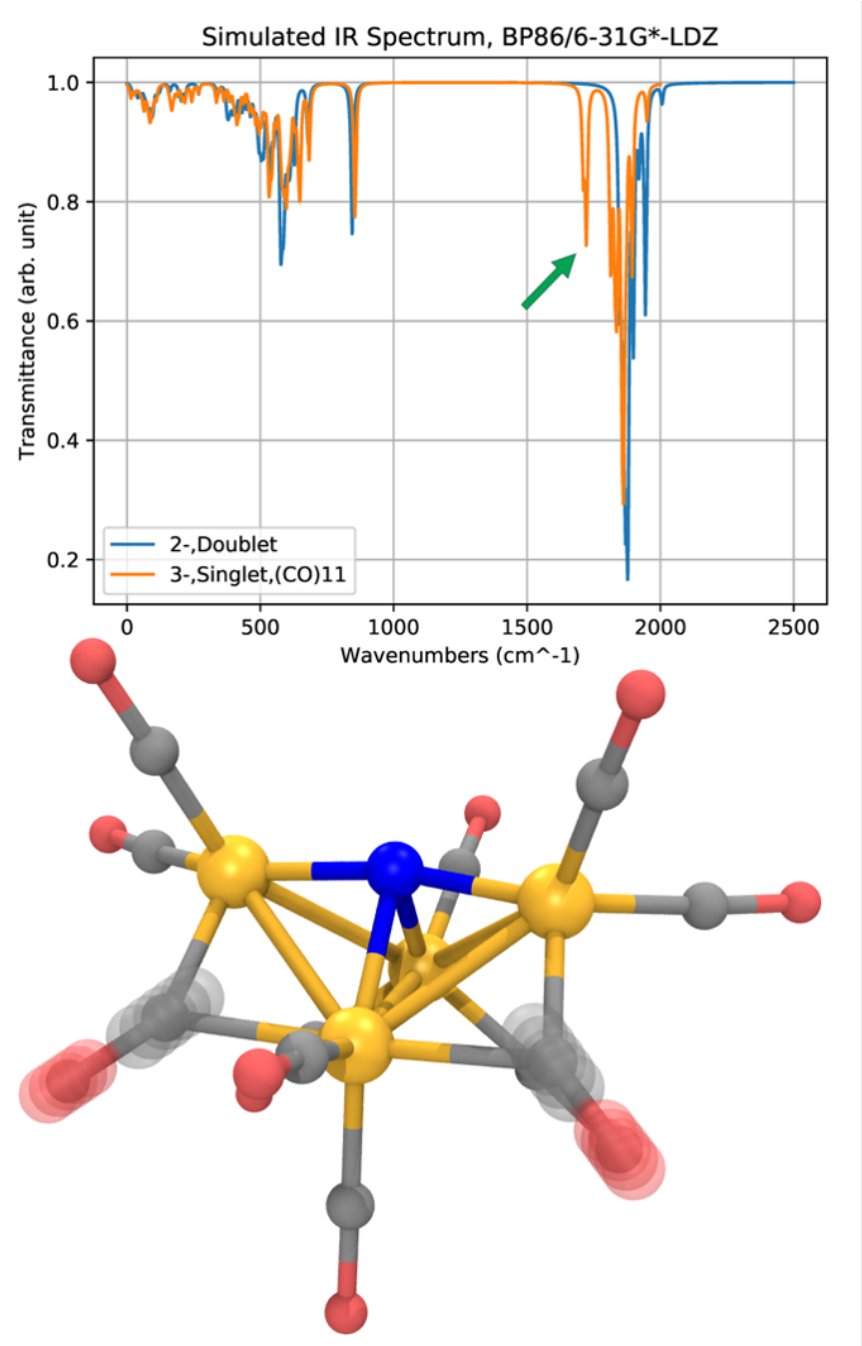


Figure 7: **Comparison of vibrational spectra calculated at the BP86/6-31G*-LDZ level for $\mathbf{1}^{2-}$ (blue) and $\mathbf{2}^{3-}$ (orange).** The IR spectrum of $\mathbf{2}^{3-}$ has a distinct peak red-shifted from the main band of CO–stretches by about 150 cm (green arrow) corresponding to a symmetric and antisymmetric stretch of the bridging CO ligands (bottom). An artificial Lorentzian broadening of 10 cm is used.

2.3 Conclusions

In this study, we calculated the redox properties of the CO_2 reduction $[Fe_4N(CO)_{12}]^-$ ($\mathbf{1}^-$) and investigated the possibility of CO dissociation from the twice-reduced state, $\mathbf{1}^{3-}$. Our calculated redox potentials show close agreement with experimentally measured values. The structure of the product of CO dissociation ($\mathbf{2}^{3-}$) was predicted and found to be in close agreement with the experimental X-ray crystal structure. The CO dissociation pathway from $\mathbf{1}^{3-}$ is energetically accessible at ambient conditions (in kcal mol^{-1}: $\Delta E = +10.6$, $E_a = 18.8$; $\Delta G_{\text{sep}} = +1.4$, $\Delta G^{\ddagger} = +17.4$). The analogous CO dissociation from $\mathbf{1}^{2-}$ has a higher reaction energy and similar activation free energy (in kcal mol^{-1}: $\Delta E = 22.1$, $E_a = 22.4$; $\Delta G_{\text{sep}} = +12.8$, $\Delta G^{\ddagger} = +18.1$) with a nearly barrierless dissociation curve. Vibrational analysis of $\mathbf{2}^{3-}$ shows a distinct CO stretching peak red-shifted from the main CO stretching band, indicating a possible vibrational signature of CO dissociation. Our calculations indicate that the BP86 semilocal functional gives more reliable results than the B3LYP hybrid functional in the study of this system. Future studies will focus on the potentially important role of counterions in stabilizing redox intermediates, as well as the strong solvent dependence in the selectivity of this catalyst for H_2 evolution vs. CO_2 reduction.

3 *respyte*: Modernized Implementation of RESP for the Development of the Next Generation of ESP-based Charge Model

3.1 Introduction

Most molecular mechanics force fields use point charges to approximate the charge distribution around atoms in a system. Since atomic partial charges are not physical observable, many different models have been proposed to define atomic partial charges in a system. The most widely used method for fixed charge force fields is the restrained electrostatic potential (RESP) method, where atomic partial charges are fitted against quantum chemical electrostatic potentials (ESP) with restraints.[32]

While RESP has seen broad application over the past 20 years, many unresolved challenges remain, including significant dependence of the fitted charges on many heuristic choices; for example, (1) it deliberately uses a low level of theory, Hartree-Fock (HF)[116,117] with 6-31G* basis set[105] in the gas phase, to overestimate the gas-phase polarity of molecules to yield appropriate polarity of hydrated molecules, polarized by the solvent reaction field[32,118,119]. However, the over-polarization of HF/6-31G* appears to be inconsistent across different molecules and to underestimate the polarization typically induced by water. Therefore, there have been several studies that explored whether ESPs computed with higher-level QM methods could provide more accurate charges and thus more accurate simulation[120–125]; (2) The original implementation samples ESP points on Merz-Singh-Kollman (MSK) shells[126] with inner and outer radii of 1.4 R_i and 2.0 R_i and 0.2 R_i spacing between point layers where R_i is the van der Waals radius (Bondi radii), and with a density of 1 points/$Å^2$ in each layer.[32] While the fitted charges heavily depend on the choices involved in sampling

grid points, the question of "what is the optimal grid-point sampling scheme?" is still not completely resolved. Because of the many considerations involved in RESP, while there have been several implementations of RESP, they produce different results for the same input and/or are not fully featured.

Meanwhile, due to the oversimplification of the model, the fixed point charge models are no longer valid in describing short range interactions, where electron clouds of adjacent atoms start to overlap. Charge penetration is the change in the electrostatic interaction between two atoms and the associated loss of nuclear screening induced by the overlap of electron clouds. To properly describe the phenomenon, more physically realistic charge models, such as AMOEBA charge penetration model[127], have been proposed.

As a first step to develop the next generation of the ESP-based charge model, we developed an open-source software implementation of RESP in Python, *respyte*. This new tool is capable of exactly reproducing the original implementation and is also easily extensible for developing improved ESP-based charge models.

3.2 Computational Methods

respyte is implemented in Python, licensed under BSD 3-clause License, and depends on the following Python modules: Psi4[128], SciPy[129], NetworkX[130], PyYAML[131], SymPy[132] and RDKit[133]. *respyte* consists of two main functions: (1) grid point selection and QM calculation, and (2) charge fitting (Figure 8). Details on how to use the package, as well as inputs and outputs can be found in Appendix B.

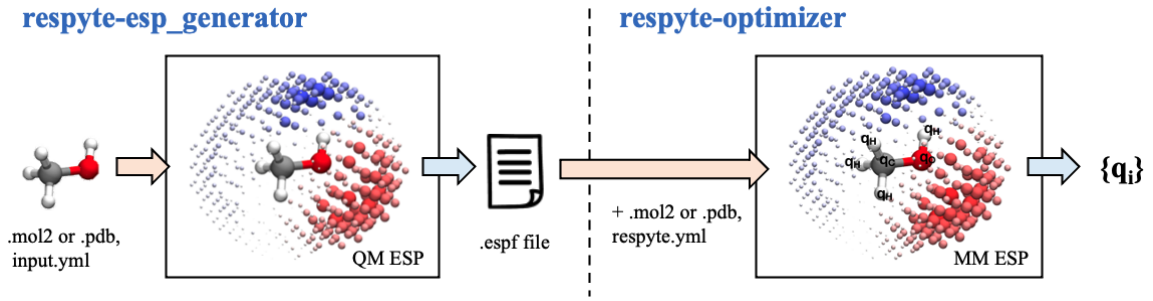


Figure 8: **Data flow diagram of *respyte*.** "respyte-esp_generator" executable takes molecule coordinate files (with .pdb or .mol2 extension) along with the input file ('input.yml'), where the user specifies the grid generation scheme and QM level of theory, and performs grid point selection and QM calculation using the Psi4 package. "respyte-optimizer" executable then takes .espf files, which store the QM data from the previous step, along with the input file ('respyte.yml'), where the user specifies charge model and many options involved in the fitting. It performs charge fitting and returns fitted charges ($\{q_i\}$). (In the visualization of QM (MM) ESPs computed around methanol molecule, the size of the spheres around the molecule represents the magnitude of ESP value at each grid point and the color of the spheres represents the sign of the value, ranging from negative (red), zero (white) through positive (blue).)

The package is carefully designed (1) to support various charge models, (2) to enable easy implementation of new models, and (3) to support a user-friendly interface.

Each charge model has its own expression of ESP generated by a single particle. For the simple point charge model, the ESP due to a single particle with the point charge q is given by

$$V(r) = \frac{q}{r} \tag{11}$$

For fuzzy charge model, a charge penetration model proposed by Paul Nerenberg group that consists of a positive point charge, q_{core}, at the particle center and a smeared negative charge, q – q_{core}, the ESP due to a single particle is given by

$$V(r) = \frac{q_{\text{core}}}{r} + \frac{q - q_{\text{core}}}{r}(1 - e^{-\alpha r}) \tag{12}$$

where α is the smearing parameter for the particle's charge distribution. *respyte* eval-

uates ESP for each grid point, for a given charge model (currently the simple point charge model and the fuzzy charge model are supported.) and construct an objective function, a sum of squared differences between QM ESP and MM ESP, which is given by

$$\chi^2_{\mathrm{ESP}}(q_1, \ldots, q_n) = \sum_{i=1}^{m} \left(V_{i,\mathrm{QM}} - V_{i,\mathrm{MM}}(q_1, \ldots, q_n) \right)^2 \tag{13}$$

When fitting along with electric field (EF), the objective contribution from EF is given by

$$\chi^2_{\mathrm{EF}}(q_1, \ldots, q_n) = \sum_{i=1}^{m} \left(E_{i,\mathrm{QM}} - E_{i,\mathrm{MM}}(q_1, \ldots, q_n) \right)^2 \tag{14}$$

and the objective function, $\chi^2(q_1, \ldots, q_n)$ is expressed as a linear combination of ESP and EF contributions.

$$\chi^2(q_1, \ldots, q_n) = \omega_{\mathrm{ESP}} \chi^2_{\mathrm{ESP}}(q_1, \ldots, q_n) + \omega_{\mathrm{EF}} \chi^2_{\mathrm{EF}}(q_1, \ldots, q_n) \tag{15}$$

where ω_{ESP} and ω_{EF} are relative weights of each component.

In RESP, to increase transferability between conformers and to remove problematic large charges, it applies restraints in the form of a penalty function to suppress undesirable large charges to lower magnitudes. Harmonic penalty function for charges is given by

$$\chi^2_{p,\mathrm{rstr}} = a \sum_{j=1}^{n} (p_j - p_{0j})^2 \tag{16}$$

where a is a restraint weight and p_{0j} is the target parameter for the restraint. Hyperbolic penalty function for charges is given by

$$\chi^2_{p,\mathrm{rstr}} = a \sum_{j=1}^{n} ((p_j^2 + b^2)^{1/2} - b) \tag{17}$$

where a is a restraint weight, b is a tightness of the hyperbola around the minimum.

Hyperbolic restraint function is more recommended than a harmonic function for charges because in such a function the well-determined polar charges are not overly penalized based on their magnitudes.[32]

The overall objective function with the penalty function is given by

$$\chi^2 = \omega_{\mathrm{ESP}}\chi^2_{\mathrm{ESP}} + \omega_{\mathrm{EF}}\chi^2_{\mathrm{EF}} + \chi^2_{q,\mathrm{rstr}} \tag{18}$$

and if a charge model has one or more additional parameter types, smearing parameters for the fuzzy charge model for example, additional penalty functions are added to the overall objective function. The gradient and hessian of the objective function are computed to run the Newton-Raphson method, which are given by

$$\mathbf{g}_i^{(t)} = \frac{\partial \chi^2}{\partial q_i}(q_1^{(t)}, \ldots, q_n^{(t} \tag{19}$$

$$\mathbf{H}_{ij}^{(t)} = \frac{\partial^2 \chi^2}{\partial q_i \partial q_j}(q_1^{(t)}, \ldots, q_n^{(t)}) \tag{20}$$

In the method, the next set of parameters is given by

$$\mathbf{q}^{(t+1)} = \mathbf{q}^{(t)} - \left[\mathbf{H}^{(t)}\right]^{-1}\mathbf{g}^{(t)} \tag{21}$$

where $\mathbf{q}^{(t)}$, $\mathbf{g}^{(t)}$, $\mathbf{H}^{(t)}$ are the set of parameters, gradient and Hessian of the objective function evaluated at iteration number t. It returns the parameter set once the convergence criterion is met.

3.3 Conclusions and Future Directions

We implemented *respyte*, an open-source version of RESP in Python. The package is capable of flexible grid point generation and QM ESP and EF calculation with

a direct interface to the Psi4 package. The charge fitting part of the package is not only able to reproduce the original implementation, but also carefully designed to be extensible for developing the next generation of ESP-based charge models. Fuzzy charge model, a charge penetration model proposed by the Paul Nerenberg group has been implemented and the preliminary study showed that the fuzzy charge model gives better reproduction of QM ESP over the simple point charge model.

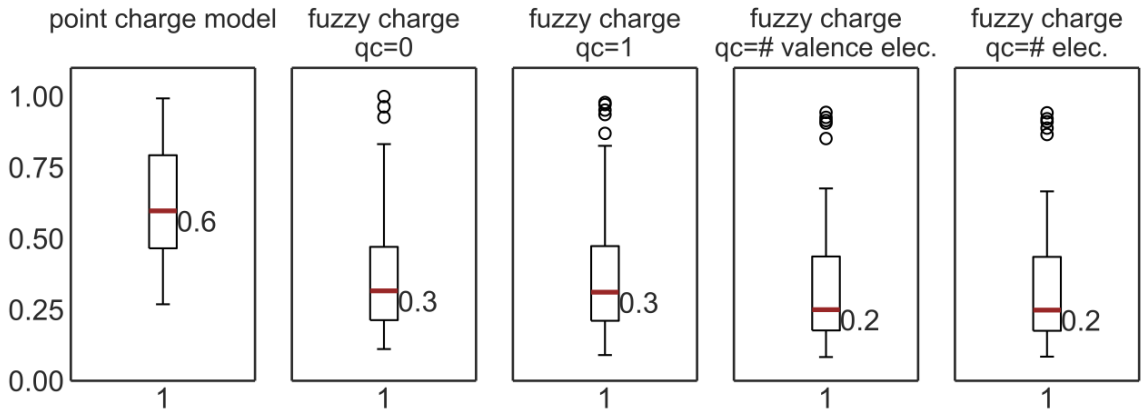


Figure 9: **Comparisons of the goodness of fit to the QM ESP for several charge models; the simple point charge model and the fuzzy charge models with different core charges (qc).** Box plots of the relative root mean square error (RRMS = $\{\chi^2_{\mathrm{ESP}}/\sum_{i=1}(V_i)^2\}^{1/2}$) of the charge models for rigid small molecules. (Results of normalized two-stage fittings with a=0.00001, b=0.1, c=0.000005, α_0 = 3. c and α_0 are for the fuzzy charges, where c is the harmonic restraint weight for α and α_0/r_{vdW} is the target α for the restraint.)

Further research will focus on the evaluation of the fuzzy charge model for the ability to reproduce experimental observables, such as the densities and heats of vaporization of pure organic liquids.

4 Development of an Open Small Molecule Force Field

4.1 Introduction

Developing a high-quality general force field for the application to the wide range of small organic molecules of interest in biology and drug discovery is challenging due to the vastness of the chemical space that must be covered. While there are several small molecule force fields widely used today, such as the general AMBER force field (GAFF)[134], the CHARMM general force field (CGenFF)[135], and the optimized potentials for liquid simulations force field (OPLS)[136], there is still much room for improvement in current general small molecule force fields. Important properties where the current generations of small molecule force fields lack accuracy include hydration free energies, partition coefficients, and protein-ligand binding free energies.[137–139]

Current efforts in improving force fields mostly involve either a small number of specialized research groups who have in-house knowledge and methods inherited or commercial closed-source efforts. Open Force Field Consortium is an academia-industry partnered open-source effort to develop the science and infrastructure for the development of the next generation of small molecule and biomolecular force fields, which aims to develop and release an automated, open, sustainable, extensible, and well-supported infrastructure and open datasets for producing and benchmarking force fields so that anyone can access and contribute to the force field development.

In our previous study, we showed our initial progress, the development of the Open Force Field (OpenFF) toolkit, an open-source software package for the development and application of force fields, and the first application of the infrastructure to create a small molecule force field, OpenFF1.0.0, code-named Parsley, the first optimized

SMIRKS-native Open Force Field (SMIRNOFF) force field. Benchmarking of Parsley showed improved accuracy in optimized geometries and conformational energetics.[33]

As a short description of SMIRNOFF, while atom-typing has been a standard way of force field parameter assignment due to its simplicity, it has several technical problems, including complexity in force field specification and proliferation of redundant parameters. As an alternative to atom-typing, SMIRNOFF formalism has been developed.[140] SMARTS is a language built on Simplified Molecular Input Line Entry Specification (SMILES) for defining chemical substructures. (`http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`) One simple example of SMARTS string is [OH]c1ccccc1, which can be used to search for phenol-containing molecules from a database. SMIRKS language (`http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html`) extends SMARTS with the numerical labels on atoms in the substructure. While it has been originally designed to specify chemical reactions, the numerical atom labels can be used to match specific atoms involved in force field terms to assign force field parameters. [#6X4:1]-[#7X3:2]-[#6X3]=[#8X1+0] is an example SMIRKS pattern to match the single bond(-) connecting the carbon with four bonds([#6X4:1]) and the nitrogen with three bonds ([#7X3:2]) in an amide group (:1 and :2 in the pattern are the numerical atom labels.) In the SMIRNOFF formalism, each force field term has a hierarchy of parameter definitions, where each definition consists of a SMIRKS pattern and numerical parameters attached to it. During a substructure search, if a SMIRKS pattern is matched to a substructure in a given molecule, the parameters attached to the pattern are assigned to the mapped atoms in the substructure. A big benefit of this approach is that since each force field term is independent of each other, an extension of parameter definitions is much simpler compared to atom-typing. The first force field adapted the SMIRNOFF formalism, SMIRNOFF99Frosst showed that it retains the accuracy

of its parent force field, AMBER parm99[141] and Merck's parm@Frosst[142] for several important thermodynamic properties, such as hydration free energies and host-guest binding thermodynamics, with only $\sim$ 5 percent of lines of parameters for GAFF.[143]

Here the present study describes further improvements in Parsley, which include modifications and additions of parameter definitions to improve performance in certain chemical spaces, and a more careful design of QM reference data used to re-optimize the SMIRNOFF force field bonded parameters. The successive benchmark results showed improved performance of the reoptimized force field, OpenFF 1.2.0, in reproducing QM optimized geometries over its predecessor, especially for phosphonate-containing molecules and exocyclic divalent nitrogen-trivalent nitrogen bond containing molecules.

4.2 Method

4.2.1 Training the Parsley Force Field

4.2.1.1 Refitted Parameters

We reoptimized the valence parameters present in modified SMIRNOFF99Frosst. Each parameter definition is uniquely identified by an interaction type (e.g., bond stretching) and a SMIRKS pattern (e.g., [#6X4:1]-[#6X3:2]) with one or more physical values attached (e.g., the equilibrium bond length and the force constant). Modification of some initial parameters and the addition of new parameter definitions have been applied to improve performance in certain chemical spaces, where SMIRNOFF99 Frosst fails to properly describe. The change can be summarized as follows:

- Modification of the angle force constant k of bond angles in three-membered rings (a3).

- Addition of a34a, t155b to properly describe nonlinear R=S=O and nitrogen-phosphorus double bond rotation, respectively.

- Three new bond and angle terms, a22a, b14a, and b36a were added to effectively describe conjugation effect of N=C=S, resonance structure of a single bond between *sp2* carbon and oxygen with negative 1 charge, and resonance structure of a double bond between nitrogen with positive 1 charge and nitrogen with negative 1 charge, respectively.

- Periodicities of nitrogen-nitrogen single bond rotations (t128, t129, t130, and t131) have been modified to properly describe the optimized geometries of nitrogen-nitrogen single bond containing molecules.

- New improper torsions i2a, i3a, and i3b were added to describe planar trivalent N centers connected to pi-bond forming S or P, planar trivalent N centers connected to a pi-bond-forming N, and planar trivalent N centers inside the 5-membered hetero-aromatic ring, respectively. And new torsions t51a, t51b, t51c, associated with the new improper torsions, were added to describe *sp3* C connected to trivalent N which is connected to N=C, *sp3* C connected to trivalent N which is connected to N=N or N=O, *sp3* C and trivalent N bond connected to 5-membered hetero-aromatic ring, respectively.

The full list of parameter definitions, which can be viewed in the published force field XML file, openff-1.2.0.offxml, can be summarized as follows:

- Harmonic bond stretch: 88 equilibrium bond lengths and force constants.

- Harmonic angle bend: 36 equilibrium angles and 40 force constants. These two numbers differ because four linear angles were kept linear during fitting.

- Proper torsions: Each of the 163 torsion types is associated with an N-term Fourier series of potential energy contributions, where N $\leq$ 6, and each term, i, is of the form of $E_i = k_i(1 + \cos(\text{periodicity}_i\,\theta - \text{phase}_i))$. We optimized all of the amplitudes that were defined in SMIRNOFF99Frosst, comprising 160, 67, 25, 5, 5, 3 values of k_1, k_2, k_3, k_4, k_5, and k_6 respectively, for a total of 264 parameters. Parameters t156, t157, t158 represent torsion angles containing a linear angle, and their values of k_1 were kept at 0.0 during fitting. The phase parameters, periodicity_i and the selection of Fourier terms used for each torsion were not optimized in this release.

- Improper torsions: The 7 improper terms were kept unmodified, to avoid overfitting. All of the above parameters were fitted simultaneously against all QM data.

4.2.1.2 Compound Sets Used in Training

Six sets of small organic molecules were used to generate the quantum chemical datasets used in fitting. The first compound set is the Roche Set, which contains 468 fragment-like molecules with one to three rotatable bonds. The set was generated using the MOE software[144]. The second set is the Coverage Set, which contains 80 molecules selected from eMolecules database[145]. The set was generated using a greedy algorithm to maximize the parameter coverage with the minimum number of molecules. These two compound sets were used for the previous training set generation, but additional four compound sets were added to extend the coverage of chemical space. The third set is the Pfizer discrepancy set, which contains 100 fragment-sized molecules, where their QM torsional profiles computed with HF/MINIX[146] followed by B3LYP/6-31G*//B3LYP/B-31G** are significantly different from those generated using the OPLS3e force field. Relevant code is open on GitHub[147]. The fourth set is

the eMolecules discrepancy set, which contains 2802 fragment-sized molecules, where their energy-minimized geometries from SMINOFF99Frosst 1.0.8 significantly differ from those generated from GAFF, GAFF2, MMFF94, and MMFF94s.[148] The fifth set is the Bayer set, which contains 5054 actual pharmaceutical compounds provided by Katharina Meier from Bayer. Lastly, the supplemental molecule set contains 15 molecules manually selected from eMolecules database[145] to achieve the full coverage of torsion parameters.

4.2.1.3 Selection of Quantum Chemistry Methodology

Quantum chemical calculations were performed using the MolSSI QCFractal[149] distributed quantum chemistry engine, and the results are deposited in the MolSSI QCArchive Server (MQCAS)[150,151], which allows open access to all data. For this work, a single level of theory was used for all QM calculations, B3LYP-D3(BJ) / DZVP[22,152–154]. We described the rationale for the choice of the methodology in our previous work[33], confirming that B3LYP-D3(BJ) reproduces the reference energies with RMSEs of < 1 kcal mol^{-1} when very large basis sets (e.g., def2-QZVP[155]) are used, and DZVP-DFT basis set gives the best compromise between accuracy and computational cost for the molecule set including amino acids, small to medium-sized peptides, and macrocycles.

4.2.1.4 Generation of Quantum Chemical Data for Compound Datasets

		Roche set	Coverage set	Pfizer discrepancy set
Opt. Geom.	Cases	298	356	197
	MQCAS Dataset	OpenFF Gen 2 Opt Set 1 Roche	OpenFF Gen 2 Opt Set 2 Coverage	OpenFF Gen 2 Opt Set 3 Pfizer Discrepancy
Vib. Freq.	Cases	201	101	86
	MQCAS Dataset	OpenFF Gen 2 Opt Set 1 Roche	OpenFF Gen 2 Opt Set 2 Coverage	OpenFF Gen 2 Opt Set 3 Pfizer Discrepancy
Tors. Scans	Cases	124	121	68
	MQCAS Dataset	OpenFF Gen 2 Torsion Set 1 Roche 2	OpenFF Gen 2 Torsion Set 2 Coverage 2	OpenFF Gen 2 Torsion Set 3 Pfizer Discrepancy 2
		eMolecules discrepancy set	Bayer set	Supplemental set
Opt. Geom.	Cases	2143	1751	-
	MQCAS Dataset	OpenFF Gen 2 Opt Set 4 eMolecules Discrepancy	OpenFF Gen 2 Opt Set 5 Bayer	-
Vib. Freq.	Cases	358	443	-
	MQCAS Dataset	OpenFF Gen 2 Opt Set 4 eMolecules Discrepancy	OpenFF Gen 2 Opt Set 5 Bayer	-
Tors. Scans	Cases	234	144	19
	MQCAS Dataset	OpenFF Gen 2 Torsion Set 4 eMolecules Discrepancy 2	OpenFF Gen 2 Torsion Set 5 Bayer 2	OpenFF Gen 2 Torsion Set 6 supplemental 2

Table 4: **Summary of quantum chemical calculations used to fit the force field valence parameters in this work.**

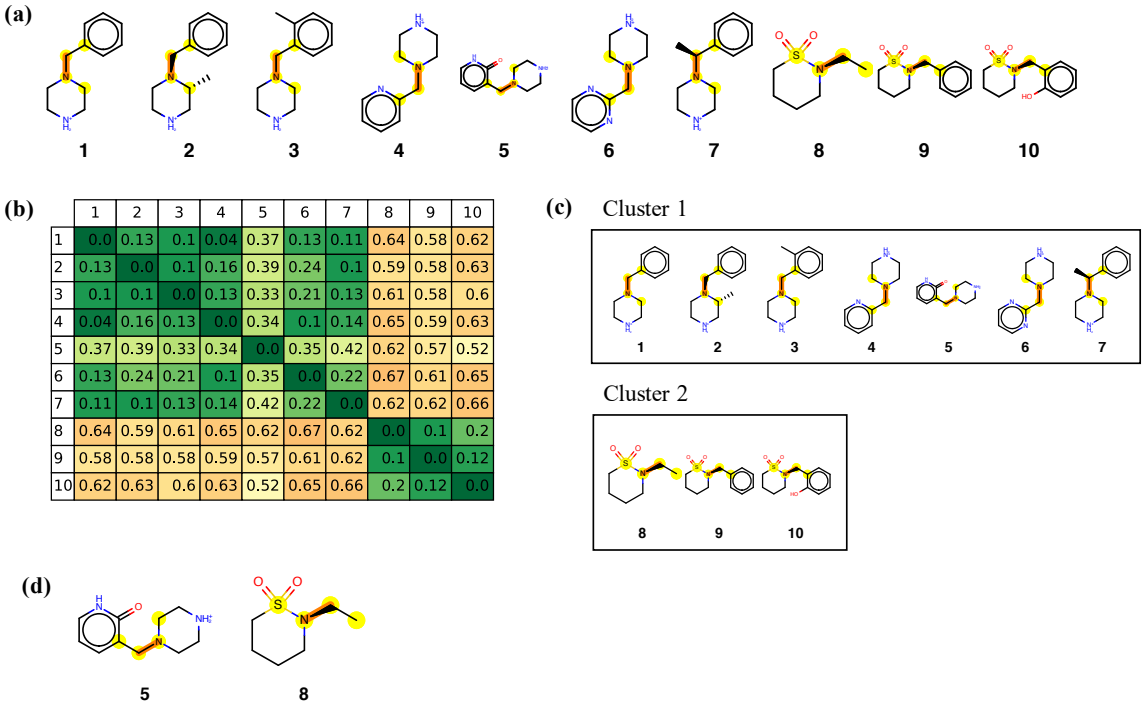
We curated the new training set to accurately model a broad range of chemistries, aiming to improve the generalizability of the force field. The three main goals of the new training set are (1) All parameter definitions are used for the force field; (2) All parameter definitions are used a reasonable amount of time ($\sim$ 5 times); (3) Each parameter definition is used in diverse chemical environments. The approach chosen to achieve the goals can be described as follows:

1. For each compound set, we labeled each molecule with bond, angle, and torsion parameter definitions, using the OpenFF toolkit.

2. For each bond/angle parameter definition, (1) we listed molecules having substructure matching to the SMIRKS pattern of the parameter definition; (2) calculated a distance matrix using MACCS keys fingerprint and Tanimoto score

available in the OpenEye package[156]; (3) performed DBSCAN clustering from the distance matrix using the scikit-learn software package[157]; (4) then selected randomly a molecule from each cluster under the assumption that the molecule was a representative of unique chemistry within that cluster. For torsion parameters, to minimize the coupled torsion problem, it used a slightly more complicated approach. First, each rotation has been converted into a directed graph representation. In the representation, each node is one torsion parameter definition, and each rotation is represented as a set of edges, whose starting point is the torsion parameter found from the rotation in step 1 and ending points are other torsion parameters sharing the central bond with the torsion parameter. After converting all rotations in the compound sets into the graph representation, it chose the most independent sets of edges (i.e., rotations) by (1) randomly selecting one set of edges for each torsion parameter, (2) calculating the number of overlapping edges then (3) looping over until it finds near-minimum overlaps.

3. We combined all selected molecules from each compound set. Prior to running quantum chemical calculations, the selected molecules were expanded to tautomeric and isomeric states, using the Fragmenter software package.[158]



	1	2	3	4	5	6	7	8	9	10
1	0.0	0.13	0.1	0.04	0.37	0.13	0.11	0.64	0.58	0.62
2	0.13	0.0	0.1	0.16	0.39	0.24	0.1	0.59	0.58	0.63
3	0.1	0.1	0.0	0.13	0.33	0.21	0.13	0.61	0.58	0.6
4	0.04	0.16	0.13	0.0	0.34	0.1	0.14	0.65	0.59	0.63
5	0.37	0.39	0.33	0.34	0.0	0.35	0.42	0.62	0.57	0.52
6	0.13	0.24	0.21	0.1	0.35	0.0	0.22	0.67	0.61	0.65
7	0.11	0.1	0.13	0.14	0.42	0.22	0.0	0.62	0.62	0.66
8	0.64	0.59	0.61	0.65	0.62	0.67	0.62	0.0	0.1	0.2
9	0.58	0.58	0.58	0.59	0.57	0.61	0.62	0.1	0.0	0.12
10	0.62	0.63	0.6	0.63	0.52	0.65	0.66	0.2	0.12	0.0

Figure 10: **Procedure of selecting torsion scans for training t51 ([*:1]-[#6X4:2]-[#7X3:3]-[*:4]) from the Roche set.** (a) is a list of 10 molecules having a substructure matching to the t51's SMIRKS pattern. (b) is the color distance matrix of the 10 molecules generated using MACCS keys fingerprint and Tanimoto score. The closer the element is to 0, the more similar the two structures are. (c) are the two clusters generated from DBSCAN clustering of the distance matrix and (d) are the two selected molecules from each cluster.

4.2.1.5 Application of ForceBalance

The parameter optimization was carried out with ForceBalance[159], a Python software package for force field optimization in a systematic and reproducible manner.[159,160] ForceBalance v1.7.1[161] was used to minimize the objective function. The OpenFF Toolkit v0.6.0[162] and the commercial OpenEye toolkit version 2019.10.2[156] were used to support the OpenFF force field and set up OpenFF simulations. The detailed optimization algorithm, convergence criteria, objective function with regularization can be found in our previous work.[33]

4.2.2 Testing the Parsley Force Field

Once the parameters had been trained as detailed in Section 4.2.1, we tested the resulting force field, Parsley 1.2.0, against the quantum chemical test set, termed the Full Benchmark Set, that we used for our previous work[33]. The test set contains two data types: optimized geometries, and energy differences among conformers of a given molecule.

4.3 Results and Discussion

This section first describes the consequences of parameter optimization for accuracy over the training set and then benchmarks Parsley on the separate test set compounds and properties. The test set results should be indicative of Parsley's accuracy in new applications.

4.3.1 Improvement in Accuracy Over Training Set Data

4.3.1.1 Optimization of the Objective Function

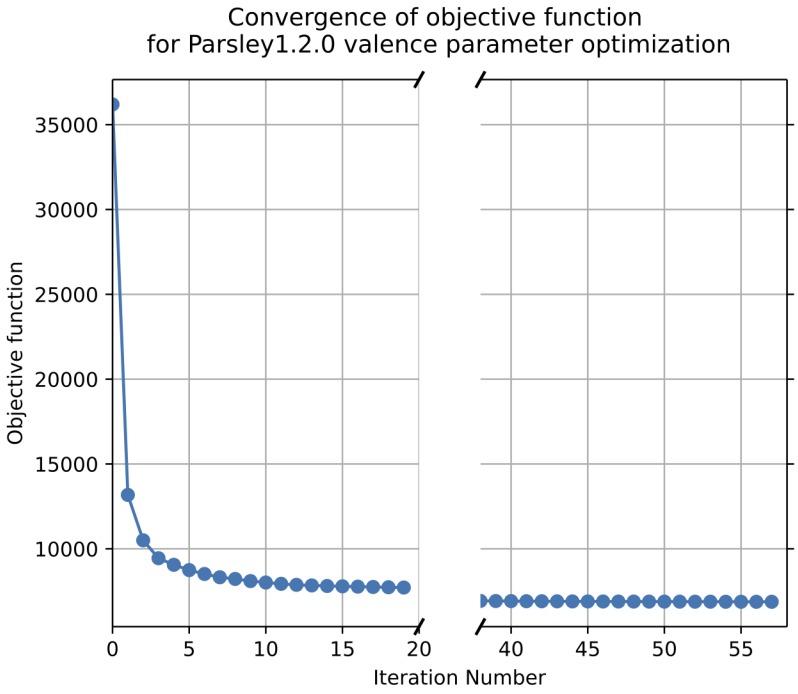


Figure 11: **Objective, or loss, function, as a function of number of ForceBalance iterations.**

The accuracy of the force field for the training data has increased dramatically during the parameter fitting process as anticipated. The dimensionless objective function – the weighted sum of squared differences between QM and MM values – decreased dramatically in the fitting, from 3.619e+04 to 6.877e+03 in 57 steps (Figure 11). The objective function is a sum of accuracies of optimized geometries, vibrational spectra, and torsion energy profiles. The summary of improvement in these components can be found in Table 5 and Figure 13, and the details are provided in the following subsections. Full fitting details can be found in the release package.[163]

		Training set			Full test set		
Data class		initial	final	% change	initial	final	% change
Geometry optimization	Bond lengths (RMSE, Angstrom)	0.0455	0.0220	-51.6 %	0.0401	0.0209	-47.8 %
	Bond angles (RMSE, degree)	4.15	2.41	-41.9 %	3.79	2.71	-28.5 %
	Improper dihedrals (RMSE, degree)	7.27	4.81	-33.9 %	6.83	4.31	-36.8 %
Vibrational spectra	Frequencies (RMSE, cm^{-1})	104.	33.8	-67.4 %	ND	ND	ND
Torsion energy profiles	Energies (RMSE, kcal/mol)	2.96	2.00	-32.4 %	ND	ND	ND
Relative energies	Energies (RMSE, kcal/mol)	ND	ND	ND	1.76	1.54	-12.3 %

Table 5: **Overall change in root-mean-squared error (RMSE) metrics vs. the quantum chemical result calculated for four types of properties, using the initial and optimized force field, for training set and test set.** ND = No Data.

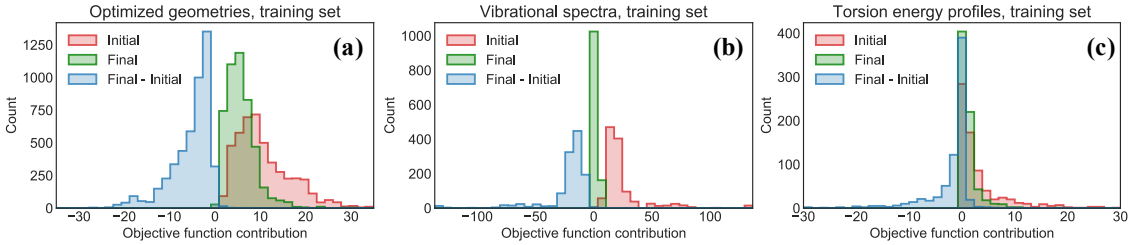


Figure 12: **Improvement in components of the training set.** Red histogram shows performance with our initial force field, green histogram shows performance with the optimized force field and blue histogram shows the distribution of changes in objective function contribution of each target (individual molecules/ geometries contributing to the objective function) due to the parameter optimization.

4.3.1.2 Optimized Geometries

For the geometric component, the objective function contribution measured the devi-

ation of the bond lengths, bond angles and improper torsion angles in MM optimized geometries from the values in corresponding QM optimized geometries. The objective function contribution for a single conformer can be given by

$$L_{\text{optgeo}}(\theta) = \sum_{i \in \text{ICs}} \left(\frac{x_i^{\text{MM}}(\theta) - x_i^{\text{QM}}}{d_i} \right)^2 \tag{22}$$

where θ is the force field parameters used in the MM optimization, d_i is a scaling factor of 0.05 Å, 8 degrees and 20 degrees for bond lengths, bond angles, and improper torsion angles, respectively. Please note that the deviations of proper torsion angles were not considered in the geometric component, to be considered comprehensively in torsional energy profiles.

The fitting led to overall improvement (Figure 12a, red to green) in the accuracy of the optimized geometries in the training set. In the blue histogram of improvements, the portion on the negative/ positive x-axis indicates the ratio of targets where accuracy is improved/ reduced, respectively. For a tiny portion of geometries, the accuracy was degraded, and this can be explained by compromises that had to be made for some geometries to improve the accuracy of other geometries which share the same parameters. Table 5 provides a physically interpretable perspective of these results, which shows that the RMSEs of bond-lengths, bond-angles, and improper torsion angles in the MM optimized geometries relative to the QM ones in the training set dropped by $34 \sim 52$ %.

Especially, molecules having deprotonated phosphonate group showed significant improvement in v1.2.0, fixing a known issue with protonated phosphorus-connected oxygens which have plagues AMBER-family force fields. In the optimized geometry from the v1.1.0 (transparent orange in the Figure 13), the hydroxyl hydrogen is located much closer to the negatively charged oxygen (1.10 Å) than in the QM opti-

mized geometry (> 2.4 Å), forming an overly strong intramolecular hydrogen bond. The v1.2.0 corrects this error and gives a closer agreement with QM optimized geometry by having a larger equilibrium angle for phosphorus-centered angle term (a38: [*:1] [#15:2] [*:3]).

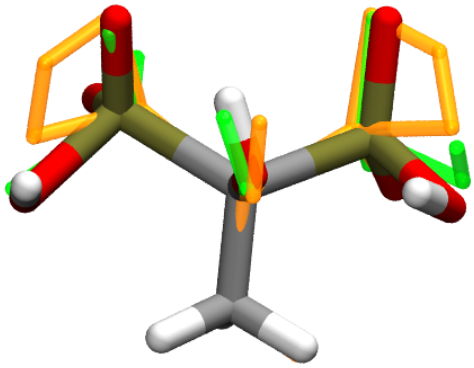


Figure 13: **QM optimized geometry of CC(O)(P@@(O)[O-])P@(O)[O-].** (Transparent orange: MM optimized geometry with v1.1.0, transparent green: v1.2.0. Gray: C; Red: O; White: H; Olive: P)

Benchmark analysis done by Victoria Lim[164] shows that v1.1.0 fails to describe (1) molecules having a single bond between divalent nitrogen and trivalent nitrogen (#7X2-#7X3); (2) azetidines; and (3) octahydrotetracenes. We speculated that the poor performances on the chemical moieties would be due to the poor chemical coverage of the training set data used for the previous fittings (< v1.2.0). In v1.1.0 release, periodicities for some nitrogen-nitrogen bond rotations were modified to properly describe the conjugated effect of the bonds. The parameter set change included the modification of the periodicity of t128. However, since there were no exocyclic divalent nitrogen-trivalent nitrogen rotations in the training set, a proper fitting of

t128 could not be carried out, resulting in inaccurate optimized force constants of the torsion parameter.

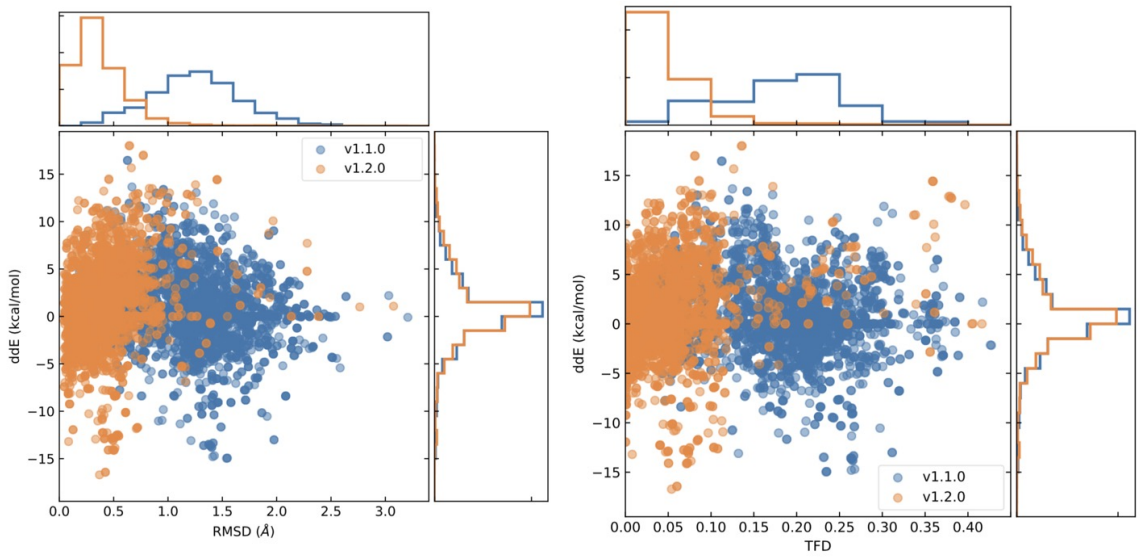


Figure 14: **ddE w.r.t. RMSD and ddE w.r.t. TFD of divalent and trivalent nitrogen-containing molecules.** ddE_i = [FF energy(i) – FF energy(0)] – [QM energy(i) - QM energy(0)], where 0th conformer is defined as the one having the lowest QM energy. Torsion fingerprint deviation (TFD) is a sum of Gaussian-weighted differences of torsion angles between two conformations.[165]

Our new training set data includes the exocyclic #7X2-#7X3 rotations and the new parameter set fitted to the training set showed clear improvement in lowering TFD and RMSD which indicates it is improved in reproducing QM optimized geometries while showing slight sacrifice in ddE (figure 14). Clear improvement in reproducing optimized geometries of exocyclic #7X2-#7X3 containing molecules can be visualized in an aligned overlay of QM optimized geometry and MM optimized geometries from v1.2.0 and older version. The overlay of optimized geometries of Cc1c(sc(n1)N/N=C/c2ccccn2)C (Figure 15) shows v1.2.0 successfully reproduces planar geometry of the molecule, fixing the bad performance in the older versions.

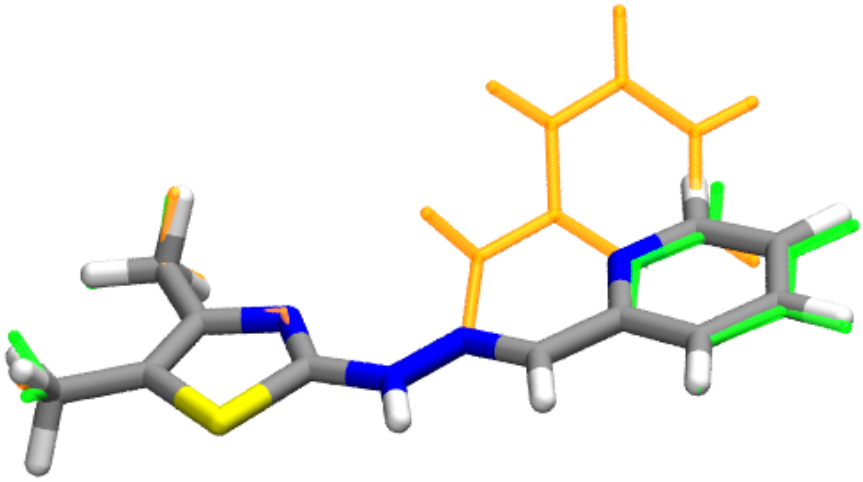


Figure 15: **QM optimized geometry of Cc1c(sc(n1)N/N=C/c2ccccn2)C.** (Transparent orange: MM optimized geometry with v1.1.0, transparent green: MM optimized geometry with v1.2.0, Gray: C; White: H; Blue: N; Yellow: S)

4.3.1.3 Vibrational Frequencies

For the vibrational component, QM and MM vibrational frequencies computed by normal mode analysis were sorted in ascending order to generate the sorted sequences $\nu_{QM,i}$ and $\nu_{MM,i}$, respectively. The objective function contribution measured the sum of squared differences of the corresponding frequencies, with a scaling factor of $d_{\text{vib}} = 200\text{cm}^{-1}$, which can be given by

$$L_{\text{vib}}(\theta) = \sum_i \left(\frac{\nu_{QM,i} - \nu_{MM,i}}{d_{\text{vib}}} \right)^2 \tag{23}$$

The fitting led to substantially improved accuracy of the vibrational frequencies in the training set (Figure 12b, red to green), and the improvement is more markable than that of optimized geometries (Figure 12a). The blue distribution in Fig 12b also shows substantial improvement and Table 5 shows the corresponding result of a 67.4 % drop in the RMSEs of MM vibrational frequencies relative to the QM ones (from

104 to 33.8 cm^{-1}).

4.3.1.4 Torsional energy profiles

For torsional energy profiles, to get the MM torsional energy profile, each structure along the QM torsional profile was partially relaxed using the force field with the harmonic restraint on each atom. During the restrained relaxations, the four atoms determining the torsion were fixed and harmonic restraint with force constant 1 kcal mol^{-1} $Å^{-2}$ was applied on each atom to avoid any large deviation of MM structures from the QM structures. The objective function contribution measured the weighted sum of squared differences between QM and MM energies along the torsional energy profiles.

$$L_{\mathrm{tor}}(\theta) = \frac{1}{{d_E}^2} \frac{\sum_{i \in N(\mathrm{gridpoints})} w\left(E_{\mathrm{QM}}(\mathbf{x}_i)\right)\left(E_{\mathrm{QM}}(\mathbf{x}_i) - E_{\mathrm{MM}}(\mathbf{x}_i;\theta)\right)^2}{\sum_{i \in N(\mathrm{gridpoints})} w\left(E_{\mathrm{QM}}(\mathbf{x}_i)\right)} \tag{24}$$

where $E(\mathbf{x}_i)$ is a relative energy at the grid point i, d_E is a scaling factor of 1.0 kcal mol^{-1}, and w is a weight of each grid point, which is calculated by an equation with two cutoffs, 1.0 kcal mol^{-1} and 5.0 kcal mol^{-1}.

$$w(E) = \begin{cases} 1 & E < 1.0\mathrm{kcal/mol} \\ \left(1 + (E-1)^2\right)^{-\frac{1}{2}} & 1.0 \leq E < 5.0\mathrm{kcal/mol} \\ 0 & E \geq 5.0\mathrm{kcal/mol} \end{cases} \tag{25}$$

The fitting led to improved accuracy of the torsional energy profiles in the training data (Figure 12c, red to green), although the improvement is less dramatic than for the other two components. Table 5shows that the RMSEs of MM torsional energy profiles relative to the QM ones decreased by 32.4 % (from 2.96 to 2.00 kcal mol^{-1}).

4.3.2 Test Set Result

Testing with data outside the training set gives an indication of the transferability of the new parameters and hence of the accuracy that may be expected in actual use. Here, we test the performance of the fitted parameters on optimized gas-phase geometries outside the training set, relative conformational energies of gas-phase molecules, using the same test set, we used for testing v1.0.0.

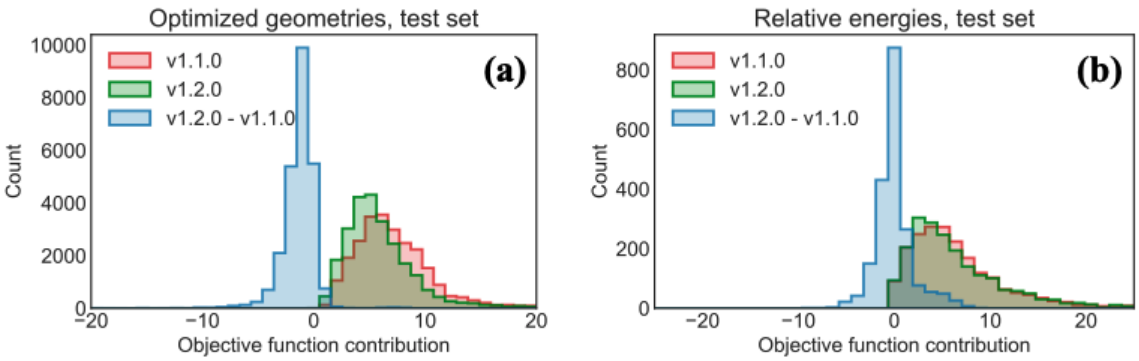


Figure 16: **Improvement in components of the test set.**

The overall objective function for the test set is lower for 1.2.0 (16712) than for the initial force field (29174), and v1.1.0 (20096). The distribution of improvements over the test set compounds shows the overall improvement in the accuracy of the MM optimized geometries in the test data set, relative to the reference QM results, while it shows no notable improvement in the accuracy of the MM relative conformer energies when compared to v1.1.0.

4.4 Conclusions and Directions

We described a methodology to generate OpenFF 1.2.0, code-name Parsley, a SMIRNOFF force field with bonded terms fitted against gas-phase QM reference data. With the careful design of the expanded training set and the proper modification of the param-

eter set, Parsley 1.2.0 provides more accurate molecular geometries in general over the older versions and especially shows significant improvement in optimized geometries of molecules having deprotonated phosphonate group and molecules having single bond between divalent nitrogen and trivalent nitrogen.

Iterative force field improvement has been included in the subsequent releases (OpenFF 1.2.1, 1.3.0, and 1.3.1), and it is noteworthy that v1.3 includes an important fix of amide-related issues; (1) poor performance of v1.2 in reproducing amide torsional energy profiles and (2) absence of appropriate torsion parameters for tertiary amides in v1.2.

We aim to extend the optimization to nonbonded interaction parameters, which is already underway. Also, we plan to address issues related to vibrational frequencies fitting. In vibrational frequencies fitting in ForceBalance, force field Hessians are computed by locally minimizing the QM geometries using the force field, followed by evaluation of forces with numerical displacements. it carries out a normal mode analysis and sorts the QM and MM frequencies from lowest to highest to yield the sorted sequences $\nu_{\mathrm{QM},j}$ and $\nu_{\mathrm{MM},j}$, respectively. The objective function contribution for each normal mode set is computed as the sum of squared differences between corresponding frequencies, scaled by a factor of $d_{\mathrm{vib}} = 200$ cm^{-1}, as:

$$L_{\mathrm{vib}}(\theta) = \sum_i \left(\frac{\nu_{\mathrm{QM},i} - \nu_{\mathrm{MM},i}}{d_{\mathrm{vib}}} \right)^2 \tag{26}$$

While the approach was chosen due to its computational efficiency, since QM and FF vibrational frequencies are sorted simply in ascending order, it carries a potential risk of mismatching between QM and MM vibrational modes. To remove the conceptual problem, the idea of replacing vibrational frequencies fitting to internal coordinate hessian fitting has been studied, and wait for a more in-depth tests and analyses.

Internal coordinate hessian fitting computes force field hessian in the same way as in the vibrational frequencies fitting, followed by converting QM and MM hessians into primitive redundant internal coordinates. The objective function contribution is computed as a difference between QM and MM internal coordinate Hessians. This approach is free from the matching issue, and the preliminary study showed that the force field optimized against the QM internal coordinate hessian gives a moderate improvement in replicating the QM vibrational frequencies and normal modes compared to the force field optimized against the QM vibrational frequencies.

A Supporting Information for Chapter 2: Quantum chemical studies of redox properties and conformational changes of a four-center iron CO_2 reduction electrocatalyst

State:	**1** (3–), Singlet		**2** (2–), Doublet	
IRC Method & Basis	BP86 6-31G*-LDZ	B3LYP 6-31G*-LDZ	BP86 6-31G*-LDZ	B3LYP 6-31G*-LDZ
Energies:	BP86/6-31G*-LDZ		BP86/6-31G*-LDZ	
Initial	**0.0**	0.0	**0.0**	0.0
TS	**18.8**	18.7	**22.5**	47.4
Final	**10.6**	10.7	**22.1**	46.9
	BP86/TZVP-LTZ		BP86/TZVP-LTZ	
Initial	0.0	0.0	0.0	0.0
TS	18.3	18.0	–	–
Final	10.6	10.0	22.7	49.2
	B3LYP/6-31G*-LDZ		B3LYP/6-31G*-LDZ	
Initial	0.0	**0.0**	0.0	**0.0**
TS	24.8	**25.3**	19.8	**25.1**
Final	11.9	**12.2**	19.6	**24.7**
	B3LYP/TZVP-LTZ		B3LYP/TZVP-LTZ	
Initial	0.0	0.0	0.0	0.0
TS	24.3	24.3	20.1	24.6
Final	11.7	11.8	20.0	23.9

Table S1: **Relative energies (kcal mol^{-1}) along CO dissociation pathway.** Each column refers to structures from an optimized IRC using the indicated method/basis. Each set of four rows contains energies calculated using the specified method/basis. Numbers in bold indicate energies taken directly from the optimized IRC; the other numbers are estimated using the IRCMax approach – i.e. taking differences of single-point energies along the pathway. Blank cells indicate that the final energy is the highest and there is no "TS" energy.

Note: Although B3LYP appears to predict a lower barrier for CO dissociation along with the BP86 IRC (highlighted in red), this result is misleading because the initial structure is not stable.

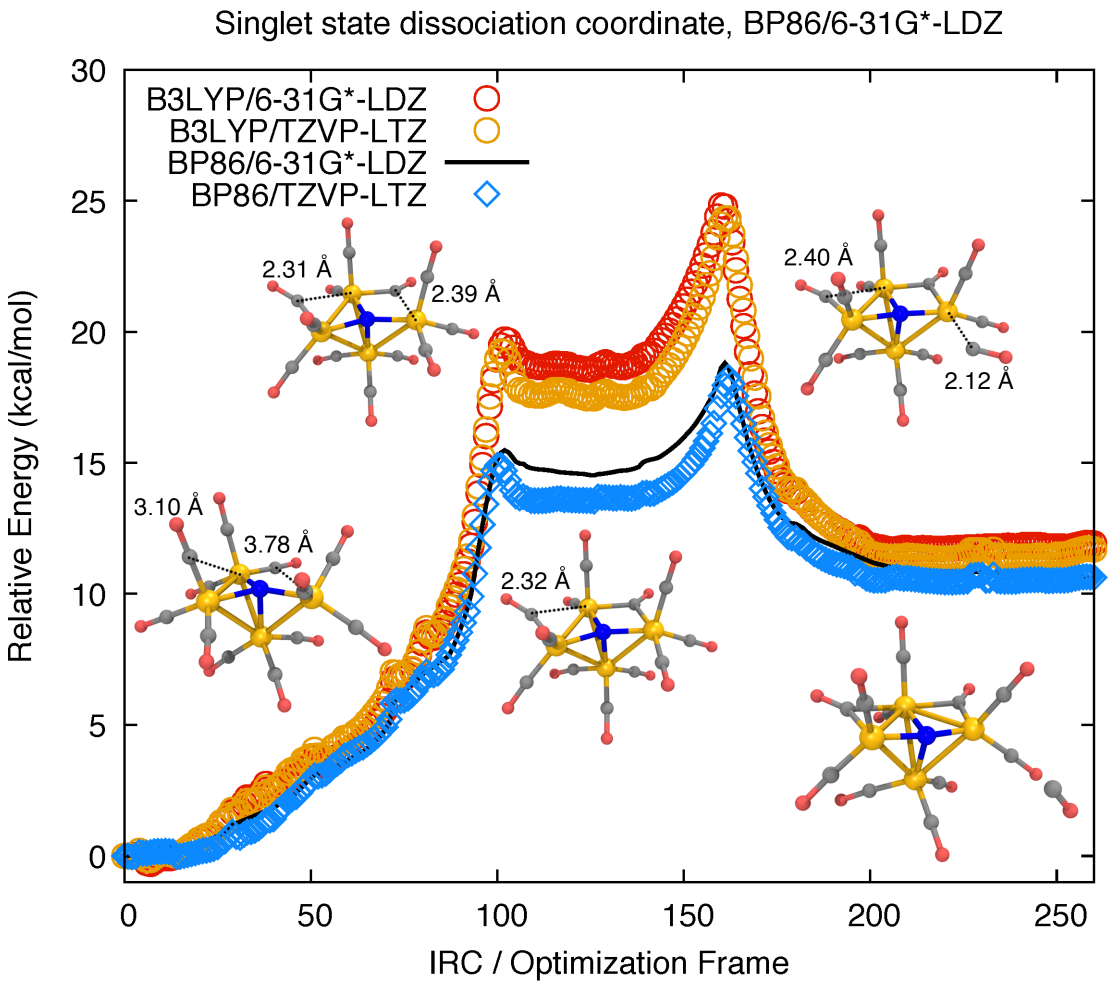


Figure S1: **Energies along the intrinsic reaction coordinate for dissociation of CO from 1^{3-} computed at the BP86/6-31G*-LDZ level (black), along with the energies computed along the path using different methods and basis sets (colored symbols).** Key Fe-C distances are labeled using dotted lines. The Fe-C bonds are drawn using a distance criterion of 2.1Å

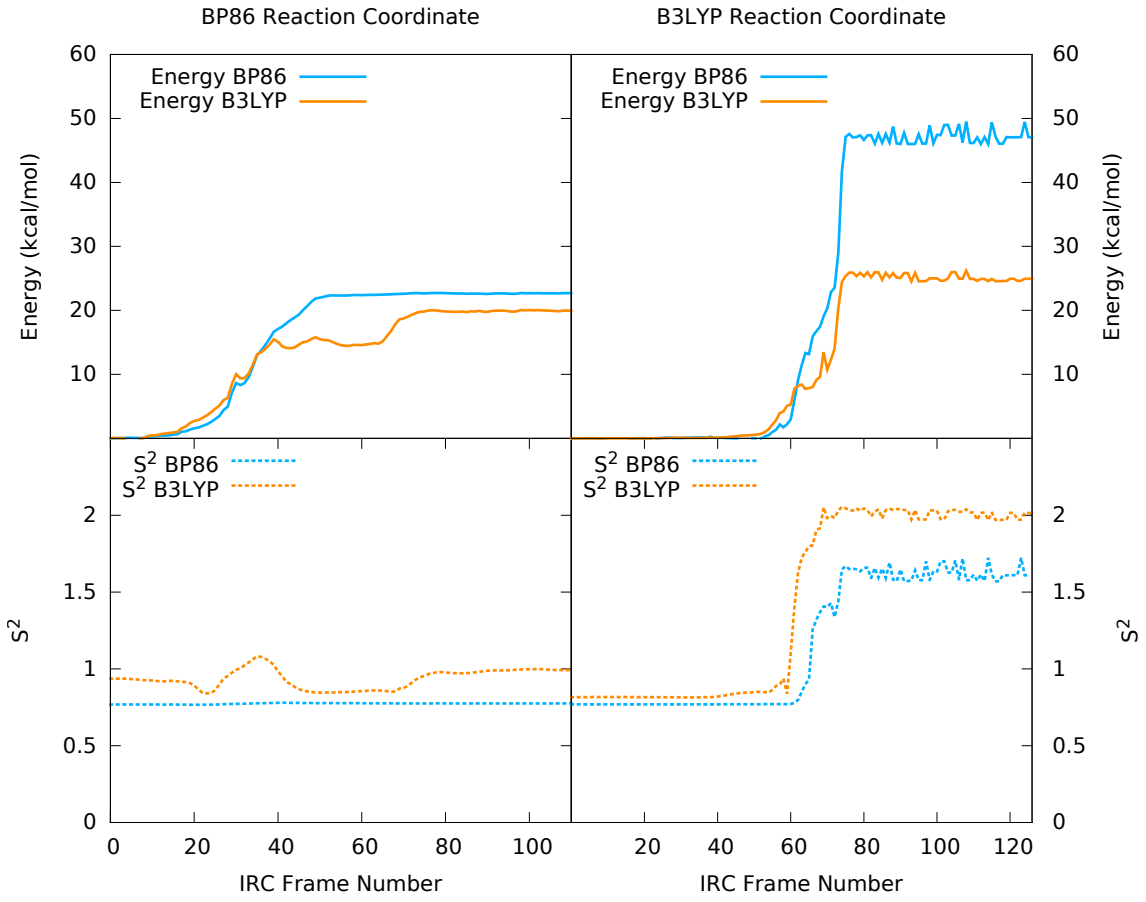


Figure S2: **Calculations of the relative energy and the expectation value of the S^2 operator for CO dissociation in 1^{2-} along the optimized IRC of BP86/6-31G*-LDZ (left column) and B3LYP/6-31G*-LDZ (right column).** The B3LYP calculations contain more spin contamination – i.e. $\langle S^2 \rangle$ greater than 0.75 – along both reaction pathways. Additionally, both BP86 and B3LYP calculations have significant spin contamination along the B3LYP IRC.

Note: Although B3LYP appears to predict a lower barrier for CO dissociation along with the BP86 IRC, the initial structure is not stable.

B Supporting Information for Chapter 3: *respyte*: Modernized Implementation of RESP for the Development of Next Generation of ESP-based Charge Model

Here Section B.1 details input/output formats and the options available for the grid point selection and QM calculation. Then Section B.2 describes the details of the charge fitting part, including the structure of the system, a short description of each class, and options available for the fitting part. An example of inputs and outputs can be found in `https://github.com/lpwgroup/respyte/tree/re-formatting/respyte/data/input.sample`.

B.1 Grid Point Selection and QM Calculation

Executable, 'respyte-esp_generator' takes an input directory, which consists of (1) 'molecules' directory, containing one or more sub-directories, each of which contains a molecule coordinate file in PDB or MOL2 file format, and (2) the input file, 'input.yml', in which the user specifies (a) the name, net charge and the number of conformers of each molecule; (b) the grid-point sampling scheme; the grid type (Merz-Singh-Kollman (MSK) grid, Face-Centered Cubic (FCC) grid), the inner and outer boundaries and spacing between grid points; and (c) the QM level of theory, inclusion/exclusion of a PCM description of the solvent. The input should follow the specific directory structure, which is shown in Figure S3.

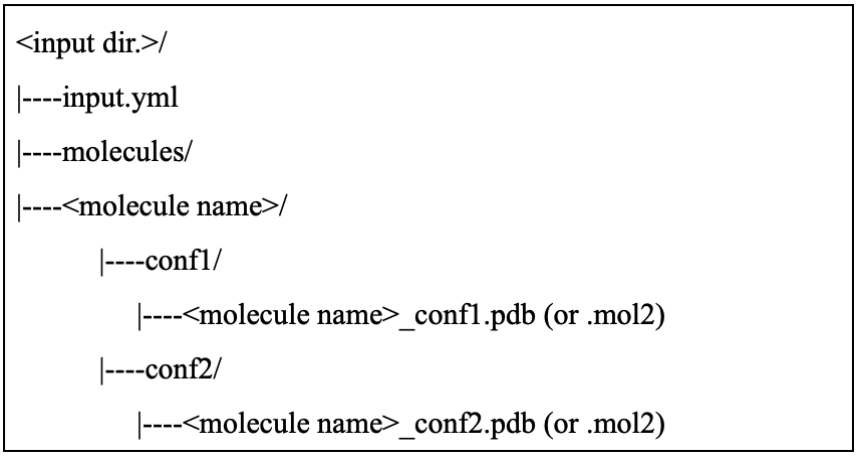


Figure S3: **Input directory structure for grid point selection and QM calculation.**

Once grid points are generated around each molecule, the 3D coordinate information of the generated points is saved in a Psi4 readable 'grid.dat' file. Then QM ESP and EF at each grid point are evaluated using the Psi4 package and saved in the output file (.espf extension), under each subdirectory.

B.2 Charge Fitting

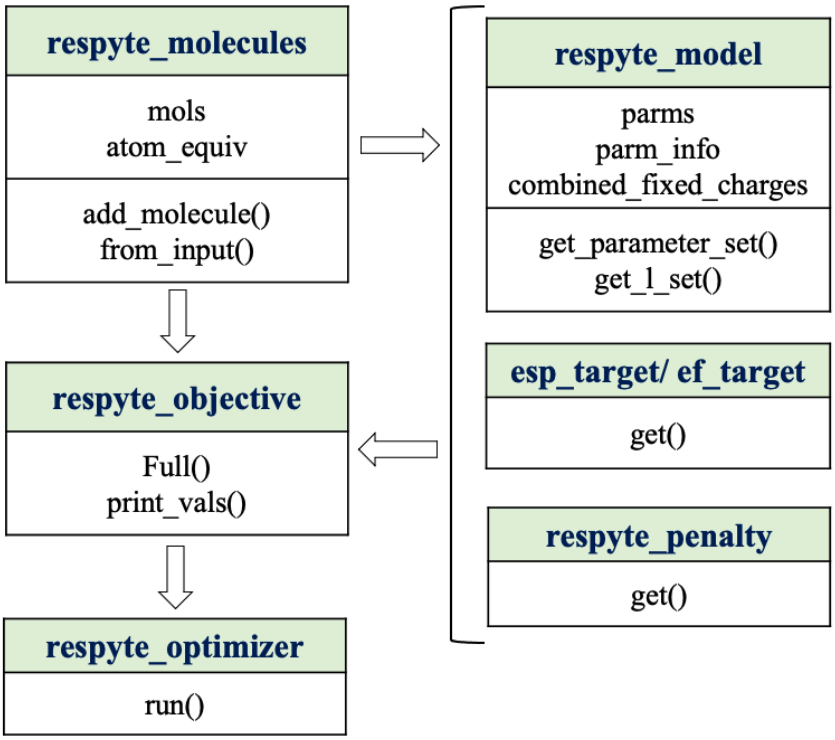


Figure S4: **Class diagrams of charge fitting part.** The middle and bottom compartments contain key attributes and operations of each class.

respyte_molecule class parses a coordinate file of a single conformation and identifies equivalent atoms in different equivalence levels, polar atoms, and polar hydrogens using RDKit. There are 5 different equivalence levels currently supported: 'connectivity', 'relaxed_connectivity', 'nosym', 'symbol', and 'symbol2'. In 'connectivity', it forces symmetry on chemically identical atoms, e.g., all hydrogens on a methyl group are forced to be equivalent in this level. In 'relaxed_connectivity', it forces symmetry on polar atoms only, i.e., two hydroxyl oxygens in ethylene glycol are considered equivalent, while the hydrogens on a methyl group are considered inequivalent to each other in the level. In 'nosym' (no forced symmetry), all individual atoms are considered inequivalent. In 'symbol', it forces symmetry based on chemical elements, e.g., all

carbons in a molecule are enforced to be the same. 'symbol2' forces the same symmetry with 'symbol' but separates polar and nonpolar H atoms. For a given molecule, it labels each atom with a unique integer, in the similar spirit as the atom-typing scheme in many traditional force fields, for each equivalence level. Also, it parses espf files and stores the QM properties.

respyte_molecules class combines one or more respyte_molecule objects and creates a single system. As a new respyte_molecule is added, it automatically re-labels all the atoms in the system so that it forces symmetry across molecules during the charge fitting. respyte_model class takes respyte_molecules object and builds a list of parameters to-be-fitted for the system. Currently supported charge models in the class are (1) the simple point charge model and (2) the fuzzy charge model. esp_target (ef_target) class takes the molecules and model objects and constructs an objective function, a sum of squared differences between QM ESP(EF) and MM ESP(EF). respyte_penalty class constructs a penalty function for the objective function. Two types of penalty functions are currently supported: hyperbolic function (L1) and harmonic function (L2). respyte_objective combines the target objects and the penalty objects to build an overall objective function of the system, evaluates the objective function, gradient, and hessian of the objective function at a given point in the parameter space. respyte_optimizer takes the respyte_objective object and runs the Newton-Raphson method using the SciPy package and returns the values once the convergence criterion is met.

The command line executable for this step is 'respyte-optimizer', which takes an input directory, which consists of (1) 'molecules' directory, containing one or more sub-directories, each of which contains a molecule coordinate file in pdb or mol2 file format and an espf file generated from the first step, and (2) the input file, 'respyte.yml', in which the user specifies (a) the charge model type (the simple point

charge model, the fuzzy charge model are currently supported), (b) targets: whether to use ESP, EF or both in the fitting, (c) the penalty function: the functional form, the restraint weight and the tightness of the hyperbola around the target value, (d) the equivalence level of each parameter type ('connectivity', 'relaxed_connectivity', 'nosym', 'symbol', and 'symbol2' are currently supported), (e) the residue charges, fixed atomic charges, (f) the equivalent atoms, etc. Once the optimization converges, under the 'resp_output' directory, a .txt file for each conformation with the optimized parameters is generated.

C Supporting Information for Chapter 4: Development of An Open Small Molecule Force Field

This document provides key additional details relating to the Parsley force field, including information on datasets and tools used in training and testing the force field as well as details on how to access these datasets and reproduce the calculations done in training and testing. Much of this information is provided in software/scripts available on GitHub and datasets available in QCArchive and elsewhere, as we detail below.

C.1 Compound Sets Used in Training

SMI and PDF of the training dataset is available at `https://github.com/openforcefield/openforcefield-forcebalance/releases/download/v1.2.0/training_unique_molecules.smi` and `https://github.com/openforcefield/openforcefield-forcebalance/releases/download/v1.2.0/training_unique_molecules.pdf`

SDF and PDF of Roche Optimization dataset is available at `https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-1-Roche/optimization_inputs.sdf` and `https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-1-Roche/optimization_inputs.pdf` . PDF of Roche Torsiondrive dataset is available at `https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-23-OpenFF-Gen-2-Torsion-Set-1-Roche-2/roche_2_selected_torsions.pdf`.

A full list of SMILE strings of the Coverage Optimization dataset is available at

https://github.com/openforcefield/openforcefield-forcebalance/raw/release-1/1_valence_parameter_fitting/1_dataset_generation/coverage_set/2019-06-25-smirnoff99Frost-coverage/chosen_supplemented.smi. SDF and PDF of Coverage Optimization dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-2-Coverage/optimization_inputs.sdf and https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-2-Coverage/optimization_inputs.pdf. PDF of Coverage Torsiondrive dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-23-OpenFF-Gen-2-Torsion-Set-2-Coverage-2/coverage_2_selected_torsions.pdf.

Initial automated selection of the Coverage Set is described in a subdirectory of the openforcefields GitHub repository, https://github.com/openforcefield/open-forcefield-data/tree/master/Utilize-All-Parameters, and additional molecules were added manually as described in https://github.com/openforcefield/open-forcefield-data/tree/master/Utilize-All-Parameters/supplement_molecules to cover remaining gaps.

SDF and PDF of Pfizer Discrepancy optmization dataset are available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-3-Pfizer-Discrepancy/optimization_inputs.sdf andhttps://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-3-Pfizer-Discrepancy/optimization_inputs.pdf. PDF of Pfizer Discrepancy Torsiondrive dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-23-OpenFF

-Gen-2-Torsion-Set-3-Pfizer-Discrepancy-2/pfizer_2_selected_torsions.pdf.

SDF and PDF of eMolecules Discrepancy optmization dataset are available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-4-eMolecules-Discrepancy/optimization_inputs.sdf andhttps://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-4-eMolecules-Discrepancy/optimization_inputs.pdf. PDF of eMolecules Discrepancy Torsiondrive dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-23-OpenFF-Gen-2-Torsion-Set-4-eMolecules-Discrepancy-2/emolecules_2_selected_torsions.pdf.

SDF and PDF of Bayer optmization dataset are available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-5-Bayer/optimization_inputs.sdf andhttps://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-5-Bayer/optimization_inputs.pdf. PDF of Bayer Torsiondrive dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-26-OpenFF-Gen-2-Torsion-Set-5-Bayer-2/bayer_2_selected_torsions.pdf.

PDF of supplemental Torsiondrive dataset is available at https://github.com/openforcefield/qca-dataset-submission/blob/master/submissions/2020-03-26-OpenFF-Gen-2-Torsion-Set-6-supplemental-2/supplemental_2_selected_torsions.pdf.

C.2 Generation of Quantum Chemical Data for Compound Datasets

An example of working with several QCArchive datasets is available at `https://github.com/openforcefield/openforcefield-forcebalance/raw/release-1/2_benchmarking/test_dataset_generation/divide_sets.ipynb`. Details of molecules which were removed can be found in `fb-fit/targets/error_mol2s` in the release package available at `github.com/openforcefield/openforcefield-forcebalance/releases/tag/v1.2.0`.

Details of how to download a dataset from the QCArchive server, filter, and generate ForceBalance-readable targets can be found here: `https://github.com/openforcefield/openforcefield-forcebalance/raw/release-1/utils/optimized_geo/make_fb_optgeo_target.py`

Details of how to download a dataset from the QCArchive server, filter, and generate ForceBalance-readable targets can be found here: `https://github.com/openforcefield/openforcefield-forcebalance/raw/release-1/utils/vib_freq_target/make_vib_freq_target.py`

Details of how to download a dataset from the QCArchive server, filter, and generate ForceBalance-readable targets can be found here: `https://github.com/openforcefield/openforcefield-forcebalance/raw/release-1/utils/torsion_target/make_torsion_target_new.py`

References

[1]M. Peplow, Chemistry's grand challenges, (accessed: 07.17.2021).

[2]V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis, and E. Kazlauskas, "Binding affinity in drug design: experimental and computational techniques", Expert Opin. Drug Discov. **14**, PMID: 31146609, 755–768 (2019).

[3]F. Jensen, Introduction to computational chemistry, 2nd ed. (John Wiley Sons, 2007).

[4]C. David Sherrill and H. F. Schaefer, "The configuration interaction method: advances in highly correlated approaches", in, Vol. 34, edited by P.-O. Löwdin, J. R. Sabin, M. C. Zerner, and E. Brändas, Advances in Quantum Chemistry (Academic Press, 1999), pp. 143–269.

[5]T. D. Crawford and H. F. Schaefer III, "An introduction to coupled cluster theory for computational chemists", in Reviews in computational chemistry (John Wiley Sons, Ltd, 2000), pp. 33–136.

[6]C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems", Phys. Rev. **46**, 618–622 (1934).

[7]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Phys. Rev. **140**, A1133–A1138 (1965).

[8]U. von Barth and L. Hedin, "A local exchange-correlation potential for the spin polarized case. i", Journal of Physics C: Solid State Physics **5**, 1629–1642 (1972).

[9]D. Langreth and J. Perdew, "The gradient approximation to the exchange-correlation energy functional: a generalization that works", Solid State Communications **31**, 567–571 (1979).

[10]D. C. Langreth and M. J. Mehl, "Beyond the local-density approximation in calculations of ground-state electronic properties", Phys. Rev. B **28**, 1809–1834 (1983).

[11]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", Phys. Rev. A **38**, 3098–3100 (1988).

[12]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", Phys. Rev. Lett. **77**, 3865–3868 (1996).

[13]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", Phys. Rev. A **38**, 3098–3100 (1988).

[14]C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density", Phys. Rev. B **37**, 785–789 (1988).

[15]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", Phys. Rev. Lett. **77**, 3865–3868 (1996).

[16]A. D. Becke, "A new inhomogeneity parameter in density-functional theory", J. Chem. Phys. **109**, 2092–2098 (1998).

[17]J. P. Perdew, S. Kurth, A. š. Zupan, and P. Blaha, "Accurate density functional with correct formal properties: a step beyond the generalized gradient approximation", Phys. Rev. Lett. **82**, 2544–2547 (1999).

[18]J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, "Climbing the density functional ladder: nonempirical meta–generalized gradient approximation designed for molecules and solids", Phys. Rev. Lett. **91**, 146401 (2003).

[19]J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, "Workhorse semilocal density functional for condensed matter physics and quantum chemistry", Phys. Rev. Lett. **103**, 026403 (2009).

[20]J. Sun, B. Xiao, and A. Ruzsinszky, "Communication: effect of the orbital-overlap dependence in the meta generalized gradient approximation", J. Chem. Phys. **137**, 051101 (2012).

[21]J. Sun, R. Haunschild, B. Xiao, I. W. Bulik, G. E. Scuseria, and J. P. Perdew, "Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence", J. Chem. Phys. **138**, 044113 (2013).

[22]A. D. Becke, "Density-functional thermochemistry. iii. the role of exact exchange", J. Chem. Phys. **98**, 5648–5652 (1993).

[23]J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations", J. Chem. Phys. **105**, 9982–9985 (1996).

[24]M. Ernzerhof and G. E. Scuseria, "Assessment of the perdew–burke–ernzerhof exchange-correlation functional", J. Chem. Phys. **110**, 5029–5036 (1999).

[25]C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: the pbe0 model", J. Chem. Phys. **110**, 6158–6170 (1999).

[26]F. Furche, "Molecular tests of the random phase approximation to the exchange-correlation energy functional", Phys. Rev. B **64**, 195120 (2001).

[27]J. Harl, L. Schimka, and G. Kresse, "Assessing the quality of the random phase approximation for lattice constants and atomization energies of solids", Phys. Rev. B **81**, 115126 (2010).

[28]M. Tachikawa and E. L. Muetterties, "Metal clusters. 25. a uniquely bonded c-h group and reactivity of a low-coordinate carbidic carbon atom", J. Am. Chem. Soc. **102**, 4541–4542 (1980).

[29]M. Tachikawa, J. Stein, E. L. Muetterties, R. G. Teller, M. A. Beno, E. Gebert, and J. M. Williams, "Metal clusters with exposed and low-coordinate nitride nitrogen atoms", J. Am. Chem. Soc. **102**, 6648–6649 (1980).

[30]M. D. Rail and L. A. Berben, "Directing the reactivity of [hfe4n(co)12] toward h+ or co2 reduction by understanding the electrocatalytic mechanism", J. Am. Chem. Soc. **133**, PMID: 22032761, 18577–18579 (2011).

[31]H. Jang, Y. Qiu, M. E. Hutchings, M. Nguyen, L. A. Berben, and L.-P. Wang, "Quantum chemical studies of redox properties and conformational changes of a four-center iron co2 reduction electrocatalyst", Chem. Sci. **9**, 2645–2654 (2018).

[32]C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model", J. Phys. Chem. **97**, 10269–10280 (1993).

[33]Y. Qiu, D. Smith, S. Boothroyd, H. Jang, J. Wagner, C. C. Bannan, and T. e. a. Gokey, "Development and benchmarking of open force field v1.0.0, the parsley small molecule force field", This content is a preprint and has not been peer-reviewed., `10.26434/chemrxiv.13082561.v2` (2020).

[34]T. P. Senftle and E. A. Carter, "The holy grail: chemistry enabling an economically viable co2 capture, utilization, and storage strategy", Acc. Chem. Res. **50**, PMID: 28945424, 472–475 (2017).

[35]Y. Hori, A. Murata, and R. Takahashi, "Formation of hydrocarbons in the electrochemical reduction of carbon dioxide at a copper electrode in aqueous solution", J. Chem. Soc. Faraday Trans. 1 **85**, 2309–2326 (1989).

[36]X. Feng, K. Jiang, S. Fan, and M. W. Kanan, "Grain-boundary-dependent co2 electroreduction activity", J. Am. Chem. Soc. **137**, PMID: 25835085, 4606–4609 (2015).

[37]X. Min and M. W. Kanan, "Pd-catalyzed electrohydrogenation of carbon dioxide to formate: high mass activity at low overpotential and identification of the deactivation pathway", J. Am. Chem. Soc. **137**, PMID: 25812119, 4701–4708 (2015).

[38]M. Asadi, K. Kim, C. Liu, A. V. Addepalli, P. Abbasi, P. Yasaei, P. Phillips, A. Behranginia, J. M. Cerrato, R. Haasch, P. Zapol, B. Kumar, R. F. Klie, J. Abiade, L. A. Curtiss, and A. Salehi-Khojin, "Nanostructured transition metal dichalcogenide electrocatalysts for co2 reduction in ionic liquid", Science **353**, 10.1126/science.aaf4767 (2016).

[39]D. R. Kauffman, D. Alfonso, C. Matranga, H. Qian, and R. Jin, "Experimental and computational investigation of au25 clusters and co2: a unique interaction and enhanced electrocatalytic activity", J. Am. Chem. Soc. **134**, PMID: 22616945, 10237–10243 (2012).

[40]S. Zhang, P. Kang, and T. J. Meyer, "Nanostructured tin catalysts for selective electrochemical reduction of carbon dioxide to formate", J. Am. Chem. Soc. **136**, PMID: 24417470, 1734–1737 (2014).

[41]J. Rosen, G. S. Hutchings, Q. Lu, S. Rivera, Y. Zhou, D. G. Vlachos, and F. Jiao, "Mechanistic insights into the electrochemical reduction of co2 to co on nanostructured ag surfaces", ACS Catalysis **5**, 4293–4299 (2015).

[42]J.-M. Savéant, "Molecular catalysis of electrochemical reactions. mechanistic aspects", Chem. Rev. **108**, PMID: 18620367, 2348–2378 (2008).

[43]E. E. Benson, C. P. Kubiak, A. J. Sathrum, and J. M. Smieja, "Electrocatalytic and homogeneous approaches to conversion of co2 to liquid fuels", Chem. Soc. Rev. **38**, 89–99 (2009).

[44]C. Costentin, M. Robert, and J.-M. Savéant, "Catalysis of the electrochemical reduction of carbon dioxide", Chem. Soc. Rev. **42**, 2423–2436 (2013).

[45]A. D. Nguyen, M. D. Rail, M. Shanmugam, J. C. Fettinger, and L. A. Berben, "Electrocatalytic hydrogen evolution from water by a series of iron carbonyl clusters", Inorg. Chem. **52**, PMID: 24116898, 12847–12854 (2013).

[46]A. Taheri, E. J. Thompson, J. C. Fettinger, and L. A. Berben, "An iron electrocatalyst for selective reduction of co2 to formate in water: including thermochemical insights", ACS Catal. **5**, 7140–7151 (2015).

[47]A. Taheri, N. D. Loewen, D. B. Cluff, and L. A. Berben, "Considering a possible role for [h-fe4n(co)12]2– in selective electrocatalytic co2 reduction to formate by [fe4n(co)12]", Organometallics **37**, 1087–1091 (2018).

[48]H. Wang, Y. Xie, R. B. King, and H. F. Schaefer, "Unsaturation in binuclear cyclopentadienyliron carbonyls", Inorg. Chem. **45**, PMID: 16602798, 3384–3392 (2006).

[49]H. Wang, Y. Xie, R. B. King, and H. F. Schaefer, "Unsaturation in binuclear cyclobutadiene iron carbonyls: triplet structures, four-electron bridging carbonyl groups, and perpendicular structures", Organometallics **27**, 3113–3123 (2008).

[50]H. Wang, Y. Xie, R. B. King, and H. F. Schaefer, "Binuclear iron carbonyl nitrosyls: bridging nitrosyls versus bridging carbonyls", Inorg. Chem. **47**, PMID: 18335979, 3045–3055 (2008).

[51]Z. Zhang, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer, "Trinuclear iron carbonyl thiocarbonyls: the preference for four- and six-electron donor bridging thiocarbonyl groups over metalmetal multiple bonding, while satisfying the 18-electron rule", Inorg. Chem. **48**, PMID: 19472988, 6167–6177 (2009).

[52]Z. Zhang, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer, "Iron carbonyl thiocarbonyls: effect of substituting a thiocarbonyl group for a carbonyl group in mononuclear and binuclear iron carbonyl derivatives", Inorg. Chem. **48**, 1974–1988 (2009).

[53]H. Wang, Z. Sun, Y. Xie, R. B. King, and H. F. Schaefer, "Unsaturation and variable hapticity in binuclear azulene iron carbonyl complexes", Organometallics **29**, 630–641 (2010).

[54]L. Xu, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer, "Prospects for making organometallic compounds with bf ligands: fluoroborylene iron carbonyls", Inorg. Chem. **49**, PMID: 20041690, 1046–1055 (2010).

[55]L. Xu, Q.-s. Li, R. B. King, and H. F. Schaefer, "Coupling of fluoroborylene ligands to give a viable cyclopentadienyliron carbonyl complex of difluorodiborene (fbbf)", Organometallics **30**, 5084–5087 (2011).

[56]Y. Zeng, S. Wang, H. Feng, Y. Xie, R. B. King, and H. F. Schaefer III, "Open chains versus closed rings: comparison of binuclear butadiene iron carbonyls with their cyclobutadiene analogues", New J. Chem. **35**, 920–929 (2011).

[57]J. Chen, S. Chen, and L. e. a. Zhong, "Binuclear dimethylaminoborole iron carbonyls: iron–iron multiple bonding versus nitrogen → iron dative bonding", Theor. Chem. Acc. **131**, 10.1007/s00214-012-1090-5 (2012).

[58]Z. Zhang, Q.-s. Li, R. B. King, and H. F. Schaefer III, "New structural features in tetranuclear iron carbonyl thiocarbonyls: exotriangular iron atoms and six-electron-

donating thiocarbonyl groups bridging four iron atoms", Eur. J. Inorg. Chem. **2012**, 1104–1113 (2012).

[59]Q. Fan, H. Feng, W. Sun, Y. Xie, R. B. King, and H. F. Schaefer, "The umbrella-shaped trimethylenemethane ligand in iron carbonyl chemistry: comparison with butadiene and cyclobutadiene analogues", Organometallics **31**, 3610–3619 (2012).

[60]R. Jin, X. Chen, Q. Du, H. Feng, Y. Xie, R. B. King, and H. F. Schaefer, "Nine-electron donor bridging indenyl ligands in binuclear iron carbonyls", Organometallics **31**, 5005–5017 (2012).

[61]R. Jin, X. Chen, Q. Du, H. Feng, Y. Xie, R. B. King, and H. F. Schaefer, "Binuclear iron carbonyl complexes of thialene", RSC Adv. **6**, 82661–82668 (2016).

[62]X. Gong, L. Zhu, J. Yang, X. Gao, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer III, "From spiropentane to butterfly and tetrahedral structures in tetranuclear iron carbonyl carbide chemistry", New J. Chem. **38**, 3762–3769 (2014).

[63]H. Li, H. Feng, W. Sun, Y. Xie, R. B. King, and H. F. Schaefer, "Alkyne dichotomy: splitting of bis(dialkylamino)acetylenes, dimethoxyacetylene, bis(methylthio)acetylene, and their heavier congeners to give carbyne ligands in iron carbonyl derivatives", Organometallics **32**, 88–94 (2013).

[64]G. Li, L. Zhou, X. Zhai, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer III, "Binuclear methylaminobis(difluorophosphine) iron carbonyls: phosphorus–nitrogen bond cleavage in preference to iron–iron multiple bond formation", New J. Chem. **37**, 3294–3302 (2013).

[65]J. Deng, Q.-s. Li, Y. Xie, R. B. King, and H. F. Schaefer III, "Binuclear hexafluorocyclopentadiene iron carbonyls: bis(dihapto) versus trihapto–monohapto bonding in iron–iron bonded structures", New J. Chem. **37**, 2902–2910 (2013).

[66]T. T. Shi, Q.-S. Li, Y. Xie, R. B. King, and H. F. Schaefer III, "Neutral homoleptic tetranuclear iron carbonyls: why haven't they been synthesized as stable molecules?", New J. Chem. **34**, 208–214 (2010).

[67]H. Wang, Y. Xie, R. B. King, and H. F. Schaefer, "Remarkable aspects of unsaturation in trinuclear metal carbonyl clusters: the triiron species fe3(co)n (n = 12, 11, 10, 9)", J. Am. Chem. Soc. **128**, PMID: 16939260, 11376–11384 (2006).

[68]G. Wang, J. Cui, C. Chi, X. Zhou, Z. H. Li, X. Xing, and M. Zhou, "Bonding in homoleptic iron carbonyl cluster cations: a combined infrared photodissociation spectroscopic and theoretical study", Chem. Sci. **3**, 3272–3279 (2012).

[69]C. Chi, J. Cui, Z. H. Li, X. Xing, G. Wang, and M. Zhou, "Infrared photodissociation spectra of mass selected homoleptic dinuclear iron carbonyl cluster anions in the gas phase", Chem. Sci. **3**, 1698–1706 (2012).

[70]M. Mirmohades, S. Pullen, M. Stein, S. Maji, S. Ott, L. Hammarström, and R. Lomoth, "Direct observation of key catalytic intermediates in a photoinduced proton reduction cycle with a diiron carbonyl complex", J. Am. Chem. Soc. **136**, PMID: 25419868, 17366–17369 (2014).

[71]S. Kuppuswamy, J. D. Wofford, C. Joseph, Z.-L. Xie, A. K. Ali, V. M. Lynch, P. A. Lindahl, and M. J. Rose, "Structures, interconversions, and spectroscopy of iron carbonyl clusters with an interstitial carbide: localized metal center reduction by overall cluster oxidation", Inorg. Chem. **56**, PMID: 28441025, 5998–6012 (2017).

[72]M.-H. Baik and R. A. Friesner, "Computing redox potentials in solution: density functional theory as a tool for rational design of redox agents", J. Phys. Chem. A **106**, 7407–7412 (2002).

[73]A. V. Marenich, J. Ho, M. L. Coote, C. J. Cramer, and D. G. Truhlar, "Computational electrochemistry: prediction of liquid-phase reduction potentials", Phys. Chem. Chem. Phys. **16**, 15068–15106 (2014).

[74]L.-P. Wang and T. Van Voorhis, "A polarizable qm/mm explicit solvent model for computational electrochemistry in water", J. Chem. Theory Comput. **8**, PMID: 26596609, 610–617 (2012).

[75]J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas", Phys. Rev. B **33**, 8822–8824 (1986).

[76]L.-P. Wang, R. T. McGibbon, V. S. Pande, and T. J. Martinez, "Automated discovery and refinement of reactive molecular dynamics pathways", J. Chem. Theory Comput. **12**, PMID: 26683346, 638–649 (2016).

[77]M. S. Gordon, M. W. Schmidt, G. M. Chaban, K. R. Glaesemann, W. J. Stevens, and C. Gonzalez, "A natural orbital diagnostic for multiconfigurational character in correlated wave functions", J. Chem. Phys. **110**, 4199–4207 (1999).

[78]P. Pulay and T. P. Hamilton, "Uhf natural orbitals for defining and starting mc-scf calculations", J. Chem. Phys. **88**, 4926–4933 (1988).

[79]R. G. Parr and W. Yang, "Density-functional theory of the electronic structure of molecules", Annu. Rev. Phys. Chem. **46**, PMID: 24341393, 701–728 (1995).

[80]A. Klamt, "Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena", J. Phys. Chem. **99**, 2224–2235 (1995).

[81]I. S. Ufimtsev and T. J. Martínez, "Quantum chemistry on graphical processing units. 1. strategies for two-electron integral evaluation", J. Chem. Theory Comput. **4**, PMID: 26620654, 222–231 (2008).

[82]I. S. Ufimtsev and T. J. Martinez, "Quantum chemistry on graphical processing units. 3. analytical energy gradients, geometry optimization, and first principles molecular dynamics", J. Chem. Theory Comput. **5**, PMID: 26631777, 2619–2628 (2009).

[83]I. S. Ufimtsev and T. J. Martinez, "Quantum chemistry on graphical processing units. 2. direct self-consistent-field implementation", J. Chem. Theory Comput. **5**, PMID: 26609609, 1004–1015 (2009).

[84]C. Song, L.-P. Wang, T. Sachse, J. Preiß, M. Presselt, and T. J. Martínez, "Efficient implementation of effective core potential integrals and gradients on graphical processing units", J. Chem. Phys. **143**, 014114 (2015).

[85]C. Song, L.-P. Wang, and T. J. Martínez, "Automated code engine for graphical processing units: application to the effective core potential integrals and gradients", J. Chem. Theory Comput. **12**, PMID: 26586267, 92–106 (2016).

[86]F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez, "Quantum chemistry for solvated molecules on graphical processing units using polarizable continuum models", J. Chem. Theory Comput. **11**, PMID: 26575750, 3131–3144 (2015).

[87]L.-P. Wang and C. Song, "Geometry optimization made simple with translation and rotation coordinates", J. Chem. Phys. **144**, 214108 (2016).

[88]H. Reiss and A. Heller, "The absolute potential of the standard hydrogen electrode: a new estimate", J. Phys. Chem. **89**, 4207–4213 (1985).

[89]W. A. Donald, R. D. Leib, J. T. O'Brien, M. F. Bush, and E. R. Williams, "Absolute standard hydrogen electrode potential measured by reduction of aqueous nanodrops in the gas phase", J. Am. Chem. Soc. **130**, PMID: 18288835, 3371–3381 (2008).

[90]F. Furche and J. P. Perdew, "The performance of semilocal and hybrid density functionals in 3d transition-metal chemistry", J. Chem. Phys. **124**, 044103 (2006).

[91]C. J. Cramer and D. G. Truhlar, "Density functional theory for transition metals and transition metal chemistry", Phys. Chem. Chem. Phys. **11**, 10757–10816 (2009).

[92]J. M. Galbraith and H. F. Schaefer, "Concerning the applicability of density functional methods to atomic and molecular negative ions", J. Chem. Phys. **105**, 862–864 (1996).

[93]A. A. Jarecki and E. R. Davidson, "Density functional theory calculations for f", Chem. Phys. Lett. **300**, 44–52 (1999).

[94]F. Jensen, "Describing anions by density functional theory: fractional electron affinity", J. Chem. Theory Comput. **6**, PMID: 26616074, 2726–2735 (2010).

[95]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[96]F. Weigend, "Accurate coulomb-fitting basis sets for h to rn", Phys. Chem. Chem. Phys. **8**, 1057–1065 (2006).

[97]L. E. Roy, P. J. Hay, and R. L. Martin, "Revised basis sets for the lanl effective core potentials", J. Chem. Theory Comput. **4**, PMID: 26636355, 1029–1031 (2008).

[98]J. Zheng, X. Xu, and D. G. Truhlar, "Minimally augmented karlsruhe basis sets", Theor. Chem. Acc. **128**, 295–305 (2011).

[99]A. Laio and M. Parrinello, "Escaping free-energy minima", PNAS **99**, 12562–12566 (2002).

[100]J. Pfaendtner and M. Bonomi, "Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics", J. Chem. Theory Comput. **11**, PMID: 26574304, 5062–5067 (2015).

[101]F. Pietrucci and W. Andreoni, "Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale", Phys. Rev. Lett. **107**, 085504 (2011).

[102]L. Wang and R. e. a. Titov A.and McGibbon, "Discovering chemistry with an ab initio nanoreactor", Nat. Chem. **6**, 1044–1048 (2014).

[103]L. Xie, Q. Zhao, K. F. Jensen, and H. J. Kulik, "Direct observation of early-stage quantum dot growth mechanisms with high-temperature ab initio molecular dynamics", J. Phys. Chem. C **120**, 2472–2483 (2016).

[104]N. Goldman, E. J. Reed, L. E. Fried, I.-F. William Kuo, and A. Maiti, "Synthesis of glycine-containing complexes in impacts of comets on early earth", Nat. Chem. **2**, 949–954 (2010).

[105]W. J. Hehre, R. Ditchfield, and J. A. Pople, "Self—consistent molecular orbital methods. xii. further extensions of gaussian—type basis sets for use in molecular orbital studies of organic molecules", J. Chem. Phys. **56**, 2257–2261 (1972).

[106]P. J. Hay and W. R. Wadt, "Ab initio effective core potentials for molecular calculations. potentials for k to au including the outermost core orbitals", J. Chem. Phys. **82**, 299–310 (1985).

[107]I. Mayer, "Bond order and valence indices: a personal account", J. Comput. Chem. **28**, 204–221 (2007).

[108]U. Bozkaya, J. M. Turney, Y. Yamaguchi, and H. F. Schaefer, "The lowest-lying electronic singlet and triplet potential energy surfaces for the hno–noh system: energetics, unimolecular rate constants, tunneling and kinetic isotope effects for the isomerization and dissociation reactions", J. Chem. Phys. **136**, 164303 (2012).

[109]Y. Zhao and D. G. Truhlar, "A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions", J. Chem. Phys. **125**, 194101 (2006).

[110]Y. Zhao and D. G. Truhlar, "The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals", Theor. Chem. Acc. **120**, 215–241 (2008).

[111]A. Schäfer, C. Huber, and R. Ahlrichs, "Fully optimized contracted gaussian basis sets of triple zeta valence quality for atoms li to kr", J. Chem. Phys. **100**, 5829–5835 (1994).

[112]A. Schäfer, H. Horn, and R. Ahlrichs, "Fully optimized contracted gaussian basis sets for atoms li to kr", J. Chem. Phys. **97**, 2571–2577 (1992).

[113]D. K. Malick, G. A. Petersson, and J. A. Montgomery, "Transition states for chemical reactions i. geometry and classical barrier height", J. Chem. Phys. **108**, 5704–5713 (1998).

[114]D. Schweinfurth, M. G. Sommer, M. Atanasov, S. Demeshko, S. Hohloch, F. Meyer, F. Neese, and B. Sarkar, "The ligand field of the azido ligand: insights into bonding parameters and magnetic anisotropy in a co(ii)–azido complex", J. Am. Chem. Soc. **137**, PMID: 25588991, 1993–2005 (2015).

[115]F. Neese, "The orca program system", Wiley Interdiscip. Rev. Comput. Mol. Sci. **2**, 73–78 (2012).

[116]D. R. Hartree, "The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods", Math. Proc. Cambridge Philos. **24**, 89–110 (1928).

[117]J. C. Slater, "The self consistent field and the structure of atoms", Phys. Rev. **32**, 339–348 (1928).

[118]H. A. Carlson, T. B. Nguyen, M. Orozco, and W. L. Jorgensen, "Accuracy of free energies of hydration for organic molecules from 6-31g*-derived partial charges", J. Comput. Chem. **14**, 1240–1249 (1993).

[119]B. H. Besler, K. M. Merz Jr., and P. A. Kollman, "Atomic charges derived from semiempirical methods", J. Comput. Chem. **11**, 431–439 (1990).

[120]Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations", J. Comput. Chem. **24**, 1999–2012 (2003).

[121]D. S. Cerutti, J. E. Rice, W. C. Swope, and D. A. Case, "Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization", J. Phys. Chem. B **117**, PMID: 23379664, 2328–2338 (2013).

[122]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "Ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb", J. Chem. Theory Comput. **11**, PMID: 26574453, 3696–3713 (2015).

[123]H. S. Muddana, N. V. Sapra, A. T. Fenley, and M. K. Gilson, "The sampl4 hydration challenge: evaluation of partial charge sets with explicit-water molecular dynamics simulations", J. Comput. Aided Mol. Des. **28**, 277–287 (2014).

[124]P. G. Karamertzanis, P. Raiteri, and A. Galindo, "The use of anisotropic potentials in modeling water and free energies of hydration", J. Chem. Theory Comput. **6**, PMID: 26615693, 1590–1607 (2010).

[125]M. Schauperl, P. S. Nerenberg, H. Jang, L.-P. Wang, C. I. Bayly, D. L. Mobley, and M. K. Gilson, "Non-bonded force field model with advanced restrained electrostatic potential charges (resp2)", Commun. Chem. **3**, 44 (2020).

[126]R. H. Henchman and J. W. Essex, "Generation of opls-like charges from molecular electrostatic potential using restraints", J. Comput. Chem. **20**, 483–498 (1999).

[127]J. A. Rackers, Q. Wang, C. Liu, J.-P. Piquemal, P. Ren, and J. W. Ponder, "An optimized charge penetration model for use with the amoeba force field", Phys. Chem. Chem. Phys. **19**, 276–291 (2017).

[128]D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.4: open-source software for high-throughput quantum chemistry", J. Chem. Phys. **152**, 184108 (2020).

[129]P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern,

E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", Nature Methods **17**, 261–272 (2020).

[130]A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx", in Proceedings of the 7th python in science conference, edited by G. Varoquaux, T. Vaught, and J. Millman (2008), pp. 11–15.

[131]Pyyaml: a full-featured yaml framework for the python programming language.

[132]A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, Š. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz, "Sympy: symbolic computing in python", PeerJ Computer Science **3**, e103 (2017).

[133]Rdkit: open-source cheminformatics.

[134]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", J. Comput. Chem. **25**, 1157–1174 (2004).

[135]K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields", J. Comput. Chem. **31**, 10.1002/jcc.21367 (2010).

[136]W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids", J. Am. Chem. **118**, 11225–11236 (1996).

[137]C. J. Fennell, K. L. Wymer, and D. L. Mobley, "A fixed-charge model for alcohol polarization in the condensed phase, and its role in small molecule hydration", J. Phys. Chem. B **118**, PMID: 24702668, 6438–6446 (2014).

[138]N. M. Henriksen, A. T. Fenley, and M. K. Gilson, "Computational calorimetry: high-precision calculation of host–guest binding thermodynamics", J. Chem. Theory Comput. **11**, PMID: 26523125, 4377–4394 (2015).

[139]J. Yin, N. M. Henriksen, H. S. Muddana, and M. K. Gilson, "Bind3p: optimization of a water model based on host–guest binding data", J. Chem. Theory Comput. **14**, PMID: 29874074, 3621–3632 (2018).

[140]D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman, "Escaping atom types in force fields using direct chemical perception", J. Chem. Theory Comput. **14**, PMID: 30351006, 6076–6092 (2018).

[141]J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules?", J. Comput. Chem. **21**, 1049–1074 (2000).

[142]C. Bayly, D. McKay, and J. Truchon, http://www.ccl.net/cca/data/parm_at_Frosst/, 2010.

[143]D. R. Slochower, N. M. Henriksen, L.-P. Wang, J. D. Chodera, D. L. Mobley, and M. K. Gilson, "Binding thermodynamics of host–guest systems with smirnoff99frosst

1.0.5 from the open force field initiative", J. Chem. Theory Comput. **15**, 6225–6242 (2019).

[144]S. Vilar, G. Cozza, and S. Moro, "Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery", Curr. Top. Med. Chem. **8**, 1555–1572 (2008).

[145]Emolecules plus database download, 2013.

[146]R. Sure and S. Grimme, "Corrected small basis set hartree-fock method for large systems", J. Comput. Chem. **34**, 1672–1685 (2013).

[147]V. Sresht and B. Rai, Pfizerrd torsional-strain, calculation of torsional strain energy, 2019.

[148]J. N. Ehrman, V. T. Lim, C. C. Bannan, N. Thi, D. Y. Kyu, and D. L. Mobley, "Improving small molecule force fields by identifying and characterizing small molecules with inconsistent parameters", en, J. Comput. Aided Mol. Des. **35**, 271–284 (2021).

[149]Molssi qcfractal documentation, 2019.

[150]Molssi qcarchive web page, 2020.

[151]D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard, and T. D. Crawford, "The molssi qcarchive project: an open-source platform to compute, organize, and share quantum chemistry data", Wiley Interdiscip. Rev. Comput. Mol. Sci. **n/a**, e1491.

[152]N. Godbout, D. R. Salahub, J. Andzelm, and E. Wimmer, "Optimization of gaussian-type basis sets for local spin density functional calculations. part i. boron through neon, optimization technique and validation", Can. J. Chem. **70**, 560–571 (1992).

[153]S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu", J. Chem. Phys. **132**, 154104 (2010).

[154]S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", J. Comput. Chem. **32**, 1456–1465 (2011).

[155]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[156]Openeye toolkits 2019.10.2 openeye scientific software, santa fe, nm.

[157]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python", Journal of Machine Learning Research **12**, 2825–2830 (2011).

[158]C. D. Stern, C. I. Bayly, D. G. A. Smith, J. Fass, L.-P. Wang, D. L. Mobley, and J. D. Chodera, "Capturing non-local through-bond effects when fragmenting molecules for quantum chemical torsion scans", bioRxiv, `10.1101/2020.08.27.270934` (2020).

[159]L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building force fields: an automatic, systematic, and reproducible approach", J. Phys. Chem. Lett. **5**, PMID: 26273869, 1885–1891 (2014).

[160]L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez, and V. S. Pande, "Building a More Predictive

Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15", J. Phys. Chem. B **121**, 4023–4039 (2017).

[161]Forcebalance v1.7.1 used for 1.2.0 valence parameter fitting, 2020.

[162]J. Wagner, D. L. Mobley, J. Chodera, C. Bannan, A. Rizzi, Camila, C. Bayly, N. M. Lim, V. Lim, S. Sasmal, J. Rodríguez-Guerra, Y. Zhao, and L.-P. Wang, openforcefield/openforcefield: 0.6.0 Library Charges, version 0.6.0, Nov. 2019.

[163]H. Jang, Openff "parsley" 1.0.0-rc2 release package, 2020.

[164]V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly, and D. Mobley, "Benchmark assessment of molecular geometries and energies from small molecule force fields", June 2020.

[165]T. Schulz-Gasch, C. Schärfer, W. Guba, and M. Rarey, "Tfd: torsion fingerprints as a new measure to compare small molecule conformations", J. Chem. Inf. Model **52**, PMID: 22670896, 1499–1512 (2012).